Anytime-Valid Generalized Universal Inference on Risk Minimizers

Neil Dey ndey3@ncsu.edu
Ryan Martin rgmarti3@ncsu.edu
Jonathan P. Williams jwilli27@ncsu.edu
Department of Statistics
North Carolina State University
Raleigh, NC 27607-6698, USA

Abstract

A common goal in statistics and machine learning is estimation of unknowns. Point estimates alone are of little value without an accompanying measure of uncertainty, but traditional uncertainty quantification methods, such as confidence sets and p-values, often require strong distributional or structural assumptions that may not be justified in modern problems. The present paper considers a very common case in machine learning, where the quantity of interest is the minimizer of a given risk (expected loss) function. For such cases, we propose a generalized universal procedure for inference on risk minimizers that features a finite-sample, frequentist validity property under mild distributional assumptions. One version of the proposed procedure is shown to be anytime-valid in the sense that it maintains validity properties regardless of the stopping rule used for the data collection process. We show how this anytime-validity property offers protection against certain factors contributing to the replication crisis in science.

Keywords: e-process, e-value, empirical risk minimization, Gibbs posterior, learning rate, machine learning, replication crisis

1. Introduction

In statistics and machine learning applications, the goal is often to use available data to estimate an unknown quantity. To this end, sophisticated and computationally efficient procedures have been developed for estimating high- or even infinite-dimensional unknowns, with strong theoretical support in the form of asymptotic convergence rates. It is important, however, to accompany these estimates with an appropriate measure of uncertainty, typically in the form of a confidence set or in the form of p-values associated with a relevant hypotheses test. Ensuring the validity or reliability of one's uncertainty quantification is a major challenge, largely because the available theory generally requires distributional assumptions or structural simplifications that the data scientist is reluctant or unable to make. As such, there is a need for general strategies that provide provably valid uncertainty quantification in modern, high-dimensional problems.

This paper considers a broad class of statistical learning problems where the quantity of interest is defined as a *risk minimizer*, i.e., the minimizer of a risk (expected loss) function. This includes, for example, the typical regression and classification problems common in machine learning. It also covers the common cases in the statistics literature in which the quantity of interest is the parameter of either a well- or incorrectly-specified statistical

model. Much of the extant statistical literature on uncertainty quantification for risk minimizers comes from the broad area of robust statistics (e.g., Huber, 1981; Hampel et al., 2011), and in particular the well-studied area of M-estimation (e.g., Huber, 1981; Maronna et al., 2006; Boos and Stefanski, 2018). The classical results in this area, however, impose regularity conditions and achieve only asymptotically valid frequentist inference. More recent results relevant to this problem (e.g., Hudson et al., 2021; Cella and Martin, 2022) require fewer regularity conditions but still only yield asymptotic validity.

An important recent development was the so-called "universal inference" framework of Wasserman et al. (2020). They present simple and elegant procedures that offer provably valid uncertainty quantification (e.g., confidence sets and p-values) under virtually no conditions and without the need for asymptotic approximations. One of their main results is based on a clever use of data-splitting to construct a "split likelihood ratio" for which finite-sample distributional bounds on error rates can be obtained under no regularity conditions. These developments are limited, however, to settings in which a likelihood function is available—that is, in problems characterized by a correctly specified statistical model. This limitation is partially addressed by the results in Park et al. (2023) that allow for misspecification of the statistical model and construction of valid inference for the parameter value that minimizes the Kullback–Leibler divergence of the posited statistical model from the true data-generating distribution. However, as mentioned above, assuming a statistical model is a non-trivial restriction in many practical applications, so there is a practical need for provably valid methods beyond model-based settings. Our main contribution here is an extension of the developments in Wasserman et al. (2020) that covers many learning problems beyond those determined by a statistical model.

Our proposed *generalized universal inference* framework replaces the log-likelihood function in the model-based universal inference framework in Wasserman et al. (2020) with the empirical risk function, an essential ingredient in the learning problem. Wasserman et al.'s developments leaned on the simple, well-known fact that likelihood ratios have expected value 1. In our present context, however, there is no likelihood ratio and, moreover, no direct analogue of Wasserman et al.'s key property to apply. We overcome this obstacle by identifying a single regularity condition—namely, the "strong central condition" in van Erven et al. (2015); Grünwald and Mehta (2020), Syring and Martin (2023), etc.—from which we can show that Wasserman et al.'s (anytime-)validity property also holds for our proposed generalized universal inference, but in a much broader class of learning problems. Beyond (anytime-)validity, we also establish results concerning the efficiency of our proposal (e.g., its asymptotic power).

The remainder of this paper is organized as follows. After some more detailed background and problem setup in Section 2, we present our generalized universal inference framework in Section 3 and state its theoretical validity and efficiency properties. An important practical detail is the choice of a suitable *learning rate* parameter, and we present our recommended data-driven selection strategy in Section 4. Then, in Section 5, we compare our method to that of Waudby-Smith and Ramdas (2023), which is designed specifically for (anytime-valid) nonparametric inference on the mean of an unknown distribution, and demonstrate our method's superior efficiency. In Section 6, we present simulation studies that demonstrate the dual validity and efficiency of our proposed approach in a variety of challenging settings; in particular, we show how it addresses various

factors contributing to the *replication crisis* in science, as well as to the problem of estimating centroids in the K-means problem with unbalanced populations. In Section 7, we demonstrate how our approach performs in a classical real data example, namely, Millikan's experiment to measure the charge of an electron. We finish with concluding remarks in Section 8. Proofs of all the theorems can be found in the appendix. Code for reproducing the simulation experiments presented in this paper is available at `https://github.com/neil-dey/universal-inference`.

2. Problem setup and related work

Suppose that the observable data $Z^n := (Z_1, \ldots, Z_n)$ are i.i.d. from an unknown distribution $\mathcal{D}$ over a set $\mathbb{Z}$. A loss function $\ell : \Theta \times \mathbb{Z} \to \mathbb{R}^+$ is chosen by a practitioner that measures how well a parameter $\theta \in \Theta$ conforms with an observed data point $Z \in \mathbb{Z}$; small $\ell(\theta, z)$ values indicate greater conformity between z and θ. Write $R(\theta) := \mathrm{E}_{Z\sim\mathcal{D}}\{\ell(\theta; Z)\}$ for the *risk* or expected loss function. Our objective is to find the risk minimizer

$$\theta^* := \arg\min_{\theta\in\Theta} R(\theta).$$

This is impossible without knowledge of the distribution of $\mathcal{D}$, but we can estimate θ^* using the data Z^n from $\mathcal{D}$. To this end, the *empirical risk minimizer* (ERM) is

$$\widehat{\theta}_n := \arg\min_{\theta\in\Theta} \widehat{R}_n(\theta),$$

where $\widehat{R}_n(\theta) := n^{-1}\sum_{i=1}^n \ell(\theta; Z_i)$ is the *empirical risk* function. The intuition is that $\widehat{R}_n$ should be close to R, at least for large n, so the ERM $\widehat{\theta}_n$ should be close to θ^*.

