Information Theoretic Guarantees for Empirical Risk Minimization with Applications to Model Selection and Large-Scale Optimization

Ibrahim Alabdulmohsin [1]

Abstract

In this paper, we derive bounds on the mutual information of the empirical risk minimization (ERM) procedure for both 0-1 and strongly-convex loss classes. We prove that under the Axiom of Choice, the existence of an ERM learning rule with a vanishing mutual information is equivalent to the assertion that the loss class has a finite VC dimension, thus bridging information theory with statistical learning theory. Similarly, an asymptotic bound on the mutual information is established for strongly-convex loss classes in terms of the number of model parameters. The latter result rests on a central limit theorem (CLT) that we derive in this paper. In addition, we use our results to analyze the excess risk in stochastic convex optimization and unify previous works. Finally, we present two important applications. First, we show that the ERM of strongly-convex loss classes can be *trivially* scaled to big data using a naïve parallelization algorithm with provable guarantees. Second, we propose a simple information criterion for model selection and demonstrate experimentally that it outperforms the popular Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC).

1. Introduction

Learning via the empirical risk minimization (ERM) of stochastic loss is a ubiquitous framework that has been widely applied in machine learning and statistics. It is often regarded as the default strategy to use, due to its simplicity, generality, and statistical efficiency (Shalev-Shwartz et al., 2009; Koren & Levy, 2015; Vapnik, 1999; Shalev-Shwartz & Ben-David, 2014). Given a fixed hypothesis space $\mathcal{H}$, a domain $\mathcal{Z}$, and a loss function on the product space $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, the ERM learning rule selects the hypothesis $\hat{\mathbf{h}}$ that minimizes the empirical risk:

$$\hat{\mathbf{h}} = \arg\min_{h \in \mathcal{H}} \Big\{ F_S(h) = \frac{1}{m} \sum_{i=1}^{m} f(h,\, z_i) \Big\}, \qquad (1)$$

where $S = (z_1, \ldots, z_m) \in \mathcal{Z}^m$ is a sample of m observations drawn i.i.d. from some unknown probability distribution $\mathcal{D}$. To simplify notation, we omit the dependence of $\hat{\mathbf{h}}$ on the sample S. By contrast, the true risk minimizer is denoted $\mathbf{h}^\star$:

$$\mathbf{h}^\star = \arg\min_{h \in \mathcal{H}} \Big\{ F(h) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}} \left[f(h,\, \mathbf{z}) \right] \Big\}. \qquad (2)$$

Hence, learning via ERM is justified if and only if $F(\hat{\mathbf{h}}) \leq F(\mathbf{h}^\star) + \epsilon$, for some provably small ϵ.

In some applications, such as classification, the stochastic loss of interest is often the 0-1 loss. However, minimizing the 0-1 loss is, in general, computationally intractable even for simple hypothesis spaces, such as linear classifiers (Feldman et al., 2009). To circumvent this difficulty, a convex, surrogate loss is used instead, which is justified by consistency results when the loss is calibrated (Bartlett et al., 2006). When the stochastic loss is convex, ERM has occasionally been referred to by various names in the literature, such as stochastic convex optimization and stochastic average approximation (Shalev-Shwartz et al., 2009; Feldman, 2016). Examples of the latter framework include least squares, logistic regression, and SVM.

In the literature, several approaches have been proposed for analyzing the generalization risk and, consequently, the consistency of empirical risk minimization. The most dominant approach is *uniform convergence*, with seminal results for 0-1 loss classes dating to the early work of Vapnik and Chervonenkis in the 1970s (Shalev-Shwartz & Ben-David, 2014; Abu-Mostafa et al., 2012; Vapnik, 1999). The Fundamental Theorem of Statistical Learning states that a hypothesis space $\mathcal{H}$ is agnostic PAC-learnable via ERM if and only if it is PAC-learnable at all, and that this occurs if and only if $\mathcal{H}$ has a finite VC dimension (Shalev-Shwartz & Ben-David, 2014). This elegant characterization shows that the generalization risk for 0-1 loss classes can be bounded by computing their VC dimensions.

[1]Saudi Aramco, Dhahran 31311, Saudi Arabia. Correspondence to: Ibrahim Alabdulmohsin <ibrahim.alabdulmohsin@kaust.edu.sa>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Similarly, when the stochastic loss is both convex and of the *generalized* linear form, where $f(h, z_i) = r(h) + g(\langle h, \phi(z_i)\rangle, z_i)$ and $r : \mathcal{H} \to \mathbb{R}$ is a strongly-convex regularizer, then, under mild additional conditions of smoothness and boundedness, it has been shown that the true risk sub-optimality $F(h) - F(\mathbf{h}^\star)$ is bounded *uniformly* in $\mathcal{H}$ by a constant multiple of the empirical risk sub-optimality $F_S(h) - F_S(\hat{\mathbf{h}})$ plus an $O(1/m)$ term (Sridharan et al., 2009). In general, it can be shown that for strongly-convex stochastic losses, the risk of the empirical risk minimizer converges to the optimal risk with rate $O(1/m)$, which justifies learning via the empirical risk minimization (Shalev-Shwartz et al., 2009). Hence, the consistency of empirical risk minimization for both 0-1 loss classes and strongly-convex loss functions is, generally, well-understood.

However, many information-theoretic approaches have been recently proposed for analyzing machine learning algorithms. These include generalization bounds that are based on max-information (Dwork et al., 2015), leave-one-out information (Raginsky et al., 2016), and the mutual information (Alabdulmohsin, 2015; Russo & Zou, 2016). They have found applications in important areas, such as in privacy and adaptive data analysis, since they naturally lead to guaranteed stability, bounded information leakage, and robustness against post-processing. Compared to other information-theoretic approaches, uniform generalization bounds using the *variational information*, also called T-information (Raginsky et al., 2016), yield the tightest results for bounded loss functions, as deduced by the Pinsker inequality (Cover & Thomas, 1991). It was proposed in (Alabdulmohsin, 2015), who used it to prove that uniform generalization, algorithmic stability, and bounded information were, in fact, equivalent conditions on the learning algorithm. A similar notion, called "robust generalization" was later proposed in (Cummings et al., 2016), who analyzed its significance in the adaptive learning setting and showed that it could be achieved using sample compression schemes, finite description lengths, and differential privacy. However, these two notions of "uniform" and "robust" generalization turned out to be equivalent (Alabdulmohsin, 2017).

