Stopping criterion for active learning based on deterministic generalization bounds

Hideaki Ishibashi
Kyushu Institute of Technology

Hideitsu Hino
The Institute of Statistical Mathematics/RIKEN AIP

Abstract

Active learning is a framework in which the learning machine can select the samples to be used for training. This technique is promising, particularly when the cost of data acquisition and labeling is high. In active learning, determining the timing at which learning should be stopped is a critical issue. In this study, we propose a criterion for automatically stopping active learning. The proposed stopping criterion is based on the difference in the expected generalization errors and hypothesis testing. We derive a novel upper bound for the difference in expected generalization errors before and after obtaining a new training datum based on PAC-Bayesian theory. Unlike ordinary PAC-Bayesian bounds, though, the proposed bound is deterministic; hence, there is no uncontrollable trade-off between the confidence and tightness of the inequality. We combine the upper bound with a statistical test to derive a stopping criterion for active learning. We demonstrate the effectiveness of the proposed method via experiments with both artificial and real datasets.

1 Introduction

In supervised learning problems, increasing the number of training samples can improve prediction performance. However, to train a predictor, annotated datasets are required, and annotation frequently requires the knowledge of experts or the conduction of experiments with a high cost, such as large-scale experiments or long-term experiments, for example, agri-cultural examinations. Active learning (AL) (Settles, 2009) is a framework in which learners can select data that improve the prediction accuracy. Although various methods have been proposed for selecting new data, only a few studies have reported on the criteria for stopping learning (Vlachos, 2008; Ertekin et al., 2007; Paisley et al., 2010; Krause and Guestrin, 2007), which is a critical aspect to make active learning practical. Furthermore, most existing criteria depend on a specific learning tasks, and parameters such as the threshold need to be appropriately defined.

The purpose of this study is to develop a versatile stopping criterion for active learning, that is, a criterion independent of task or loss function. Realization of this goal faces two challenges. The first issue is determining the measure to use for the stopping criterion. The simplest approach is to evaluate the convergence of learning by monitoring the generalization error of the prediction model using the test dataset. However, active learning is often applied in circumstances in which obtaining a sufficient amount of test data is unreasonable. Therefore, it is necessary to evaluate the convergence of learning without using the test dataset. The second issue is the development of a concrete algorithm to determine the stop timing of learning. The simplest approach for this is to set a threshold and stop learning when a value defined based on a certain measure exceeds the threshold. However, in general, the appropriate threshold may not be known in advance.

In this study, the generalization error is evaluated without using the test data by employing the upper bound of the difference in the expected generalization errors based on the PAC-Bayesian framework (McAllester, 1999, 2003; Langford, 2005; Catoni, 2007). Furthermore, the sequence of the difference in the expected generalization errors is regarded as time series data, and the stop timing of learning is determined automatically via a runs test (Wald and Wolfowitz, 1940).

The major contributions of our work are as follows:

A versatile stopping criterion for active learning is pro-

posed. The proposed method is applicable to arbitrary cost functions and can be applied to both classification and regression tasks. Moreover, the proposed criterion can evaluate performance of a model at low calculation cost. In the stopping criterion of AL, we need to consider a trade-off between the labeling cost, computational cost for both learning the model and determining the criterion as well as the predictive performance of the model. The proposed method assume the cost of labeling is dominant, and the predictive performance of the model is of importance. The computational cost for evaluating the stopping criterion is of the constant order, hence it does not increase the over all computational cost.

Theoretical guarantees for the criterion are derived based on the PAC-Bayesian theory. The proposed criterion is the upper bound of the optimal value in terms of the PAC-Bayesian framework. It is notable that even though the proposed criterion is derived from the PAC-Bayesian theory, it is a deterministic bound; therefore, there is no trade-off between the confidence and tightness of the bound. We combine the proposed bound with statistical test to realize a statistically sound criterion for stopping active learning.

2 Active learning and its stopping criteria

Let $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ be the input variable and the corresponding output variable, respectively, in certain domains $\mathcal{X}$ and $\mathcal{Y}$. Supervised learning considers the problem of estimating the predictor $f : x \mapsto \mathbb{E}[y|x]$. For problems with high annotation cost or acquisition cost, active learning (Settles, 2009) is a practical method for constructing a useful prediction model with the minimum number of annotations or labeling for the output variable. Specifically, active learning yields a new datum by repeating two processes: (i) estimating the predictor from the acquired training data and (ii) determining the new input datum x^* to maximize the *acquisition function* $a(x|f)$ based on the current predictor:

$$x^* = \arg\max_{x \in \mathcal{X}} a(x|f). \tag{1}$$

In practical application of active learning algorithms, the timing to terminate learning is decided based on a predetermined budget, which is called the "fixed-budget" approach. However, there are two problems in this standard approach. (1) Even within the fixed budget, it is possible that the learner has enough training data, and the fixed-budget approach oversamples in this situation, which misses possible saving of budget. In addition, if we knew that the learner has much room for improvement when we have reached the bud-

get limit, we can claim that we should spend more of the budget. (2) Active learning can be used for outlier removal (Kobayashi and Sugiyama, 2012; Cohan and Naderiparizi, 2018), but the fixed-budget approach cannot rule out sampling outliers. Developing a stopping criterion for active learning is in these sense important.

The existing stopping criteria for active learning can be classified into accuracy- and uncertainty-based approaches. Methods pertaining to the accuracy-based approach evaluate predictive errors. A typical method is to evaluate the predictive error by using selected unlabeled data or pool data (Zhu, 2007; Zhu et al., 2008a,b; Laws and Schütze, 2008). Another popular method is based on the stability of agreement of multiple predictors (Bloodgood and Vijay-Shanker, 2009; Bloodgood and Grothendieck, 2013; Altschuler and Bloodgood, 2019; Olsson and Tomanek, 2009). Methods pertaining to the uncertainty-based approach evaluate the uncertainty of prediction by using the pooled data. The margin of support vector machine (SVM) classifier is used for measuring the uncertainty (Schohn and Cohn, 2000; Vlachos, 2008). Criteria based on the convergence of the margin or other related quantities evaluated by using the pooled dataset are proposed in (Laws and Schütze, 2008; Krause and Guestrin, 2007). However, most of those criteria depend on the learning models, acquisition function, and problem settings such as the use of classification or regression.

