Learning and 1-bit Compressed Sensing under Asymmetric Noise

Pranjal Awasthi PRANJAL.AWASTHI@RUTGERS.EDU
Rutgers University

Maria-Florina Balcan NINAMF@CS.CMU.EDU
Nika Haghtalab NHAGHTAL@CS.CMU.EDU
Hongyang Zhang HONGYANZ@CS.CMU.EDU
Carnegie Mellon University

Abstract

We study the approximate recovery problem under noise: Given corrupted 1-bit measurements of the form $\text{sign}(w^* \cdot x_i)$, recover a vector w with a small 0/1 loss w.r.t. $w^* \in \mathbb{R}^d$. In learning theory, this is known as the problem of *learning halfspaces with noise*, and in signal processing, as *1-bit compressed sensing*, in which there is an additional assumption that w^* is t-sparse. Direct formulations of the approximate recovery problem are non-convex and are NP-hard to optimize. In this paper, we propose adaptively solving a sequence of convex optimizations to mitigate the issue. Our algorithms output solutions with error as small as the information-theoretic limit under bounded and adversarial noise models. We also show that the usual one-shot approach of minimizing a convex surrogate fails to achieve this goal for a large family of loss functions.

1. Introduction

Designing noise-tolerant algorithms is a fundamental problem in machine learning, statistics, and signal processing (Cristianini and Shawe-Taylor, 2000; Freund and Schapire, 1997; Kearns and Vazirani, 1994; Valiant, 1984; Vapnik, 1998). In machine learning and statistics, study of such algorithms has led to significant advances in both the theory and practice of robust prediction and regression. In signal processing, these algorithms are used to recover sparse signals via a few noisy measurements. This is known as noisy compressed sensing or sparse recovery. In both cases, the problem can be stated as recovering a vector $w^* \in \mathbb{R}^d$ given noisy information about $w^* \cdot x_i$ or $\text{sign}(w^* \cdot x_i)$, where the x_i's are drawn from a distribution D. The focus of this work is on the latter setting known as classification or 1-bit compressed sensing in the respective communities. The goal is to output a solution that has a small 0/1 error, or equivalently, that minimizes the non-convex object function $\Pr_{x \sim D}[\text{sign}(w \cdot x) \neq \text{sign}(w^* \cdot x)]$.

Despite a large amount of work on designing noise-tolerant algorithm, many fundamental questions remain unresolved. In learning theory, one of the long-standing questions is designing efficient noise-tolerant learning algorithms that can approximate the unknown target vector w^* to arbitrary accuracy. In the absence of noise, the recovery problem can be solved efficiently via linear programming. However when measurements are noisy, this problem becomes more challenging in both its classification and 1-bit compressed sensing forms. This is due to the fact that direct formulations of the approximate recovery problem are non-convex and are NP-hard to optimize (Guruswami and Raghavendra, 2006). There is significant evidence to indicate that without assumptions on the noise and the distribution of x_i, such recovery might not be computationally possible (Daniely, 2015; Klivans and Kothari, 2014).

Due to the difficulty of the most general form of the problem, most positive results for obtaining arbitrarily good approximation have focused on the case of symmetric noise. A noise process is called *symmetric* if the probability that $\text{sign}(w^* \cdot x_i)$ is corrupted only depends on the magnitude $|w^* \cdot x_i|$ (Plan and Vershynin, 2013). Symmetric noise has many structural properties that one can exploit. For instance, when the marginal distribution over the instance space is symmetric, it can be shown that the sign weighted average of the samples is a good approximation to w^*. This is the main insight behind some of the existing works on classification and 1-bit compressed sensing that use a one-time application of convex optimization for recovering w^* (Servedio, 2001; Plan and Vershynin, 2013). However, when the noise is asymmetric, it can be shown that a one-shot application of convex optimization fails to recover the

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

true signal for a wide range of convex loss functions. Therefore, necessitating the design of non-convex methods.

