# An information-theoretic route from generalization in expectation to generalization in probability

**Ibrahim Alabdulmohsin**
Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

## Abstract

One fundamental goal in any learning algorithm is to mitigate its risk for overfitting. Mathematically, this requires that the learning algorithm enjoys a small generalization risk, which is defined either in expectation or in probability. Both types of generalization are commonly used in the literature. For instance, generalization in expectation has been used to analyze algorithms, such as ridge regression and SGD, whereas generalization in probability is used in the VC theory, among others. Recently, a third notion of generalization has been studied, called *uniform generalization*, which requires that the generalization risk vanishes uniformly in expectation across all bounded parametric losses. It has been shown that uniform generalization is, in fact, *equivalent* to an information-theoretic stability constraint, and that it recovers classical results in learning theory. It is achievable under various settings, such as sample compression schemes, finite hypothesis spaces, finite domains, and differential privacy. However, the relationship between uniform generalization and concentration remained unknown. In this paper, we answer this question by proving that, while a generalization in expectation does not imply a generalization in probability, a *uniform* generalization in expectation does imply concentration. We establish a chain rule for the uniform generalization risk of the composition of hypotheses and use it to derive a large deviation bound. Finally, we prove that the bound is tight.

## 1 INTRODUCTION

One of the central questions in statistical learning theory is to establish the conditions for generalization from a finite collection of observations to the future. Mathematically, this is formalized by bounding the difference between the empirical and the true risks of a given learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$, where $\mathcal{Z}$ is the observation space and $\mathcal{H}$ is the hypothesis space.

Informally, suppose we have a learning algorithm $\mathcal{L}$ that receives a sample $S_m = \{Z_1, \ldots, Z_m\}$, which comprises of $m$ i.i.d. observations $Z_i \sim p(z)$, and uses $S_m$ to select a hypothesis $H \in \mathcal{H}$. Because $H$ is selected *based on* the sample $S_m$, its *empirical* risk on $S_m$ is a biased estimator of its *true* risk with respect to the distribution of observations $p(z)$. The difference between the two risks, referred to as the *generalization* risk, determines the prospect of over-fitting in the learning algorithm.

In the literature, generalization bounds are often expressed either in expectation or in probability. Let $L(\cdot; H) : \mathcal{Z} \to [0, 1]$ be some parametric loss function that satisfies the Markov chain $S_m \to H \to L(\cdot; H)$. We write $R_{true}(H)$ and $R_{emp}(H; S_m)$ to denote, respectively, the true and the empirical risks of the hypothesis $H$ w.r.t. $L(\cdot; H)$:

$$R_{emp}(H; S_m) = \frac{1}{m} \sum_{Z_i \in S_m} L(Z_i;\, H)$$

$$R_{true}(H) = \mathbb{E}_{Z \sim p(z)} \big[L(Z;\, H)\big] \qquad (1)$$

Then, generalization in expectation and generalization in probability are defined as follows:

**Definition 1** (Generalization in Expectation)**.** *The expected generalization risk of a learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ with respect to a parametric loss $L(\cdot; H) : \mathcal{Z} \to [0, 1]$ is defined by:*

$$R_{gen}(\mathcal{L}) = \mathbb{E}_{S_m, H|S_m} \big[R_{emp}(H; S_m) - R_{true}(H)\big], \quad (2)$$

*where the empirical risk $R_{emp}(H; S_m)$ and the true risk $R_{true}(H)$ are given by Eq. (1). A learning algorithm $\mathcal{L}$*

Proceedings of the $20^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

*generalizes in expectation if $R_{gen}(\mathcal{L}) \to 0$ as $m \to \infty$ for all distributions $p(z)$.*

**Definition 2** (Generalization in Probability)**.** *A learning algorithm $\mathcal{L}$ generalizes in probability if for any positive constant $\epsilon > 0$, we have:*

$$p\Big\{\big|R_{true}(H) - R_{emp}(H; S_m)\big| > \epsilon\Big\} \to 0 \quad \text{as } m \to \infty,$$

*where the probability is evaluated over the random choice of $S_m$ and the internal randomness of $\mathcal{L}$.*

Clearly, for bounded loss functions, a generalization in probability implies a generalization in expectation but the converse is not generally true.

In general, both types of generalization have been studied in the literature. For instance, generalization in probability is used in the Vapnik-Chervonenkis (VC) theory, the covering numbers, and the PAC-Bayesian framework, among others (Vapnik, 1999; Blumer et al., 1989; McAllester, 2003; Bousquet and Elisseeff, 2002; Bartlett and Mendelson, 2002; Bousquet et al., 2004; Audibert and Bousquet, 2007). Generalization in expectation, on the other hand, was used to analyze learning algorithms, such as the stochastic gradient descent (SGD), differential privacy, and ridge regression (Hardt et al., 2016; Dwork et al., 2015; Shalev-Shwartz and Ben-David, 2014; Raginsky et al., 2016). Its common tool is a replace-one averaging lemma, similar to the Luntz-Brailovsky theorem (Luntz and Brailovsky, 1969; Vapnik and Chapelle, 2000), which relates generalization to algorithmic stability (Hardt et al., 2016; Shalev-Shwartz and Ben-David, 2014). Generalization in expectation is often simpler to analyze, but it provides a weaker performance guarantee.