Apart from the studies on active learning, some other studies have considered the optimal stopping timing for learning algorithms. In the nonparametric regression and neural networks literature, *early stopping* (Prechelt, 2012; Raskutti et al., 2011) and its variants (Wang and Yan, 2018; Raskutti et al., 2011) have been widely used in practice to reduce computational time and overfitting. In the framework of Bayesian optimization, several heuristics have been devised (Lorenz et al., 2015; Desautels et al., 2014), although a theoretically supported method with practical utility is yet to be developed. In the multi-armed bandit literature, the best arm identification has been considered as a problem of finding the best model in the minimum number of trials (Kaufmann et al., 2016; Even-Dar et al., 2006; Audibert and Bubeck, 2010; Aziz et al., 2018). Determining an appropriate stopping timing is also a critical issue when running Markov chain Monte Carlo (MCMC) algorithms, in which the sequential fixed-width confidence interval is one of the most popular methods (Chow and Robbins, 1965; Gong and Flegal, 2016; Jones et al., 2006). However, the methods developed in the studies related to MCMC algorithms focus on the convergence to a stationary distribution, and thus they cannot be directly

applied in active learning because the aim of active learning is not to obtain a certain probability model but to train an accurate predictor.

In this work, we focus on a stopping criterion that monitors the difference in the generalization errors before and after a new training sample is obtained. If sufficient test data are available, it is easy to evaluate the difference in the generalization errors. However, this is not the case when active learning is involved. To alleviate this problem, we adopt the PAC-Bayesian framework (McAllester, 1999, 2003; Langford, 2005; Catoni, 2007; Germain et al., 2016), in which the generalization error is bounded by using only the training dataset.

3 Evaluation of difference between generalization errors

Let $S = (X, Y) = \{(x_i, y_i)\}_{i=1}^t$, $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ be the observed dataset. We assume that (X, Y) is generated by a probability distribution $\mathcal{D}$. Let $l : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to [a, b]$ be a loss function, where $\mathcal{F}$ is a set of predictors and $0 \le a < b < \infty$. In the Bayesian framework, we assume that the predictor f is a function-valued random variable and consider its prior and posterior distributions. The expected risk $\mathcal{L}_\mathcal{D}(f)$ and empirical risk $\mathcal{L}_S(f)$ can be defined as follows:

$$\mathcal{L}_\mathcal{D}(f) = \mathbb{E}_\mathcal{D}[l(f, x, y)], \quad \mathcal{L}_S(f) = \frac{1}{t} \sum_{i=1}^t l(f, x_i, y_i).$$

PAC-Bayesian theory binds the expected generalization error to any posterior $q(f)$ of predictor f (McAllester, 2003; Catoni, 2007), that is, $\mathbb{E}_{q(f)}[\mathcal{L}_\mathcal{D}(f)]$. So far, various upper bounds in classification have been proposed (McAllester, 1999, 2003; Langford, 2005; Catoni, 2007). Recently, the upper bound of the expected generalization error in the regression problem has also been proposed (Alquier et al., 2016; Germain et al., 2016).

From the viewpoint of active learning, the greatest advantage of the PAC bound is its universality because it is derived without assuming any specific loss function. Based on the PAC-Bayesian approach, not only different learning models but also both classification and regression problems can be addressed using the unified framework.

3.1 Upper bound for difference between generalization errors

Let $q(f|S)$ be the posterior distribution of predictor $f \in \mathcal{F}$ given a dataset $S = (X, Y)$. The posterior is calculated by using Bayes' theorem as fol-

lows: Assuming that the loss function is the negative log-likelihood (Banerjee, 2006), that is, $l(f, x, y) = -\log p(y|f)$, we have $p(y|f, x) = e^{-l(f, x, y)}$. If a vector of function value f for input X is denoted by $\mathbf{f}_X$, the posterior can be obtained using

$$q(\mathbf{f}_X|S) = p(Y|\mathbf{f}_X, X)p(\mathbf{f}_X)/p(Y|X) \quad (2)$$

$$= e^{-t\mathcal{L}_S(f)}p(\mathbf{f}_X)/p(Y|X), \quad (3)$$

where $p(\mathbf{f}_X)$ is a prior. Let X_* and $\mathbf{f}_{X_*}$ be the test input and its corresponding vector of the function value, respectively. Then, the predictive distribution $q(\mathbf{f}_{X_*}|S)$ is given by $q(\mathbf{f}_{X_*}|S) = \int p(\mathbf{f}_{X_*}|\mathbf{f}_X)q(\mathbf{f}_X|S)d\mathbf{f}_X$, where $p(\mathbf{f}_{X_*}|\mathbf{f}_X)$ is the conditional probability w.r.t. the prior distribution $p(\mathbf{f}_{X_*}, \mathbf{f}_X)$. The difference between the expected generalization error w.r.t $q(f|S)$ and that w.r.t $q(f|S')$ is denoted by $\mathcal{R}(q(f|S), q(f|S'))$, that is,

$$\mathcal{R}(q(f|S), q(f|S')) = \mathbb{E}_{q(f|S)}[\mathcal{L}_\mathcal{D}(f)] - \mathbb{E}_{q(f|S')}[\mathcal{L}_\mathcal{D}(f)].$$

Note that the sample sizes of S and S' can be different.

