Information-theoretic analysis of generalization capability of learning algorithms

Aolin Xu†  Maxim Raginsky‡

Abstract

We derive upper bounds on the generalization error of a learning algorithm in terms of the mutual information between its input and output. The upper bounds provide theoretical guidelines for striking the right balance between data fit and generalization by controlling the input-output mutual information of a learning algorithm. The results can also be used to analyze the generalization capability of learning algorithms under adaptive composition, and the bias-accuracy tradeoffs in adaptive data analytics. Our work extends and leads to nontrivial improvements on the recent results of Russo and Zou.

1 Introduction

A learning algorithm can be viewed as a randomized mapping, or a channel in information-theoretic language, which takes a training dataset as input and generates a hypothesis as output. The generalization error is the difference between the empirical risk of the output hypothesis on the training data and its population risk, and measures how much the learning algorithm suffers from overfitting. Recently, motivated by improving the accuracy of adaptive data analytics, Russo and Zou [1] showed that the mutual information between the collection of empirical risks of the available hypotheses and the final output of the algorithm can be effectively used to analyze and control the bias of data analysis, which is equivalent to the generalization error in learning problems. The method of analysis proposed by Russo and Zou is simpler than the existing methods based on differential privacy, e.g., by Dwork et al. [2,3] and Bassily et al. [4]. It can handle unbounded loss functions, and provides elegant information-theoretic insights on designing answer-generating mechanisms with reduced bias and improved accuracy. In a similar information-theoretic spirit, Alabdulmohsin [5,6] proposed to bound the generalization error using the total-variation information between a random instance in the dataset and the output hypothesis, but his analysis applies only to bounded loss functions.

In this paper, we follow the information-theoretic framework proposed by Russo and Zou [1] to derive upper bounds on the generalization error of learning algorithms. We extend the results in [1] to the situation where the hypothesis space is uncountably infinite, and provide improved upper bounds on the expected absolute generalization error. We also prove high-probability bounds on the generalization error, which were not given in [1]. While the main quantity examined in [1] is the

---

*This work was supported in part by the NSF CAREER award CCF-1254041 and in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

†aolinxu2@illinois.edu

‡maxim@illinois.edu
mutual information between the collection of empirical risks of the hypotheses and the output of the algorithm, we mainly focus on relating the generalization error to the mutual information between the input dataset and the output of the algorithm, which formalizes the intuition that the less information a learning algorithm can extract from the input dataset, the less it will overfit. This viewpoint provides theoretical guidelines for striking the right balance between data fit and generalization by controlling the input-output mutual information of a learning algorithm. For example, we show that regularizing the empirical risk minimization algorithm with the input-output mutual information leads to the well-known Gibbs algorithm. Another benefit of focusing on the input-output mutual information is the ease of analyzing learning algorithms obtained from adaptive composition of constituent algorithms, where we show that, for binary classification, the generalization error of certain boosting and bagging techniques can be conveniently related to the VC dimension of the base hypothesis spaces. When applied back to adaptive data analytics, our work leads to nontrivial improvements on the results in [1], and provides a high-probability bound complementing the results in [2–4]. In the rest of this section, we formally define the quantities that will be used in the paper.

In the standard framework of statistical learning theory [7], we have an instance space $Z$, a hypothesis space $W$, and a nonnegative loss function $\ell : W \times Z \to \mathbb{R}^+$. A learning algorithm characterized by a Markov kernel $P_{W|S}$ takes as input a dataset of size $n$, i.e., an $n$-tuple $S = (Z_1, \ldots, Z_n)$ of i.i.d. random elements of $Z$ with some unknown distribution $\mu$, and picks a random element $W$ of $W$ as the output hypothesis according to $P_{W|S}$. The population risk of a hypothesis $w \in W$ on $\mu$ is

$$L_\mu(w) \equiv \mathbb{E}[\ell(w, Z)] = \int_Z \ell(w, z)\mu(dz).$$

The goal of learning is to ensure that the population risk of the output hypothesis $W$ is small, either in expectation or with high probability, under any data generating distribution $\mu$. The excess risk of $W$ is the difference $L_\mu(W) - \inf_{w \in W} L_\mu(w)$, and its expected value is denoted as $R_{\text{excess}}(\mu, P_{W|S})$. Since $\mu$ is unknown, the learning algorithm cannot directly compute $L_\mu(w)$ for any $w \in W$, but can instead compute the empirical risk of $w$

$$L_S(w) \equiv \frac{1}{n} \sum_{i=1}^{n} \ell(w, Z_i)$$

as a proxy. For a learning algorithm characterized by $P_{W|S}$, the generalization error on $\mu$ is the difference $L_\mu(W) - L_S(W)$, and its expected value is denoted as

$$\text{gen}(\mu, P_{W|S}) \equiv \mathbb{E}[L_\mu(W) - L_S(W)],$$

where the expectation is taken with respect to the joint distribution $P_{S,W} = \mu^\otimes n \otimes P_{W|S}$. The expected population risk can then be decomposed as

$$\mathbb{E}[L_\mu(W)] = \mathbb{E}[L_S(W)] + \text{gen}(\mu, P_{W|S}),$$

where the first term reflects how well the output hypothesis fits the dataset; while the second term reflects how well the output hypothesis generalizes. It is generally impossible to minimize the two terms simultaneously, and any learning algorithm faces a trade-off between the empirical risk and the generalization error. In what follows, we will show how the generalization error can be related to the mutual information between the input and output of the learning algorithm.
2 Algorithmic Stability in Input-output Mutual Information

As discussed in Sec. 1, having a small generalization error is crucial for a learning algorithm to produce an output hypothesis with a small population risk. It turns out that the generalization error of a learning algorithm can be determined by its stability properties. Traditionally, a learning algorithm is said to be stable if a small change of the input to the algorithm does not change the output of the algorithm much. Examples include the notion of on-average stability defined in Shalev-Shwartz et al. [8] and the notion of uniform stability defined in Bousquet and Elisseeff [9]. In recent years, information-theoretic stability notions, such as those measured by differential privacy [3], KL divergence [4, 10], total-variation information [5], and erasure mutual information [11], have been proposed. All existing notions of stability show that the generalization capability of a learning algorithm hinges on how sensitive the output of the algorithm is to local modifications of the input dataset. It implies that the less dependent the output hypothesis \( W \) is on the input dataset \( S \), the better the learning algorithm generalizes. From an information-theoretic point of view, the dependence between \( S \) and \( W \) can be naturally measured by the mutual information between them, which prompts the following information-theoretic definition of stability. We say that a learning algorithm is \((\varepsilon, \mu)\)-stable in input-output mutual information if, under the data-generating distribution \( \mu \),

\[
I(S; W) \leq \varepsilon. \tag{6}
\]

Further, we say that a learning algorithm is \( \varepsilon \)-stable in input-output mutual information if

\[
\sup_\mu I(S; W) \leq \varepsilon. \tag{7}
\]

According to these definitions in (6) and (7), the less information the output of a learning algorithm can provide about its input dataset, the more stable it is. Interestingly, if we view the learning algorithm \( P_{W|S} \) as a channel from \( Z^n \) to \( W \), the quantity \( \sup_\mu I(S; W) \) can be viewed as the information capacity of the channel, under the constraint that the input distribution is of a product form. The definition in (7) means that a learning algorithm is more stable if its information capacity is smaller. We mainly focus on studying the consequences of this notion of stability in the rest of this paper.

3 Upper-bounding Generalization Error via \( I(S; W) \)

Now we derive various generalization guarantees for learning algorithms that are stable in input-output mutual information.