Recently, however, a third notion of generalization has been introduced in Alabdulmohsin (2015), which is called *uniform* generalization. It also expresses generalization bounds in expectation, but it is stronger than the notion of generalization in Definition 1 because it requires that the generalization risk vanishes uniformly across all bounded parametric loss functions. Importantly, uniform generalization is shown to be *equivalent* to an information-theoretic algorithmic stability constraint, and that it recovers classical results in learning theory. It has been connected to the VC dimension as well (Alabdulmohsin, 2015). Moreover, many conditions can be shown to be sufficient for uniform generalization. These include differential privacy, sample compression schemes, perfect generalization, robust generalization, typical generalization, finite description lengths, or finite domains. Indeed, we prove in Appendix A that all such conditions are sufficient for uniform generalization to hold.

Unfortunately, uniform generalization bounds hold only in expectation without any concentration guarantees. This sheds doubt on the utility of the notion of uniform generalization and its information-theoretic approach of analyzing learning algorithms. For instance, we will later construct a learning algorithm that generalizes *perfectly* in expectation w.r.t. a *specific* parametric loss even though it does not generalize almost surely over the random draw of the sample $S_m$. Hence, generalization in expectation is insufficient to ensure that a generalization will take place in practice.

Nevertheless, we will establish in this paper that a *uniform* generalization in expectation is, in fact, sufficient for a generalization in probability to hold. Moreover, we will derive a tight concentration bound. Hence, all of the uniform generalization bounds, such as the ones derived in (Alabdulmohsin, 2015), hold, not only in expectation but with a high probability as well. Besides, our result provides, as far as we know, the first strong connection between the two forms of generalization in the literature. We present examples of how our concentration bound can be used to deduce concentration results for important classes of learning algorithms, such as those guaranteeing differential privacy.

The proof of our concentration bound rests on a chain rule that we derive for uniform generalization, which is analogous to the chain rule of mutual information in information theory (Cover and Thomas, 1991). Using the chain rule, we show that learning algorithms that generalize uniformly in expectation are amenable to non-adaptive composition, which is analogous to earlier results using differential privacy, sample compression schemes, and perfect generalization (Dwork and Roth, 2013; Cummings et al., 2016). Moreover, the implications of the chain rule are consistent with the *information budget* framework, which was recently proposed for controlling the bias of estimators in the adaptive setting using information theory (Russo and Zou, 2016).

The rest of the paper is structured as follows. We will, first, briefly outline the terminology and notation used in this paper and review the existing literature. Next, we recount the main results pertaining to uniform generalization and algorithmic stability and describe how uniform generalization differs from uniform convergence. Finally, we derive the concentration bound for uniform generalization, prove its tightness, and discuss some of its implications afterward.

## 2 Terminology and Notation

Throughout this paper, we will always write $\mathcal{Z}$ to denote the space of observations (a.k.a. *domain*) and write $\mathcal{H}$ to denote the hypothesis space (a.k.a. *range*). A learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ is formally treated as a stochastic map, where the hypothesis $H \in \mathcal{H}$ can be a deterministic or a randomized function of the sample $S_m \in \mathcal{Z}^m$.

We consider the general setting of learning introduced by Vapnik (Vapnik, 1999). In this setting, the observations $Z_i \in \mathcal{Z}$ can be instance-label pairs $Z_i = (X_i,\, Y_i)$ as in supervised learning or they can comprise of instances only as in unsupervised learning. The distinction between the two learning paradigms is irrelevant. Moreover, we allow the hypothesis $H$ to be an arbitrary random variable. For instance, $H$ can be a classifier, a regression function, a statistical query, a set of centroids, a density estimate, or an enclosing sphere. Only the relationship between the two random variables $S_m$ and $H$ matters in our analysis.

Moreover, if $Z \sim p(z)$ is a random variable drawn from $\mathcal{Z}$ and $f(Z)$ is a function of $Z$, we write $\mathbb{E}_{Z\sim p(Z)}\, f(Z)$ to denote the expected value of $f(Z)$ with respect to the distribution $p(z)$. Occasionally, we omit $p(z)$ from the subscript if it is clear from the context. If $Z$ takes its values from a finite set $S$ uniformly at random, we write $Z \sim S$ to denote this fact. We write $\mathbb{E}_{A|B} f(A)$ to denote the expectation of $f(A)$ under the *conditional* distribution $p(A|B)$. If $X$ is a boolean random variable, then $\mathbb{I}\{X\} = 1$ if and only if $X$ is true, otherwise $\mathbb{I}\{X\} = 0$.

Finally, given two probability measures $P$ and $Q$ defined on the same space, we will write $\langle P,\ Q\rangle$ to denote the *overlapping coefficient* between $P$ and $Q$. That is, $\langle P,\ Q\rangle = 1 - ||P,\ \ Q||_{\mathcal{T}}$, where $||P,\ \ Q||_{\mathcal{T}} = \frac{1}{2}\big|\big|P - Q\big|\big|_1$ is the total variation distance.

## 3 Related Work

Generalization can be rightfully considered as an extension to the *law of large numbers*, which is one of the earliest and most important results in probability theory and statistics. Suppose we have $m$ i.i.d. observations $S_m = \{Z_1, \ldots, Z_m\} \in \mathcal{Z}^m$ and let $f : \mathcal{Z} \to [0, 1]$ be an arbitrary function. If $f$ is fixed *independently* of $S_m$, then $\mathbb{E}_{Z_i\sim S_m}[f(Z_i)] \to \mathbb{E}_{Z\sim p(z)}[f(Z)]$ a.s. as $m \to \infty$. This law is generally attributed to Jacob Bernoulli, who wrote an extensive treatise on the subject published posthumously in 1713 (Stigler, 1986). Modern proofs include low-confidence guarantees, e.g. the Chebychev inequality, and high confidence bounds, e.g. the Chernoff method (Boucheron et al., 2004).

