Information Theoretic Guarantees for Empirical Risk Minimization

Xuechao Wang

May 8, 2019

Abstract

In this project, we investigated the tight relations between mutual information and the empirical risk minimization (ERM) algorithm. First, we introduce a precise information-theoretic characterization for the uniform generalization risk of a learning algorithm. Then it is proved that under the Axiom of Choice, an ERM learning rule that has a vanishing learning capacity if and only if the 0-1 loss class has a finite VC dimension, while the ERM of strongly-convex stochastic loss generalizes uniformly in expectation as well. Finally, an application to large-scale convex optimization is discussed.

1 Backgrounds

In this section, we give a brief introduction to empirical risk minimization (ERM) and the concept of uniform generalization for a learning algorithm.

1.1 ERM

Due to the simplicity,generality, and statistical efficiency, learning via ERM of stochastic loss has been widely applied in many learning problems. Given a hypothesis space $\mathcal{F}$, a domain $\mathcal{Z}$, and a loss function on the product space $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$, the ERM learning rule selects the hypothesis $\hat{\mathbf{f}}$ that minimizes the empirical risk:

$$\hat{\mathbf{f}} = \arg\min_{f\in\mathcal{F}} \left\{ L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell\left(f, z_i\right) \right\},$$

where $S = (z_1, \ldots, z_m) \in \mathcal{Z}^m$ and each z_i is independently drawn according to some unknown probability distribution $\mathcal{P}$. By contrast, the true risk minimizer is denoted as $\mathbf{f}^\star$ and given by

$$\mathbf{f}^\star = \arg\min_{f\in\mathcal{F}} \{L(h) = \mathbb{E}_{z\sim\mathcal{P}}[\ell(f, z)]\}.$$

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

The Fundamental Theorem of Statistical Learning states that a hypothesis space $\mathcal{F}$ is agnostic PAC-learnable via ERM if and only if it is PAC-learnable at all, and that this occurs if and only if $\mathcal{F}$ has a finite VC dimension[1].

1.2 Uniform Generalization

Suppose we have a learning algorithm $\mathcal{A} : \mathcal{Z}^m \to \mathcal{F}$, which selects a hypothesis $\mathbf{f} \in \mathcal{F}$ according to a training sample $S \in \mathcal{Z}^m$. Then the generalization risk of $\mathcal{A}$ w.r.t. some bounded loss function $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$ can be defined by

$$R_{gen}(\mathcal{A}) = \mathbb{E}_{S\sim\mathcal{P}^m,\mathbf{f}} \left[L(\mathbf{f}) - L_S(\mathbf{f})\right],$$

where the expectation is taken over the random choice of the sample and the internal randomness in the learning algorithm. Theorem 10.3 in the lecture notes implies that stability on average, or equivalently, generalization on average,is sufficient for an ERM algorithm to be consistent.

Further, we can give the definition of uniform generalization of a learning algorithm $\mathcal{A}$ as follow.

Definition 1.1. (Uniform Generalization) A learning algorithm $\mathcal{A} : \mathcal{Z}^m \to \mathcal{F}$ generalizes uniformly with rate $\epsilon \geq 0$ if for all bounded parametric losses $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$, we have $|R_{gen}(\mathcal{A})| \leq \epsilon$.

2 Main Theorem

The main theorem in the paper[2] provides a precise information theoretic characterization for the uniform generalization risk.

Theorem 2.1. *Given a fixed $0 \leq \epsilon \leq 1$ and a learning algorithm $\mathcal{A} : \mathcal{Z}^m \to \mathcal{F}$ that selects a hypothesis $\mathbf{f} \in \mathcal{F}$ according to a training sample $S \in \mathcal{Z}^m$, where $z_i \sim \mathcal{P}$ are i.i.d., then $\mathcal{A}$ generalizes uniformly with rate ϵ if and only if $\mathcal{J}(\mathbf{f}; \hat{z}) \leq \epsilon$, where $\hat{z} \sim S$ is a single random training example, $\mathcal{J}(\mathbf{x}; \mathbf{y}) = \|p(\mathbf{x})p(\mathbf{y}), p(\mathbf{x}, \mathbf{y})\|_{\mathcal{T}}$, and $\|q_1, q_2\|_{\mathcal{T}}$ is the total variation distance between the probability measures q_1 and q_2.*

Informally, we will call $\mathcal{J}(\mathbf{x}; \mathbf{y})$ the "variational information" between the random variables $\mathbf{x}$ and $\mathbf{y}$.