3.1 A Decoupling Estimate

We start with a digression from the statistical learning problem to a more general problem, which may be of independent interest. Consider a pair of random variables \( X \) and \( Y \) with joint distribution \( P_{X,Y} \). Let \( \tilde{X} \) be an independent copy of \( X \), and \( \tilde{Y} \) an independent copy of \( Y \), such that \( P_{X,Y} = P_X \otimes P_Y \). For an arbitrary real-valued function \( f : X \times Y \to \mathbb{R} \), we have the following upper bound on the absolute difference between \( \mathbb{E}[f(X, Y)] \) and \( \mathbb{E}[f(\tilde{X}, \tilde{Y})] \).
We are often interested in analyzing the absolute generalization error. Suppose we are interested in the generalization error of a learning algorithm. We need to develop stronger tools to tackle this problem, by setting

\[ \lambda_{W}(s, w) = L_{S}(w) \]

where the joint distribution of \( S \) and \( W \) is \( P_{S,W} = \mu^{\otimes n} \otimes P_{W|S} \). If \( \ell(w, Z) \) is \( \sigma \)-subgaussian for all \( w \in W \), then \( f(S, w) \) is \( \sigma/\sqrt{n} \)-subgaussian due to the i.i.d. assumption of \( Z_i \)'s, hence \( f(\bar{S}, \bar{W}) \) is \( \sigma/\sqrt{n} \)-subgaussian. This, together with Lemma 1, implies the following theorem.

**Theorem 1.** Suppose \( \ell(w, Z) \) is \( \sigma \)-subgaussian under \( \mu \) for all \( w \in W \), then

\[
|\text{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2\sigma^2}{n}} I(S; W).
\]

(10)

Theorem 1 suggests that, by controlling the mutual information between the input and the output of a learning algorithm, we can control the generalization error of the learning algorithm.

Russo and Zou [1] considered the same problem setup with the restriction that the hypothesis space \( W \) is finite, and showed that \( |\text{gen}(\mu, P_{W|S})| \) can be upper-bounded in terms of \( I(\Lambda_{W}(S); W) \), where

\[
\Lambda_{W}(S) \triangleq (L_{S}(w))_{w \in W}
\]

is the collection of empirical risks of the hypotheses in \( W \). Using Lemma 1 by setting \( X = \Lambda_{W}(S) \), \( Y = W \), and \( f(\Lambda_{W}(s), w) = L_{S}(w) \), we immediately recover the result by Russo and Zou even when \( W \) is uncountably infinite:

**Theorem 2 (Russo and Zou [1]).** Suppose \( \ell(w, Z) \) is \( \sigma \)-subgaussian under \( \mu \) for all \( w \in W \), then

\[
|\text{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2\sigma^2}{n}} I(\Lambda_{W}(S); W).
\]

(12)

It should be noted that Theorem 1 can be obtained as a consequence of Theorem 2 because

\[
I(\Lambda_{W}(S); W) \leq I(S; W)
\]

(13)

due to the Markov chain \( \Lambda_{W}(S) - S - W \), as for each \( w \in W \), \( L_{S}(w) \) is a function of \( S \). However, if the output \( W \) of the learning algorithm depends on \( S \) only through the empirical risks \( \Lambda_{W}(S) \), in other words, when the Markov chain \( S - \Lambda_{W}(S) - W \) holds, then Theorem 1 and Theorem 2 are equivalent. The advantage of Theorem 1 is that \( I(S; W) \) can be much easier to evaluate than \( I(\Lambda_{W}(S); W) \), and can provide better insights to guide the algorithm design. We will elaborate on this when we discuss the Gibbs algorithm and the adaptive composition of learning algorithms.

Theorem 1 and Theorem 2 only provide upper bounds on the expected generalization error. We are often interested in analyzing the absolute generalization error \( |L_{\mu}(W) - L_{S}(W)| \), e.g., its expected value or the probability for it to be small. We need to develop stronger tools to tackle these problems, which is the subject of the next two subsections.

---

1Recall that a random variable \( U \) is \( \sigma \)-subgaussian if \( \log \mathbb{E}[e^{\lambda(U - \mathbb{E}[U])}] \leq \lambda^2 \sigma^2/2 \) for all \( \lambda \in \mathbb{R} \).
3.3 High-probability Bound on $|L_\mu(W) - L_S(W)|$

For any fixed $w \in W$, if the loss function $\ell(w, Z)$ is $\sigma$-subgaussian under $\mu$, the Chernoff-Hoeffding bound gives $P[|L_\mu(w) - L_S(w)| > \alpha] \leq 2e^{-\alpha^2 n/2\sigma^2}$. It implies that, if $S$ and $W$ are independent, then a sample size of $n = (2\sigma^2/\alpha^2)\log(2/\beta)$ suffices to guarantee that

$$P[|L_\mu(W) - L_S(W)| > \alpha] \leq \beta. \quad \text{(14)}$$

The following results show that, even when $W$ is a random element from $W$ and is dependent on $S$, as long as $I(S; W)$ is sufficiently small, a sample complexity polynomial in $1/\alpha$ and logarithmic in $1/\beta$ still suffices to guarantee (14), where the probability now is taken with respect to the joint distribution $P_{S,W} = \mu^{\otimes n} \otimes P_W|S$.

**Theorem 3** (proved in Appendix B). Suppose $\ell(w, Z)$ is $\sigma$-subgaussian under $\mu$ for all $w \in W$. If a learning algorithm satisfies $I(\Lambda_W(S); W) \leq \varepsilon$, then for any $\alpha > 0$ and $0 < \beta \leq 1$, (14) can be guaranteed by a sample complexity of

$$n \geq \frac{8\sigma^2}{\alpha^2} \left( \frac{\varepsilon}{\beta} + \log \frac{2}{\beta} \right). \quad \text{(15)}$$

In view of (13), any learning algorithm that is $(\varepsilon, \mu)$-stable in input-output mutual information satisfies the condition $I(\Lambda_W(S); W) \leq \varepsilon$. The proof of Theorem 3 is based on Lemma 1 and an adaptation of the “monitor technique" proposed by Bassily et al. [4]. While the high-probability bounds of [2–4] based on differential privacy are for bounded loss functions and for functions with bounded differences, the result in Theorem 3 only requires $\ell(w, Z)$ to be subgaussian. We have the following corollary of Theorem 3.

**Corollary 1.** Under the conditions in Theorem 3, if $\varepsilon \leq (g(n) - 1)\beta \log(2/\beta)$ for some function $g(n) \geq 1$, then a sample complexity that satisfies $n/g(n) \geq (8\sigma^2/\alpha^2)\log(2/\beta)$ guarantees (14).

For example, taking $g(n) = 2$, Corollary 1 implies that if $\varepsilon \leq \beta \log(2/\beta)$, then (14) can be guaranteed by a sample complexity of $n = (16\sigma^2/\alpha^2)\log(2/\beta)$, which is on the same order of the sample complexity when $S$ and $W$ are independent. As another example, taking $g(n) = \sqrt{n}$, Corollary 1 implies that if $\varepsilon \leq (\sqrt{n} - 1)\beta \log(2/\beta)$, then $n = (64\sigma^2/\alpha^4)\log(2/\beta)^2$ guarantees (14).

3.4 Upper Bound on $\mathbb{E}|L_\mu(W) - L_S(W)|$

A byproduct of the proof of Theorem 3 (setting $m = 1$ in the proof) is an upper bound on the expected absolute generalization error.