When the function $f$ depends on $S_m$, the law of large numbers is no longer valid because $f(Z_i)$ are not independent random variables. One remedy is to look into the function $F(S_m) = \mathbb{E}_{Z_i\sim S_m} f(Z_i)$. For instance, the Efron-Stein-Steele lemma might be used to bound the variance of $F$, which, in turn, can be translated into a concentration bound using the Chebychev inequality (Boucheron et al., 2004; Bousquet and Elisseeff, 2002). Alternatively, if $F$ satisfies the *bounded-difference* property, then McDiarmid's inequality yields a high-confidence guarantee (Boucheron et al., 2004; Bousquet and Elisseeff, 2002).

In this paper, the same question is being addressed. However, we address it in an information-theoretic manner. We will show that if the function $f : \mathcal{Z} \to [0, 1]$ (as a random variable instantiated after observing the sample $S_m$) carries little *information* about any individual observation $Z_i \in S_m$, then the difference between $\mathbb{E}_{Z_i\sim S_m}[f(Z_i)]$ and $\mathbb{E}_{Z\sim p(z)}[f(Z)]$ will be small with a high probability. The measure of information used here is given by the notion of *variational information* $\mathcal{J}(X;\, Y) = 1 - S(X; Y)$ between the random variables $X$ and $Y$, where $S(X; Y)$ is the mutual stability introduced in Alabdulmohsin (2015). Variational information, also called $T$-information (Raginsky et al., 2016), is an instance of the class of *informativity* measures using $f$-divergences, for which an axiomatic basis has been proposed (Csiszár, 1972, 2008).

The information-theoretic approaches of analyzing the generalization risks of learning algorithms, such as the one pursued in this paper, have found applications in adaptive data analysis. This includes the method of Dwork et al. (2015) using the *max-information*, the method of Russo and Zou (2016) using the *mutual information*, and the method of Raginsky et al. (2016) using the *leave-one-out* information. For bounded losses, uniform generalization bounds using the *variational information* yield tighter results, as deduced by the Pinsker inequality (Reid and Williamson, 2009). In this paper, we prove that these bounds hold not only in expectation but with a high probability as well.

As a consequence of our main theorem, concentration bounds for a given learning algorithm can be immediately deduced once we recognize that the algorithm generalizes uniformly in expectation. Examples of when this holds include having (1) a finite average description length of the hypothesis, (2) a finite VC dimension of the *induced concept class*, (3) differential privacy, (4) bounded mutual information, (5) sample compression schemes, and (6) finite domains. We briefly describe these settings and others that have been previously studied in the literature, and prove their connections to uniform generalization in Appendix A. We also present connections between uniform generalization and learnability in Appendix B. A second consequence of our work is establishing the equivalence between the notion of uniform generalization studied by Alabdulmohsin (2015) and the notion of *robust generalization* considered more recently by Cummings et al. (2016).

Besides deriving a concentration bound, we also establish that our bound is tight. This tightness result is inspired by the work of Bassily et al. (2016) (Lemma 7.4) and Shalev-Shwartz et al. (2010) (Example 3), where similar results are established for differential privacy and learnability respectively. In Section 5.6, we combine techniques from both works to show that our concentration bound is indeed tight.

# 4 Uniform Generalization

First, we review the main results pertaining to uniform generalization and algorithmic stability. We only mention the key results here for completeness. The reader is referred to Alabdulmohsin (2015) for details.

## 4.1 Uniform Generalization vs. Uniform Convergence

The main result of Alabdulmohsin (2015) is the *equivalence* between algorithmic stability and *uniform* generalization in expectation across all bounded parametric loss functions on the product space $\mathcal{Z}\times\mathcal{H}$.

**Definition 3** (Uniform Generalization)**.** *A learning algorithm $\mathcal{L}:\cup_{m=1}^{\infty}\mathcal{Z}^m\to\mathcal{H}$ generalizes uniformly if for any $\epsilon>0$, $\exists m_0(\epsilon)>0$ such that for all distributions $p(z)$ on $\mathcal{Z}$, all parametric losses, and all sample sizes $m>m_0(\epsilon)$, we have $\big|R_{gen}(\mathcal{L})\big|\le\epsilon$, where $R_{gen}(\mathcal{L})$ is given in Eq. (2).*

**Definition 4.** *A learning algorithm $\mathcal{L}$ generalizes uniformly with rate $\epsilon>0$ if the expected generalization risk satisfies $|R_{gen}(\mathcal{L})|\le\epsilon$ for all distributions $p(z)$ on $\mathcal{Z}$ and all parametric losses.*

With some abuse of terminology, we will occasionally say that a learning algorithm generalizes uniformly when it generalizes uniformly according to Definition 4 for some provably small $\epsilon$. Whether we are referring to Definition 3 or 4 will be clear from the context.