Consider the case with a finite hypothesis space $|\mathcal{F}| < \infty$, by a classical argument with the union bound[1], the uniform generalization risk in this case is $\tilde{O}(\sqrt{\log |\mathcal{F}|/m})$. Meanwhile, we can use some information theoretic inequalities to derive the same bound for the variational information $\mathcal{J}(\mathbf{f}; \hat{z})$:

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

where $I(\mathbf{x}, \mathbf{y})$ is the mutual information. Here, the first inequality is called Pinsker's inequality in information theory, the second inequality holds because z_i's are i.i.d., and the last inequality follows because the mutual information is bounded by the entropy and the entropy is maximized by uniform distribution.

Definition 2.2. (Capacity) The capacity of a learning algorithm $\mathcal{A}$ is defined by

$$C(\mathcal{A}) = \sup_{p(z)} \left\{\mathcal{J}(\mathbf{f}; \hat{z}) = \mathbb{E}_{\hat{z}\sim p(z)} \|p(\mathbf{f}), p(\mathbf{f}|\hat{z})\|_{\mathcal{T}}\right\},$$

where the supremum is taken over all possible distributions.

Therefore, the generalization risk of $\mathcal{A}$ is bounded by $C(\mathcal{A})$ for any probability distribution and any bounded loss function. In the next section, we will give several bounds on the capacity of ERM algorithms.

3 Bounds on Capacity

3.1 ERM of 0-1 Loss Classes

First, we will recall a fundamental result in modern set theory.

Definition 3.1. (Well-ordered) A non-empty set Q is said to be well-ordered if Q is endowed with a total order $\preceq$ such that every non-empty subset of Q contains a least element. And $\preceq$ is called a well-ordering on Q.

A fundamental theorem proved by Ernst Zermelo in 1904 states that under the Axiom of Choice, every non-empty subset can be well-ordered. Based on this fundamental result, the following upper bounds on capacity of ERM algorithm can be proved[3].

Theorem 3.2. *Given a well-ordered hypothesis space $\mathcal{F}$ endowed with $\preceq$, a domain $\mathcal{Z}$, and a 0-1 loss $\ell : \mathcal{F} \times \mathcal{Z} \to \{0, 1\}$. Let $\mathcal{A} : \mathcal{Z}^m \to \mathcal{F}$ be the learning rule that outputs the "least" empirical risk minimizer to the training sample $S \in \mathcal{Z}^m$ according to $\preceq$. Then, $C(\mathcal{A}) \to 0$ as $m \to \infty$ if $\mathcal{F}$ has a finite VC dimension d. In particular:*

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

Consider the standard binary classification setting, a lower bound for all ERM rules can be proved.

Theorem 3.3. *In any fixed domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, let the hypothesis space $\mathcal{F}$ be a concept class on $\mathcal{X}$ and let $\ell(f, x, y) = \mathbb{I}\{y \neq f(x)\}$ be the 0-1 loss. Then, any ERM learning rule $\mathcal{A}$ w.r.t. ℓ has a learning capacity $C(\mathcal{A})$ 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 $\mathcal{F}$.*

From the above theorems, we can directly obtain a characterization of the VC dimension of concept classes in term of information theory, which establishes tight relations between uniform generalization and the ERM algorithm. This will allow us to bridge information theory with statistical learning theory.

Theorem 3.4. *Given a fixed domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, let the hypothesis space $\mathcal{F}$ be a concept class on $\mathcal{X}$ and let $\ell(f, x, y) = \mathbb{I}\{y \neq f(x)\}$ be the 0-1 loss. Let m be the training sample size. Then, the following statements are equivalent under the Axiom of Choice:*

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

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

Theorem 3.4 implies that for an ERM learning rule of 0-1 loss classes with finite VC dimensions, no adversary can post-process the hypothesis and causes over-fitting to occur. Equivalently, the empirical performance of the ERM algorithm on the sample S is a faithful approximation to its true risk, regardless of how that performance is measured.

3.2 ERM of Strongly-Convex Loss Classes

In this section, the ERM learning rule for strongly-convex loss classes is analyzed. We now further assume that $\ell(f,z)$ is γ-strongly convex, L-Lipschitz, and twice-differentiable for each z. Moreover, the hypothesis space is $\mathbb{R}^d$ for some finite $d < \infty$.

First, a central limit theorem (CLT) can be established for the ERM learning rule in this setting. In this section, we will simplify the notation by writing $\ell_i(f) = \ell(f, z_i)$.

Theorem 3.5. *If the distribution $\mathcal{P}$ is supported on γ-strongly convex, L-Lipschitz, and twice-differentiable loss functions, then $\sqrt{m}\left(\hat{\mathbf{f}} - \mathbf{f}^\star\right) \to \mathcal{N}(0, \Sigma)$ as $m \to \infty$, where*

$$\Sigma = \left(\mathbb{E}_{\ell\sim\mathcal{P}}\left[\nabla^2 \ell\left(\mathbf{f}^\star\right)\right]\right)^{-1} \cdot \mathrm{Cov}\left(\nabla \ell\left(\mathbf{f}^\star\right)\right) \cdot \left(\mathbb{E}_{\ell\sim\mathcal{P}}\left[\nabla^2 \ell\left(\mathbf{f}^\star\right)\right]\right)^{-1}$$

Theorem 3.5 shows that the sample complexity of stochastic convex optimization depends on the curvature of the risk $\mathbb{E}_{\ell\sim\mathcal{P}}[\ell(f)]$ at its minimizer $\mathbf{f}^\star$. Further, we can establish the following "conditional" version of the central limit theorem.