**Theorem 4.** Suppose $\ell(w, Z)$ is $\sigma$-subgaussian under $\mu$ for all $w \in W$. If a learning algorithm satisfies that $I(\Lambda_W(S); W) \leq \varepsilon$, then

$$\mathbb{E}|L_S(W) - L_\mu(W)| \leq \sqrt{\frac{2\sigma^2}{n}}(\varepsilon + \log 2). \quad \text{(16)}$$

This result improves [1, Prop. 3.2], which states that $\mathbb{E}|L_S(W) - L_\mu(W)| \leq \sigma/\sqrt{n} + 36\sqrt{2\sigma^2 \varepsilon/n}$. Theorem 4 together with Markov’s inequality implies that (14) can be guaranteed by a sample complexity of $n = \frac{2\sigma^2}{\alpha^2 \beta^2}(\varepsilon + \log 2)$, but it has a worse dependence on $\beta$ as compared to the sample complexity given by Theorem 3.
4 Learning Algorithms with Input-output Mutual Information Stability

In this section, we analyze several learning problems/algorithms from the viewpoint of input-output mutual information stability. We first consider two learning problems where the input-output mutual information of a learning algorithm can be directly upper-bounded via the properties of the hypothesis space. Then we propose two learning algorithms with controlled input-output mutual information stability. We also discuss other methods to induce input-output mutual information stability, and the stability of learning algorithms obtained from adaptive composition of constituent algorithms.

4.1 Countable Hypothesis Space

When a hypothesis space is countable, the input-output mutual information can be directly upper-bounded by \( H(W) \), the entropy of \( W \), which is at most \( \log k \) if \( |W| = k \). From Theorem 1, for any learning algorithm \( P_{W|S} \), if \( \ell(w, Z) \) is \( \sigma \)-subgaussian for all \( w \in W \), then

\[
|\text{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2\sigma^2 H(W)}{n}} \tag{17}
\]

in this case. The empirical risk minimization (ERM) algorithm satisfies

\[
\mathbb{E}[L_S(W_{\text{ERM}})] = \mathbb{E}\left[ \inf_{w \in W} L_S(w) \right] \leq \inf_{w \in W} \mathbb{E}[L_S(w)] = \inf_{w \in W} L_\mu(w). \tag{18}
\]

Therefore, for the ERM algorithm, the upper bound in (17) also holds for the expected excess risk.

4.2 Binary Classification

For the problem of binary classification, \( Z = X \times Y, Y = \{0, 1\} \), and \( W \) is a collection of classifiers \( w : X \to Y \), which could be uncountably infinite. The loss function takes the form \( \ell(w, z) = 1\{w(x) \neq y\} \). Define the set of length-\( n \) binary vectors

\[
A_W(S) \triangleq \{(\ell(w, Z_1), \ldots, \ell(w, Z_n)), w \in W\}. \tag{19}
\]

By bounding the cardinality of \( A_W(S) \) in terms of the VC dimension of the hypothesis space, we have the following result.

**Theorem 5** (proved in Appendix C). For binary classification, any learning algorithm satisfies

\[
I(A_W(S); W) \leq V(W) \log(n + 1), \tag{20}
\]

where \( V(W) \) is the VC dimension of the hypothesis space \( W \). Moreover, for any learning algorithm such that the Markov chain \( S - A_W(S) - W \) holds, the above upper bound also holds for \( I(S; W) \).

Theorem 5 together with Theorem 2 \footnote{For binary classification, since \( \ell(w, z) \in \{0, 1\}, \ell(w, Z) \) is \( 1/2 \)-subgaussian for all \( w \in W \).} implies that for any learning algorithm for binary classification,

\[
|\text{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{V(W) \log(n + 1)}{2n}}, \tag{21}
\]
which is consistent with the result obtained by the symmetrization argument and Rademacher complexity [12]. For the ERM algorithm, this upper bound also holds for the expected excess risk due to (18). Russo and Zou [1] also considered binary classification. However, their result [1, Prop. 4.1] is restricted to the case where the features \((x_1, \ldots, x_n)\) are fixed and only the labels \((Y_1, \ldots, Y_n)\) are random. Theorem 5 holds for the general problem of binary classification, where both the features and the labels are random.

4.3 Gibbs Algorithm

The decomposition of the expected population risk in (5) suggests that a good learning algorithm should output a hypothesis with both small empirical risk and small generalization error. As Theorem 1 shows that the generalization error can be upper-bounded in terms of the mutual information \(I(S; W)\), it is natural to consider an algorithm that minimizes the empirical risk regularized by \(I(S; W)\):

\[
P^*_W|S = \arg \inf_{P_W|S} \left( \mathbb{E}[L_S(W)] + \frac{1}{\beta} I(S; W) \right),
\]

(22)

where \(\beta > 0\) is a parameter that balances fitting and generalization. To deal with the issue that \(\mu\) is unknown to the learning algorithm, we can relax the above optimization problem by replacing \(I(S; W)\) with an upper bound \(D(P_W|S \parallel Q| P_S) = I(S; W) + D(P_W \parallel Q)\) on it, where \(Q\) is an arbitrary distribution on \(W\), so that the solution of the relaxed optimization problem does not depend on \(\mu\). It turns out that the well-known Gibbs algorithm solves the relaxed optimization problem.

**Theorem 6** (proved in Appendix D). *The solution to the optimization problem*

\[
P^*_W|S = \arg \inf_{P_W|S} \left( \mathbb{E}[L_S(W)] + \frac{1}{\beta} D(P_W|S \parallel Q| P_S) \right)
\]

(23)

*is the Gibbs algorithm, which satisfies*

\[
P^*_W|S=s(\text{dw}) = \frac{e^{-\beta L_s(w)}Q(\text{dw})}{\mathbb{E}_Q[e^{-\beta L_s(W)}]} \quad \text{for each } s \in \mathbb{Z}^n.
\]

(24)

The Gibbs algorithm can thus be interpreted as a way to stabilize the ERM algorithm by controlling the input-output mutual information. The parameter \(\beta > 0\) controls how well the Gibbs algorithm approximates the ERM algorithm. Using Theorem 1, we can analyze the generalization error of the Gibbs algorithm by upper-bounding \(I(S; W)\):

**Corollary 2** (proved in Appendix E). *If the loss function \(\ell\) takes values in \([0, 1]\), then the expected generalization error of the Gibbs algorithm \(P^*_W|S\) described in (24) satisfies*

\[
|\text{gen}(\mu, P^*_W|S)| \leq \sqrt{\frac{\beta}{n}}.
\]

(25)

Moreover, if \(|W| = k\), choosing \(Q\) as the uniform distribution on \(W\) and \(\beta = 2\sqrt{n \log k}\), the expected excess risk satisfies \(R_{\text{excess}}(\mu, P^*_W|S) \leq \sqrt{\frac{\log k}{n}}\).

The proof of (25) is based on the fact that \(P^*_W|S\) is differentially private, and by upper-bounding \(I(S; W)\) using the group property of differential privacy [13]. The mutual information \(I(\Lambda_W(S); W)\), on the other hand, is not as straightforward to evaluate as \(I(S; W)\).
4.4 Noisy Empirical Risk Minimization

Another algorithm with guaranteed input-output mutual information stability is the noisy empirical risk minimization algorithm, where independent noise $N_w, w \in W$, is added to the empirical risk of each hypothesis, and the algorithm outputs a hypothesis that minimizes the noisy empirical risks:

$$W = \arg \min_{w \in W} (L_S(w) + N_w).$$  \hfill (26)

For the case of adding Gaussian noise, we have the following corollary of Theorem 1.