Uniform generalization is different from the classical notion of uniform convergence. To see the difference, we note that a parametric loss $L(Z;H):\mathcal{Z}\times\mathcal{H}\to[0,1]$ is a function of the two random variables $Z\in\mathcal{Z}$ and $H\in\mathcal{H}$. This parametric loss on the product space $\mathcal{Z}\times\mathcal{H}$, sometimes called the *loss class* (Bousquet et al., 2004), is a family of loss functions on $Z$ indexed by $H$. Uniform convergence, such as by using the union bound or the growth function, establishes sufficient conditions for uniform convergence to hold within the family of loss functions indexed by $H$ for *a single* parametric loss. These uniform convergence guarantees are often independent of how $\mathcal{L}$ works.

By contrast, suppose that the learning algorithm $\mathcal{L}$ produces a hypothesis $H$ given a sample $S_m\in\mathcal{Z}^m$ with probability $p_{\mathcal{L}}(H|S_m)$, where $p_{\mathcal{L}}(H|S_m)$ can be degenerate in deterministic algorithms. Then, in principle, one can compute the expected generalization risk $R_{gen}(\mathcal{L})$, defined in Eq. (2), for every possible parametric loss. This is the *average* loss within each possible family of bounded loss functions indexed by $H$, averaged over the random choice of $S_m$ and the internal randomness of $\mathcal{L}$. Uniform generalization establishes the conditions for $|R_{gen}(\mathcal{L})|$ to go to zero uniformly across all parametric loss functions. Unlike uniform convergence, which depends on the loss class alone, uniform generalization is determined by all aspects of $\mathcal{L}$, such as the nature of its domain $\mathcal{Z}$, its hypothesis space $\mathcal{H}$, and how $\mathcal{L}$ selects its hypothesis.

## 4.2 Previous Results

The main result proved in Alabdulmohsin (2015) is that uniform generalization is equivalent to an information-theoretic stability constraint on $\mathcal{L}$.

**Definition 5** (Variational Information)**.** *The variational information $\mathcal{J}(X;Y)$ between the random variables $X$ and $Y$ is defined by $\mathcal{J}(X;Y)=||p(X)p(Y)\,,\;p(X,Y)||_{\mathcal{T}}$, where $||P\,,\;Q||_{\mathcal{T}}$ is the total variation distance. The mutual stability is defined by $S(X;Y)=1-\mathcal{J}(X;Y)$.*

Informally speaking, $\mathcal{J}(X;Y)$ measures the influence of observing the value of $X$ on the distribution of $Y$. The rationale behind this definition is revealed next[1]

**Definition 6** (Algorithmic Stability)**.** *Let $\mathcal{L}$ be a learning algorithm that uses $S_m=\{Z_i\}_{i=1,..,m}\sim p^m(z)$ to produce a hypothesis $H\in\mathcal{H}$. Let $Z_{trn}\sim S_m$ be a random variable whose value is drawn uniformly at random from the sample $S_m$. Then, the algorithmic stability of $\mathcal{L}$ is defined by: $\mathbb{S}(\mathcal{L})=\inf_{p(z)}\;S(H;\,Z_{trn})$, where the infimum is taken over all possible distributions of observations $p(z)$. A learning algorithm is called stable if $\lim_{m\to\infty}\;\mathbb{S}(\mathcal{L})=1$.*

Intuitively, a learning algorithm is stable if the influence of a *single* training example vanishes as $m\to\infty$.

**Theorem 1** (Alabdulmohsin, 2015)**.** *For any learning algorithm $\mathcal{L}:\cup_{m=1}^{\infty}\mathcal{Z}^m\to\mathcal{H}$, algorithmic stability (Definition 6) is both necessary and sufficient for uniform generalization (Definition 3). In addition, $\big|R_{gen}(\mathcal{L})\big|\le\mathcal{J}(H;\,Z_{trn})\le1-\mathbb{S}(\mathcal{L})$, with $R_{gen}(\mathcal{L})$ defined in Eq. (2).*

Theorem 1 reveals that uniform generalization has, at least, three *equivalent* interpretations:

[1] In Definition 6 and the rest of the paper, $Z_{trn}$ is a single training example drawn uniformly at random from the training sample. In Definition 6, algorithmic stability quantifies how much information this single example reveals about the hypothesis.

1. *Statistical Interpretation*: A learning algorithm generalizes uniformly if and only if its expected generalization risk $R_{gen}(\mathcal{L})$ vanishes as $m \to \infty$ uniformly across all bounded parametric losses.

2. *Information-Theoretic Interpretation*: A learning algorithm generalizes uniformly if and only if its hypothesis $H$ reveals a *vanishing* amount of information about any *single* observation in $S_m$ as $m \to \infty$. This, for example, is satisfied if $H$ has a finite description length.

3. *Algorithmic Interpretation*: A learning algorithm generalizes uniformly if and only if the contribution of any *single* observation on the hypothesis $H$ vanishes as $m \to \infty$. That is, a learning algorithm generalizes uniformly if and only if it is algorithmically stable.

Other results have also been established in Alabdulmohsin (2015) including the data processing inequality, the information-cannot-hurt inequality, and the uniform generalization bound in the finite hypothesis space setting. Some of those results will be used in our proofs in this paper.

## 5 Generalization in Expectation vs. Generalization in Probability

The main contribution of this paper is to prove that a uniform generalization in expectation implies a generalization in probability and to derive a tight concentration bound. By contrast, a non-uniform generalization in expectation does not imply that a generalization will actually take place in practice. In addition, we will also establish a *chain rule* for variational information and prove that our large-deviation bound is tight. Interestingly, our proof reveals that uniform generalization is a *robust* property of learning algorithms. Specifically, adding a *finite* amount of information (in bits) to a hypothesis $H$ that generalizes uniformly cannot remove its uniform generalization property.