The quantity $\mathcal{R}(q(f|S_t), q(f|S_{t+1}))$ represents the reduction of the expected generalization error by the data acquisition. Substituting $p(f)$ to $q(f|S_0)$, we have $\mathcal{R}(p(f), q(f|S_t)) = \sum_{i=1}^t \mathcal{R}(q(f|S_{i-1}), q(f|S_i))$. This represents the cumulative reduction of the expected generalization error by adding t samples. Because $\mathbb{E}_{p(f)}[\mathcal{L}_\mathcal{D}(f)]$ is constant for added samples, we can assume it is a constant d. Then, $\mathcal{R}(p(f), q(f|S_t)) = d - \mathbb{E}_{q(f|S_t)}[\mathcal{L}_\mathcal{D}(f)]$; hence, the convergence of $\mathcal{R}(p(f), q(f|S_t))$ is equivalent to the convergence of $\mathbb{E}_{q(f|S_t)}[\mathcal{L}_\mathcal{D}(f)]$, which is the rationale behind the definition of $\mathcal{R}$. We evaluate the convergence of $\mathcal{R}(q(f|S_t), q(f|S_{t+1}))$ in the framework of a statistical test.

The KL divergence between $p(f)$ and $q(f)$ can be defined as $D_{KL}[p(f)||q(f)] = \mathbb{E}_{p(f)}\left[\log \frac{dp(f)}{dq(f)}\right]$. Then, the following theorem holds:

Theorem 1. *Let $q(f|S)$ and $q(f|S')$ be the posteriors w.r.t. predictor $f \in \mathcal{F}$ given S and S'. For any measurable function $\mathcal{L}_\mathcal{D}(f)$[1], the following inequality holds:*

$$\mathcal{R}(q(f|S), q(f|S')) \le D_{KL}[q(f|S)||q(f|S')] + C, \quad (4)$$

where $C = 2\log \frac{e^a + e^b}{2} - a - b$. We denote the upper bound by $\tilde{\mathcal{R}}(q(f|S), q(f|S'))$.

---

[1]With a probability space $(\mathcal{F}, \Sigma, q(\cdot|S))$ and measurable space $([a, b], T)$, suppose $\mathcal{L}_\mathcal{D} : \mathcal{F} \to [a, b]$ is a measurable function. Then, $\forall E \in T$, and we have $\mathcal{L}_\mathcal{D}^{-1}(E) \in \Sigma$; hence, there exists a probability distribution p such that $p(E) = q(\mathcal{L}_\mathcal{D}^{-1}(E)|S)$, where $\mathcal{L}_\mathcal{D}^{-1}$ is the inverse correspondence. Now, we have $\mathbb{E}_{q(f|S)}\left[e^{\mathcal{L}_\mathcal{D}(f)}\right] = \mathbb{E}_{p(\mathcal{L})}\left[e^{\mathcal{L}}\right]$, and we can use Jensen's inequality in the proofs of Lemma 1 and Theorem 1.

If the two posterior probabilities $q(f|S)$ and $q(f|S')$ share a common prior distribution, the following corollary holds:

Corollary 1. *Let $q(f|S)$ and $q(f|S')$ be the posteriors of f given S and S', respectively. We assume that a prior of $q(f|S)$ and that of $q(f|S')$ are the same probability distribution. Then, for any measurable function $\mathcal{L}_{\mathcal{D}}(f)$, the following inequality holds:*

$$\mathcal{R}(q(f|S), q(f|S')) \le D_{KL}\left[q(\mathbf{f}_{X_+}|S))||q(\mathbf{f}_{X_+}|S'))\right] + C. \tag{5}$$

Here, $\mathbf{f}_{X_+}$ is a random variable vector of the function values for inputs $X_+ := X \cup X'$, and the posterior distributions of $\mathbf{f}_{X_+}$ given S and S' are denoted by $q(\mathbf{f}_{X_+}|S)$ and $q(\mathbf{f}_{X_+}|S')$, respectively.

Corollary 1 implies that $\mathcal{R}(q(f|S), q(f|S'))$ can be evaluated by the upper bound, which is computable by using the observed data.

3.2 KL divergence between Gaussian processes

As a specific example, we consider active learning with the Gaussian process (GP) as the prediction function. We note that the proposed method for stopping active learning is applicable as long as we can estimate the KL divergence between posterior distributions. In the GP, the loss function is assumed to be the negative log likelihood of the Gaussian distribution, and the prior distribution can be obtained as

$$\begin{bmatrix} \mathbf{y} \\ f(x) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \mu(x) \end{bmatrix}, \begin{bmatrix} \mathbf{K} + \beta^{-1}\mathbf{I} & \mathbf{k}(x) \\ \mathbf{k}^{\mathrm{T}}(x) & k(x,x) \end{bmatrix}\right).$$

Let $S_t = \{(x_i, y_i)\}_{i=1}^t$ be the observed dataset of size t. Then, the posterior can be defined as

$$f(x)|x, S_t \sim \mathcal{N}(\mu(x), \sigma(x,x)), \tag{6}$$

where $\mu(x) = \mathbf{k}^{\mathrm{T}}(x)(\mathbf{K}+\beta^{-1}\mathbf{I})^{-1}\mathbf{y}$, $\sigma(x,x) = k(x,x) - \mathbf{k}^{\mathrm{T}}(x)(\mathbf{K}+\beta^{-1}\mathbf{I})^{-1}\mathbf{k}(x)$, and $\mathbf{y} = (y_1, y_2, \cdots, y_t) \in \mathbb{R}^t$.

Let $q(f|S_t)$ and $q(f|S_{t+1})$ be the posteriors of f given $S_t = \{(x_i, y_i)\}_{i=1}^t$ and $S_{t+1} = \{(x_i, y_i)\}_{i=1}^{t+1}$, respectively. We assume that the prior of $q(f|S_t)$ is the same as that of $q(f|S_{t+1})$. Then, it is easy to calculate $\mathcal{R}(q(f|S_t), q(f|S_{t+1}))$ in the case of a GP posterior. Let μ_t and σ_t be the mean and covariance functions of $q(f|S_{t+1})$, respectively. Then, from Corollary 1, the following equality holds:

$$\begin{aligned} &\tilde{\mathcal{R}}(q(f|S_t), q(f|S_{t+1})) \\ =&D_{KL}\left[\mathcal{N}(\mathbf{f}_{X_{t+1}}|\mu_t, \boldsymbol{\Sigma}_t)||\mathcal{N}(\mathbf{f}_{X_{t+1}}|\mu_{t+1}, \boldsymbol{\Sigma}_{t+1})\right] + C \\ =&\frac{1}{2}\beta\sigma_t(x_{t+1}, x_{t+1}) - \frac{1}{2}\log|1 + \beta\sigma_t(x_{t+1}, x_{t+1})| \\ &+ \frac{1}{2}\frac{\beta\sigma_t(x_{t+1}, x_{t+1})}{\sigma_t(x_{t+1}, x_{t+1}) + \beta^{-1}}(y_{t+1} - \mu_t(x_{t+1}))^2 + C. \end{aligned} \tag{7}$$

Details of the derivation is described in the supplementary material.