To describe uniform generalization in informal terms, suppose we have a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$, which selects a hypothesis $\mathbf{h} \in \mathcal{H}$ according to a training sample $S \in \mathcal{Z}^m$. Let the generalization risk of $\mathcal{L}$ w.r.t. some bounded loss function $l : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ be defined by:

$$R_{gen}(\mathcal{L}) = \mathbb{E}_{S\sim\mathcal{D}^m, \mathbf{h}}\left[L(\mathbf{h}) - L_S(\mathbf{h})\right], \qquad (3)$$

where $L(h) = \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[l(h, \mathbf{z})]$ and L_S is its empirical counterpart $L_S(h) = \mathbb{E}_{\mathbf{z}\sim S}[l(h, \mathbf{z})]$. In Eq. (3), the expectation is taken over the random choice of the sample and the internal randomness (if any) in the learning algorithm. Note that l in Eq. (3) can be different from the loss f in Eq. (1) that is optimized during the learning stage, e.g. l can be a 0-1 misclassification error rate while f is some convex, surrogate loss. Then, $\mathcal{L}$ is said to generalize *uniformly* with rate $\epsilon > 0$ if $|R_{gen}(\mathcal{L})| \leq \epsilon$ is guaranteed to hold *independently* of the choice of l. That is, the generalization guarantee holds *uniformly in expectation* across all parametric loss functions, hence the name.

Definition 1 (Uniform Generalization)**.** *A learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ generalizes uniformly with rate $\epsilon \geq 0$ if for all bounded parametric losses $l : \mathcal{H} \times \mathcal{Z} \to [0, 1]$, we have $|R_{gen}(\mathcal{L})| \leq \epsilon$, where $R_{gen}(\mathcal{L})$ is given in Eq. (3).*

Informally, Definition 1 states that once a hypothesis $\mathbf{h}$ is selected by a learning algorithm $\mathcal{L}$ that achieves uniform generalization, then no "adversary" can post-process the hypothesis in a manner that causes over-fitting to occur (Alabdulmohsin, 2015; Cummings et al., 2016). Equivalently, uniform generalization implies that the empirical performance of $\mathbf{h}$ on the sample S is a faithful approximation to its true risk, regardless of how that performance is measured. For example, the loss function l in Eq. (3) can be the misclassification error rate, a cost-sensitive error rate in fraud detection and medical diagnosis (Elkan, 2001), or it can be the Brier score in probabilistic predictions (Kull & Flach, 2015). The generalization guarantee would hold in any case.

The main theorem of (Alabdulmohsin, 2015) states that the uniform generalization risk has a precise information-theoretic characterization:

Theorem 1 (Alabdulmohsin, 2015)**.** *Given a fixed $0 \leq \epsilon \leq 1$ and a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ that selects a hypothesis $\boldsymbol{h} \in \mathcal{H}$ according to a training sample $S = \{z_1, \ldots, z_m\}$, where $z_i \sim \mathcal{D}$ are i.i.d., then $\mathcal{L}$ generalizes uniformly with rate ϵ if and only if $\mathcal{J}(\boldsymbol{h}; \hat{\boldsymbol{z}}) \leq \epsilon$, where $\hat{\boldsymbol{z}} \sim S$ is a single random training example, $\mathcal{J}(\boldsymbol{x}; \boldsymbol{y}) = ||p(\boldsymbol{x})p(\boldsymbol{y}) \;,\; p(\boldsymbol{x}, \boldsymbol{y})||_{\mathcal{T}}$, and $||q_1 \;,\; q_2||_{\mathcal{T}}$ is the total variation distance between the probability measures q_1 and q_2.*

Throughout this paper, we will adopt the terminology used in (Alabdulmohsin, 2017) and call $\mathcal{J}(\mathbf{x}; \mathbf{y})$ the "variational information" between the random variables $\mathbf{x}$ and $\mathbf{y}$. Variational information is an instance of the class of *informativity measures* using f-divergences, for which an axiomatic basis has been proposed (Csiszár, 1972; 2008).

To illustrate Theorem 1, consider the case of a finite hypothesis space $|\mathcal{H}| < \infty$. Then, a classical argument using the union bound (Shalev-Shwartz & Ben-David, 2014) can be used to show that the generalization risk w.r.t. any fixed bounded loss $l : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ is $\tilde{O}(\sqrt{\log|\mathcal{H}|/m})$. This follows from the fact that the union bound argument does not make any additional assumptions on the loss l beyond the fact that it has a bounded range. Hence, the *uniform* generalization risk is $\tilde{O}(\sqrt{\log|\mathcal{H}|/m})$. However, Theorem 1 states that this bound must also hold for the variational information $\mathcal{J}(\mathbf{h}; \hat{\mathbf{z}})$ as well. This claim can, in fact, be ver-

ified readily using information theoretic inequalities since we have (Alabdulmohsin, 2015):

$$\mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}}) \leq \sqrt{\frac{I(\mathbf{h};\hat{\mathbf{z}})}{2}} \leq \sqrt{\frac{I(\mathbf{h};S)}{2m}} \leq \sqrt{\frac{\log|\mathcal{H}|}{2m}},$$

where $I(\mathbf{x};\mathbf{y})$ is the Shannon mutual information measured in nats (i.e. using natural logarithms). Here, the first inequality is due to Pinsker (Cover & Thomas, 1991), the second inequality follows because z_i are i.i.d., and the last inequality holds because the Shannon mutual information is bounded by the entropy. Note that in the latter case, only information-theoretic inequalities were used to recover this classical result without relying on the union bound.

The generalization guarantee in Theorem 1 depends on the probability distribution $\mathcal{D}$. It can be made distribution-free by defining the *capacity* of a learning algorithm $\mathcal{L}$ by:

$$C(\mathcal{L}) = \sup_{p(z)} \{\mathcal{J}(\mathbf{h};\, \hat{\mathbf{z}}) = \mathbb{E}_{\hat{\mathbf{z}}\sim p(\mathbf{z})}||p(\mathbf{h})\,,\;\; p(\mathbf{h}|\hat{\mathbf{z}})||_{\mathcal{T}}\}, \quad (4)$$

where the supremum is taken over all possible distributions of observations. This is analogous to the capacity of communication channels in information theory (Cover & Thomas, 1991). Then, the generalization risk of $\mathcal{L}$ is bounded by $C(\mathcal{L})$ for any distribution $\mathcal{D}$ and any bounded loss function $l : \mathcal{H} \times \mathcal{Z} \rightarrow [0, 1]$. In many cases, such as in finite description lengths, countable domains, mean estimation, differential privacy, and sample compression schemes, the capacity of $C(\mathcal{L})$ can be tightly bounded (Alabdulmohsin, 2015; 2017). In this paper, we derive new bounds for the capacity of the empirical risk minimization (ERM) procedure.