### 5.1 Non-Uniform Generalization

We begin by showing why a non-uniform generalization in expectation does not imply concentration.

**Proposition 1.** *There exists a learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ and a parametric loss $L(\cdot; H) : \mathcal{Z} \to [0,1]$ such that the expected generalization risk is $R_{gen}(\mathcal{L}) = 0$ for all $m \geq 1$, but for all $m \geq 1$ we have $p\Big\{\big|R_{true}(H) - R_{emp}(H; S_m)\big| = \frac{1}{2}\Big\} = 1$, where the probability is evaluated over the random choice of $S_m$ and the internal randomness of $\mathcal{L}$.*

*Proof.* Here is a proof outline. Let $\mathcal{X}$ be a continuum with a continuous marginal density, $\mathcal{Y} = \{-1, +1\}$ be a target set, and let the hypothesis $H$ be $(S_m, B)$, where $B \in \{0, 1\}$ is a Bernoulli r.v. with $p(B = 1) = \frac{1}{2}$. We can define a loss function, parameterized by the hypothesis $(S_m, B)$, such that the algorithm overfits when $B = 1$ and underfits when $B = 0$, while its generalization risk remains zero *on average.* □

Proposition 1 shows that a generalization in expectation does not imply a generalization in probability. Importantly, it is crucial to observe that the learning algorithm constructed in the proof of Proposition 1 does not, in fact, generalize *uniformly* in expectation. Indeed, this latter observation is not a coincidence as will be proved later in Theorem 4.

### 5.2 Robustness of Uniform Generalization

Next, we prove that uniform generalization is a robust property of learning algorithms. We will use this fact later to prove that a uniform generalization in expectation implies a generalization in probability. In order to achieve this, we begin with the following chain rule.

**Definition 7** (Conditional Variational Information). *The conditional variational information between the two random variables $A$ and $B$ given $C$ is defined by:*

$$\mathcal{J}(A;\, B\,|\,C) = \mathbb{E}_C\big[||p(A, B\,|\,C)\,,\;\, p(A|C)\cdot p(B|C)||_{\mathcal{T}}\big],$$

*which is analogous to the conditional mutual information in information theory (Cover and Thomas, 1991).*

**Theorem 2** (Chain Rule). *Let $(H_1, \ldots, H_k)$ be a sequence of random variables. Then, for any random variable $Z$, we have: $\mathcal{J}(Z;\, (H_1, \ldots, H_k)) \leq \sum_{t=1}^{k} \mathcal{J}(Z;\, H_t\,|\,(H_1, \ldots, H_{t-1}))$*

Although the chain rule above provides an upper bound, the upper bound is tight in the following sense:

**Proposition 2.** *For any random variables $A, B$, and $C$, we have $\Big|\mathcal{J}(A;\, (B, C)) - \mathcal{J}(A;\, C\,|\,B)\Big| \leq \mathcal{J}(A;\, B)$ and $\Big|\mathcal{J}(A;\, (B, C)) - \mathcal{J}(A;\, B)\Big| \leq \mathcal{J}(A;\, C\,|\,B)$.*

In other words, the inequality in the chain rule $\mathcal{J}(A;\, (B, C)) \leq \mathcal{J}(A;\, B) + \mathcal{J}(A;\, C\,|\,B)$ becomes an equality if $\min\{\mathcal{J}(A;\, B),\ \mathcal{J}(A;\, C\,|\,B)\} = 0$.

The chain rule provides a recipe for computing the bias of estimators for a composition of hypotheses $(H_1, \ldots, H_k)$, whether this composition is adaptive or non-adaptive. Recently, Russo and Zou (2016) proposed an *information budget* framework for controlling the bias of estimators by controlling the mutual information between $H$ and the sample $S_m$. The proposed framework rests on the chain rule for mutual information. Here, we note that the argument for the information budget framework also holds when using the variational information due to the chain rule above.

Next, we use the chain rule in Theorem 2 to prove that uniform generalization is a robust property of learning algorithms. More precisely, if $K$ has a finite domain, then a hypothesis $H$ generalizes uniformly in expectation if and only if the pair $(H, K)$ generalizes uniformly in expectation. Hence, adding any finite amount of information (in bits) to a hypothesis cannot alter its uniform generalization property[2].

**Theorem 3.** *Let $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm whose hypothesis is $H \in \mathcal{H}$, which is obtained from a sample $S_m$. Let $K \in \mathcal{K}$ be a different hypothesis that is obtained from the same sample $S_m$. If $Z_{trn} \sim S_m$ is a random variable whose value is drawn uniformly at random from $S_m$, then:*

$$\mathcal{J}(Z_{trn};\,(H,K)) \le (2 + \frac{|\mathcal{K}|}{2}) \cdot \mathcal{J}(Z_{trn};\,H) + \sqrt{\frac{\log|\mathcal{K}|}{2m}}$$

### 5.3 Uniform Generalization Implies Concentration

Theorem 3 shows that adding a finite amount of information (in bits) cannot remove the uniform generalization property of learning algorithms. We will use this fact, next, to prove that a uniform generalization in expectation implies a generalization in probability.