3.3 Conventional PAC-Bayesian bounds

In this section, we present the typical PAC-Bayesian bounds derived in the literature. The most famous bound is McAllester's bound (McAllester, 1999), which binds the expectation of the generalization error by the training dataset $S = \{(x_i, y_i)\}_{i=1}^t$ as follows:

$$\begin{aligned} \underset{S \sim D}{Pr}\Bigg(&\forall q: \underset{q(f)}{\mathbb{E}}\left[\mathcal{L}_{\mathcal{D}}(f)\right] \le \underset{q(f)}{\mathbb{E}}\left[\mathcal{L}_S(f)\right] \\ &+ \sqrt{\left(D_{KL}\left[q(f)||p(f)\right] + \log\left(2\sqrt{t}/\delta\right)\right)/2t}\Bigg) \ge 1 - \delta, \end{aligned} \tag{8}$$

where $\delta \in (0, 1]$ is the confidence parameter, and $p(f)$ and $q(f)$ are the prior and posterior distributions, respectively. McAllester's bound is only applicable to classification problems.

For regression problems, Alquier et al. (2016) derived the following bound:

$$\begin{aligned} \underset{S \sim D}{Pr}\Bigg(&\forall q: \underset{q(f)}{\mathbb{E}}\left[\mathcal{L}_{\mathcal{D}}(f)\right] \le \underset{q(f)}{\mathbb{E}}\left[\mathcal{L}_S(f)\right] \\ &+ t^{-1}\left(D_{KL}\left[q(f)||p(f)\right] - \log\delta\right) + 1/2(b-a)^2\Bigg) \ge 1 - \delta, \end{aligned} \tag{9}$$

in which the range of loss function l is restricted to $[a, b]$. Germain et al. (2016) proposed another bound that does not have any restrictions on the range of l; however, this bound is derived by assuming a specific form of the loss function.

The most notable difference between the existing and proposed bounds is that the proposed bound is a *deterministic* bound. We can guarantee a gap between the posterior distributions before and after adding a new sample without any confidence parameter. This reliability is a particularly important characteristic when the bound is to be applied to a measure of the stopping criterion.

We propose to determine whether to stop the learning or not by testing the convergence of sequence of $\mathcal{R}(q(f|S_t), q(f|S_{t+1}))$. Our aim is to develop a reliable criterion for stopping active learning in the framework of a statistical test. Ordinary PAC-Bayesian bounds have parameters δ (similar to $(\alpha, 1 - \beta)$, the significance level and power, for a statistical test). Since the confidence parameter appears both inside and outside of the probability function, it is not straightforward to cast the PAC-Bayesian bound in a standard statistical test framework. In contrast, using the proposed novel deterministic bound makes it possible to develop a tractable statistical test, as introduced in the next section.

4 Convergence test

4.1 Wald–Wolfowitz runs test

The Wald–Wolfowitz runs test (Wald and Wolfowitz, 1940) is a nonparametric test of the randomness hypothesis of a given binary sequence. It was originally proposed as a one-sample test but has been extended to two-sample tests by Barton and David (1957). Usually, the null hypothesis is set to be $H_0 : p(E_1, E_2, \cdots, E_T) = \Pi_{t=1}^{T} p(E_t)$, and under this hypothesis, we assume that the data at time t is $E_t \in \{0, 1\}$ and the probability that sequence $(E_1, E_2, \cdots, E_T)$ is generated is $p(E_1, E_2, \cdots, E_T)$. It is known that the power of the runs test is superior to that of the Kolmogorov–Smirnov test when the difference in location between the two sequences is small with a large difference in the variance (Magel and Wibowo, 1997).

In the runs test, the sequence of the same number (zero or one) is called the run, and the length of runs are treated as random variables. We denote the random variable for the total number of runs as U. Let t_0 and t_1 be the numbers of zeros and ones, respectively, and let T be the length of the sequence, that is, $T = t_0 + t_1$. We assume that $t_1 \geq t_0$. Then, the probability distribution of U under the null hypothesis is as follows:

$$p(U = 2t) = 2(_{t_0-1}C_{t-1}\,{}_{t_1-1}C_{t-1})/_T C_{t_0},$$
$$p(U = 2t + 1)$$
$$= (_{t_0-1}C_{t-1}\,{}_{t_1-1}C_{t-2} + _{t_0-1}C_{t-2}\,{}_{t_1-1}C_{t-1})/_T C_{t_0},$$

where $t = 1, 2, \cdots, t_0$. $p(U)$ is shown to be a normal distribution with average $\mu = 1 + (2t_0 t_1)/T$ and variance $\sigma^2 = 2t_0 t_1 (2t_0 t_1 - T)/(T^2(T-1))$. Then, the randomness of the sequential data is tested by using the test statistic $Z = (U - \mu)/\sigma$.