A variety of approaches are available to quantify the uncertainty in $\widehat{\theta}_n$. As mentioned in Section 1, classical solutions offer asymptotic frequentist guarantees under rather strong regularity conditions. It is demonstrated in Wasserman et al. (2020), on the other hand, that with a well-specified model $\{P_\theta \mid \theta \in \Theta\}$ featuring a likelihood function $L(\theta; Z^n)$, one can obtain confidence sets for θ^* with no regularity conditions. One of their proposed strategies is *sample splitting*. That is, partition the sample Z^n into sub-samples $Z^{(1)}$ and $Z^{(2)}$ and compute the maximum likelihood estimator $\widehat{\theta}^{(1)}$ on $Z^{(1)}$; then a $1-\alpha$ level confidence set for θ^* is given by $\{\theta \in \Theta \mid T(\theta) \le \alpha^{-1}\}$, where

$$T(\theta) = T(\theta; Z^{(1)}, Z^{(2)}) := \frac{L(\widehat{\theta}^{(1)}; Z^{(2)})}{L(\theta; Z^{(2)})}$$

is called the "split likelihood-ratio" for obvious reasons. It is notable that $T(\theta)$ is an example of an *e-value*, defined by the property that $\mathrm{E}_{\theta^*}\{T(\theta^*)\} \le 1$, where E_{θ^*} denotes the expected value under the assumption that the data Z^n was generated from P_{θ^*}.

The notion of an e-value is classical, dating at least as far back as Wald (1945, 1947); however, e-values have become an object of great interest more recently (e.g., Vovk and Wang, 2021; Howard et al., 2021; Xu et al., 2021; Ramdas et al., 2023) for two reasons. First, the reciprocal of an e-value is a p-value (i.e., is stochastically no smaller than a uniform random variable) by Markov's inequality, so e-values can readily be used for uncertainty quantification. Second, e-values have several benefits as "measures of evidence"

over general p-values. For example, whereas it is not straightforward how to combine two p-values from independent tests, it is clear that that taking the product of independent e-values results in another valid e-value. Furthermore, this product of independent e-values remains an e-value under optional continuation—the practice of deciding whether or not to continue collecting new data and conducting further independent tests based on the outcomes of previous tests—and so has practical use in meta-analyses (Grünwald et al., 2023). Additionally, e-values also tend to be more robust to model misspecification and dependence compared to general p-values; e.g., it is proven in Wang and Ramdas (2022) that applying the Benjamini–Hochberg procedure to e-values maintains control over false discovery rates even for arbitrary dependence between the e-values, whereas the same cannot be said for p-values. However, e-values are not a direct upgrade to p-values: their safety guarantees imply that uncertainty quantification with e-values tends to be more conservative than that with p-values.

The e-value is closely related to the *e-process*, which is defined as a non-negative supermartingale $(E_n)_{n\in\mathbb{N}}$ such that $\mathrm{E}_{\theta^*}\{E_\tau\} \leq 1$ for any stopping time τ (Shafer et al., 2011; Ramdas et al., 2023; Ruf et al., 2023). It is clear that any stopped e-process is also an e-value and thus inherits the relevant benefits. Additionally, the definition of an e-process yields an "anytime-validity" property: If $(E_n)_{n\in\mathbb{N}}$ is an e-process, the reciprocal of $\max_{n=1,\ldots,\tau} E_n$ remains a p-value for any stopping time τ (Ramdas et al., 2023). That is, the sample size need not be fixed ahead of time, and one may even choose whether or not to collect more data based on what has been observed "up to that point." This is in stark contrast to a standard p-value, which generally depends on fixing a sample size ahead of time and prohibits any sort of data-snooping. Because peeking at the data to decide whether to stop or continue data-collection is common in science, the use of anytime-valid measures of evidence such as e-processes is highly desirable.

How does one construct an e-process? Like the e-value described above, these take the general form of likelihood ratios but with a sequential flavor (e.g., Wald, 1947, Eq. 10.10). A general proposal was given in Wasserman et al. (2020, Sec. 8) and particular instantiations have been put forward in, e.g., Gangrade et al. (2023) and Dixit and Martin (2023); see, also, the survey in Ramdas et al. (2023). In particular, as an alternative to sample splitting described above, consider lagged estimators

$$\widehat{\theta}_k = \arg\max_{\theta\in\Theta} L(\theta; Z^k), \quad k = 1, 2, \ldots, \quad \text{with } \widehat{\theta}_0 \text{ a fixed constant},$$

and the corresponding "running likelihood-ratio" test statistic

$$M_n(\theta) := \frac{\prod_{i=1}^n L(\widehat{\theta}_{i-1}; Z^i)}{\prod_{i=1}^n L(\theta; Z^i)}. \tag{1}$$

Then $\{M_n(\theta^*)\}_{n\in\mathbb{N}}$ is an e-process and, therefore, provides anytime-valid inference on θ^*.

3. Generalized universal inference

If the data-generating distribution $\mathcal{D}$ is unknown, or if the quantity of interest is not defined as the parameter that determines a statistical model (and is rather defined as the minimizer of a more general risk function), then the approach of Wasserman et al. (2020) is not directly

applicable. To deal with the general statistical learning problem, we propose the following generalized universal inference framework. To start, define the online *generalized universal e-value* (GUe-value, pronounced "gooey-value"):

$$G^{(\omega)}_{n,\text{on}}(\theta) := \exp\Big[-\omega \sum_{i=1}^{n} \{\ell(\widehat{\theta}_{i-1}; Z_i) - \ell(\theta; Z_i)\}\Big], \tag{2}$$

where $\widehat{\theta}_k$ is the ERM on the first k sample elements, with $\widehat{\theta}_0$ a pre-specified constant, and $\omega > 0$ is a *learning rate* discussed in detail in Section 4. The right-hand side of the above display is analogous to the running likelihood-ratio (1) in that it makes use of lagged ERMs, but it does not require a correctly specified likelihood.