1.1. Our Contribution

This paper is a summary of some of our recent works (Awasthi et al., 2014; 2015; 2016). We tackle the problem of approximate recovery under asymmetric noise and advance the state-of-the-art algorithms. In particular, we show how a sequence of carefully chosen convex optimization subproblems leads to a good approximation for w^*. We also show that a one-shot application of convex optimization does not achieve a desirable accuracy for a wide range of convex surrogate loss functions. We believe that our techniques have the potential to be applied to more general settings for solving hard non-convex optimization problems in machine learning.

Formally, we study a natural asymmetric noise model known as the *bounded noise (a.k.a Massart noise)* model. In this model, the probability of corrupting the $\mathrm{sign}(w^* \cdot x_i)$ is upper bounded by a constant $\frac{1}{2} - \frac{\beta}{2}$, i.e., an adversary flips the label of each point x_i with probability $\eta(x_i) \leq \frac{1}{2} - \frac{\beta}{2}$ (Boucheron et al., 2005; Sloan, 1996). In this work, we introduce a novel algorithm that efficiently approximates linear separators to arbitrary accuracy ϵ for any constant value of $\beta > 0$ in time $\mathrm{poly}(d, \frac{1}{\epsilon})$, when the marginal distribution is isotropic log-concave in $\mathbb{R}^d$. We also introduce an attribute-efficient variation of this algorithm and perform 1-bit compressed sensing with number of samples scaling only polynomially in the sparsity parameter and polylogarithmic in the ambient dimension.

We also consider the more challenging *adversarial (a.k.a agnostic)* noise model. Here, an adversary can flip any τ fraction of labels and no other assumption is made about the nature of the noise. As a result, even information theoretically, approximate recovery within arbitrarily small error is not possible (Kearns and Li, 1988). However, one could ask for recovering a w that satisfies $\Pr_{x\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq c\tau + \epsilon$, where $\epsilon > 0$ can be arbitrarily small. One would like to keep c as small as possible, ideally a constant. We provide a polynomial time algorithm that can approximately recover w^* in this model with $c = O(1)$ and the dependence on the number of samples is $O(\frac{t}{\epsilon^3}\log(\frac{1}{\epsilon})\mathrm{polylog}(d))$. This improves on the best known result of (Plan and Vershynin, 2013) in three ways: a) We improve the c to a constant, almost matching the information theoretic limit, as opposed to $\sqrt{\log\frac{1}{\tau}}$ used in the previous work; b) We improve the dependence on the number of samples to $\frac{1}{\epsilon^3}$ as opposed to $\frac{1}{\epsilon^6}$ in previous work. Furthermore, our algorithm is an active learning algorithm by design with improved label complexity of $O(\frac{t}{\epsilon^2}\log(\frac{1}{\epsilon})\mathrm{polylog}(d))$; And c) Our results hold when the distribution of x_i is any isotropic log-concave distribution. Prior work on 1-bit compressed sensing only handles the special case when the distribution is Gaussian.

Upper Bound: We discuss the case when the marginal distribution over x_i's is isotropic log-concave. We show that our algorithm for learning linear separators in $\mathbb{R}^d$ outputs a solution that can be arbitrarily close to the true classifier under bounded noise model.

Theorem 1 (Bounded Noise, Learning Problem)**.** *Let the optimal Bayes classifier be a halfspace denoted by w^*. Assume that the bounded noise condition holds for some constant $\beta \in (0, 1]$. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that there is an Algorithm with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, $\nu = \frac{3C_1}{8CC_2}$, $e_{KKMS} = \beta(\nu/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \nu\,\gamma_{k-1}/(4c_1 + 4c_2 + 2)$ running in polynomial time, proceeding in $s = O(\log\frac{1}{\epsilon})$ rounds, where in round k it takes $n_k = \mathrm{poly}(d, \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = \mathrm{poly}(d, \log(s/\delta))$ labels and with probability $1 - \delta$ returns a vector $w \in \mathbb{R}^d$ such that $\Pr_{x\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$.*

For the 1-bit compressed sensing, our algorithm outputs a solution that can be arbitrarily close to the underlying separator with fewer samples than the learning problem under the bounded noise model.