The intuition behind the proof is as follows. Suppose we have a hypothesis $H$ that generalizes uniformly in expectation but, for the purpose of obtaining a contradiction, suppose that there exists a parametric loss $L(\cdot; H)$ that does not generalize in probability. Then, *adding* little information to the hypothesis $H$ will allow us to construct a *different* parametric loss that does not generalize in expectation. In particular, we will only to need to know whether the empirical risk w.r.t. $L(\cdot; H)$ is greater than, approximately equal to, or is less than the true risk w.r.t. the same loss. This is described in, at most, two bits. Knowing this additional information, we can define a new parametric loss that does not generalize in expectation, which contradicts the statement of Theorem 3. This line of reasoning is formalized in the following theorem.

**Theorem 4.** *Let $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm, whose risk is evaluated using a parametric loss function $L(\cdot; H) : \mathcal{Z} \to [0, 1]$. Then:*

$$p\Big\{\big|R_{emp}(H; S_m) - R_{true}(H)\big| \ge t\Big\}$$
$$\le \frac{7}{2t}\Big[\mathcal{J}(Z_{trn};\,H) + \sqrt{\frac{\log 3}{49m}}\Big] \le \frac{7}{2t}\Big[1 - \mathbb{S}(\mathcal{L}) + \sqrt{\frac{\log 3}{49m}}\Big],$$

*where $\mathbb{S}(\mathcal{L})$ is the algorithmic stability of $\mathcal{L}$ given in Definition 6, and the probability is evaluated over the random choice of $S_m$ and the internal randomness of $\mathcal{L}$. In particular, if $\mathcal{L}$ generalizes uniformly in expectation, i.e. $\mathbb{S}(\mathcal{L}) \to 1$ as $m \to \infty$, it generalizes in probability for any chosen parametric loss.*

The same proof technique used in Theorem 4 also implies the following concentration bound, which is useful when $I(H; S_m) = o(m)$. The following bound compares well with the bound derived in Russo and Zou (2016) using properties of sub-Gaussian loss functions.

**Proposition 3.** *Let $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ be a learning algorithm, whose risk is evaluated using a parametric loss function $L(\cdot; H) : \mathcal{Z} \to [0, 1]$. Then:*

$$p\Big\{\big|R_{emp}(H; S_m) - R_{true}(H)\big| \ge t\Big\} \le \frac{1}{t}\sqrt{\frac{I(S_m; H) + 2}{2m}},$$

Note that having a bounded mutual information, i.e. $I(S_m; H) = o(m)$, which is the setting recently considered in the work of Russo and Zou (2016), is sufficient for uniform generalization to hold.

### 5.4 Implications

#### 5.4.1 Concentration

In Alabdulmohsin (2015), it was shown that the notion of uniform generalization allows us to reason about learning algorithms in pure information-theoretic terms. This is because uniform generalization is equivalent to an information-theoretic algorithmic stability constraint on learning algorithms. For example, the data processing inequality implies that one can improve the uniform generalization risk by either post-processing the hypothesis, such as sparsification or decision tree pruning, or by pre-processing training examples, such as by introducing noise. Needless to mention, both are common techniques in machine learning. In addition, uniform generalization recovers classical results in learning theory, such as the generalization bounds in the finite hypothesis space setting and finite domains (Alabdulmohsin, 2015). However, such conclusions previously held only *in expectation*.

The most important implication of Theorem 4 is to establish that such conclusions actually hold with a high probability as well. In addition, the concentration bound derived in Theorem 4 shows that algorithmic

[2] Note, by contrast, that the proof of Proposition 1 illustrates an example where a hypothesis $H$ may generalize perfectly in expectation w.r.t. a fixed parametric loss, but a single bit of information suffices to destroy this generalization advantage. This never occurs when $H$ generalizes uniformly since uniform generalization is a robust property.

stability $\mathbb{S}(\mathcal{L})$ not only controls the generalization risk of $\mathcal{L}$ in expectation, i.e. due to its equivalence with uniform generalization, but it also controls the rate of convergence in probability. This brings us to the following important remark:

**Remark 1.** *By improving algorithmic stability, we improve both the expectation of the generalization risk* and *its variance.*

Besides, Theorem 4 can be useful in deriving new concentrations bounds for important classes of learning algorithms once we recognize the existence of uniform generalization. We illustrate this technique on *differential privacy* next.

### 5.5 Differential Privacy

Differential privacy addresses the goal of obtaining useful information about the sample $S_m$ as a whole without revealing a lot of information about each individual observation in the sample (Dwork and Roth, 2013). It closely resembles the notion of algorithmic stability proposed in Alabdulmohsin (2015) because a learning algorithm is stable according to the latter definition if and only if the posterior distribution of an individual observation $Z_{trn}$ in the sample $S_m$ becomes arbitrarily close, in the total variation distance, to the prior distribution $p(Z_{trn})$ as $m \to \infty$. Indeed, differential privacy is a stronger privacy guarantee.

**Definition 8** (Dwork & Roth, 2013)**.** *A randomized learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ is $(\epsilon, \delta)$ differentially private if for any $\mathcal{O} \subseteq \mathcal{H}$ and any two samples $S$ and $S'$ that differ in one observation only, we have:*

$$p(H \in \mathcal{O} \mid S) \le e^{\epsilon} \cdot p(H \in \mathcal{O} \mid S') + \delta$$

Concentration bounds for differential privacy have been derived, such as the recent work of Bassily et al. (2016). Nevertheless, we remark here that Theorem 4 can be used to derive a new concentration bound for differential privacy. Comparing our bound with the lower bound of Lemma 7.4 in Bassily et al. (2016) reveals that the dependence on $\delta$ and $t$ is tight up to a constant factor.