The online GUe-value requires computation of n-many ERMs, which may be expensive. As an alternative, define the *offline* GUe-value

$$G^{(\omega)}_{n,\text{off}}(\theta) \equiv G^{(\omega)}_{S,\text{off}}(\theta) := \exp\Big[-\omega n_2 \{\widehat{R}_{S_2}(\widehat{\theta}_{S_1}) - \widehat{R}_{S_2}(\theta)\}\Big], \tag{3}$$

where $S_1 \sqcup S_2$ is a partition of the sample S into two sub-samples of size n_1 and n_2, respectively. Again, this is in analogy to the split likelihood-ratio of Wasserman et al. (2020). Because the online and offline GUe-values share many properties, we write $G^{(\omega)}_n$ when distinguishing between the two is unnecessary and refer to simply the GUe-value.

The intuition for the GUe-value is that $G^{(\omega)}_n(\theta)$ is large if and only if a suitable empirical risk function at θ is large, suggesting that θ is highly inconsistent with the data compared to the estimators. It is also interesting to note that the offline GUe-value can be written as the ratio of Gibbs posterior probability density functions (e.g., Zhang, 2006; Bissiri et al., 2016; Grünwald and Mehta, 2020; Martin and Syring, 2022) when using the (possibly improper) uniform prior. Hence, the offline GUe-value is analogous to the Bayes factor between θ and the ERM $\widehat{\theta}_{S_1}$. These intuitions thus suggest that the GUe-value at θ^* should be rather small, and that we ought to be able to say that only values θ such that $G^{(\omega)}_n(\theta)$ is sufficiently small could be plausible values for θ^*. It turns out that this intuition is indeed sound, as we explain below.

We should not refer to these as "e-values" or "e-processes" without first demonstrating that they satisfy the respective defining properties. Unlike in the context of a well-specified statistical model, it is not possible to do this demonstration without imposing certain mild conditions on the data-generating process $\mathcal{D}$ and the loss function ℓ. It turns out that the *strong central condition* advanced in van Erven et al. (2015), commonly used in the analysis of ERMs and Gibbs posteriors, is sufficient for our purposes as well.

Strong Central Condition. *A learning problem determined by a data-generating process $\mathcal{D}$ on $\mathbb{Z}$ and a loss function $\ell : \Theta \times \mathbb{Z} \to \mathbb{R}^+$ satisfies the strong central condition if there exists $\bar{\omega} > 0$ such that*

$$\mathrm{E}_{Z\sim\mathcal{D}} \exp\big[-\omega\{\ell(\theta; Z) - \ell(\theta^*; Z)\}\big] \leq 1 \quad \textit{for all } \theta \in \Theta \textit{ and all } \omega \in [0, \bar{\omega}).$$

The strong central condition is effectively a bound on the moment generating function of $\ell(\theta^*; Z) - \ell(\theta; Z)$ in a small positive interval $[0, \bar{\omega})$ containing the origin. As discussed in detail in van Erven et al. (2015), this condition holds in a number of practically relevant

cases; see, also, Grünwald and Mehta (2020). First, if the learning problem is determined by a well-specified statistical model, as in Wasserman et al. (2020) and many other papers, where the loss ℓ is the negative log-likelihood, then it follows from Hölder's inequality that the strong central condition holds with $\bar{\omega} = 1$. Even if the statistical model is incorrectly specified, under certain convexity conditions (e.g., Kleijn and van der Vaart, 2006; De Blasi and Walker, 2013; Ramamoorthi et al., 2015), one can often satisfy the strong central condition for some $\bar{\omega} < 1$; see, e.g., de Heide et al. (2020) for an application to misspecified generalized linear models. More generally, outside the context of a posited statistical model, the strong central condition holds for any bounded loss when the parameter space is convex. In particular, it holds for the L^p losses when the parameter space is bounded and convex; while the parameter space is rarely bounded in theory, it is typical for some reasonable bounds on the parameter space to exist in practice. For the particularly important special case of the L^2 loss, the relatively weak conditions necessary for the strong central condition to hold are discussed in van Erven et al. (2015); we provide an example of this in Appendix B where the strong central condition holds with the L^2 loss for i.i.d. data having finite third moment.

The following results demonstrate that, under the strong central condition, the online and offline GUe-values are e-processes and e-values, respectively.

Lemma 1. *Under the strong central condition, the online GUe-value $G^{(\omega)}_{n,on}(\theta^*)$ in (2) is an e-process for all sufficiently small $\omega > 0$.*

Proof We first show that $E_n := G^{(\omega)}_{n,\mathrm{on}}(\theta^*)$ is a non-negative supermartingale. For convenience of notation, define $\Delta_i := \ell(\widehat{\theta}_{i-1}; Z_i) - \ell(\theta^*; Z_i)$, for $i = 1, 2, \ldots$, where, again, $\widehat{\theta}_0$ is a fixed constant. Then

$$\begin{aligned}
\mathrm{E}(E_n \mid Z^{n-1}) &= \mathrm{E}\left\{ \exp\left(-\omega \sum_{i=1}^{n} \Delta_i \right) \,\middle|\, Z^{n-1} \right\} \\
&= \mathrm{E}\left\{ \exp\left(-\omega \sum_{i=1}^{n-1} \Delta_i \right) \cdot \exp(-\omega \Delta_n) \,\middle|\, Z^{n-1} \right\} \\
&= E_{n-1} \cdot \mathrm{E}\left\{ \exp(-\omega \Delta_n) \mid Z^{n-1} \right\},
\end{aligned}$$

where the last equality follows because $\sum_{i=1}^{n-1} \Delta_i$ is a measurable function of Z^{n-1}. Since Z_n and Z^{n-1} are independent and $\widehat{\theta}_{n-1}$ is a measurable function of Z^{n-1}, the latter conditional expectation in the above display can be re-expressed as

$$\mathrm{E} \exp\bigl[-\omega \{ \ell(\vartheta; Z) - \ell(\theta^*; Z) \} \bigr], \quad \text{for some fixed } \vartheta \in \Theta,$$

and is bounded by 1, by the strong central condition; thus, $E_n = G^{(\omega)}_{n,\mathrm{on}}(\theta^*)$ is a non-negative supermartingale. Finally, since

$$\mathrm{E}(E_1) = \mathrm{E} \exp\Bigl[-\omega \{ \ell(\widehat{\theta}_0; Z_1) - \ell(\theta^*; Z_1) \} \Bigr] \leq 1,$$

again by the strong central condition, it follows by a variant of the optional stopping theorem (e.g., Durrett, 2019, Theorem 4.8.4) that $E_n = G^{(\omega)}_{n,\mathrm{on}}(\theta^*)$ is an e-process. ■