**Corollary 3** (proved in Appendix F). Suppose $|W| = k$ and $\ell(w, Z)$ is $\sigma$-subgaussian under $\mu$ for all $w \in W$. For the noisy ERM algorithm $P_{W|S}$ where $N_w$’s are i.i.d. $\mathcal{N}(0, \sigma^2_N)$,

$$|\text{gen}(\mu, P_{W|S})| \leq \frac{\sigma^2 \sqrt{k}}{n \sigma_N} \wedge \sqrt{\frac{2\sigma^2 \log k}{n}}.$$  \hfill (27)

Moreover, choosing $\sigma^2_N = \sigma^2/4n$, the expected excess risk satisfies $R_{\text{excess}}(\mu, P_{W|S}) \leq 2\sqrt{\frac{2\sigma^2 \log k}{n}}$.

Instead of adding noise, the empirical risks can be quantized and the hypothesis minimizing the quantized empirical risks can be picked. The generalization error can be analyzed similarly.

4.5 Other Methods to Induce Input-Output Mutual Information Stability

In addition to the Gibbs algorithm and the noisy ERM algorithm, many other methods can be used to control the input-output mutual information of the learning algorithm. One method is to preprocess the dataset $S$ to obtain $\tilde{S}$, and then run a learning algorithm on the preprocessed dataset $\tilde{S}$. The preprocessing can be adding noise to the data or erasing some of the instances in the dataset, etc. In any case, we have the Markov chain $S \rightarrow \tilde{S} \rightarrow W$, which implies $I(S; W) \leq \min \{I(S; \tilde{S}), I(\tilde{S}; W)\}$. Another method is the postprocessing of the output of a learning algorithm. For example, the weights $\tilde{W}$ generated by a neural network training algorithm can be quantized or perturbed by noise. This gives rise to the Markov chain $S \rightarrow \tilde{W} \rightarrow W$, which implies $I(S; W) \leq \min \{I(\tilde{W}; W), I(S; \tilde{W})\}$. Moreover, strong data processing inequalities [14] may be used to sharpen these upper bounds on $I(S; W)$. Preprocessing of the dataset and postprocessing of the output hypothesis are among numerous regularization methods used in the field of deep learning [15, Ch. 7.5]. Other regularization methods may also be interpreted as ways to induce the input-output mutual information stability of a learning algorithm, and this would be an interesting direction of future research.

4.6 Adaptive Composition of Learning Algorithms

Beyond analyzing the generalization error of individual learning algorithms, examining the input-output mutual information is also useful for analyzing the generalization capability of learning algorithms under adaptive composition. Consider the situation where $k$ learning algorithms are sequentially executed. The output of the $j$th algorithm may depend on the dataset $S$, as well as on the outputs $W^{j-1}$ of the previously executed learning algorithms. The final output $W_k$ can be viewed as obtained from an adaptive composition of the $k$ constituent learning algorithms $P_{W_k|S, W^{j-1}}$, $j = 1, \ldots, k$. A common example with $k = 2$ is model selection followed by a learning algorithm using
We can thus control the generalization error of the final output by controlling the conditional mutual information,

\[ I(S; W_k) \leq I(S; W^k) = \sum_{j=1}^{k} I(S; W_j|W^{j-1}). \]  

(28)

We can thus control the generalization error of the final output by controlling the conditional mutual information \( I(S; W_j|W^{j-1}) \) at each step of the composition.

Various boosting techniques in machine learning can be viewed as instances of adaptive composition. We may use (28) to analyze their generalization performance. Taking binary classification as an example, suppose the first \( k \) hypotheses are picked from a base hypothesis space \( \mathcal{W} \), and the final hypothesis is an aggregation of the first \( k \) hypotheses and lies in a larger hypothesis space \( \overline{\mathcal{W}} \). If, conditionally on \( W^{j-1} \), the Markov chain \( S - \mathcal{A}_W(S) - W_j \) holds for \( j = 1, \ldots, k \), and if the Markov chain \( S - W^k - W \) holds for the final hypothesis \( W \), then as a result of (28) and Theorem 5,

\[ I(S; W) \leq I(S; W^k) \leq kV(W) \log(n + 1), \]

(29)

and

\[ |\text{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{kV(W) \log(n + 1)}{2n}}. \]

(30)

The same results hold for the bagging technique, which can be viewed as a special case of the above example where the first \( k \) hypotheses are generated non-adaptively based on \( S \). For the above problem, \( I(\Lambda_{\overline{W}}(S); W) \) can be upper-bounded by \( \sum_{j=1}^{k} I(\Lambda_{\overline{W}}(S); W_j|W^{j-1}) \) as well; however, each term in the summation cannot be directly bounded in terms of the VC dimension of the base hypothesis space. This shows that \( I(S; W) \) can be much more handy to use than \( I(\Lambda_{\overline{W}}(S); W) \) in analyzing adaptively composed algorithms.

5 Application to Adaptive Data Analytics

The results derived so far for analyzing the generalization capability of learning algorithms can also be applied to analyzing the bias in adaptive data analytics. In data analytics, there is an unknown distribution \( \mu \) on \( Z \), and a random dataset \( S \in \mathbb{Z}^n \) drawn from \( \mu^{\otimes n} \). Given a query space \( \mathcal{W} \) and a function \( \ell: \mathcal{W} \times Z \rightarrow \mathbb{R}^+ \), the data analyst picks some query \( w \in \mathcal{W} \) and wishes to evaluate the quantity \( L_\mu(w) = \mathbb{E}[\ell(w, Z)] \) under \( Z \sim \mu \). There is an answer-generating mechanism holding the dataset \( S \), which accepts the query \( w \) and returns an answer \( Y \) to the data analyst. By the law of large numbers (assuming the function \( \ell \) is bounded), there is a strong and uniform guarantee that the answer given as the empirical mean \( L_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, Z_i) \) can well approximate \( L_\mu(w) \) for all \( \mu \). In practice, however, data analysis is often performed in multiple rounds in an adaptive manner: in the \( j \)th round, the data analyst issues a query \( W_j \) based on the previously issued queries \( W^{j-1} \) as well as the answers \( Y^{j-1} \) received so far; a new answer \( Y_j \) is then generated based on the dataset \( S \) and the query \( W_j \). In this case, the queries \( W_j \) for \( j \geq 2 \) are no longer independent of the dataset \( S \); hence, the empirical mean \( L_S(W_j) \) can severely deviate from the true expectation \( L_\mu(W_j) \). The difference \( L_\mu(W_j) - L_S(W_j) \) is called the bias of \( W_j \) on \( S \). An important problem in adaptive data analytics is to design answer-generating mechanisms such that the answers \( Y_j \) are close to \( L_\mu(W_j) \) under multiple rounds of interaction. Recently, ideas in differential privacy have been brought to bear on the problem of designing answer-generating mechanisms for adaptive data analytics [2–4].
these works, the bias analysis is based on deriving generalization guarantees of differentially private algorithms. The accuracy of the answers can then be determined by combining the upper bound on the bias and the accuracy guarantee of the privacy-inducing mechanism. Compared to existing works based on differential privacy, the analysis based on mutual information pioneered by Russo and Zou [1] is simpler and provides information-theoretic insights on designing good answer-generating mechanisms that reduce bias and improve accuracy. Following the information-theoretic framework in [1], we develop a proof using Theorem 3 and Theorem 4, which leads to an improved bound on the expected absolute error and a sample complexity bound for the absolute error to be small with high probability. Our proof also removes the jointly Gaussian assumption on the empirical risks.