**Corollary 1.** *If a learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$ is $(\epsilon, \delta)$ differentially private, then: $p\Big\{\big|R_{emp}(H; S_m) - R_{true}(H)\big| \ge t\Big\} \le \frac{7}{2t}\Big[\frac{e^{\epsilon}-1+\delta}{2} + \sqrt{\frac{\log 3}{49m}}\Big]$.*

Not surprisingly, the differential privacy parameters $(\epsilon, \delta)$ control the generalization risk of differential privacy, with the quantity $(e^{\epsilon} - 1 + \delta)$ acting a role that is analogous to the role of the standard error.

#### 5.5.1 Equivalence with Robust Generalization

Another implication of the concentration bound in Theorem 4 is establishing the *equivalence* between the notion of uniform generalization and the notion of *robust generalization* studied in Cummings et al. (2016).

**Definition 9** (Robust Generalization)**.** *A learning algorithm $\mathcal{L}$ is $(\epsilon, \delta)$ robustly generalizing if for all distribution $p(z)$ on $\mathcal{Z}$ and any binary-valued parametric loss function $L(\cdot; H) : \mathcal{Z} \to \{0, 1\}$ that satisfies the Markov chain $S_m \to H \to L(\cdot; H)$, we have with a probability of at least $1 - \zeta$ over the choice of $S$ that:*

$$p\Big\{\big|\mathbb{E}_{Z\sim p(z)} L(Z; H) - \frac{1}{m}\sum_{Z_i \in S} L(Z_i; H)\big| \le \epsilon\Big\} \ge 1 - \gamma,$$

*for some $\gamma, \zeta$ such that $\delta = \gamma + \zeta$*[3]*.*

In the following theorem, we prove that robust generalization is equivalent to uniform generalization.

**Corollary 2.** *If a learning algorithm $\mathcal{L}$ is $(\epsilon, \delta)$ robustly generalizing, then it generalizes uniformly at the rate $\epsilon + \delta$. Conversely, if a learning algorithm generalizes uniformly with rate $\tau$, then it is $(\epsilon, \gamma)$ robustly generalizing with $\gamma = (7/2)(\tau + \sqrt{\log 3/(49m)})/\epsilon$. Moreover, if $\mathbb{S}(\mathcal{L}) \to 0$ as $m \to \infty$, then both $\gamma$ and $\epsilon$ can be made arbitrarily close to zero using a sufficiently large sample size $m$.*

### 5.6 Tightness Result

Finally,we note that the concentration bound has a linear dependence on the algorithmic stability term $1 - \mathbb{S}(\mathcal{L})$ or, in a distribution-dependent manner, on the variational information $\mathcal{J}(Z_{trn}; H)$. Typically, $\mathcal{J}(Z_{trn}; H) = O(1/\sqrt{m})$, which holds, for instance, when $\mathcal{Z}$ or $\mathcal{H}$ are countable sets. By contrast, the VC bound provides an exponential decay for supervised classification tasks (Vapnik, 1999; Shalev-Shwartz et al., 2010). This raises the question of whether or not the concentration bound in Theorem 4 can be improved. In this section, we prove that the bound is actually tight. The following theorem is inspired by the work of (Bassily et al., 2016) (Lemma 7.4) and Shalev-Shwartz et al. (2010) (Example 3), who established similar results for differential privacy and learnability respectively.

**Theorem 5.** *For any rational $0 < t < 1$, there exists a learning algorithm $\mathcal{L} : \cup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{H}$, a sample size $m$, a distribution $p(z)$, and a parametric loss*

[3] The original definition proposed in Cummings et al. (2016) states that the probability is evaluted over any "adversary" that takes the hypothesis $H$ as input to produce a loss function $L(\cdot; H)$. However, this is equivalent to the Markov chain $S_m \to H \to L(\cdot; H)$.

$L(\cdot; H) : \mathcal{Z} \to [0, 1]$ *such that $\mathcal{L}$ generalizes uniformly in expectation and it satisfies:*

$$p\Big\{\big|R_{emp}(H; S_m) - R_{true}(H)\big| = t\Big\} = \frac{\mathcal{J}(Z_{trn};\, H)}{t}$$

Theorem 5 shows that, without making any additional assumptions beyond that of uniform generalization, the concentration bound in Theorem 4 is tight up to constant factors. Essentially, the only difference between the upper and the lower bounds is a vanishing $O(1/\sqrt{m})$ term that is *independent* of $\mathcal{L}$.

## 6 Conclusions

Uniform generalization in expectation is a notion of generalization that is equivalent to an information-theoretic algorithmic stability constraint on learning algorithms. In this paper, we proved that whereas generalization in expectation does not imply a generalization in probability, a uniform generalization in expectation implies a generalization in probability and we derived a tight concentration bound. The bound reveals that algorithmic stability improves both the expectation of the generalization risk and its variance. Hence, by constraining the "amount" of information that a hypothesis can carry about any *individual* training example or, equivalently, by limiting the "size" of the contribution of any individual training example on the final hypothesis, the learning algorithm is guaranteed to generalize well with a high probability. Furthermore, we proved a chain rule for variational information, which revealed that uniform generalization is a robust property of learning algorithms. Finally, we proved that the concentration bound is tight.