Lemma 2. *Under the strong central condition, the offline GUe-value* $G^{(\omega)}_{n,off}(\theta^*)$ *in* (3) *is an e-value for sufficiently small* ω*.*

Proof For convenience of notation, define $\Delta_i := \ell(\widehat{\theta}_{S_1}; Z_i) - \ell(\theta^*; Z_i)$, for $i = 1, 2, \ldots, n_2$ where each $Z_i \in S_2$. Note that $\widehat{\theta}_{S_1}$ is a measurable function of S_1, so the strong central condition implies that

$$\mathrm{E}\{\exp(-\omega\Delta_i) \mid S_1\} \leq 1, \quad \text{for each } i = 1, 2, \ldots.$$

We hence have that

$$\mathrm{E}\{G^{(\omega)}_{S,\mathrm{off}}(\theta^*) \mid S_1\} = \mathrm{E}\left\{\exp\left(-\omega\sum_{i=1}^{n_2}\Delta_i\right)\,\middle|\, S_1\right\} = \prod_{i=1}^{n_2}\mathrm{E}\left\{\exp(-\omega\Delta_i)\,\middle|\, S_1\right\} \leq 1$$

since the Δ_i are independent given S_1. The law of iterated expectations gives

$$\mathrm{E}\{G^{(\omega)}_{S,\mathrm{off}}(\theta^*)\} = \mathrm{E}\,\mathrm{E}\{G^{(\omega)}_{S,\mathrm{off}}(\theta^*) \mid S_1\} \leq 1,$$

and so the offline GUe-value is indeed an e-value. ■

We can now begin to see the trade-off between the online and offline GUe-values: the online GUe-value is an e-process and hence has stronger error rate control properties, as described in our main result, Theorem 3. The offline GUe-value is only an e-value, so its properties are generally weaker (e.g., combining offline GUe-values only maintains validity under optional continuation for independent offline GUe-values, whereas combining online GUe-values can maintain the anytime-valid property even if the online GUe-values are dependent), but it is typically far less expensive to compute compared to the online variant that requires evaluation of the lagged ERMs.

Theorem 3. *Suppose that the strong central condition holds and take* $\omega > 0$ *sufficiently small. Fix a desired significance level* $\alpha \in (0, 1)$*. Then the test that rejects* $H_0 : \theta^* \in \Theta_0$ *in favor of* $H_1 : \theta^* \notin \Theta_0$ *if and only if*

$$G^{(\omega)}_n(\Theta_0) := \inf_{\theta\in\Theta_0} G^{(\omega)}_n(\theta) \geq \alpha^{-1},$$

controls the frequentist Type I error at level α*, i.e.,*

$$\Pr\{G^{(\omega)}_n(\Theta_0) \geq \alpha^{-1}\} \leq \alpha, \quad \textit{for all } \Theta_0 \textit{ that contain } \theta^*.$$

Also, the set estimator

$$C_\alpha(Z^n) := \left\{\theta \in \Theta : G^{(\omega)}_n(\theta) < \alpha^{-1}\right\}$$

has frequentist coverage probability at least $1 - \alpha$*, i.e.,*

$$\Pr\{C_\alpha(Z^n) \ni \theta^*\} \geq 1 - \alpha.$$

Furthermore, for the online GUe-value specifically, the above tests and confidence sets are anytime-valid, i.e., for any stopping time τ*,*

$$\Pr\{G^{(\omega)}_\tau(\Theta_0) \geq \alpha^{-1}\} \leq \alpha \quad \textit{and} \quad \Pr\{C_\alpha(Z^\tau) \ni \theta^*\} \geq 1 - \alpha.$$

Proof Since Θ_0 contains θ^*, it follows that $G_n^{(\omega)}(\Theta_0) \leq G_n^{(\omega)}(\theta^*)$. Then Markov's inequality and Theorem 2 gives

$$\Pr\{G_n^{(\omega)}(\Theta_0) \geq \alpha^{-1}\} \leq \Pr\{G_n^{(\omega)}(\theta^*) \geq \alpha^{-1}\} \leq \alpha \, \mathrm{E}\{G_n^{(\omega)}(\theta^*)\} \leq \alpha,$$

which proves the first claim. The coverage probability claim follows since $C_\alpha(Z^n) \not\ni \theta^*$ if and only if $G_n^{(\omega)}(\theta^*) \geq \alpha^{-1}$, and the latter event has probability $\leq \alpha$ as just shown. The final two claims follow from the same arguments given above, thanks to the fact that $G_n^{(\omega)}(\theta^*)$ is an e-process, as shown in Theorem 1. ■

Now the trade-off between the online and offline GUe-values is more clear. While both constructions lead to tests and confidence sets with finite-sample control of frequentist error rates, the online version is anytime-valid; i.e., the bounds hold uniformly over all stopping rules, but generally with a higher computational cost. The advantage of anytime-validity, again, is that the method is robust to the common practice of making within-study decisions about whether to proceed with further data collection and analysis.

The above results do not rely specifically on $\widehat{\theta}_{i-1}$ and $\widehat{\theta}_{S_1}$ being ERMs. Our primary motivation for choosing ERMs is for the sake of efficiency: ERMs are often consistent estimators of the risk minimizer, and this property leads to analogous large-sample consistency results for the above tests and confidence regions. The next two theorems present successively stronger results along these lines.

Theorem 4. *Suppose* $\sup_{\theta \in \Theta} |\widehat{R}_n(\theta) - R(\theta)| \xrightarrow{p} 0$ *as* $n \to \infty$. *Then*

$$\lim_{n \to \infty} \Pr\big\{G_{n,off}^{(\omega)}(\theta) \geq \alpha^{-1}\big\} = 1, \quad \textit{for any } \theta \textit{ with } R(\theta) > \inf_{\vartheta \in \Theta} R(\vartheta). \tag{4}$$

In addition, if the ERM mapping $z^k \mapsto \widehat{\theta}(z^k)$ *is leave-one-out stable in the sense that*

$$|\ell(\widehat{\theta}(z^{n-1}); z_n) - \ell(\widehat{\theta}(z^n); z_n)| = o(1), \quad \textit{for all } (z_1, z_2, \ldots) \in \mathbb{Z}^\infty, \tag{5}$$

then the consistency result (4) *also holds for the online GUe-value* $G_{n,on}^{(\omega)}$.