Consider the $k$-round adaptive data analysis: at the $j$th round, the query $W_j$ is drawn according to the kernel $P_{W_j|W_{j-1},Y_{j-1}}$, and the answer $Y_j$ is drawn according to the kernel $P_{Y_j|S,W_j}$. To upper-bound the bias, we first bound the mutual information $I(S;W_j)$ for $j = 1,\ldots,k$.

**Lemma 2** (proved in Appendix G). For the $k$-round adaptive data analytics,

$$I(S;W_j) \leq \sum_{i=1}^{j-1} I(S;Y_i|W_i) \quad \text{for } j = 1,\ldots,k.$$  

With Lemma 2, we can analyze the accuracy in the special case where the $j$th answer $Y_j$ is generated by adding Gaussian noise to $L_S(W_j)$.

**Theorem 7** (proved in Appendix H). Suppose for $j = 1,\ldots,k$, the answer $Y_j$ is generated by $Y_j = L_S(W_j) + N_j$, where $N_j$’s are i.i.d. zero-mean Gaussian with variance $\sigma^2_j = \sigma^2\sqrt{j/n}$. If $\ell(w,Z)$ is $\sigma$-subgaussian under $\mu$ for all $w \in W$, then

$$\max_{j \in [k]} \mathbb{E}|Y_j - L_\mu(W_j)| \leq \frac{3\sigma k^{1/4}}{\sqrt{n}}.$$  

It follows that a sample complexity of

$$n = \frac{9\sigma^2\sqrt{k}}{\alpha^2}$$  

guarantees $\max_{j \in [k]} \mathbb{E}|Y_j - L_\mu(W_j)| \leq \alpha$. Moreover, a sample complexity of

$$n = \frac{96\sigma^2\sqrt{k}}{\alpha^2\beta}$$  

guarantees $\max_{j \in [k]} \mathbb{P}[|Y_j - L_\mu(W_j)| > \alpha] \leq \beta$.

The upper bound in (31) improves the result in [1, Prop. 5.1], which states that $\max_{j \in [k]} \mathbb{E}|Y_j - L_\mu(W_j)| \leq c\sigma k^{1/4}/\sqrt{n}$ with some constant $c < 53$. Moreover, the result in [1] is obtained under the assumption that $(L_S(w_j))_{j \in [k]}$ are jointly Gaussian with variance $\sigma^2/n$ for all $w_j \in W, j \in [k]$, which is not needed by Theorem 7. Under the same jointly Gaussian assumption on the empirical risks and a richness assumption on the query space, Wang et al. [16] obtained the minimax rate of the mean squared error for the $k$-fold adaptive data analytics which matches the upper bound in (31).

For the high probability bound, Dwork et al. [2, Coro. 16] showed that, if the loss function takes values in $[0,1]$, then, for the Laplacian noise-adding mechanism, a sample complexity of
\( n = O(\sqrt{k}\log^{3/2}(1/\beta)/\alpha^{3/2}) \) suffices to ensure that \( \max_{j \in [k]} \Pr[|Y_j - L_\mu(W_j)| > \alpha] \leq \beta \). This sample complexity was later improved in Bassily et al. [4, Coro. 6.1], who have shown that, if the loss function is \( O(1/n) \)-sensitive and \( Z \) is a finite set, then there is a differentially private answer generating mechanism such that \( n = O(\sqrt{k}\log\log k\log^{3/2}(1/\alpha\beta)/\alpha^2) \) provides a stronger guarantee of \( \Pr[\max_{j \in [k]} |Y_j - L_\mu(W_j)| > \alpha] \leq \beta \). Compared to the above two results, the sample complexity given in (33) has a worse dependence on \( \beta \); nevertheless, Theorem 7 only requires that the loss function is subgaussian (potentially unbounded), and does not require \( Z \) to be a finite set.
A Proof of Lemma 1

Just like Russo and Zou [1], we exploit the Donsker–Varadhan variational representation of the relative entropy [17, Corollary 4.15]: for any two probability measures \( \pi, \rho \) on a common measurable space \((\Omega, \mathcal{F})\),

\[
D(\pi \| \rho) = \sup_F \left\{ \int_{\Omega} F \, d\pi - \log \int_{\Omega} e^F \, d\rho \right\},
\]

where the supremum is over all measurable functions \( F: \Omega \to \mathbb{R} \), such that \( e^F \in L^1(\rho) \). From (A.1), we know that for any \( \lambda \in \mathbb{R} \),

\[
D(P_{X,Y} \| P_X \otimes P_Y) \geq \mathbb{E}[\lambda f(X,Y)] - \log \mathbb{E}[e^{\lambda f(X,Y)}]
\]

\[
\geq \lambda (\mathbb{E}[f(X,Y)] - \mathbb{E}[f(X,\bar{Y})]) - \frac{\lambda^2 \sigma^2}{2},
\]

where the second step follows from the subgaussian assumption on \( f(X,\bar{Y}) \):

\[
\log \mathbb{E}[e^{\lambda (f(X,\bar{Y}) - \mathbb{E}[f(X,\bar{Y})])}] \leq \frac{\lambda^2 \sigma^2}{2} \quad \forall \lambda \in \mathbb{R}.
\]

Inequality (A.2) gives a nonnegative parabola in \( \lambda \), whose discriminant must be nonpositive, which implies

\[
|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(X,\bar{Y})]| \leq \sqrt{2\sigma^2 D(P_{X,Y} \| P_X \otimes P_Y)}.
\]

The result follows by noting that \( I(X;Y) = D(P_{X,Y} \| P_X \otimes P_Y) \).

B Proof of Theorem 3

To prove Theorem 3, we need the following two lemmas.

**Lemma B.1.** Consider the parallel execution of \( m \) independent copies of \( P_{W|S} \) on independent datasets \( S_1, \ldots, S_m \): for \( t = 1, \ldots, m \), an independent copy of \( P_{W|S} \) takes \( S_t \sim \mu^\otimes n \) as input and outputs \( W_t \). Define \( S^m \triangleq (S_1, \ldots, S_m) \). If under \( \mu \), \( P_{W|S} \) satisfies that \( I(\Lambda_W(S);W) \leq \varepsilon \), then the overall algorithm \( P_{W^m|S^m} \) satisfies \( I(\Lambda_W(S_1), \ldots, \Lambda_W(S_m); W^m) \leq m\varepsilon \).

**Proof.** The proof is based on the independence among \((S_t, W_t), t = 1, \ldots, m\), and the chain rule of mutual information. \( \square \)

**Lemma B.2.** Let \( S^m \triangleq (S_1, \ldots, S_m) \), where \( S_t \sim \mu^\otimes n \). If an algorithm \( P_{W,T;R|S^m}: Z^{m \times n} \to \mathcal{W} \times [m] \times \{\pm 1\} \) satisfies \( I(\Lambda_W(S_1), \ldots, \Lambda_W(S_m); W, T, R) \leq \varepsilon \), and if \( \ell(w,Z) \) is \( \sigma \)-subgaussian for all \( w \in \mathcal{W} \), then

\[
\mathbb{E}[R(L_{S^c}(W) - L_{\mu}(W))] \leq \sqrt[2]{\frac{2\sigma^2 \varepsilon}{n}}.
\]

**Proof.** The proof is based on Lemma 1. Let \( X = (\Lambda_W(S_1), \ldots, \Lambda_W(S_m)), Y = (W, T, R) \), and

\[
f((\Lambda_W(s_1), \ldots, \Lambda_W(s_m)), (w, t, r)) = r L_{st}(w).
\]
If \( \ell(w, Z) \) is \( \sigma \)-subgaussian under \( Z \sim \mu \) for all \( w \in \mathcal{W} \), then 
\[
\frac{1}{n} \sum_{t=1}^{n} \ell(w, Z_t, i) \text{ is } \sigma/\sqrt{n} \text{-subgaussian for all } w \in \mathcal{W}, t \in [m] \text{ and } r \in \{\pm 1\}, \text{ and hence } f(\tilde{X}, \tilde{Y}) \text{ is } \sigma/\sqrt{n} \text{-subgaussian.}
\]
Lemma 1 implies that
\[
\mathbb{E}[R(L_{S_{r}}(W)) - \mathbb{E}(RL_{\mu}(W))] \leq \sqrt{2 \sigma^2 I(\Lambda_{W}(S_1), \ldots, \Lambda_{W}(S_m); W, T, R) / n}
\]
and proves the claim. \( \Box \)

Note that the upper bound in Lemma B.2 does not depend on \( m \). With these lemmas, we can prove Theorem 3.