Finally, we point out that even though uniform generalization guarantees are originally defined only in expectation as shown in Eq. (3), it has recently been established that these guarantees of *uniform* generalization in expectation also implied a generalization with a *high probability* as well (Alabdulmohsin, 2017). This is in stark contrast with traditional generalization in expectation guarantees that do not imply concentration.

2. Contributions

The first contribution of this paper is to establish tight relations between uniform generalization and the empirical risk minimization (ERM) procedure. This will allow us to bridge information theory with statistical learning theory. More specifically, we will prove that under the Axiom of Choice, an ERM learning rule always *exists* that has a vanishing learning capacity $C(\mathcal{L})$ if and only if the 0-1 loss class has a finite VC dimension.

Second, we prove that the empirical risk minimization of strongly-convex stochastic loss also generalizes uniformly in expectation. To establish the latter result, we prove the asymptotic normality of the empirical risk minimizer over the random choice of the training sample, which is a useful result in its own right, such as for uncertainty quantification and hypothesis testing. In the machine learning literature, central limit theorems have been employed to analyze online learning algorithms. For instance, (Polyak & Juditsky, 1992) derived a central limit theorem (CLT) when using an averaging method to accelerate convergence. More recently, (Mandt et al., 2016) derived a central limit theorem for SGD with a fixed learning rate under simplifying assumptions. Then, (Mandt et al., 2016) used their result to select the learning rate such that the trajectory of SGD generates samples from the posterior distribution. Unlike the work in (Mandt et al., 2016), we prove our CLT without relying on some unnecessary simplifying assumptions, such as the Gaussianity of noise. As a consequence of our work, we present a generalization bound for stochastic convex optimization, which only depends on the number of model parameters and applies to any learning task, such as regression, multi-class classification, and ranking. Next, we use our results to analyze the excess risk in stochastic convex optimization and unify previous works.

Finally, we demonstrate two important applications. First, we show that the ERM of strongly-convex loss classes can be *trivially* distributed and scaled to big data using a naïve parallelization algorithm whose performance is provably equivalent to that of the ERM learning rule. Unlike previous works, our parallelization algorithm does not rely on advanced treatments, such as the ADMM procedure (Boyd et al., 2011). Second, we use our results to propose a simple information criterion for model selection in terms of the number of model parameters and demonstrate experimentally that it outperforms the popular Akaike's information criterion (AIC) (Akaike, 1998; Bishop, 2006) and Schwarz's Bayesian information criterion (BIC) (Schwarz, 1978).

3. Notation

Our notation is fairly standard. Some exceptions that may require further clarification are as follows. First, when $\mathbf{x}$ is a random variable whose value is drawn uniformly at random from a finite set S, we will write $\mathbf{x} \sim S$ to denote this fact. Second, if $\mathbf{x}$ is a predicate, then $\mathbb{I}\{\mathbf{x}\} = 1$ if and only if $\mathbf{x}$ is true, otherwise $\mathbb{I}\{\mathbf{x}\} = 0$. Third, $A \preceq B$ is a linear matrix inequality (LMI), i.e. $A \preceq B$ is equivalent to the assertion that $B - A$ is positive semidefinite.

Moreover, we will use the *order in probability* notation for real-valued *random* variables. Here, we adopt the notation used in (Janson, 2011; Tao, 2012). In particular, let $\mathbf{x} = \mathbf{x}_n$ be a real-valued random variable that depends on some parameter $n \in \mathbb{N}$. Then, we will write $\mathbf{x}_n = O_p(f(n))$ if for any $\delta > 0$, there exists absolute constants C and n_0 such that for any fixed $n \geq n_0$, the inequality $|\mathbf{x}_n| < C\,|f(n)|$

holds with a probability of, at least, $1-\delta$. In other words, the ratio $\mathbf{x}_n/f(n)$ is *stochastically* bounded (Janson, 2011). Similarly, we write $\mathbf{x}_n = o_p(f(n))$ if $\mathbf{x}_n/f(n)$ converges to zero in probability. As an example, if $\mathbf{x} \sim \mathcal{N}(0, I_d)$ is a standard multivariate Gaussian vector, then $||\mathbf{x}||_2 = O_p(\sqrt{d})$ even though $||\mathbf{x}||_2$ can be arbitrarily large. Intuitively, the probability of the event $||\mathbf{x}||_2 \geq d^{\frac{1}{2}+\epsilon}$ when $\epsilon > 0$ goes to zero as $d \to \infty$ so $||\mathbf{x}||_2$ is *effectively* of the order $O(\sqrt{d})$.

Finally, let $\mathcal{H}$ be a fixed hypothesis space and let $\mathcal{Z}$ be a fixed domain of observations. Given a 0-1 loss function $f : \mathcal{H} \times \mathcal{Z} \to \{0, 1\}$, we will abuse terminology slightly by speaking about the "VC dimension" of $\mathcal{H}$, when we actually mean the VC dimension of the loss class $\{f(h, \cdot) : h \in \mathcal{H}\}$. Because the 0-1 loss f will be clear from the context, this should not cause any ambiguity.

4. Empirical Risk Minimization of 0-1 Loss Classes

We begin by analyzing the ERM learning rule for 0-1 loss classes. Before we establish that a finite VC dimension is sufficient to guarantee the existence of ERM learning rules that generalize uniformly in expectation, we first describe why ERM by itself is not sufficient even when the hypothesis space has a finite VC dimension[1].

Proposition 1. *For any sample size $m \geq 1$ and a positive constant $\epsilon > 0$, there exists a hypothesis space $\mathcal{H}$, a domain $\mathcal{Z}$, and a 0-1 loss function $f : \mathcal{H} \times \mathcal{Z} \to \{0, 1\}$ such that: (1) $\mathcal{H}$ has a VC dimension $d = 1$, and (2) a learning algorithm $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ exists that outputs an empirical risk minimizer $\hat{\boldsymbol{h}}$ with $\mathcal{J}(\hat{\boldsymbol{h}}; \hat{\boldsymbol{z}}) \geq 1 - \epsilon$, where $\hat{\boldsymbol{z}} \sim S$ is a single random training example.*

Proposition 1 shows that one cannot obtain a non-trivial bound on the uniform generalization risk of an ERM learning rule in terms of the VC dimension d and the sample size m without imposing some additional restrictions. Next, we prove that an ERM learning rule exists that satisfies the uniform generalization property if the hypothesis space has a finite VC dimension.