Theorem 3.6. *Let $\hat{\ell} \sim \mathcal{P}$ be a fixed instance of the stochastic loss, and let the training sample be $S = \{\hat{\ell}\} \cup \{\ell_2, \ell_3, \ldots, \ell_m\}$ with $\ell_i \sim \mathcal{P}$ drawn i.i.d. and independently of $\hat{\ell}$. Then, under the conditions of Theorem 3.5, we have*

$$p(\hat{\mathbf{f}}|\hat{\ell}) \to \mathcal{N}\left(\tilde{\mu}, \frac{1}{m-1}\Sigma\right),$$

where $\tilde{\mu} = \arg\min_{f\in\mathbb{R}^d}\left\{\mathbb{E}_{\ell\sim\mathcal{P}}\left[\ell(f) + \frac{1}{m-1}\hat{\ell}(f)\right]\right\}$ and Σ is the covariance matrix given by Theorem 3.5.

Theorem 3.6 implies that a single realization of the stochastic loss shifts the expectation of the empirical risk minimizer $\hat{\mathbf{f}}$ and rescales its covariance. Using Theorem 3.5 and Theorem 3.6, an upper bound on the capacity of ERM algorithms for strongly-convex loss classes can be derived.

Theorem 3.7. *Suppose that normality as given by Theorem 3.5 and Theorem 3.6 holds for the ERM learning rule $\mathcal{A}$, Then the capacity of the ERM algorithm satisfies*

$$C(\mathcal{A}) \le \sqrt{\frac{d}{2m}} + o\left(\frac{1}{\sqrt{m}}\right).$$

Theorem 3.7 implies that the capacity of the ERM learning rule of stochastic, strongly-convex loss classes satisfies $C(\mathcal{A}) \to 0$ as $m \to \infty$.

4 Application to Large-Scale Optimization

During the last decade, the data sizes have grown faster than the speed of processors. In this context, the capabilities of statistical machine learning methods is limited by the computing time rather than the sample size. The true goal behind stochastic convex optimization in the machine learning setting is to estimate $\mathbf{f}^\star$. The ERM rule provides such an estimate $\hat{\mathbf{f}}$. However, a different estimator can be constructed, which is as effective as the empirical risk minimizer $\hat{\mathbf{f}}$.

Theorem 4.1. *Under the conditions of Theorem 3.5, let* $S = \{\ell_1, \ldots, \ell_m\}$ *be* m *i.i.d. realizations of the stochastic loss* $\ell \sim \mathcal{P}$ *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* $\hat{\mathbf{f}}_j$ *to be the empirical risk minimizer for* S_j *only. Then,* $\tilde{\mathbf{f}} = \frac{1}{K}\sum_{j=1}^{K} \hat{\mathbf{f}}_j$ *is asymptotically normally distributed around* $\mathbf{f}^{\star}$ *with covariance* $(1/m)\Sigma$, *where* Σ *is given by Theorem 3.5.*

Theorem 4.1 implies that in the machine learning setting, one can trivially scale the empirical risk minimization procedure to big data using a naive parallelization algorithm. The alternating direction method of the multiplier (ADMM) is a popular procedure in distributed learning, which produces a distributed algorithm with message passing for minimizing the empirical risk by reformulating stochastic convex optimization problem into a "global consensus problem". However, theorem 4.1, by contrast, presents a much simpler algorithm that achieves the desired goal.

5 Conclusion

In conclusion, we investigate the tight relations between variational information and uniform generalization risk of the ERM algorithm. Several bounds have been derived on capacity of the ERM rule for both 0-1 and strongly-convex loss classes. It is proved that under the Axiom of Choice, an ERM learning rule that has a vanishing learning capacity if and only if the 0-1 loss class has a finite VC dimension, while the ERM of strongly-convex stochastic loss generalizes uniformly in expectation as well. After that, it is proved that the ERM learning rule for strongly-convex loss classes can be trivially scaled to big data.

References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[2] I. M. Alabdulmohsin, "Algorithmic stability and uniform generalization," in *Advances in Neural Information Processing Systems*, pp. 19–27, 2015.

[3] I. M. Alabdulmohsin, "Information theoretic guarantees for empirical risk minimization with applications to model selection and large-scale optimization," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 149–158, 2018.