**Proof of Theorem 3.** The proof is an adaptation of a “monitor technique” proposed by Bassily et al. [4]. First, let \( P_{W|m|S} \) be the parallel execution of \( m \) independent copies of \( P_{W|S} \): for \( t = 1, \ldots, m \), an independent copy of \( P_{W|S} \) takes an independent \( S_t \sim \mu^{\otimes n} \) as input and outputs \( W_t \). Given \( S^m \) and \( W^m \), let the output of the “monitor” be a sample \((W^*, T^*, R^*) \) drawn from \( W \times [m] \times \{\pm 1\} \) according to
\[
(T^*, R^*) = \arg \max_{t \in [m], r \in \{\pm 1\}} r(L_{\mu}(W_t) - L_{S_t}(W_t)) \text{ and } W^* = W_{T^*}.
\]
(B.3)

This gives
\[
R^*(L_{\mu}(W^*) - L_{S_{T^*}}(W^*)) = \max_{t \in [m]} |L_{\mu}(W_t) - L_{S_t}(W_t)|.
\]

Taking expectation on both sides, we have
\[
\mathbb{E}[R^*(L_{\mu}(W^*) - L_{S_{T^*}}(W^*))] = \mathbb{E}\left[ \max_{t \in [m]} |L_{\mu}(W_t) - L_{S_t}(W_t)| \right].
\]
(B.4)

Note that conditional on \( W^m \), the tuple \((W^*, T^*, R^*)\) can take only \( 2m \) elements, which means that
\[
I(\Lambda_{W}(S_1), \ldots, \Lambda_{W}(S_m); W^*, T^*, R^* | W^m) \leq \log(2m).
\]
(B.5)

In addition, since \( P_{W|S} \) is assumed to satisfy \( I(\Lambda_{W}(S); W) \leq \varepsilon \), Lemma B.1 implies that
\[
I(\Lambda_{W}(S_1), \ldots, \Lambda_{W}(S_m); W^*, T^*, R^*) \leq m \varepsilon.
\]

Therefore, by the chain rule of mutual information and the data processing inequality, we have
\[
I(\Lambda_{W}(S_1), \ldots, \Lambda_{W}(S_m); W^*, T^*, R^*) \leq I(\Lambda_{W}(S_1), \ldots, \Lambda_{W}(S_m); W^m, W^*, T^*, R^*) \leq m \varepsilon + \log(2m).
\]

By Lemma B.2 and the assumption that \( \ell(w, Z) \) is \( \sigma \)-subgaussian,
\[
\mathbb{E}[R^*(L_{S_{T^*}}(W^*) - L_{\mu}(W^*))] \leq \sqrt{\frac{2 \sigma^2}{n}(m \varepsilon + \log(2m))}.
\]
(B.6)

Combining (B.6) and (B.4) gives
\[
\mathbb{E}\left[ \max_{t \in [m]} |L_{S_t}(W_t) - L_{\mu}(W_t)| \right] \leq \sqrt{\frac{2 \sigma^2}{n}(m \varepsilon + \log(2m))}.
\]
(B.7)
The rest of the proof is by contradiction. Choose \( m = \lfloor 1/\beta \rfloor \). Suppose the algorithm \( P_{W|S} \) does not satisfy the claimed generalization property, namely,
\[
P(\|L_S(W) - L_\mu(W)\| > \alpha) > \beta. \tag{B.8}
\]
Then by the independence among the pairs \((S_t, W_t), t = 1, \ldots, m\),
\[
P(\max_{t \in [m]} |L_{S_t}(W_t) - L_\mu(W_t)| > \alpha) > 1 - (1 - \beta)^{1/\beta} > 1/2.
\]
Thus
\[
E \left[ \max_{t \in [m]} |L_{S_t}(W_t) - L_\mu(W_t)| \right] > \alpha/2. \tag{B.9}
\]
Combining (B.7) and (B.9) gives
\[
\frac{\alpha}{2} < \sqrt{\frac{2\sigma^2}{n} \left( \frac{\varepsilon}{\beta} + \log \frac{2}{\beta} \right)}. \tag{B.10}
\]
The above inequality implies that
\[
n < \frac{8\sigma^2}{\alpha^2} \left( \frac{\varepsilon}{\beta} + \log \frac{2}{\beta} \right), \tag{B.11}
\]
which contradicts the condition in (15). Therefore, under the condition in (15), the assumption in (B.8) cannot hold. This completes the proof. \( \square \)

C Proof of Theorem 5

First note that
\[
I(\Lambda_W(S); W) \leq I(A_W(S); W)
\]
because \( \Lambda_W(S) \) can be written as a function of \( A_W(S) \). Define the class of binary-valued functions
\[
A_W \triangleq \{ \ell(w, \cdot) : Z \to \{0, 1\}, w \in W \}.
\]
We have
\[
I(A_W(S); W) \leq \log S_n(A_W),
\]
where
\[
S_n(A_W) \triangleq \sup_{s \in \mathbb{Z}^n} |A_W(s)|
\]
is the \( n \)th shatter coefficient of \( A_W \). By Sauer’s lemma,
\[
S_n(A_W) \leq (n + 1)^{V(A_W)},
\]
where \( V(A_W) \) is the VC dimension of the function class \( A_W \). Due to the identity \( V(A_W) = V(W) \) [18, Theorem 13.1], we know that
\[
I(\Lambda_W(S); W) \leq V(W) \log(n + 1),
\]
which proves the first claim. Moreover, for any learning algorithm such that \( S - A_W(S) - W \) form a Markov chain, we have
\[
I(S; W) \leq I(A_W(S); W) \leq V(W) \log(n + 1).
\]
D Proof of Theorem 6

To solve the relaxed optimization problem in (23), first note that

$$\inf_{P_{W|S}} \left( \mathbb{E}[L_S(W)] + \frac{1}{\beta}D(P_{W|S}||P_S) \right)$$

$$= \inf_{P_{W|S}} \int_{Z^n} \mu^{\otimes n}(ds) \left( \mathbb{E}[L_s(W)|S = s] + \frac{1}{\beta}D(P_{W|S=s}||Q) \right)$$

$$= \int_{Z^n} \mu^{\otimes n}(ds) \inf_{P_{W|S=s}} \left( \mathbb{E}[L_s(W)|S = s] + \frac{1}{\beta}D(P_{W|S=s}||Q) \right).$$

It follows that for each \( s \in Z^n \), the algorithm \( P^*_{W|S=s} \) that minimizes (23) satisfies

$$P^*_{W|S=s} = \arg\inf_{P_{W|S=s}} \left( \mathbb{E}[L_s(W)|S = s] + \frac{1}{\beta}D(P_{W|S=s}||Q) \right). \quad (D.12)$$

The solution to (D.12) for each \( s \in Z^n \) turns out to be the Gibbs algorithm [19] as described in (24), which does not depend on \( \mu \).