We begin by recalling a fundamental result in modern set theory. A non-empty set $\mathcal{Q}$ is said to be *well-ordered* if $\mathcal{Q}$ is endowed with a total order $\preceq$ such that every non-empty subset of $\mathcal{Q}$ contains a least element. The following fundamental result is due to Ernst Zermelo.

Theorem 2 (Well-Ordering Theorem)**.** *Under the Axiom of Choice, every non-empty subset can be well-ordered.*

Theorem 2 was proved by Zermelo in 1904 (Kolmogorov & Fomin, 1970).

[1]Detailed formal proofs are available in the supplementary materials.

Theorem 3. *Given a hypothesis space $\mathcal{H}$, a domain $\mathcal{Z}$, and a 0-1 loss $f : \mathcal{H} \times \mathcal{Z} \to \{0, 1\}$, let $\preceq$ be a well-ordering on $\mathcal{H}$ and let $\mathcal{L} : \mathcal{Z}^m \to \mathcal{H}$ be the learning rule that outputs the "least" empirical risk minimizer to the training sample $S \in \mathcal{Z}^m$ according to $\preceq$. Then, $C(\mathcal{L}) \to 0$ as $m \to \infty$ if $\mathcal{H}$ has a finite VC dimension. In particular:*

$$C(\mathcal{L}) \leq \frac{3}{\sqrt{m}} + \sqrt{\frac{1 + d \log \frac{2em}{d}}{m}},$$

where $C(\mathcal{L})$ is given by Eq. (4) and d is the VC dimension of $\mathcal{H}$, provided that $m \geq d$.

Next, we prove a converse statement. Before we do this, we present a learning problem that shows why a converse to Theorem 3 is not generally possible without making some additional assumptions. Hence, our converse will be later established for the binary classification setting only.

Example 1 (Integer Subset Learning Problem)**.** *Let $\mathcal{Z} = \{1, 2, 3, \ldots, d\}$ be a finite set of positive integers. Let $\mathcal{H} = 2^{\mathcal{Z}}$ and define the loss of a hypothesis $h \in \mathcal{H}$ to be $f(h, z) = \mathbb{I}\{z \notin h\}$. Then, the VC dimension is d. However, the learning rule that outputs $h = \mathcal{Z}$ is always an ERM learning rule that generalize uniformly with rate $\epsilon = 0$ regardless of the sample size and the distribution of observations.*

The previous example shows that a converse to Theorem 3 is not possible without imposing some additional constraints. In particular, in the Integer Subset Learning Problem, the VC dimension is not a useful measure of the complexity of the hypothesis space $\mathcal{H}$ because many hypotheses dominate others (i.e. perform better across all distributions of observations). For example, the hypothesis $h' = \{1, 2, 3\}$ dominates $h'' = \{1\}$ because there is no distribution on observations in which h'' outperforms h'. Even worse, the hypothesis $h = \mathcal{Z}$ dominates all other hypotheses.

Consequently, in order to prove a lower bound for all ERM rules, we consider the standard binary classification setting.

Theorem 4. *In any fixed domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, let the hypothesis space $\mathcal{H}$ be a concept class on $\mathcal{X}$ and let $f(h, x, y) = \mathbb{I}\{y \neq h(x)\}$ be the misclassification error. Then, any ERM learning rule $\mathcal{L}$ w.r.t. f has a learning capacity $C(\mathcal{L})$ that is bounded from below by $C(\mathcal{L}) \geq \frac{1}{2}\left(1 - \frac{1}{d}\right)^m$, where m is the training sample size and d is the VC dimension of $\mathcal{H}$.*

Using both Theorem 3 with Theorem 4, we arrive a crisp characterization of the VC dimension of concept classes in terms of information theory.

Theorem 5. *Given a fixed domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, let the hypothesis space $\mathcal{H}$ be a concept class on $\mathcal{X}$ and let $f(h, x, y) = \mathbb{I}\{y \neq h(x)\}$ be the misclassification error. Let m be the sample size. Then, the following statements are equivalent under the Axiom of Choice:*

1. *$\mathcal{H}$ admits an ERM learning rule $\mathcal{L}$ whose learning capacity $C(\mathcal{L})$ satisfies $C(\mathcal{L}) \to 0$ as $m \to \infty$.*

2. *$\mathcal{H}$ has a finite VC dimension.*

Proof. The lower bound in Theorem 4 holds for all ERM learning rules. Hence, an ERM learning rule exists that generalize uniformly with a vanishing rate across all distributions only if $\mathcal{H}$ has a finite VC dimension. However, under the Axiom of Choice, $\mathcal{H}$ can always be well-ordered by Theorem 2 so, by Theorem 3, a finite VC dimension is also sufficient to guarantee the existence of a learning rule that generalize uniformly. □

Theorem 5 presents a crisp characterization of the VC dimension in terms of information theory. According to the theorem, an ERM learning rule can be constructed that does not encode the training sample *if and only if* the hypothesis space has a finite VC dimension.

Remark 1. *In (Cummings et al., 2016), it has been argued that uniform generalization, called "robust generalization" in the paper, is important in the adaptive learning setting because it implies that no adversary can post-process the hypothesis and causes over-fitting to occur. In other words, no adversary can use the hypothesis to infer new conclusions, which do not themselves generalize. It was shown in (Cummings et al., 2016) that this robust generalization guarantee was achievable by sample compression schemes and differential privacy. Theorem 3 shows that an ERM learning rule of 0-1 loss classes with finite VC dimensions always exists that satisfies this robust generalization property.*

Remark 2. *One method of constructing a well-ordering on a hypothesis space $\mathcal{H}$ is to use the fact that computers are equipped with finite precisions. Hence, in practice, every hypothesis space is enumerable, from which the normal ordering of the integers is a valid well-ordering.*

5. Empirical Risk Minimization of Strongly-Convex Loss Classes

Next, we analyze the ERM learning rule for strongly-convex loss classes. To recall, the ERM learning rule selects the hypothesis $\hat{\mathbf{h}}$ given by Eq. (1), which is identical to what was assumed before. However, we now further assume that $f(h,\, z)$ is γ-strongly convex, L-Lipschitz, and twice-differentiable on its first argument. Moreover, the hypothesis space is $\mathbb{R}^d$ for some finite $d < \infty$.

5.1. Central Limit Theorem

We begin by establishing a central limit theorem (CLT). Throughout this section, we will simplify notation by writing $f_i(h) = f(h, z_i)$.

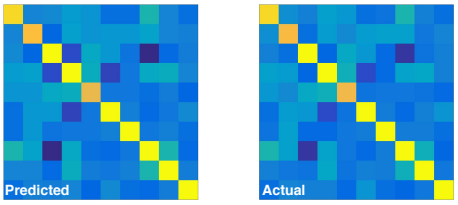

Figure 1. This figure presents the predicted (left) and actual (right) covariance matrices of the empirical risk minimizer to an ℓ_2-regularized logistic regression problem, where $d = 10$ and $m = 1000$. The colormap used is `parula`, and similar colors correspond to similar values.