Theorem 2 (Bounded Noise, 1-bit Compressed Sensing)**.** *Let the optimal Bayes classifier be a halfspace denoted by w^* such that $\|w^*\|_0 = t$. Assume that the bounded noise condition holds for some constant $\beta \in (0, 1]$. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that there is an Algorithm with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, $\nu = \frac{3C_1}{8CC_2}$, $e_{KKMS} = \beta(\nu/(4c_1 + 4c_2 + 2))^4$, and $\tau_k = \nu\,\gamma_{k-1}/(4c_1 + 4c_2 + 2)$ running in polynomial time, proceeding in $s = O(\log\frac{1}{\epsilon})$ rounds, where in round k it takes $n_k = \mathrm{poly}(t\log(d), \exp(k), \log(\frac{1}{\delta}))$ unlabeled samples and $m_k = \mathrm{poly}(t, \log(sd/\delta), \exp(k))$ labels and with probability $1 - \delta$ returns a vector $w \in \mathbb{R}^d$ such that $\Pr_{x\sim D}[\mathrm{sign}(w \cdot x) \neq \mathrm{sign}(w^* \cdot x)] \leq \epsilon$.*

For the adversarial noise model, we study 1-bit compressed sensing and show that there is an algorithm that outputs a solution with error as small as the information-theoretic limit $O(\tau) + \epsilon$

Theorem 3 (Upper Bound for Adversarial Noise, 1-bit Compressed Sensing)**.** *Assume that the noise is adversarial and let the optimal linear classifier be a halfspace denoted by w^* such that $\|w^*\|_0 = t$. Let $\tau > 0$ be the error of w^*. For any $\epsilon > 0$, $\delta > 0$, there exist absolute constants e_0, C, C_1, C_2, c_1, c_2 such that there is an Algorithm with parameters $r_k = \frac{e_0}{C_1 2^k}$, $\gamma_k = Cr_k$, $\nu = \frac{3C_1}{8CC_2}$, and $\tau_k = \nu\,\gamma_{k-1}/(4c_1 + 4c_2 + 2)$ running in polynomial time, proceeding in $s = \log\frac{1}{\epsilon}$ rounds, where in round k it takes $n_k = \mathrm{poly}(t, \log(d), \exp(k), \log(\frac{1}{\delta}))$ unlabeled*

samples and $m_k = O(t\ \mathrm{polylog}(sd/\delta)2^{2k})$ *labels and with probability* $1-\delta$ *returns a vector* $w \in \mathbb{R}^d$ *such that* $\Pr_{x\sim D}[\mathrm{sign}(w\cdot x) \neq \mathrm{sign}(w^*\cdot x)] \leq O(\tau)+\epsilon$.

1.2. Our Technique: Iterative Convex Optimization

We use a localization technique inspired by the work of (Balcan et al., 2007). Instead of doing one-shot minimization, our algorithm chooses a sequence of optimization problems to solve in an adaptive manner. The algorithm is initialized with a classifier w_0 with an appropriate small constant *excess error*. The algorithm then proceeds in rounds, aiming to cut down the excess error by half in each round. By the properties of the noise and the log-concave distribution, excess error of a classifier is a linear function of its angle to w^*. Therefore, our algorithm aims to cut the angle by half at each round and eventually outputs a w that is close in angle to w^*.

In round $k-1$, consider w_{k-1} with angle $\leq \alpha_k$ to w^*. It can be shown that for a band of width $\gamma_{k-1} = \Theta(\alpha_k)$ around the separator w_{k-1}, w_{k-1} makes most of its error in this band. Therefore, improving the accuracy of w_{k-1} in the band significantly improves the accuracy of w_{k-1} overall. When considering vectors that are at angle $\leq \alpha_k$ to w_{k-1}, it can be shown that any vector w_k that achieves *a small enough constant excess error with respect to the distribution in the band*, indeed, enjoys a much stronger guarantee of having *excess error that is half of* w_{k-1} *overall.* Therefore, if such a vector w_k can be found efficiently in the presence of noise, a classifier of excess error ϵ can be learned in $O(\log(\frac{1}{\epsilon}))$ steps. In order to make the above method work we need to achieve two goals: a) achieve a constant excess error while tolerating noise rate of $\frac{1}{2}-\frac{\beta}{2}$ in the band and b) the hypothesis output should be a halfspace.