Theorem 4 is a statement that under regularity conditions, the power function for the GUe-value test of the point null $H_0 : \theta^* = \theta$ converges to 1 for each θ that is not a risk minimizer. Hence, for any non-risk minimizing θ, we see that the associated $(1-\alpha)$-level confidence set for θ^* shrinks to eventually exclude θ with high probability as more data is collected. The uniform convergence of the empirical risk to the risk is a standard condition, as it is sufficient for $\widehat{R}_n(\widehat{\theta}_n) \to R(\theta^*)$ in the first place. Similarly, it is typical to require some form of estimator stability in the online setting in order to learn θ^* (e.g., Bousquet and Elisseeff, 2002; Rakhlin et al., 2005; Shalev-Shwartz et al., 2010). For example, the ERM is leave-one-out stable in the sense of (5) [with rate n^{-1}, see (6)] if the loss function is smooth and strongly convex over Euclidean space (Zhang, 2023, Theorem 7.10).

Vanishing Type II error probability under fixed alternatives is a relatively weak property. A more refined analysis considers alternatives θ_n that are different from but converging to the risk minimizer. Then the relevant question is: How fast can the alternative θ_n converge to θ^* and still the GUe-value can distinguish the two? The following theorem gives an

answer to this question, effectively bounding the radius of the GUe-value confidence set: For the β defined below, the confidence set contains a point more than $\gtrsim n^{-\beta}$ distance away from θ^* with vanishing probability.

Theorem 5. *Fix $\beta \in (0,1)$ and let $(\theta_n)_{n\in\mathbb{N}}$ be a sequence in Θ such that $R(\theta_n) - \inf_\vartheta R(\vartheta) \gtrsim n^{-\beta}$. Then we have the following rate results for the GUe-value:*

1. *Suppose that $\sup_\theta |\widehat{R}_n(\theta) - R(\theta)| = o_p(n^{-\beta})$. If the ERM mapping $z^k \mapsto \widehat{\theta}(z^k)$ is leave-one-out stable at rate $n^{-\beta}$, i.e.,*

$$|\ell(\widehat{\theta}(z^{n-1}); z_n) - \ell(\widehat{\theta}(z^n); z_n)| = o(n^{-\beta}), \quad \text{for all } (z_1, z_2, \ldots) \in \mathbb{Z}^\infty, \tag{6}$$

   then $\lim_{n\to\infty} \Pr\{G^{(\omega)}_{n,on}(\theta_n) \geq \alpha^{-1}\} = 1$.

2. *Suppose that $\sup_\theta |\widehat{R}_{S_2}(\theta) - R(\theta)| = o_p(n_2^{-\beta})$. If $R(\widehat{\theta}_{S_1}) \xrightarrow{p} \inf_\vartheta R(\vartheta)$ as $n_1 \to \infty$, then*

$$\lim_{(n_1,n_2)\to(\infty,\infty)} \Pr\{G^{(\omega)}_{n,off}(\theta_{n_2}) \geq \alpha^{-1}\} = 1.$$

3. *Suppose that $\sup_\theta |\widehat{R}_{S_2}(\theta) - R(\theta)| = o_p(n_2^{-\beta})$. If $R(\widehat{\theta}_{S_1}) \xrightarrow{p} \inf_\vartheta R(\vartheta)$ as $n_1 \to \infty$ and $n_1 \lesssim n_2$, then*

$$\lim_{(n_1,n_2)\to(\infty,\infty)} \Pr\left\{G^{(\omega)}_{S,off}(\theta_n) \geq \alpha^{-1}\right\} = 1.$$

Note that $R(\widehat{\theta}_{S_1}) \xrightarrow{p} \inf_\vartheta R(\vartheta)$ as $n_1 \to \infty$ holds if S_1 is identically distributed to S_2 and that $n_1 \lesssim n_2$ holds if the sample splitting occurs with a constant proportion. Thus, Theorem 5 simply states that for our power function to exhibit desirable behavior, we only require uniform convergence of the empirical risk to the risk at a reasonable rate (and, in the offline case, for our split between training and validation sub-samples to be done at random). Furthermore, the size of our confidence set decays at the same rate as the ERM converges to the infimum risk.

The rate requirement of Theorem 5 is far from restrictive: a rate of about $o_p(n^{-1/2})$ is fairly typical. As a concrete example, suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random vectors from any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, let $h : \mathcal{X} \times \Theta \to \{0,1\}$ be any measurable function, and consider the zero-one loss function $\ell\{\theta; (x,y)\} = \mathbb{I}\{y \neq h(x;\theta)\}$. If the set $\{x \mapsto h(x;\theta) \mid \theta \in \Theta\}$ is of finite VC dimension (e.g., the set is a subset of a finite-dimensional affine space), it follows from Corollary 3 of Hanneke (2016) that $\sup_\theta |\widehat{R}_n(\theta) - R(\theta)| = o_p(n^{-\beta})$ for any $\beta < 1/2$, so we see that the rate condition of the theorem holds.

We finally note that although confidence sets only make sense when the risk minimizer θ^* exists, Theorems 4 and 5 apply even if $\inf_\vartheta R(\vartheta)$ is never attained. Two instances where the infimum risk fails to be attained include models where the parameter space is not compact (such as when θ represents a variance component that lies in $(0,\infty)$) and those that use risk functions that are non-coercive (such as the cross-entropy loss). Indeed, the most common example where the infimum fails to be attained is when θ denotes the parameter in logistic regression and the population is separated—i.e. for the population P, where $P \subseteq \mathbb{R}^p \times \{0,1\}$, there exists $\beta \in \mathbb{R}^p$ such that for any $(x,y) \in P$, we have that $\beta^\top x > 0$ implies $y = 1$ and $\beta^\top x < 0$ implies $y = 0$—as separation forces at least one component of θ^* to be infinite.

Even in such cases, the theorems guarantee that the GUe-value grows large on all of Θ; consequently, the corresponding confidence sets shrink to the empty set as more data are collected. This may indicate to the practitioner that their statistical learning problem is ill-posed, if they were not aware of this already.