E Proof of Corollary 2

If \( \ell(\cdot, \cdot) \in [0, 1] \), then

$$e^{-2\beta/n} \leq \frac{dP^*_{W|S=s}}{dP^*_{W|S=s'}} \leq e^{2\beta/n}$$

for all \( s, s' \in Z^n \) such that \( d_H(s, s') \leq 1 \). This implies that the Gibbs algorithm with \( \ell(\cdot, \cdot) \in [0, 1] \) is \((2\beta/n, 0)\)-differentially private. From the group privacy property of \((2\beta/n, 0)\)-differentially private mechanisms [13, Theorem 2.2], we know that, if the loss function \( \ell(\cdot, \cdot) \in [0, 1] \), then

$$e^{-2\beta} \leq \frac{dP^*_{W|S=s}}{dP^*_{W|S=s'}} \leq e^{2\beta} \quad \forall s, s' \in Z^n,$$

which implies that for any \( \mu \)

$$I(S; W) \leq \sup_{s, s' \in Z^n} D(P^*_{W|S=s}||P^*_{W|S=s'}) \leq 2\beta.$$ 

By Theorem 1, the generalization error satisfies

$$|\text{gen}(\mu, P^*_{W|S})| \leq \sqrt{\frac{\beta}{n}}.$$ 

This proves the first claim.

In addition, from Hoeffding’s lemma and the fact that the Gibbs algorithm is \((1 - e^{-2\beta/n})\)-TV stable, it is shown in [11] that

$$|\text{gen}(\mu, P^*_{W|S})| \leq (1 - e^{-2\beta/n}) \wedge \frac{\beta}{4n} \wedge \sqrt{\frac{\beta}{n}}. \quad (E.13)$$
When the hypothesis space $W$ has cardinality $k$, for the Gibbs algorithm $P^*_W$ with $Q$ chosen as the uniform distribution on $W$, we can also bound the empirical risk of the Gibbs algorithm. Using a proof similar to that of [4, Lemma 7.1], we can show that, for any dataset $s$,

$$\mathbb{E}[L_s(W) | S = s] \leq \min_{w \in W} L_s(w) + \frac{1}{\beta} \log k.$$  \hspace{1cm} (E.14)

Choosing $\beta = 2\sqrt{n\log k}$, we have

$$\mathbb{E}[L_s(W)] \leq \mathbb{E}[L_s(W_{\text{ERM}})] + \frac{1}{2} \sqrt{\frac{\log k}{n}} + \frac{1}{2} \sqrt{\frac{\log k}{n}}.$$  

where the last step is due to (18). Combining with the second upper bound in (E.13) on the generalization error, the population risk can be upper bounded by

$$\mathbb{E}[L_{\mu}(W)] \leq \min_{w \in W} L_{\mu}(w) + \frac{1}{2} \log \frac{k}{n} + \frac{1}{2} \sqrt{\frac{\log k}{n}} + \frac{1}{2} \sqrt{\frac{\log k}{n}}.$$  

Therefore, the expected excess risk of the Gibbs algorithm in this case satisfies

$$R_{\text{excess}}(\mu, P^*_W) \leq \sqrt{\frac{\log k}{n}}.$$  

F Proof of Corollary 3

Suppose $W = \{w_1, \ldots, w_k\}$. We have the following chain of inequalities

$$I(S; W) \leq I((L_S(w_i))_{i \in [k]}; (L_S(w_i) + N_i)_{i \in [k]})$$

$$\leq \sum_{j=1}^k I(L_S(w_j); L_S(w_j) + N_j)$$

$$\leq \max_{j \in [k]} \frac{k}{2} \log \left(1 + \frac{\text{Var}[L_S(w_j)]}{\sigma^2_N}\right)$$

$$\leq \frac{k}{2} \log \left(1 + \frac{\sigma^2}{n\sigma^2_N}\right)$$

$$\leq \frac{k\sigma^2}{2n\sigma^2_N},$$

where we have used the data processing inequality for mutual information; the fact that for product channels, the mutual information between the overall input and output is upper bounded by the sum of the input-output mutual information of individual channels [20]; the formula for the capacity of input power constrained Gaussian channel; the fact that $L_S(w_j)$’s are $\sigma/\sqrt{n}$ subgaussian hence $\text{Var}[L_S(w_j)] \leq \sigma^2/n$; and the fact that $\log(1 + x) \leq x$. Also note that in this case

$$I(S; W) \leq H(W) \leq \log k.$$  

The first claim then follows from Theorem 1.

We can also analyze the expected empirical risk of the noisy ERM algorithm.
Theorem 8. The expected empirical risk of the noisy ERM algorithm satisfies
\[ \mathbb{E}[L_S(W)] \leq \mathbb{E}[L_S(W_{ER})] + \mathbb{E}[\max_{w \in W} N_w] + \mathbb{E}[\max_{w \in W} (-N_w)]. \]

Proof. We first show that for an arbitrary realization \( s \) of the dataset, the noisy ERM algorithm satisfies
\[ \mathbb{E}[L_s(W) | S = s] \leq \min_{w \in W} L_s(w) + \mathbb{E}[\max_{w \in W} N_w] + \mathbb{E}[\max_{w \in W} (-N_w)]. \] (F.15)

To prove this result, we use the similar technique in the proof of [21, Theorem 4.2]. From the definition of \( W \),
\[ \frac{1}{n} \sum_{i=1}^{n} \ell(W, z_i) + N_W = \min_{w \in W} \left( \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) + N_w \right) \leq \min_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) + \max_{w \in W} N_w. \]

It follows that
\[ \frac{1}{n} \sum_{i=1}^{n} \ell(W, z_i) \leq \min_{w \in W} \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) + \max_{w \in W} N_w + \max_{w \in W} (-N_w). \]

The result in (F.15) follows by taking expectation of \( W \) and \((N_w)_{w \in W}\) on both sides of the above inequality conditional on \( S = s \), and by noting that \((N_w)_{w \in W}\) does not depend on \( S \). The claim of Theorem 8 follows by taking expectation of \( S \) on both sides of (F.15).

If \( W \) has cardinality \( k \) and \( N_w \)'s are i.i.d. \( \mathcal{N}(0, \sigma_N^2) \), then Theorem 8 implies
\[ \mathbb{E}[L_S(W)] \leq \mathbb{E}[L_S(W_{ER})] + 2\sqrt{2\sigma_N^2 \log k} \leq \min_{w \in W} L_\mu(w) + 2\sqrt{2\sigma_N^2 \log k}. \]

Choosing \( \sigma_N^2 = \frac{\sigma^2}{4n} \), and combining with the second upper bound on the expected generalization error in (27), we have an upper bound on the expected excess risk for this case:
\[ R_{\text{excess}}(\mu, P_W|S) \leq 2\sqrt{\frac{2\sigma^2 \log k}{n}}. \]

G Proof of Lemma 2

To explore the conditional independences among the variables in the adaptive data analytics, we plot the Bayesian network of the query-answer pairs in Fig. 1 for \( k = 4 \).
From the chain rule of mutual information, we have the following chain of inequalities: for \( j = 1, \ldots, k \),

\[
I(S; W_j) \leq I(S; W_j^{j-1}, Y^j) \\
= \sum_{i=1}^{j-1} I(S; W_i, Y_{i|W_i^{i-1}, Y_{i-1}}) + I(S; Y_{i|W_i, W_i^{i-1}, Y_{i-1}}) \\
\leq \sum_{i=1}^{j-1} I(S; Y_{i|W_i}), \tag{G.16}
\]

where the last step uses the fact that \( I(S; W_i|W_i^{i-1}, Y_{i-1}) = 0 \) because of the Markov chain \( S \rightarrow W_i^{i-1}, Y_{i-1} \rightarrow W_i \), and the fact that

\[
I(S; Y_{i|W_i, W_i^{i-1}, Y_{i-1}}) \leq I(W_i^{i-1}, Y_{i-1}, S; Y_{i|W_i}) = I(S; Y_{i|W_i})
\]

because \( W_i^{i-1}, Y_{i-1} \rightarrow S \rightarrow Y_i \) form a Markov chain conditional on \( W_i \).