On one hand, efficient proper learning methods, such as surrogate loss minimization in the band, readily achieve goal (b). However, convex surrogate loss functions are only a good approximation of the 0/1 loss when the noise is small enough. Since the noise in the band can be as high as $\frac{1}{2}-\frac{\beta}{2}$, this directly restricts the noise rate that can be tolerated with such methods. Indeed, Awasthi et al. (2015) demonstrated that when hinge-loss minimization is used in the band, such a method only works if the probability of flipping the label is as small as $\approx 10^{-6}$, i.e., when β is very close to 1. On the other hand, the polynomial regression approach of (Kalai et al., 2008) learns linear separators to an arbitrary excess error of ϵ with runtime $\mathrm{poly}(d, \exp(\mathrm{poly}(\frac{1}{\epsilon}))$ when the marginal distribution is log-concave, requiring no additional assumption on noise. Since the distribution in the band is also log-concave, this method can achieve *an arbitrarily small constant* excess error in the band thereby achieving goal (a). However, this algorithm outputs the sign of a polynomial $p(\cdot)$ as a hypothesis, which is not necessarily a halfspace.

Instead, our algorithm takes a novel two-step approach to find w_k for *any amount of noise*. This is done by first finding a polynomial p_k that has a small constant excess error in the band. To obtain such a polynomial, we choose $\mathrm{poly}(d, \log(\frac{\log(1/\epsilon)}{\delta}))$ labeled samples from the distribution in the band and use the algorithm by (Kalai et al., 2005) to find a polynomial with a small enough but, importantly, *a constant excess error*, e_{KKMS}, in the band. Note that at this point p_k already satisfies goal (a) but it does not satisfy goal (b) as it is not a halfspace. At a high level, since p_k has a small excess error with respect to w^* in the band, using a structural property of the noise that connects the excess error and disagreement of a classifier with respect to w^*, we can show that p_k is also close in classification to w^*. Therefore, it suffices to agnostically learn a halfspace w_k to a constant error for samples in the band that are labeled based on $\mathrm{sign}(p(\cdot))$. To achieve this, we use localized hinge loss minimization in the band over a set of samples that are labeled based on predictions of p_k to find w_k. Therefore, w_k is close in classification to p_k in the band, which is in turn close to w^* in the band. As a result, w_k also has a small error in the band as desired. The pseudocode of the technique is displayed in Algorithm 1.

Algorithm 1 LEARNING HALFSPACES UNDER NOISE

Input: An initial classifier w_0, a sequence of values γ_k, τ_k and r_k for $k = 1, \ldots, \log(1/\epsilon)$. An error value e_{KKMS}.

1. Let w_0 be the initial classifier with small constant error.
2. For $k = 1, \ldots, \log(1/\epsilon) = s$
   (a) Take $\mathrm{poly}(d, \log(\frac{s}{\delta}))$ labeled samples from $\tilde{D}_k$, the conditional distribution within the band $\{x : |w_{k-1}\cdot x| \leq \gamma_{k-1}\}$, and place them in the set T. Run the polynomial regression algorithm (Kalai et al., 2005) over T to find a polynomial p_k such that $\mathrm{err}_{\tilde{D}_k}(\mathrm{sign}(p_k)) \leq \mathrm{err}_{\tilde{D}_k}(h_{w^*}) + e_{\mathrm{KKMS}}$.
   (b) Take $d(d+\log(k/\delta))$ unlabeled samples from $\tilde{D}_k$ and label them according to $\mathrm{sign}(p_k(\cdot))$. Call this set of labeled samples T'.
   (c) Find $v_k \in B(w_{k-1}, r_{k-1})$ that approximately minimizes the empirical hinge loss over T' using threshold τ_k, i.e., $L_{\tau_k}(v_k, T') \leq \min_{w\in B(w_{k-1}, r_{k-1})} L_{\tau_k}(w, T') + \frac{\nu}{12}$.
   (d) Let $w_k = \frac{v_k}{\|v_k\|_2}$.

Output: Return w_s that has error ϵ with probability $1-\delta$.