4. Choice of the learning rate

The choice of learning rate ω is critical for the validity and efficiency of the GUe-value hypothesis tests and confidence sets: If ω is too large, then $G_n^{(\omega)}(\theta^*)$ is smaller than it should be, the confidence sets are likewise too small and are likely to undercover. On the other hand, if ω is too small, the confidence sets for θ^* are larger than necessary, resulting in inference that is overly conservative.

We have previously noted that the offline GUe-value can be written as a ratio of Gibbs posterior densities. Hence, it is not unreasonable to apply the same learning rate selection methods used to choose a Gibbs posterior learning rate. Some proposals for choosing a Gibbs posterior learning rate include the unit information loss approach of Bissiri et al. (2016), the matching information gain idea of Holmes and Walker (2017), the asymptotic Fisher information matching approach of Lyddon et al. (2019), the R-Safe Bayes algorithm of Grünwald and van Ommen (2017), and a sample-splitting strategy of Perrotta (2020), among others. However, it is found in Wu and Martin (2023) that with the learning rate chosen according to these strategies, the corresponding Gibbs posterior credible sets generally fail to achieve the nominal frequentist coverage probability. They do, however, identify one algorithm that maintains valid frequentist coverage even under model misspecification: the general posterior calibration (GPC) algorithm of Syring and Martin (2019); see, also, Martin and Syring (2022).

The GPC algorithm proceeds by constructing a $1-\alpha$ level confidence set for θ^* using the nonparametric bootstrap, resampling from the original sample S and choosing ω such that the credal set contains $\widehat{\theta}_S$ with probability $1-\alpha$ over the bootstrap resamples. Given that the GPC algorithm does well in attaining valid confidence sets from the Gibbs posterior, it is sensible to similarly use the nonparametric bootstrap to select the learning rate for the GUe-value. This nonparametric bootstrap approach is detailed in Algorithm 1.

Because the nonparametric bootstrap chooses an appropriate learning rate in a principled manner agnostic to how the data are distributed, it still tends to be conservative in general, choosing smaller learning rates than necessary. If one is sure that the data come from a particular parametric model, one may obtain less conservative choices of learning rates by instead employing the parametric bootstrap to choose a learning rate—at the expense of possibly having confidence sets with below-nominal-level coverage if the model is actually misspecified. This algorithm is presented as Algorithm 2.

One can also often obtain exactly correct choices for the learning rate under certain distributional assumptions for fixed loss functions. Appendix B presents such results for the L^2 loss function for the mean of a random variable, given that it satisfies the strong central condition. In particular, we demonstrate that for Gaussian distributed data, one can theoretically calculate a learning rate for the GUe confidence set that obtains exactly the correct coverage; furthermore, this learning rate asymptotically yields the correct coverage rate even for non-Gaussian data, due to the central limit theorem. We also discuss the

Algorithm 1 Nonparametric Bootstrap for Learning Rate Calibration

Require: $(z_1, \ldots, z_n)$, collected data
Require: Ω, a set of candidate learning rates
Require: α, a significance level to calibrate to
Require: N, the number of bootstrap iterations to do
  Compute $\widehat{\theta}$, the ERM for $(z_1, \ldots, z_n)$
  coverages$(\omega) \leftarrow 0$ **for all** $\omega \in \Omega$
  for $\omega \in \Omega$ **do**
    for i in $1, \ldots, N$ **do**
      Draw $S_B = (z_{b(1)}, \ldots, z_{b(n)})$ uniformly from $(z_1, \ldots, z_n)$
      if $G^{(\omega)}_{S_B}(\widehat{\theta}) < 1/\alpha$ **then**
        coverages$(\omega) \leftarrow$ coverages$(\omega) + 1/N$
      end if
    end for
  end for
  return arg $\min_{\omega \in \Omega} |\text{coverages}(\omega) - (1 - \alpha)|$

Algorithm 2 Parametric Bootstrap for Learning Rate Calibration

Require: $\{\mathcal{D}_\theta\}_{\theta \in \Theta}$, a family of parametric distributions
Require: $(z_1, \ldots, z_n)$, collected data from $\mathcal{D}_\theta$ for some θ
Require: Ω, a set of candidate learning rates
Require: α, a significance level to calibrate to
Require: N, the number of bootstrap iterations to do
  Compute $\widehat{D}$, the best-fitting distribution for $(z_1, \ldots, z_n)$ from $\{\mathcal{D}_\theta\}_{\theta \in \Theta}$
  Compute $\widehat{\theta}$, the ERM for $(z_1, \ldots, z_n)$
  coverages$(\omega) \leftarrow 0$ **for all** $\omega \in \Omega$
  for $\omega \in \Omega$ **do**
    for i in $1, \ldots, N$ **do**
      Draw $S_B = (z_{b(1)}, \ldots, z_{b(n)}) \sim \widehat{D}^n$
      if $G^{(\omega)}_{S_B}(\widehat{\theta}) < 1/\alpha$ **then**
        coverages$(\omega) \leftarrow$ coverages$(\omega) + 1/N$
      end if
    end for
  end for
  return arg $\min_{\omega \in \Omega} |\text{coverages}(\omega) - (1 - \alpha)|$

existence of a learning rate that obtains at least the nominal coverage if the data are indeed i.i.d. and one either knows or has good estimates for the second and third moments of the population. We acknowledge that these results are narrow in scope—either requiring strong distributional assumptions or once again relying on asymptotics rather than guaranteeing finite-sample validity—but they provide a useful starting point for generalizing the theory of learning rate selection in future work.

5. Comparison to existing methods

Another methodology that shares similar aims as our GUe confidence sets is given by the *predictable plugin empirical Bernstein* (PrPl-EB) confidence sets of Waudby-Smith and Ramdas (2023), which are also safe confidence sets due to e-process properties, but are limited to estimating the mean of a bounded random variable. Given a sample $Z_1, \ldots, Z_n$, the $1-\alpha$ level PrPl-EB confidence interval is given by $\bigcap_{t=1}^{n} C_t$, where

$$
\begin{aligned}
C_t &:= \left(\frac{\sum_{i=1}^{t} \lambda_i Z_i}{\sum_{i=1}^{t} \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^{t} (Z_i - \widehat{\mu}_{i-1})^2 (-\log(1-\lambda_i) - \lambda_i)}{\sum_{i=1}^{t} \lambda_i} \right) \\
\lambda_t &:= \min \left(c, \sqrt{\frac{2\log(2/\alpha)}{\widehat{\sigma}_{t-1}^2 t \log(1+t)}} \right) \\
\widehat{\sigma}_t^2 &:= \frac{1/4 + \sum_{i=1}^{t} (Z_i - \widehat{\mu}_i)^2}{t+1} \\
\widehat{\mu}_t &:= \frac{1/2 + \sum_{i=1}^{t} Z_i}{t+1},
\end{aligned}
$$

and c is any reasonable value in $(0,1)$—we follow the authors' recommendation of $c = 1/2$. The width of the PrPl-EB confidence interval in the i.i.d. setting scales with the true (unknown) standard deviation, and thus obtains reasonable coverages at large samples. However, as Figure 20 of Waudby-Smith and Ramdas (2023) shows, for modest sample sizes, the PrPl-EB confidence interval tends to cover almost the entirety of the support.