Theorem 6. *Let $\hat{\boldsymbol{h}}$ and $\boldsymbol{h}^\star$ be as given in Eq. (1) and Eq. (2) respectively for some i.i.d. realizations of the stochastic loss $f \sim \mathcal{D}$. If the distribution $\mathcal{D}$ is supported on γ-strongly convex, L-Lipschitz, and twice differentiable loss functions, then $\sqrt{m}\left(\hat{\boldsymbol{h}} - \boldsymbol{h}^\star\right) \to \mathcal{N}\big(0,\, \Sigma\big)$ as $m \to \infty$, where Σ is equal to:*

$$\left(\mathbb{E}_{f\sim\mathcal{D}}[\nabla^2 f(\boldsymbol{h}^\star)]\right)^{-1} \cdot Cov(\nabla f(\boldsymbol{h}^\star)) \cdot \left(\mathbb{E}_{f\sim\mathcal{D}}[\nabla^2 f(\boldsymbol{h}^\star)]\right)^{-1}$$

Proof. Here is the outline of the proof. First, we use strong convexity, the first-order optimality condition of $\mathbf{h}^\star$, and Theorem 6 in (Shalev-Shwartz et al., 2009) to show that $||\hat{\mathbf{h}} - \mathbf{h}^\star||_2 = O_p(1/\sqrt{m})$. This allows us to estimate the error term of the second-order Taylor expansion. Next, we use the *continuous mapping theorem* (Mann & Wald, 1943) and the *matrix inversion lemma* (Hager, 1989) to show that:

$$\hat{\mathbf{h}} - \mathbf{h}^\star = \frac{1}{m}\left(\mathbb{E}_{f\sim\mathcal{D}}[\nabla^2 f(\mathbf{h}^\star)]\right)^{-1} \sum_{i=1}^{m} \nabla f_i(\mathbf{h}^\star) + o_p\Big(\frac{1}{m}\Big)$$

Because $(1/m)\sum_{i=1}^m \nabla f_i(\mathbf{h}^\star)$ is an average of i.i.d. realizations, applying the classical central limit theorem, the first-order optimality condition on $\mathbf{h}^\star$, and the affine property of the multivariate Gaussian distribution will yield the desired result. □

Remark 3. *The CLT in Theorem 6 is an asymptotic result; it holds for a sufficiently large m. When the gradient $\nabla f_i(w)$ is, additionally, R-Lipschitz, it can be shown using the Matrix-Hoeffding bound (Tropp, 2012) that the normal approximation is valid as long as $m \gg L^2/\gamma^2 + RL\log d$.*

To verify Theorem 6 empirically, we have implemented the ℓ_2-regularized logistic regression on a mixture of two Gaussians with different means and covariance matrices, one corresponding to the positive class and one corresponding to the negative class. Both the actual and predicted covariance matrices of the empirical risk minimizer $\hat{\mathbf{h}}$, whose randomness is derived from the randomness of the training sample, are depicted in Fig. 1. As shown in the figure, the actual covariance matrix matches with the one predicted by Theorem 6.

Theorem 6 shows that the sample complexity of stochastic convex optimization depends on the curvature of the risk $\mathbb{E}_{f\sim\mathcal{D}}[f(h)]$ at its minimizer $\mathbf{h}^\star$. Next, we show that the dependence of ERM on this curvature is, in fact, *optimal*.

Proposition 2. *There exists a stochastic loss $f:\mathbb{R}^d\to\mathbb{R}$ that satisfies the conditions of Theorem 6 and a distribution $\mathcal{D}$ such that if $\tilde{\boldsymbol{h}}$ is an unbiased estimator of $\boldsymbol{h}^\star = \arg\min_{h\in\mathbb{R}^d}\{\mathbb{E}_{f\sim\mathcal{D}}[f(h)]$, then the covariance of the estimator is bounded from below (in the positive semidefinite sense) by:*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[(\tilde{\boldsymbol{h}}-\boldsymbol{h}^\star)\cdot(\tilde{\boldsymbol{h}}-\boldsymbol{h}^\star)] \succeq (1/m)\,\Sigma,$$

where Σ is the covariance of ERM given by Theorem 6.

5.2. The Excess Risk

Theorem 6 states that when the loss f is strongly convex, the empirical risk minimizer $\hat{\mathbf{h}}$ is normally distributed around the population risk minimizer $\mathbf{h}^\star$ with a vanishing $O(1/m)$ covariance. This justifies learning via the empirical risk minimization rule.

We use this central limit theorem, next, to derive the asymptotic behavior of the excess risk $F(\hat{\mathbf{h}})-F(\mathbf{h}^\star)$, where $F(h)$ is defined in Eq. (2). In particular, we would like to determine the impact of *regularization* on the asymptotic behavior of the excess risk. We write:

$$f(h)=\frac{\lambda}{2}||\,h\,||_2^2+g(h), \tag{5}$$

where $\lambda\geq 0$, g is a κ-strongly-convex function for some $\kappa\geq 0$, and $\lambda+\kappa=\gamma>0$. Because f satisfies the conditions of Theorem 6, we know that $\hat{\mathbf{h}}\to\mathbf{h}^\star$ as $m\to\infty$. Hence, we can take the Taylor expansion of $F(h)$ around $\mathbf{h}^\star$ and write:

$$F(\hat{\mathbf{h}})-F(\mathbf{h}^\star)=\frac{1}{2}(\hat{\mathbf{h}}-\mathbf{h}^\star)^T\,\nabla^2F(\mathbf{h}^\star)\,(\hat{\mathbf{h}}-\mathbf{h}^\star)+o_p\Big(\frac{1}{m}\Big),$$

which holds since $\nabla F(\mathbf{h}^\star)=0$ and $||\,\hat{\mathbf{h}}-\mathbf{h}^\star\,||_2 = O_p(1/\sqrt{m})$. Using Theorem 6 and simplifying yields:

$$F(\hat{\mathbf{h}})-F(\mathbf{h}^\star)=\frac{1}{2m}\mathbf{x}^TU^T\,\nabla^2F(\mathbf{h}^\star)\,U\,\mathbf{x}+o_p\Big(\frac{1}{m}\Big), \tag{6}$$

where $\Sigma=UU^T$, $\mathbf{x}=U^{-1}(\hat{\mathbf{h}}-\mathbf{h}^\star)$, and Σ is given by Theorem 6.