1.3. On the Efficacy of One-shot Convex Surrogate

Approximating a non-convex optimization with a convex surrogate loss is a common practice in the optimization community. There are many positive results that prove the validity of such a technique. For example, in sparse coding

or LASSO, it has been shown that replacing the ℓ_0 norm with its convex hull, the ℓ_1 norm, exactly outputs a solution that is sparse under some mild conditions (Tibshirani, 1996). A similar positive phenomenon occurs when replacing the rank function with the nuclear norm in the matrix completion (Candès and Recht, 2009) or robust PCA problem (Candès et al., 2011; Zhang et al., 2015). On the other hand, there are negative results that provide counterexamples on the efficacy of one-shot convex surrogate. For example, it has been shown that nuclear norm minimization does not necessarily guarantee a solution of low rank in the subspace clustering problem (Zhang et al., 2013).

In classification problems or 1-bit compressed sensing, the goal is to output a vector w that has a small 0/1 error in presence of noise. Traditional approaches, e.g. the support vector machine, replace the 0/1 error with a convex surrogate loss, e.g., the hinge loss, and perform one-shot convex optimization. Below, we give a counterexample/lower bound showing that even in the presence of a mild noise model, such as bounded noise, the answer is negative for a large family of surrogate losses. This justifies that an adaptive sequence of convex optimizations is indispensable to achieving a small 0/1 error in presence of noise.

1.4. Lower Bound

In this section, we show that one-shot minimization does not work for a large family of loss functions that include any continuous loss with a natural property that points at the same distance from the separator have the same loss. This justifies why minimizing a sequence of carefully designed losses, as we did in this paper, is indispensable to achieving an arbitrarily small error under bounded noise.

Formally, let $\mathcal{P}_\beta$ be the class of noisy distribution $\tilde{D}$ with uniform marginal over the unit ball, and let (z_w, φ_w) represent the polar coordinate of a point P in the instance space, where φ_w represents the angle between the linear separator h_w and the vector from origin to P, and z_w is the L_2 distance of the point P and the origin. Let $\ell_+^w(z_w, \varphi_w)$ and $\ell_-^w(z_w, \varphi_w)$ denote the loss functions on point P with correct and incorrect classification by h_w, respectively. The loss functions we study satisfy the following properties.

Definition 1. *Continuous loss functions $\ell_+^w(z_w, \varphi_w)$ and $\ell_-^w(z_w, \varphi_w)$ are called* proper, *if and only if 1. $\ell_+^w(z_w, \varphi_w) = \ell_+^w(z_w, k\pi \pm \varphi_w)$ and $\ell_-^w(z_w, \varphi_w) = \ell_-^w(z_w, k\pi \pm \varphi_w)$, for $k \in N$; 2. For $z_w > 0$, $\ell_-^w(z_w, \varphi_w) \geq \ell_+^w(z_w, \varphi_w)$; The equality holds if and only if $\varphi_w = k\pi$, $\forall k \in N$.*

Intuitively, Property 1 states that the loss $\ell_+^w(z_w, \varphi_w)$ (or $\ell_-^w(z_w, \varphi_w)$) on the points with the same angle to the separator (indicated by points of the same color) are the same. Property 2 is a very natural assumption since to achieve low error, it is desirable to penalize misclassification more.

In fact, most of the commonly used loss functions (Bartlett et al., 2006) satisfy our two properties in Definition 1, e.g., the (normalized) hinge loss, logistic loss, square loss, exponential loss, and truncated quadratic loss. Furthermore, we highlight that Definition 1 covers the loss even with regularized term on w. A concrete example is traditional 1-bit compressed sensing, with loss function formulated as $\ell_+(z_w, \varphi_w) = -|z_w \sin \varphi_w| + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2$ and $\ell_-(z_w, \varphi_w) = |z_w \sin \varphi_w| + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2$. Thus our lower bound demonstrates that one-shot 1-bit compressed sensing cannot always achieve arbitrarily small excess error under the Massart noise.