In Figure 1, we compare the coverage of the PrPl-EB confidence set and the GUe confidence sets on an i.i.d. sample of size 10 from the Beta distribution, where the learning rate for the GUe confidence sets were chosen via the parametric bootstrap (with both a correctly specified and a misspecified parametric model). In agreement with the findings of Waudby-Smith and Ramdas (2023), the PrPl-EB confidence set covers the entire interval $[0,1]$ at such a small sample size, whereas the GUe confidence sets attain essentially exactly the correct coverage in the correctly-specified and misspecified settings. We also examine in Figure 2 the coverage of the offline GUe confidence set when using the learning rate suggested by equation (10) (i.e., the learning rate that yields asymptotically correct coverage discussed in Appendix B), as well as when dividing this learning rate by two (which is our suggestion to ensure correct coverage at finite sample sizes). Note that although equation (10) requires either Gaussian data or a large enough sample size for the central limit theorem to apply (neither of which is true in this case), the GUe confidence sets from this approach are still approximately calibrated—though, there is some undercoverage. On the other hand, our suggested heuristic of halving this learning rate is more than conservative enough to hit the nominal coverage in these examples.

We emphasize that in this example, our comparison uses i.i.d. data and both the PrPl-EB and GUe approaches use the full dataset to create a single confidence set, as opposed to creating the so-called "confidence sequences" examined in Ramdas et al. (2023), in which data are added one point at a time. Furthermore, we note that we are using the data to estimate the learning rate for the online GUe-value—a methodology which we have not proven to preserve the e-process properties that the PrPl-EB method provably has (though the results here and in Section 6 give empirical evidence that anytime-validity is indeed

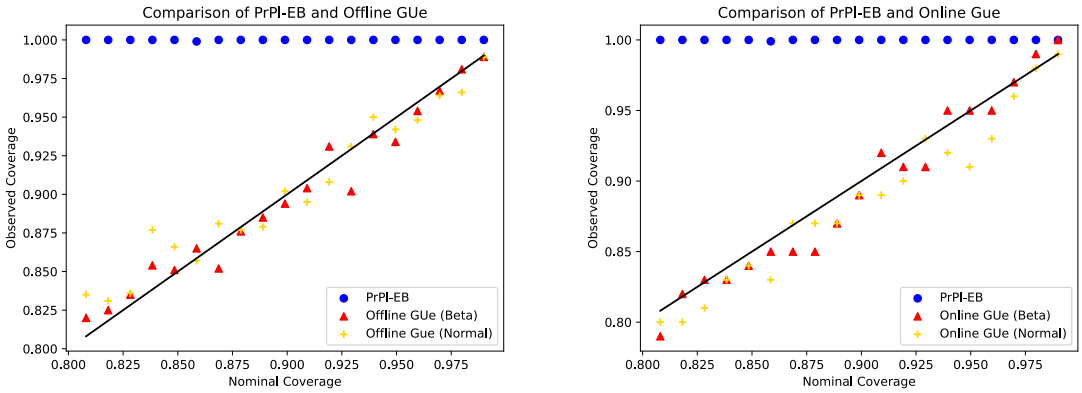

Figure 1: Nominal vs observed coverage of the PrPl-EB and GUe confidence sets on i.i.d. Beta$(5, 2)$ data. The learning rate for GUe was chosen via the parametric bootstrap, once correctly specifying the Beta model and once misspecifying a Gaussian model.

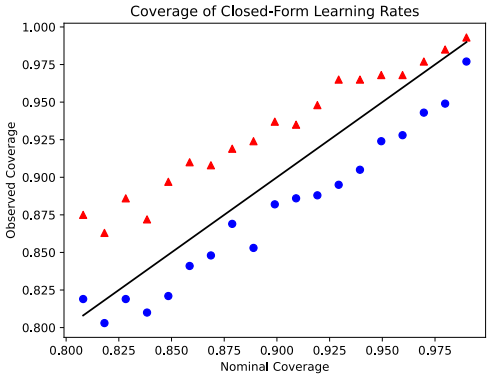

Figure 2: Nominal vs observed coverage of the offline GUe confidence sets on i.i.d. Beta$(5, 2)$ data. Blue circles indicate coverages when the learning rate is taken from equation (10), and red triangles are those when the learning rate is taken as half the value from equation (10).

maintained). The take-away message is that our online GUe-value solution, with data-driven choice of learning rate, is a powerful and promising alternative to PrPl-EB. We hope this will motivate further investigation into the interplay of data-driven learning rate choices and the desired anytime-validity property.

6. Simulation studies

6.1 Replication crisis-related applications

The replication crisis in science is a problem that has received significant attention in recent years. In this subsection, we showcase a variety of common problems that facilitate the lack of replicability of scientific experiments, and we demonstrate how these problems are mitigated by our GUe-value proposals.

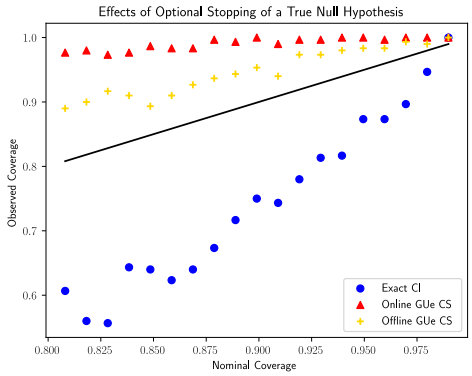

Figure 3: Nominal vs observed coverage of the "exact" and GUe confidence sets when collecting data until the null hypothesis $H_0 : (\mu_1 + \mu_2)/2 = 0$ is rejected for $\frac{1}{2}N(\mu_1, \sigma^2) + \frac{1}{2}N(\mu_2, \sigma^2)$ data, with $\mu_1 = 5$, $\mu_2 = 10$, and $\sigma^2 = 10^4$

Example 1. Consider the following simple setup: A scientist is studying two populations distributed as $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, and wants to estimate the best threshold $\mu = (\mu_1 + \mu_2)/2$ that separates the two populations.