Remark 1. *It is known that the Wald–Wolfowitz test is reasonably powerful when the alternative hypothesis has a Markov property (David, 1947). Unfortunately, in our setting, this is not the case. In future work, we will investigate the condition in which the proposed test is the most powerful.*

4.2 Proposed method for stopping active learning

We describe a specific algorithm for stopping active learning with a GP. The proposed stopping criterion is also applicable to other active learning frameworks; however, here we simply select a new input datum based on the uncertainty sampling strategy. Various measures of uncertainty exist, and entropy is one of the reasonable measures (Settles, 2009). Because we consider the GP as a predictor, selecting a new input

---

Algorithm 1 Active learning with automatic termination by testing convergence of $\mathcal{R}$

Sample $S_1 = \{(x_1, y_1)\}$ and initialize $R = \{\}$
Calculate GP posterior $q_1(f|S_1)$
for $t = 1, 2, \ldots$ **do**
    Sample by maximum uncertainty AL
        $x_{t+1} = \arg \max_x \sigma(x, x)$
    Update dataset
        $S_{t+1} \leftarrow S_t \cup \{(x_{t+1}, y_{t+1})\}$
    Calculate GP posterior $q_{t+1}(f|S_{t+1})$
    Update sequence of upper bounds
        $r_t \leftarrow \tilde{\mathcal{R}}(q_t, q_{t+1}),\ \ R = R \cup \{r_t\}$
    Calculate median of R
        $m = \mathrm{median}(R)$
    Convert from upper bounds to binary
    Initialize $E = \{\}$
    for $i = 1, 2, \ldots, t$ **do**
        $e_i \leftarrow \mathrm{sgn}(r_i - m)$
        $E = E \cup e_i$
    end for
    Convergence test E by using runs test
    if runTest(E) **then break end if**
end for

---

datum with the maximum entropy is equivalent to selecting the point with the maximum variance:

$$x^* = \arg \max_x \sigma(x, x), \qquad (10)$$

where $\sigma(x, x)$ is the covariance function of $q(f|S_{t+1})$, which serves as an acquisition function in Eq. (1). Let R be the sequence of $\tilde{\mathcal{R}}(q(f|S_t), q(f|S_{t+1}))$, i.e., $R = \{r_1, r_2, \cdots, r_T\}$ and $r_t = \tilde{\mathcal{R}}(q(f|S_t), q(f|S_{t+1}))$. Because R is a sequence of continuous values, we cannot directly perform the runs test for R. In this work, following the work by Jani (2014), each r_t is converted to 1 if $r_t \geq \mathrm{median}(R)$ and 0 if $r_t < \mathrm{median}(R)$. Algorithm 1 summarizes the procedure explained above.

Finally, we consider the computational aspect. The proposed criterion requires computation of the mean function and covariance function of the posterior $q(f|S_t)$ for evaluating $\tilde{\mathcal{R}}(q_t, q_{t+1})$. However, the computation cost for evaluating $\tilde{\mathcal{R}}(q_t, q_{t+1})$ is $\mathcal{O}(1)$ since the mean function and covariance function are already calculated during exploration of new datum. Therefore, the overall computational cost is equal to that of the runs test.

5 Experimental results

This section describes the evaluation of the effectiveness of the proposed stopping criterion via a set of regression experiments with one artificial and five real-world datasets. The real-world datasets are obtained

from the UCI machine learning repository. Every feature of these datasets is normalized so that their means are zero and their standard deviations are one[2].

5.1 Evaluation measure

Let $q(f|S_T)$ be the posterior distribution of f obtained using the complete dataset $S_T = \{(X_T, Y_T)\}$, and let $q(f|S_t)$ be the distribution obtained using the dataset S_t of size t. For quantitative evaluation of the determined stopping time, we define the *optimal* stopping time t_{opt} as the minimum data size t that satisfies $\mathbb{E}_{q(f|S_t)}[\mathcal{L}_\mathcal{D}(f)] \leq \eta$, where η is a predefined threshold. For determining the threshold η for t_{opt}, we use the complete dataset of interest and resampled 50 points for training and 1950 points for test for the artificial dataset, and 100 points for training and the remaining points for test datasets for the real-world datasets 100 times. By using these 100 pairs of training and test datasets, we calculate the empirical estimate of the expected generalization errors, and η is set to be the average + 2 sd of the generalization errors. There are two possible approaches for active learning, aggressive and conservative (Bloodgood and Vijay-Shanker, 2009). Since, basically, the aim of active learning is to save the cost for annotation, in this work we adopt aggressive approach and η is set to be average +2sd.

Because we consider GP regression, the loss function is defined as $l(f, x, y) = \frac{\beta}{2}(y - f(x))^2 + \frac{1}{2}\log(\beta/2\pi)$. By denoting the posterior of f by $q(f|S_t) = \mathcal{N}(\mu_t, \sigma_t)$ and the test dataset by $S_{\tilde{T}} = \{(x_i, y_i)\}_{i=1}^{\tilde{T}}$, the posterior average of the expected loss can be approximated by

$$\mathop{\mathbb{E}}_{q(f|S_t)}[\mathcal{L}_\mathcal{D}(f)] \approx \mathop{\mathbb{E}}_{q(f|S_t)}[\mathcal{L}_{S_{\tilde{T}}}(f)]$$
$$= \frac{\beta}{2\tilde{T}}\left\{\sum_{i=1}^{\tilde{T}}(y_i - \mu_t(x_i))^2 + \mathrm{Tr}(\Sigma_t)\right\} + \frac{1}{2}\log\frac{\beta}{2\pi},$$

With a stopping time of t_* determined by a certain criterion, we consider

$$e_{\mathrm{stop}} := |t_* - t_{\mathrm{opt}}| \tag{11}$$

as a measure of goodness for the stopping criterion.

5.2 Dataset and methods for comparison

We considered a simple one-dimensional model

$$y_i = e^{-(x_i-2)^2/2} + e^{-(x_i-6)^2/10} + (x_i^2 + 1)^{-1} + \epsilon_i$$

---

[2]Simple Python implementation for our proposed method and competing methods have been submitted as the supplementary material and will be made publicly available after the review.

with the additive Gaussian observation noise $p(\epsilon) = \mathcal{N}(0, \beta^{-1})$. From this generative model, we sampled $1,000$ pairs of inputs and outputs (x_i, y_i), where x_i are uniform i.i.d. samples in $[-5, 15]$. Among the $1,000$ pairs, 950 pairs are retained as the test dataset and the remaining 50 pairs are pooled for training the prediction model via active learning. Independent sampling of size $1,000$ is repeated 100 times, and the average and standard deviation of e_{stop} defined in Eq. (11) are reported.