By substituting Eq. (5), we arrive at the following corollary.

Corollary 1. *Let f be decomposed into the sum of a regularization term and an unregularized loss, as given by Eq. (5). Write $G(h)$ to denote the unregularized risk:*

$$G(h)=\mathbb{E}_{f\sim\mathcal{D}}[g(h)]=\mathbb{E}_{f\sim\mathcal{D}}[f(h)-(\lambda/2)||\,h\,||_2^2].$$

Then, under the conditions of Theorem 6, $G(\hat{\boldsymbol{h}})-G(\boldsymbol{h}^\star)$ converges in distribution (over the random choice of the training sample) to $\frac{\lambda}{\sqrt{m}}c^T\boldsymbol{x}+\frac{\boldsymbol{x}^TD\boldsymbol{x}}{2m}$ for some $c\in\mathbb{R}^d$ and some $D\succeq 0$ that are both independent of m, where $\boldsymbol{x}\sim\mathcal{N}(0,\,I_d)$ is a standard multivariate Gaussian random variable.

We can interpret Corollary 1 in sharp $\Theta(\cdot)$ terms. In the following remarks, both "expectation" and "probability" are taken over the random choice of the training sample:

- In the absence of regularization, the ERM learning rule of strongly convex loss enjoys a fast learning rate of $\Theta(1/m)$ *both* in expectation and in probability.
- When regularization is used, the unregularized risk converges to its infinite-data limit with a $\Theta(1/m)$ rate *only* in expectation. By contrast, it has a $\Theta(1/\sqrt{m})$ rate of convergence in probability.

Corollary 1 unifies previous results, such as those reported in (Shalev-Shwartz & Ben-David, 2014; Sridharan et al., 2009; Shalev-Shwartz et al., 2009). It also illustrates why a fast $\Theta(1/m)$ learning rate in expectation, such as those reported for Exp-Concave minimization (Koren & Levy, 2015), do not necessarily correspond to fast learning rates in practice because the same excess risk can be $\Theta(1/\sqrt{m})$ in probability even if it is $\Theta(1/m)$ in expectation.

To validate these claims experimentally, we have implemented a linear regression problem by minimizing the *Huber* loss of the residual. The mean and the standard deviation of $G(\hat{\mathbf{h}})-G(\mathbf{h}^\star)$ are plotted in Fig 2 against the sample size m. As shown in Fig 2 (a, b), we have a fast $\Theta(1/m)$ convergence both in expectation and in probability when no regularization is used. However, when regularization is used ($\gamma=0.01$ in this experiment), then the expectation of $G(\hat{\mathbf{h}})-G(\mathbf{h}^\star)$ is $O(1/m)$ but its standard deviation is $O(1/\sqrt{m})$ as shown in Fig 2 (c, d). These results are in agreement with Corollary 1.

Corollary 1 provides the excess risk between the regularized empirical risk minimizer $\hat{\mathbf{h}}$ and its *infinite data limit* $\mathbf{h}^\star$. We can use the same corollary to obtain a bound on the excess risk $G(\hat{\mathbf{h}})-\inf_{h\in\mathbb{R}^d}\{G(h)\}$ using *oracle* inequalities in a manner that is similar to (Sridharan et al., 2009; Shalev-Shwartz & Ben-David, 2014). Moreover, the regularization magnitude λ may be optimized to minimize this excess risk. Because this is quite similar to previous works, the reader is referred to (Sridharan et al., 2009; Shalev-Shwartz & Ben-David, 2014) for further details.

5.3. Uniform Generalization Bound

Theorem 6 can be a viable tool in uncertainty quantification and hypothesis testing. In this section, however, we focus on its implication for the uniform generalization risk. We begin with the following "conditional" version of the central limit theorem.

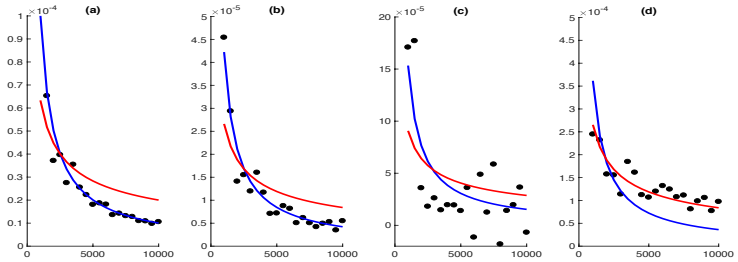

Figure 2. This figure plots the quantity $E(m) = G(\hat{\mathbf{h}}) - G(\mathbf{h}^\star)$ vs. the sample size m for a regression problem, where we minimize the Huber loss of the residual. Figures (a) and (b) display the expectation and standard deviation of $E(m)$, respectively, when $\gamma = 0$ (no regularization). By contrast, Figures (c) and (d) display the same quantities when $\gamma = 0.01$. The blue curves are the best fitted curves assuming a fast $\Theta(1/m)$ convergence while the red curves are the best fitted curves assuming a standard $\Theta(1/\sqrt{m})$ convergence.

Proposition 3. *Let $\hat{\boldsymbol{f}} \sim \mathcal{D}$ be a fixed instance of the stochastic loss, and let the training sample be $S = \{\hat{\boldsymbol{f}}\} \cup \{f_2, f_3, \ldots, f_m\}$ with $f_i \sim \mathcal{D}$ drawn i.i.d. and independently of $\hat{\boldsymbol{f}}$. Let $\hat{\boldsymbol{h}} \in \mathbb{R}^d$ be the empirical risk minimizer given by Eq. (1). Then, under the conditions of Theorem 6:*

$$p(\hat{\boldsymbol{h}} \,|\hat{\boldsymbol{f}}) \to \mathcal{N}(\tilde{\mu}, \frac{1}{m-1}\Sigma),$$

where:

$$\tilde{\mu} = \arg\min_{h \in \mathbb{R}^d} \Big\{ \mathbb{E}_{f \sim \mathcal{D}} \big[f(h) + \frac{1}{m-1} \hat{\boldsymbol{f}}(h) \big] \Big\}$$

Here, Σ is the covariance matrix given by Theorem 6.

Proposition 3 states, in other words, that a single realization of the stochastic loss shifts the expectation of the empirical risk minimizer $\hat{\mathbf{h}}$ and rescales its covariance. Both Theorem 6 and Proposition 3 imply that the capacity of the ERM learning rule of stochastic, strongly-convex loss classes satisfies $C(\mathcal{L}) \to 0$ as $m \to \infty$. We establish a more useful quantitative result in the following theorem.