## A Relations to Other Notions of Generalization & Stability

The connection between differential privacy and uniform generalization is summarized as follows.

**Proposition 4.** *Let $\mathcal{L}$ be a $(\epsilon, \delta)$-differentially private learning algorithm. Let $Z_{trn} \sim S$ be a single training example and let $H \sim p_{\mathcal{L}}(h|S)$ be the hypothesis produced by $\mathcal{L}$. Then $\mathcal{J}(Z_{trn};\, H) \le \frac{e^{\epsilon}-1+\delta}{2}$.*

Next, it can be shown that perfect generalization implies differential privacy (Cummings et al., 2016) so it implies uniform generalization. Also, sample compression implies robust generalization (Cummings et al., 2016), which in turn implies a uniform generalization. Moreover, typical stability (Bassily and Freund, 2016) is equivalent to perfect generalization when the observations are drawn i.i.d., so it implies uniform generalization as well.

The proof that a bounded mutual information, i.e. having $I(S_m; H) = o(m)$, implies uniform generalization is a direct consequence of the concentration bound in Proposition 3.

Finally, the proofs that a finite hypothesis space, a finite VC dimension in the induced concept class, and a finite domain are each sufficient for uniform generalization to hold are provided in Alabdulmohsin (2015).

## B Uniform Generalization and Learnability

### B.1 Consistency of Empirical Risk Minimization

Uniform generalization is a sufficient condition for the consistency of empirical risk minimization (ERM). Suppose we have an ERM learning algorithm, whose hypothesis is denoted $H_{ERM}$. We have by definition:

$$\begin{aligned} R_{emp}(\mathcal{L}) &= \mathbb{E}_S \mathbb{E}_{H|S}[R_{emp}(H; S)] = \mathbb{E}_S[\min_{h\in\mathcal{H}} R_{emp}(h; S)] \\ &\le \min_{h\in\mathcal{H}} \big[\mathbb{E}_S R_{emp}(h; S)\big] = \min_{h\in\mathcal{H}} R_{true}(h) \\ &= R_{true}(h^\star), \end{aligned}$$

where $h^\star$ is the optimal hypothesis. However, the true risk of $\mathcal{L}$ satisfies:

$$R_{true}(\mathcal{L}) - R_{true}(h^\star) \le R_{true}(\mathcal{L}) - R_{emp}(\mathcal{L})$$

Thus, algorithmic stability of ERM implies consistency because $\mathcal{J}(Z_{trn};\, H_{ERM})$ provides an upper bound on $\big|R_{true}(\mathcal{L}) - R_{emp}(\mathcal{L})\big|$. In fact, because $R_{true}(H_{ERM}) - R_{true}(h^\star) \ge 0$, we have by the Markov inequality:

$$p\Big\{R_{true}(H_{ERM}) - R_{true}(h^\star) \ge t\Big\} \le \frac{\mathcal{J}(Z_{trn};\, H_{ERM})}{t}$$

### B.2 Sample Compression and Learnability

Moreover, recent results on the connection between sample compression schemes and learnability (David et al., 2016) reveal that any learnable hypothesis space is learnable using an algorithm that generalizes uniformly in expectation, with only a logarithmic increase in the sample complexity. Because sample compression schemes satisfy robust generalization (Cummings et al., 2016), they generalize uniformly in expectation. Alternatively, the connection between sample compression and uniform generalization may be established more directly by noting that the generalization bounds of sample compression schemes, such as in Shalev-Shwartz and Ben-David (2014), are derived using the *union bound*, which holds *independently* of the choice of the parametric loss.

## References

Alabdulmohsin, I. (2015). Algorithmic stability and uniform generalization. In *NIPS*, pages 19–27.

Audibert, J.-Y. and Bousquet, O. (2007). Combining PAC-Bayesian and generic chaining bounds. *JMLR*, 8:863–889.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482.

Bassily, R. and Freund, Y. (2016). Typicality based stability and privacy. arXiv:1604.03336 [cs.LG].

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2016). Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pages 1046–1059.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965.

Boucheron, S., Lugosi, G., and Bousquet, O. (2004). Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240. Springer.

Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors, *Advanced Lectures on Machine Learning*, volume 3176, pages 169–207.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *JMLR*, 2:499–526.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory.* Wiley & Sons.

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

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

Cummings, R., Ligett, K., Nissim, K., Roth, A., and Wu, Z. S. (2016). Adaptive learning with robust generalization guarantees. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT).*

David, O., Moran, S., and Yehudayoff, A. (2016). Supervised learning through the lens of compression. In *NIPS.*

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (STOC)*, pages 117–126.

Dwork, C. and Roth, A. (2013). The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd international conference on Machine learning (ICML).*

Luntz, A. and Brailovsky, V. (1969). On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3(6).

McAllester, D. (2003). PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21.

Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y., and Xu, A. (2016). Information-theoretic analysis of stability and bias of learning algorithms. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 26–30. IEEE.

Reid, M. D. and Williamson, R. C. (2009). Generalised Pinsker inequalities. In *COLT.*

Russo, D. and Zou, J. (2016). Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS).*

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

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *JMLR*, 11:2635–2670.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press.

Vapnik, V. and Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036.

Vapnik, V. N. (1999). An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999.