In the experiments, we compare the proposed criterion with the following four criteria:

(1) *PAC-Bayesian criterion*: The upper bound of the generalization error is approximated by using the conventional PAC-Bayesian result. By denoting the posterior of GP used in the training dataset S_t by $q(f|S_t)$, the following upper bound can be derived (Alquier et al., 2016):

$$\mathbb{E}_{q(f|S_t)}[\mathcal{L}_\mathcal{D}(f)]$$
$$\leq \mathbb{E}_{q(f|S_t)}[\mathcal{L}_{S_t}(f)] + t^{-1}D_{KL}[q(f|S_t)||p(f)]$$
$$- t^{-1}\log\delta + (b-a)^2/2 := a_t. \tag{12}$$

To stabilize the KL divergence, $\kappa\mathbf{I}$ is added to the covariance matrices of the prior and the posterior, where $\kappa = 0.01$. When a_t is smaller than a prespecified threshold, we terminate the active learning procedure. The confidence parameter δ is set to 0.01. a and b are set to $a = 0$ and $b = \max_{y \in Y_T} y - \min_{y \in Y_T} y$.

(2) *Ground truth*: We consider the convergence of $\mathcal{R}(p(f), q(f|S_t))$, which is approximated by the test set $S_{\tilde{T}}$ and denoted by $\mathcal{R}_{\mathrm{test}}(p(f), q(f|S_t))$. We stop learning when $\mathcal{R}_{\mathrm{test}}(p(f), q(f|S_t))$ is larger than a certain prespecified threshold.

(3) *Cross-validation criterion*: We divide the samples collected in the active learning process into training and test datasets and evaluate the expected generalization error by 5-fold cross validation. When the estimated generalization error is smaller than a predetermined threshold, the learning is stopped.

(4) *Maximum variance criterion*: For classification problems, stopping criteria based on the uncertainty of class assignment have been proposed (Zhu, 2007; Zhu et al., 2008a). Herein, we consider a regression counterpart. Because we use the GP as a predictor, the posterior variance can be used as a measure of uncertainty. When the variance of all the possible or pooled data is smaller than a certain predefined threshold, learning is stopped.

5.3 Parameter settings

As a prior for the GP, we use a GP with a Gaussian kernel $k(x, x') = \exp\left(-\frac{1}{2h^2}\|x - x'\|^2\right)$. The common parameters for all the methods are scale parameter h for the kernel and variance of observation noise β^{-1}. These parameters are determined by using the marginal likelihood maximization using training datasets. The complete training dataset is not available in practice, but our aim is to set these parameter values in an objective manner and enable fair comparison.

In practical applications, it is better to update the hyperparameter at every update of the GP model, but the assumption in Corollary 1 does not hold when we change the hyperparameter. Marginalizing w.r.t. the hyperparameter is one of the reasonable approach, but again whether the assumption in Corollary 1 holds for the marginalized KL is uncertain. There are two possible practical approaches for this problem. (1) To assume that the Corollary 1 approximately holds between different hyperparameters and calculate Eq. (7). This would be reasonable because sample is added only one by one. (2) To calculate the upper bound by using a new hyper parameter in each time. In this work, we keep using the common parameters for the sake of simplicity.

For the proposed method, we must specify the significance level α for the statistical test. The level is fixed such that the type-I error rate is 0.1%. The KL divergence between GP posteriors and the range of the cost function l are calculated in the same manner as for the *PAC-Bayesian criterion*. For the *ground truth*, for each dataset, we use the complete (training and test) dataset to perform bootstrap resampling 100 times and evaluate $\mathcal{R}_{\text{test}}(p(f), q(f|S_t))$; the threshold for $\mathcal{R}_{\text{test}}(p(f), q(f|S_t))$ is set to the average $- 2 \times$sd of the bootstrap samples of $\mathcal{R}_{\text{test}}(p(f), q(f|S_t))$. The other three methods also require thresholds for termination. Because there is no universally applicable and objective method for setting the threshold, we considered one dataset, "airfoil self-noise," as the reference dataset. In particular, we select the threshold minimizing Eq. (11) from a set of thresholds for each method. The set of thresholds is generated by sampling $10,000$ points at equal intervals within a range; the ranges is set to $[0.01, 100]$, $[0.001, 10]$, and $[0.0001, 1]$ for the *PAC-Bayesian criterion, cross-validation criterion* and *maximum variance criterion*, respectively.

It should be emphasized that the method of setting the threshold for the *ground truth* is not applicable in actual situations, which is the reason the method is called the *ground truth*. For the other three methods, because there exist no standard and objective threshold determination methods, we set the threshold values by using the reference dataset. In other words, these four criteria utilize the reference dataset, which is not available in practice, and the experimental setting is thus beneficial to them.

5.4 Results

The expected generalization error evaluated by the test dataset over the number of training samples is plotted in Fig. 1 for six datasets. From Fig. 1 (d), it can be noted that the stopping time determined by using the proposed method can be considerably different from t_{opt} in some cases, although the proposed method tends to terminate active learning when the training has been converged or is about to converge. The *max variance criterion* and *cross validation criterion* also offer stable and reasonable stopping times. The *PAC-Bayesian criterion* tends to either stop learning too early or overshoot the reasonable stopping time.

Table 1 summarizes the values of the average and standard error of e_{stop} calculated in 100 runs with independent resampling of the training and test datasets. Except for the *ground truth*, the proposed method achieved the smallest average error in four cases over the six datasets. The experimental results indicate that the proposed criterion can accurately and stably determine when to stop active learning without using the test dataset for evaluating the stopping criterion or determining the threshold.