Theorem 7. *Suppose that normality holds, where the empirical risk minimizer $\hat{\boldsymbol{h}}$ has the probability density $p(\hat{\boldsymbol{h}}) = \mathcal{N}(\mu, \Sigma/m)$ with μ and Σ given by Theorem 6, and for any given single realization of the stochastic loss $\hat{\boldsymbol{f}}$, we also have $p(\hat{\boldsymbol{h}} \,|\hat{\boldsymbol{f}}) = \mathcal{N}(\tilde{\mu}, \Sigma/(m-1))$ as given by Proposition 3. Then:*

$$I(\hat{\boldsymbol{h}}; \hat{\boldsymbol{f}}) \leq \frac{d}{m} + o\Big(\frac{1}{m}\Big), \tag{7}$$

where $I(\boldsymbol{x};\boldsymbol{y})$ is the Shannon mutual information between the random variables $\boldsymbol{x}$ and $\boldsymbol{y}$.

Corollary 2. *Under the conditions of Theorem 7, the uniform generalization risk satisfies:*

$$\mathcal{J}(\hat{\boldsymbol{h}}; \hat{\boldsymbol{f}}) \leq \sqrt{\frac{d}{2m}} + o\Big(\frac{1}{\sqrt{m}}\Big) \tag{8}$$

Proof. This follows immediately from Theorem 7 and Pinsker's inequality (Cover & Thomas, 1991). □

6. Applications

6.1. Large-Scale Stochastic Convex Optimization

Theorem 6 states that the empirical risk minimizer of strongly convex stochastic loss is normally distributed around the population risk minimizer $\mathbf{h}^\star$ with covariance $(1/m)\Sigma$. This justifies learning via the empirical risk minimization procedure. However, the true goal behind stochastic convex optimization in the machine learning setting is not to compute the empirical risk minimizer $\hat{\mathbf{h}}$ *per se* but to estimate $\mathbf{h}^\star$. The empirical risk minimizer provides such an estimate. However, a different estimator can be constructed, which is as effective as the empirical risk minimizer $\hat{\mathbf{h}}$.

Theorem 8. *Under the conditions of Theorem 6, let $S = \{f_1, \ldots, f_m\}$ be m i.i.d. realizations of the stochastic loss $\boldsymbol{f} \sim \mathcal{D}$ and fix a positive integer $K \geq 1$. Let $\cup_{j=1}^{K} S_j$ be a partitioning of S into K subsets of equal size and define $\hat{\boldsymbol{h}}_j$ to be the empirical risk minimizer for S_j only. Then, $\tilde{\boldsymbol{h}} = \frac{1}{K}\sum_{j=1}^{K} \hat{\boldsymbol{h}}_j$ is asymptotically normally distributed around the population risk minimizer $\boldsymbol{h}^\star$ with covariance $(1/m)\Sigma$, where Σ is given by Theorem 6.*

Proof. By Theorem 6, every $\hat{\mathbf{h}}_j$ is asymptotically normally distributed around $\mathbf{h}^\star$ with covariance $(K/m)\Sigma$. Hence, the average of those hypotheses is asymptotically normally distributed around $\mathbf{h}^\star$ with covariance $(1/m)\Sigma$. □

Theorem 8 shows that in the machine learning setting, one can trivially scale the empirical risk minimization procedure to big data using a naïve parallelization algorithm. Indeed, the estimator $\tilde{\mathbf{h}}$ described in the theorem is not a minimizer to the empirical risk but it is as effective as the empirical risk minimizer in estimating the population risk minimizer $\mathbf{h}^\star$. Hence, both $\hat{\mathbf{h}}$ and $\tilde{\mathbf{h}}$ enjoy the same performance guarantee.

In the literature, methods for scaling machine learning algorithms to big data using distributed algorithms do not always distinguish between optimization as it is used in the traditional setting vs. the optimization that is used in

Algorithm	Test Error	Time
ERM ($m = 50,000$)	$14.8 \pm 0.3\%$	2.93s
ERM ($m = 1,000$)	$15.3 \pm 0.5\%$	0.03s
Parallelized ($K = 50, m = 1,000$)	$14.8 \pm 0.1\%$	0.03s

Table 1. In this experiment, we run the ℓ_2 regularized logistic regression on the MiniBooNE Particle Identification problem (Blake & Merz, 1998). The first row corresponds to running the algorithm on 50,000 training examples, the second row is for 1,000 examples, while the last row is for the parallelization algorithm of Theorem 8, which splits 50,000 examples into 50 separate smaller problems.

the machine learning setting despite several calls that highlighted such a subtle distinction (Bousquet & Bottou, 2008; Shalev-Shwartz et al., 2012). One popular, and often-cited, procedure that does not make such a distinction is the alternating direction method of the multiplier (ADMM) (Boyd et al., 2011). The ADMM procedure produces a distributed algorithm with message passing for minimizing the *empirical* risk by reformulating stochastic convex optimization into a "global consensus problem". However, the empirical risk is merely a proxy for the true risk that one seeks to minimize. Theorem 8, by contrast, presents a much simpler algorithm that achieves the desired goal. Table 1 validates this claim experimentally.

Remark 4. *The theoretical guarantee of Theorem 8 is established for stochastic* convex *optimization only. When the stochastic loss is non-convex, such as the case in neural networks, then the parallelization method is likely to fail. Intuitively, the average of good solutions is not necessarily a good solution itself when the loss is non-convex.*

6.2. Information Criterion for Model Selection

Theorem 7 shows that under the assumption of normality, which is justified by the central limit theorems (Theorem 6 and Proposition 3), the uniform generalization risk of the ERM learning rule of stochastic, strongly-convex loss is asymptotically bounded by $\sqrt{d/(2m)}$. Remarkably, this bound does not depend on the strong-convexity parameter γ, nor on any other properties of the stochastic loss f that is optimized during the training stage. In addition, because it is a *uniform* generalization bound, it also holds for *any* bounded loss function $l : \mathbb{R}^d \times \mathcal{Z} \to [0, 1]$. Hence, it can serve as a simple information criterion for model selection. As a result, a learning algorithm should aim at minimizing the *Uniform Information Criterion* (UIC) given by:

$$UIC = \mathbb{E}_{\mathbf{z}\sim S}[l(\hat{\mathbf{h}}, \mathbf{z})] + \sqrt{\frac{d}{2m}}, \tag{9}$$

where d is the number of learned parameters.