**H Proof of Theorem 7**

Starting with the result of Lemma 2, we have

\[
I(S; Y_i|W_i) \leq I(L_S(W_i); L_S(W_i) + N_i|W_i),
\]

since \( S - L_S(W_i) - Y_i \) form a Markov chain conditional on \( W_i \). Then, similar to the proof of Corollary 3, we make use of the subgaussian assumption of \( \ell(w, Z) \) and the capacity formula of the
power-constrained AWGN channel, and obtain

\[
I(S; W_j) \leq \sum_{i=1}^{j-1} I(S; Y_i | W_i)
\]

\[
\leq \sum_{i=1}^{j-1} I(L_S(W_i); L_S(W_i) + N_i | W_i)
\]

\[
\leq \sum_{i=1}^{j-1} \frac{1}{2} \log \left( 1 + \sup_{w \in W} \text{Var}[L_S(w)] \right)
\]

\[
\leq \sum_{i=1}^{j-1} \frac{\sigma^2}{2n \sigma_i^2}.
\] (H.17)

From Theorem 4,

\[
\mathbb{E}[Y_j - L \mu(W_j)] \leq \mathbb{E}[Y_j - L_S(W_j)] + \mathbb{E}[L_S(W_j) - L \mu(W_j)]
\]

\[
\leq \sigma_j + \sqrt{\frac{2\sigma^2}{n} (I(S; W_j) + \log 2)}
\]

\[
\leq \sigma_j + \sqrt{\frac{\sigma^4 \sum_{i=1}^{j-1} \frac{1}{\sigma_i^2} + 2\sigma^2 \log 2}{n}}.
\]

Using the assumption that \( \sigma_j^2 = \sigma^2 \sqrt{j/n} \) and the fact that \( \sum_{i=1}^{j} \frac{1}{\sqrt{i}} \leq 2\sqrt{j} \), we get

\[
\mathbb{E}[Y_j - L \mu(W_j)] \leq \frac{\sigma_j^{1/4}}{\sqrt{n}} + \sqrt{\frac{2\sigma^2 \sqrt{j} + 2\sigma^2 \log 2}{n}}
\]

\[
\leq 3\sigma_j^{1/4} \quad j = 1, \ldots, k.
\] (H.18)

From (H.18), we have

\[
\max_{j \in [k]} \mathbb{E}[Y_j - L \mu(W_j)] \leq \frac{3\sigma k^{1/4}}{\sqrt{n}}.
\] (H.19)

This proves (31).

To prove the second claim, note that

\[
P[|Y_j - L \mu(W_j)| > \alpha] = P[|Y_j - L \mu(W_j)| > \alpha]
\]

\[
\leq P[|Y_j - L_S(W_j)| + |L_S(W_j) - L \mu(W_j)| > \alpha]
\]

\[
\leq P[|Y_j - L_S(W_j)| > \alpha/2] + P[|L_S(W_j) - L \mu(W_j)| > \alpha/2].
\] (H.20)

Since

\[
P[|Y_j - L_S(W_j)| > \alpha/2] \leq 2e^{-\alpha^2/8\sigma_j^2}
\]

and \( \sigma_j^2 = \sigma^2 \sqrt{j/n} \), to make

\[
P[|Y_j - L_S(W_j)| > \alpha/2] \leq \frac{\beta}{2},
\]

19
it suffices to have

\[ n \geq \frac{8\sigma^2 \sqrt{j}}{\alpha^2} \log \frac{4}{\beta}. \] (H.21)

The second term in (H.20) can be upper-bounded by Theorem 3. From (H.17) and the assumption that \( \sigma_j^2 = \sigma^2 \sqrt{j}/n \), we have

\[ I(S;W_j) \leq \sqrt{j}. \]

By Theorem 3, to make

\[ \mathbb{P}[|L_S(W_j) - L_\mu(W_j)| > \alpha/2] \leq \frac{\beta}{2}, \]

it suffices to have

\[ n \geq \frac{32\sigma^2}{\alpha^2} \left( \frac{2\sqrt{j}}{\beta} + \log \frac{4}{\beta} \right). \] (H.22)

Taking the maximum of (H.21) and (H.22), and noting that \( \log(4/\beta) < \sqrt{j}/\beta \) for \( j \geq 2 \) and \( 0 < \beta \leq 1 \), we know that

\[ n = \frac{96\sigma^2 \sqrt{k}}{\alpha^2 \beta} \]

suffices to guarantee

\[ \max_{j \in [k]} \mathbb{P}[|Y_j - L_\mu(W_j)| > \alpha] \leq \beta. \]
References

[1] D. Russo and J. Zou, “How much does your data exploration overfit? Controlling bias via information usage,” *arXiv preprint*, 2016. [Online]. Available: https://arxiv.org/abs/1511.05219

[2] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, “Preserving statistical validity in adaptive data analysis,” in *Proc. of 47th ACM Symposium on Theory of Computing (STOC)*, 2015.

[3] ———, “Generalization in adaptive data analysis and holdout reuse,” in *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[4] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, “Algorithmic stability for adaptive data analysis,” in *Proceedings of The 48th Annual ACM Symposium on Theory of Computing (STOC)*, 2016.

[5] I. Alabdulmohsin, “Algorithmic stability and uniform generalization,” in *28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[6] ———, “An Information-Theoretic Route from Generalization in Expectation to Generalization in Probability,” in *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

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

[8] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, “Learnability, stability and uniform convergence,” *J. Mach. Learn. Res.*, vol. 11, pp. 2635–2670, 2010.

[9] O. Bousquet and A. Elisseeff, “Stability and generalization,” *J. Machine Learning Res.*, vol. 2, pp. 499–526, 2002.

[10] Y.-X. Wang, J. Lei, and S. E. Fienberg, “On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms,” in *Proceedings of the International Conference on Privacy in Statistical Databases*, 2016.

[11] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, “Information-theoretic analysis of stability and bias of learning algorithms,” in *Proceedings of IEEE Information Theory Workshop*, 2016.

[12] S. Boucheron, O. Bousquet, and G. Lugosi, “Theory of classification: a survey of some recent advances,” *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005.

[13] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, 2014.

[14] M. Raginsky, “Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels,” *IEEE Trans. Inform. Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[16] Y.-X. Wang, J. Lei, and S. E. Fienberg, “A minimax theory for adaptive data analysis,” *arXiv:1602.04287*, 2016. [Online]. Available: https://arxiv.org/abs/1602.04287
[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, 2013.

[18] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[19] T. Zhang, “Information-theoretic upper and lower bounds for statistical estimation,” *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1307 – 1321, 2006.

[20] Y. Polyanskiy and Y. Wu, “Lecture Notes on Information Theory,” Lecture Notes for ECE563 (UIUC) and 6.441 (MIT), 2012-2016. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf

[21] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.