We will add reasoning on the experimental performance for each compared methods: "PAC-Bayes" contains the training errors, and implicitly assumes the independence for training samples. Our bound does not contain training error term and remains valid in more general situation. We conjecture this is one of the reason why our method outperforms conventional PCA-Bayes. "CV" splits dataset so it uses less data for both training and validation. This would make the accuracy of the estimated test error low. Our method does not require a test data, leading to better performance. "Max variance" is directly connected to the acquisition function used for AL, while proposed method considers the divergence between posterior distributions of the predictive model. In certain ideal case, the proposed method could be considered as solely based on entropy, but before the convergence of the predictive model, both variance and mean largely affect the gap between before and after adding a new sample. This would be one of the reason why the max variance approach is nice but in many cases our method outperforms others.
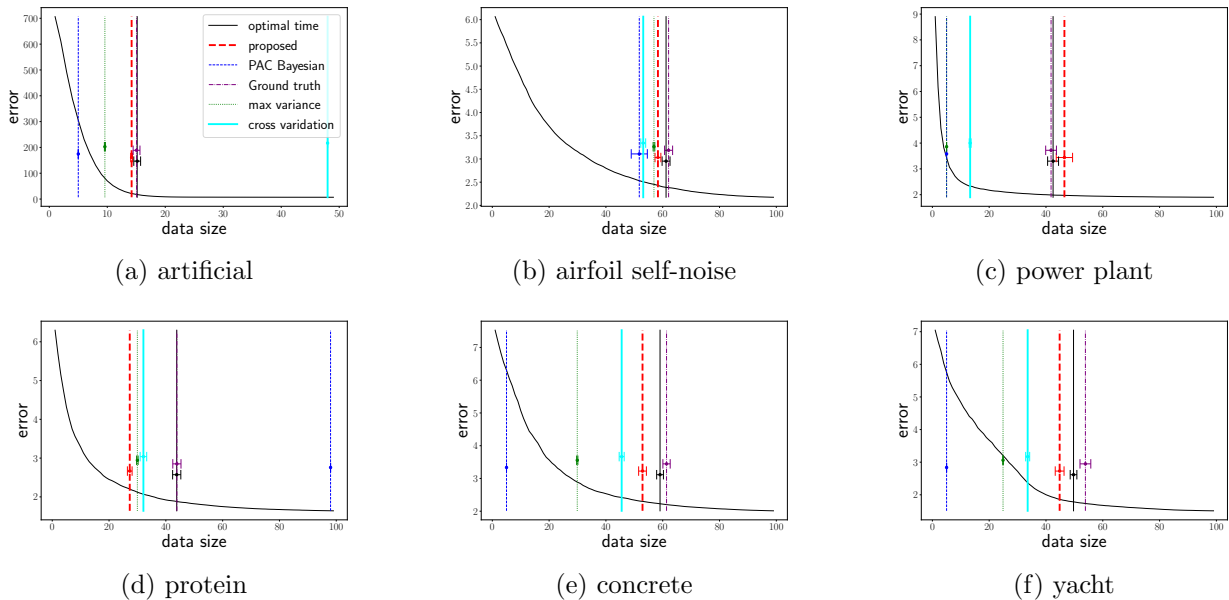
Figure 1: Averaged generalization error and training sample size. Vertical lines correspond to the optimal time and the stopping times determined by various methods.

Table 1: Average and standard error of e_{stop}.

	artificial	airfoil self-noise	power plant	protein	concrete	yacht
# of samples/features	2000/1	1503/6	10721/8	45730/9	1030/9	308/7
Ground truth	0.25 ± 0.05	1.81 ± 0.25	2.11 ± 0.39	0.06 ± 0.03	4.29 ± 0.58	15.5 ± 1.52
Proposed	$\mathbf{2.38 \pm 0.58}$	13.52 ± 1.05	$\mathbf{27.89 \pm 2.34}$	17.26 ± 1.5	$\mathbf{15.83 \pm 1.28}$	$\mathbf{16.33 \pm 1.28}$
PAC-Bayesian	10.16 ± 0.59	30.35 ± 1.8	37.5 ± 1.93	54.17 ± 1.44	54.06 ± 1.19	44.7 ± 1.18
Cross validation	32.84 ± 0.59	13.84 ± 1.39	29.2 ± 1.92	14.35 ± 1.21	17.09 ± 1.35	17.69 ± 1.42
Maximum variance	5.57 ± 0.61	$\mathbf{10.52 \pm 1.12}$	37.5 ± 1.93	$\mathbf{14.26 \pm 1.39}$	29.15 ± 1.25	24.82 ± 1.2

6   Conclusion

We proposed a criterion for stopping active learning based on the PAC-Bayesian theory and a runs test. A noteworthy fact regarding the proposed criterion is that the gap between the expected generalization errors w.r.t. the posterior distributions before and after adding a new sample is deterministically bounded. The criterion does not require the test dataset for evaluating the generalization error and enables stopping of active learning automatically in a statistically reliable manner. Moreover, although we concentrated on GP regression in this study, the criterion can be used for both classification and regression problems with an arbitrary cost function. In the experiments, the effectiveness of the criterion was demonstrated in the cases of both an artificial dataset and real-world datasets.

Our newly derived upper bound does not assume independence of observation, and applicable to any objective function or any posterior distribution of the predictor. Therefore, the proposed bound will be used for other learning frameworks such as Bayesian optimiza-

tion and online learning besides AL. The applicability of our new bound to other learning framework is one of the important future works.

References

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf.

Andreas Vlachos. A stopping criterion for active learning. *Comput. Speech Lang.*, 22(3):295–312, July 2008. ISSN 0885-2308. doi: 10.1016/j.csl.2007.

12.001. URL http://dx.doi.org/10.1016/j.csl.2007.12.001.

Seyda Ertekin, Jian Huang, and C. Lee Giles. Active learning for class imbalance problem. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 823–824, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277927. URL http://doi.acm.org/10.1145/1277741.1277927.

J. Paisley, X. Liao, and L. Carin. Active learning and basis selection for kernel-based linear models: A bayesian perspective. *IEEE Transactions on Signal Processing*, 58(5):2686–2700, May 2010. ISSN 1053-587X. doi: 10.1109/TSP.2010.2042491.

Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: An exploration-exploitation approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 449–456, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273553. URL http://doi.acm.org/10.1145/1273496.1273553.

David A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, Dec 1999. ISSN 1573-0565. doi: 10.1023/A:1007618624809. URL https://doi.org/10.1023/A:1007618624809.

David A. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, Apr 2003. ISSN 1573-0565. doi: 10.1023/A:1021840411064. URL https://doi.org/10.1023/A:1021840411064.

John Langford. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.*, 6:273–306, December 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1058111.

Olivier Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. 2007. ISBN 9780940600720. doi: 10.1214/074921707000000391.

A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *Ann. Math. Statist.*, 11(2):147–162, 06 1940. doi: 10.1214/aoms/1177731909. URL https://doi.org/10.1214/aoms/1177731909.

T. Kobayashi and M. Sugiyama. Early stopping heuristics in pool-based incremental active learning for least-squares probabilistic classifier. *IEICE Transactions on Information and Systems*, E95.D(8):2065–2073, 2012. doi: 10.1587/transinf.E95.D.2065.

S. Cohan and S. Naderiparizi. A survey on active learning. *Machine Learning II course*, 2018.

Jingbo Zhu. Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In *In Proceedings of ACL*, pages 783–790, 2007.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. Learning a stopping criterion for active learning for word sense disambiguation and text classification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008a. URL https://www.aclweb.org/anthology/I08-1048.

Jingbo Zhu, Huizhen Wang, and Eduard Hovy. Multi-criteria-based strategy to stop active learning for data annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1129–1136, Manchester, UK, August 2008b. Coling 2008 Organizing Committee. URL https://www.aclweb.org/anthology/C08-1142.

Florian Laws and Hinrich Schütze. Stopping criteria for active learning of named entity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 465–472, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 9781905593446. doi: 10.3115/1599081.1599140. URL http://portal.acm.org/citation.cfm?id=1599140.

Michael Bloodgood and K. Vijay-Shanker. A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 39–47, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W09-1107.

Michael Bloodgood and John Grothendieck. Analysis of stopping active learning based on stabilizing predictions. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 10–19, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W13-3502.

M. Altschuler and M. Bloodgood. Stopping active learning based on predicted change of f measure for text classification. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 47–54, Jan 2019. doi: 10.1109/ICOSC.2019.8665646.

Fredrik Olsson and Katrin Tomanek. An intrinsic stopping criterion for committee-based active learning. In *Proceedings of the Thirteenth Con-*

ference on Computational Natural Language Learning, CoNLL '09, pages 138–146, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-29-9. URL http://dl.acm.org/citation.cfm?id=1596374.1596398.

Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 839–846, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2. URL http://dl.acm.org/citation.cfm?id=645529.657802.

Lutz Prechelt. *Early Stopping — But When?*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_5. URL https://doi.org/10.1007/978-3-642-35289-8_5.

G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1318–1325, Sep. 2011. doi: 10.1109/Allerton.2011.6120320.

Wei Wang and Xue-feng Yan. Early stopping criterion combining probability density function with validation error for improving the generalization capability of the backpropagation neural network. *DEStech Transactions on Engineering and Technology Research*, 03 2018. doi: 10.12783/dtetr/amme2017/19507.

Romy Lorenz, Ricardo P Monti, Ines R Violante, Aldo A Faisal, Christoforos Anagnostopoulos, Robert Leech, and Giovanni Montana. Stopping criteria for boosting automatic experimental design using real-time fMRI with Bayesian optimization. pages 1–8, 2015. URL http://arxiv.org/abs/1511.07827.

Thomas Desautels, Andreas Krause, and Joel W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:4053–4103, 2014. URL http://jmlr.org/papers/v15/desautels14a.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, January 2016. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2946645.2946646.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning

problems. *J. Mach. Learn. Res.*, 7:1079–1105, December 2006. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248547.1248586.

Jean-Yves Audibert and Sébastien Bubeck. Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, page 13 p., Haifa, Israel, June 2010. URL https://hal-enpc.archives-ouvertes.fr/hal-00654404.

Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure exploration in infinitely-armed bandit models with fixed-confidence. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 3–24. PMLR, 07–09 Apr 2018. URL http://proceedings.mlr.press/v83/aziz18a.html.

Y. S. Chow and Herbert Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.*, 36(2):457–462, 04 1965. doi: 10.1214/aoms/1177700156. URL https://doi.org/10.1214/aoms/1177700156.

Lei Gong and James M. Flegal. A practical sequential stopping rule for high-dimensional markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 25(3):684–700, 2016. doi: 10.1080/10618600.2015.1044092. URL https://doi.org/10.1080/10618600.2015.1044092.

Galin L. Jones, Murali Haran, Brian S. Caffo, and Ronald Neath. Fixed-width output analysis for markov chain monte carlo. *Journal of the American Statistical Association*, 101(476):1537–1547, 2006. ISSN 01621459. URL http://www.jstor.org/stable/27639771.

Pascal Germain, Francis Bach, Alexandre Lacoste, and Simon Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 1884–1892, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157096.3157307.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(236):1–41, 2016. URL http://jmlr.org/papers/v17/15-290.html.

Arindam Banerjee. On bayesian bounds. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 81–88, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143855. URL http://doi.acm.org/10.1145/1143844.1143855.

D. E. Barton and F. N. David. Multiple runs. *Biometrika*, 44(1/2):168–178, 1957. ISSN 00063444.

Rhonda C. Magel and Sasmito H. Wibowo. Comparing the powers of the wald-wolfowitz and kolmogorov-smirnov tests. *Biometrical Journal*, 39(6):665–675, 1997. doi: 10.1002/bimj.4710390605.

F. N. David. A power function for tests of randomness in a sequence of alternatives. *Biometrika*, 34 (3-4):335–339, 12 1947. ISSN 0006-3444. doi: 10. 1093/biomet/34.3-4.335. URL https://doi.org/10.1093/biomet/34.3-4.335.

P.N. Jani. *Business Statistics: Theory and Applications*. PHI Learning, 2014.