The UIC above is analogous to the Akaike information criteria (AIC) (Akaike, 1998; Bishop, 2006) which seeks to minimize $\mathbb{E}_{\mathbf{z}\sim S}[f(\hat{\mathbf{h}}, \mathbf{z})] + d/m$. It is also similar to Schwarz's Bayesian information criterion (BIC) (Schwarz, 1978), which seeks to minimize $\mathbb{E}_{\mathbf{z}\sim S}[f(\hat{\mathbf{h}}, \mathbf{z})] + (\log m/2)\, d/m$. In all three criteria, the generalization risk is estimated by the number of learned parameters. However, the penalty for the number of model parameters is proportional to d in both AIC and BIC, which, by the discretization trick (Shalev-Shwartz & Ben-David, 2014), is known to be pessimistic. Indeed, the discretization trick suggests that the generalization risk is proportional to $O(\sqrt{d/m})$, which agrees with the UIC. Experimentally, the UIC succeeds when both AIC and BIC fail, as demonstrated in Fig. 3[2].

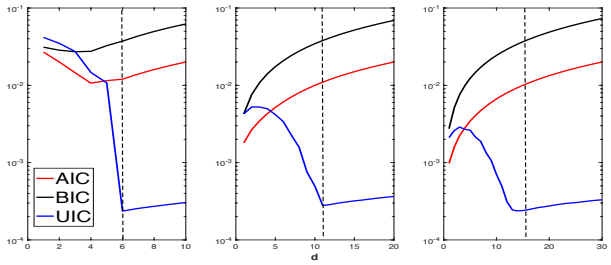

Figure 3. In this experiment, we have a least-squares polynomial regression where $\mathcal{X} = [-1, +1]$, $\mathbf{y} = f(\mathbf{x}) + \epsilon$, f is a polynomial of degree $d^\star$, and ϵ is i.i.d. Gaussian noise. The model selection problem seeks to determine the optimal polynomial degree $d^\star$. In each of the three figures above, a value of $d^\star$ is selected, which is marked by the dashed vertical line. Then, $100\, d^\star$ training examples are provided. Each curve plots an information criterion against d. Ideally, $d^\star$ minimizes the corresponding curve. As shown above, only UIC succeeds in estimating $d^\star$ in all problems. In this experiment, the empirical risk is normalized in the range $[0,\ 1]$ by dividing it by the maximum observed loss in the training sample.

7. Conclusion

In this paper, we derive bounds on the mutual information of the ERM rule for both 0-1 and strongly-convex loss classes. We prove that under the Axiom of Choice, the existence of an ERM rule with a vanishing mutual information is equivalent to the assertion that the loss class has a finite VC dimension, thus bridging information theory with statistical learning theory. Similarly, an asymptotic bound on the mutual information is established for strongly-convex loss classes in terms of the number of model parameters. The latter result uses a central limit theorem that we derive in this paper. After that, we prove that the ERM learning rule for strongly-convex loss classes can be trivially scaled to big data. Finally, we propose a simple information criterion for model selection and demonstrate experimentally that it outperforms previous works.

[2]The MATLAB codes that generate Table 1 and Fig. 3 are provided in the supplementary materials.

References

Abu-Mostafa, Yaser S, Magdon-Ismail, Malik, and Lin, Hsuan-Tien. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

Akaike, Hirotogu. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer, 1998.

Alabdulmohsin, Ibrahim M. Algorithmic stability and uniform generalization. In *NIPS*, pp. 19–27, 2015.

Alabdulmohsin, Ibrahim M. An information theoretic route from generalization in expectation to generalization in probability. In *AISTATS*, 2017.

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

Bishop, Christopher M.. *Pattern recognition and machine learning*. Springer, 2006.

Blake, C. L. and Merz, C. J. UCI repository of machine learning databases, 1998.

Bousquet, Olivier and Bottou, Léon. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pp. 161–168, 2008.

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, and Eckstein, Jonathan. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. Wiley & Sons, 1991.

Csiszár, Imre. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2:191–213, 1972.

Csiszár, Imre. Axiomatic characterizations of information measures. *Entropy*, 10:261–273, 2008.

Cummings, Rachel, Ligett, Katrina, Nissim, Kobbi, Roth, Aaron, and Wu, Zhiwei Steven. Adaptive learning with robust generalization guarantees. In *COLT*, pp. 23–26, 2016.

Dwork, Cynthia, Feldman, Vitaly, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Roth, Aaron. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pp. 117–126, 2015.

Elkan, Charles. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. Agnostic learning of monomials by halfspaces is hard. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pp. 385–394, Oct 2009. doi: 10.1109/FOCS.2009.26.

Feldman, Vitaly. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *NIPS*, pp. 3576–3584, 2016.

Hager, William W. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

Janson, Svante. Probability asymptotics: notes on notation. *arXiv preprint arXiv:1108.3924*, 2011.

Kolmogorov, Andreĭ Nikolaevich and Fomin. *Introductory real analysis*. Dover Publication, Inc., 1970.

Koren, Tomer and Levy, Kfir. Fast rates for exp-concave empirical risk minimization. In *NIPS*, pp. 1477–1485, 2015.

Kull, Meelis and Flach, Peter. Novel decompositions of proper scoring rules for classification: score adjustment as precursor to calibration. In *ECML-PKDD*, pp. 68–85. Springer, 2015.

Mandt, Stephan, Hoffman, Matthew D., and Blei, David M. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML, pp. 354–363. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045429.

Mann, H. B. and Wald, A. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943. ISSN 00034851. URL http://www.jstor.org/stable/2235800.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30 (4):838–855, July 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL http://dx.doi.org/10.1137/0330046.

Raginsky, Maxim, Rakhlin, Alexander, Tsao, Matthew, Wu, Yihong, and Xu, Aolin. Information-theoretic analysis of stability and bias of learning algorithms. In *Information Theory Workshop (ITW), 2016 IEEE*, pp. 26–30. IEEE, 2016.

Russo, Daniel and Zou, James. Controlling bias in adaptive data analysis using information theory. In *Proceedings*

of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), 2016.

Schwarz, Gideon. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Shalev-Shwartz, Shai, Shamir, Ohad, Srebro, Nathan, and Sridharan, Karthik. Stochastic convex optimization. In *COLT*, 2009.

Shalev-Shwartz, Shai, Shamir, Ohad, and Tromer, Eran. Using more data to speed-up training time. In *AISTATS*, pp. 1019–1027, 2012.

Sridharan, Karthik, Shalev-Shwartz, Shai, and Srebro, Nathan. Fast rates for regularized objectives. In *NIPS*, pp. 1545–1552, 2009.

Tao, Terence. *Topics in random matrix theory*. American Mathematical Society, 2012.

Tropp, Joel A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Vapnik, Vladimir N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5): 988–999, 1999.