Theorem 4 (Bounded Noise, Lower Bound)**.** *For every bounded noise parameter $0 \leq \beta < 1$, there exists a distribution $\tilde{D}_\beta \in \mathcal{P}_\beta$ (that is, a distribution over $\mathbb{R}^2 \times \{+1, -1\}$, where the marginal distribution on $\mathbb{R}^2$ is uniform over the unit ball, and the labels $\{+1, -1\}$ satisfies bounded noise condition with parameter β) such that any proper loss minimization is not consistent on $\tilde{D}_\beta$ w.r.t. the class of halfspaces. That is, there exists an $\epsilon \geq 0$ and a sample size $m(\epsilon)$ such that any proper loss minimization will output a classifier of excess error larger than ϵ by a high probability over sample size at least $m(\epsilon)$.*

2. Conclusion and Open Problem

Our work improves the state of the art results on classification and 1-bit compressed sensing in presence of asymmetric noise. For the general non-sparse case, our work provides the first algorithm for finding a halfspace that is arbitrarily close to w^* in presence of bounded noise for any constant maximum flipping probability. Furthermore, we extend the platform used for approximate recovery in presence of bounded or adversarial noise to 1-bit compressed sensing problem. By adaptively solving a sequence of convex optimization problems, we get a solution with error as small as the information-theoretic limit.

A family of interesting noise models lie between bounded noise model and adversarial noise model. One of them is Tsybakov model. Instead of having a constant ($< \frac{1}{2}$) maximum flipping probability for each example, the Tsybakov model allows the flipping probability to be arbitrarily close to $\frac{1}{2}$, say $\frac{1}{2} - \epsilon$, provided that the density of the region where the probability is that close decays as a polynomial function of ϵ (Tsybakov, 2004; Boucheron et al., 2005; Castro and Nowak, 2007). In the Tsybakov model, the noise level increases in the band around the true separator w^*, and hence a straightforward application of our localization technique fails. It remains a fascinating open problem to design an efficient algorithm that outputs a solution with arbitrarily small error under the Tsybakov noise model.

References

Awasthi, P., Balcan, M. F., Haghtalab, N., and Urner, R. (2015). Efficient learning of linear separators under bounded noise. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*.

Awasthi, P., Balcan, M. F., Haghtalab, N., and Zhang, H. (2016). Learning and 1-bit compressed sensing under asymmetric noise. In *To appear in Proceedings of the 29th Annual Conference on Learning Theory (COLT)*.

Awasthi, P., Balcan, M. F., and Long, P. M. (2014). The power of localization for efficiently learning linear separators with noise. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 449–458. ACM.

Balcan, M. F., Broder, A., and Zhang, T. (2007). Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:9:323–375.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Castro, R. and Nowak, R. (2007). Minimax bounds for active learning.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Daniely, A. (2015). Complexity theoretic limitations on learning halfspaces. *CoRR*, abs/1505.05800.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Guruswami, V. and Raghavendra, P. (2006). Hardness of learning halfspaces with noise. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

Kalai, A. T., Klivans, A. R., Mansour, Y., and Servedio, R. A. (2005). Agnostically learning halfspaces. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

Kalai, A. T., Mansour, Y., and Verbin, E. (2008). On agnostic boosting and parity learning. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 629–638. ACM.

Kearns, M. and Li, M. (1988). Learning in the presence of malicious errors. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC)*.

Kearns, M. and Vazirani, U. (1994). *An introduction to computational learning theory*. MIT Press, Cambridge, MA.

Klivans, A. and Kothari, P. (2014). Embedding hard learning problems into gaussian space. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, 28:793–809.

Plan, Y. and Vershynin, R. (2013). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494.

Servedio, R. A. (2001). *Efficient algorithms in computational learning theory*. Harvard University.

Sloan, R. H. (1996). PAC learning, noise, and geometry. In *Learning and Geometry: Computational Approaches*, pages 21–41. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Zhang, H., Lin, Z., and Zhang, C. (2013). A counterexample for the validity of using nuclear norm as a convex surrogate of rank. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 226–241. Springer.

Zhang, H., Lin, Z., Zhang, C., and Chang, E. Y. (2015). Exact recoverability of robust PCA via outlier pursuit with tight recovery bounds. In *AAAI*, pages 3143–3149.