If the scientist does everything correctly—collecting a single data set of independent observations of a fixed, predetermined sample size, then generates a confidence interval from this data—it is no surprise that the confidence interval works as planned: For any nominal coverage level, the practitioner shall observe precisely that level of coverage. However, this is not the reality for many scientists. What often occurs is that the scientist has a null hypothesis $H_0 : \mu = 0$ and an alternative hypothesis $H_1 : \mu \neq 0$, and funding or publication hinges on the null hypothesis being rejected. Thus, especially when gathering data is expensive, the scientist may choose to gather more data when the data set collected so far fails to reject the null, and stops collecting data once the null is rejected. Figure 3 demonstrates the effects of such a stopping rule: The confidence intervals generated by the scientist tend to be less than the nominal coverage. On the other hand, thanks to the anytime-validity property of e-processes, the online GUe confidence sets with the learning rate chosen via the nonparametric bootstrap consistently exhibit coverage above the nominal level. Moreover and quite interestingly, even though the offline GUe-value is not provably an e-process, it too exhibits coverage only slightly above the nominal level.

In some sense, the setting described in Example 1 is the "best case" scenario, where the practitioner gathers data until a null hypothesis is *correctly* rejected; a meta-analysis of replication studies could plausibly correct this issue. But what happens when publications in the literature only present *false* rejections of a null hypothesis?

Example 2. Consider what happens in the same setting as Example 1 when the null hypothesis is of the form $H_0 : \mu \geq c$ for some c, and it is indeed the case that $\mu \geq c$. Due to the difficulty in publishing negative results, the only data sets present in the literature will be those that falsely reject this null hypothesis, and the coverage of these intervals is shown in Figure 4 to be essentially zero for nearly all levels of nominal coverage. No meta-analysis can possibly correct for this issue, as all published data are simply biased

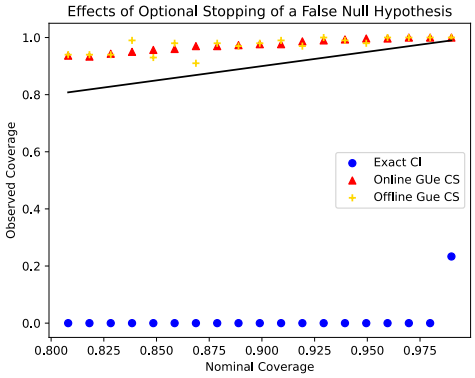

Figure 4: Nominal versus observed coverage of the "exact" and GUe confidence sets when only considering data where the null hypothesis $H_0 : (\mu_1 + \mu_2)/2 \geq -10$ is falsely rejected for $\frac{1}{2}N(\mu_1, \sigma^2) + \frac{1}{2}N(\mu_2, \sigma^2)$ data, with $\mu_1 = 5$, $\mu_2 = 10$, and $\sigma^2 = 10^4$.

towards the incorrect alternative hypothesis. However, Figure 4 clearly demonstrates that the GUe confidence set remains valid at all nominal coverage levels, even to the extent that the empirical coverage nearly matches nominal coverage at all relevant levels.

Example 3. Another common way for science to fail to be replicated is due to the unjustifiable removal of outliers. Doing so can significantly reduce the standard errors and may appear to be justifiable—after all, one should surely remove data that are the supposed result of experimental error. To illustrate the effects of cherry-picking data in this way, we simulate data that have "outliers" removed using Tukey's fences criterion for outliers with $k = 1$. Figure 5 demonstrates the effects of unwarranted removal of outliers on the validity of confidence sets for $N(0, 1)$ and $\text{Beta}(5, 2)$ data. As usual, our proposed confidence sets (with learning rates chosen via nonparametric bootstrap) maintain the correct level of coverage[1], whereas the "exact" confidence intervals fail to attain the nominal coverage level.

6.2 K-means

We now provide an illustration of the utility of the GUe-value even in the ideal sampling scenario of collecting an i.i.d. sample of fixed size. Consider the K-means algorithm, an unsupervised learning method that clusters data $Z_1, \ldots, Z_n$ into K clusters, with K fixed in advance, where each cluster has minimum within-cluster variance. Specifically, K-means aims to find a partition $\theta = (\theta_1, \ldots, \theta_K)$ of the data, where $\theta_k \subseteq \{1, \ldots, n\}$ for each

1. A caveat to the use of GUe confidence sets for this purpose is that rather than the strong central condition holding for the data-generating distribution, it must hold for the cherry-picked distribution. While removing all outliers above and below the mean as done in this experiment yields a distribution that satisfies the strong central condition, only removing outliers on one side (e.g., only outliers that are "too high") would typically cause the strong central condition to fail.

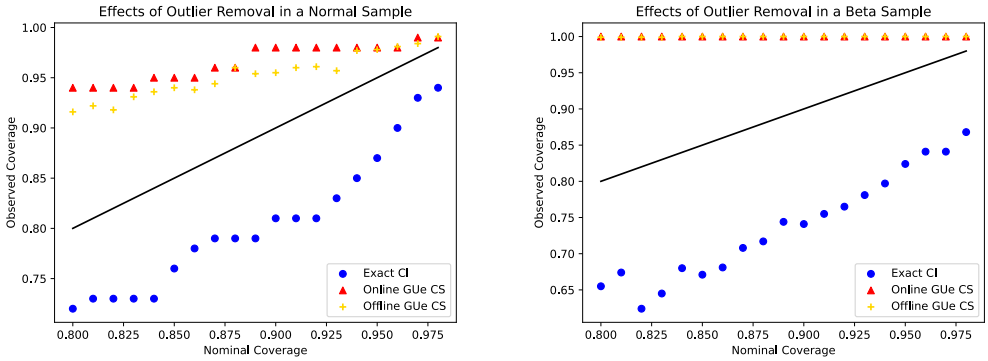

Figure 5: Coverage of the mean of $N(0,1)$ and Beta$(5,2)$ data when outliers are removed via the Tukey criterion ($k = 1$).

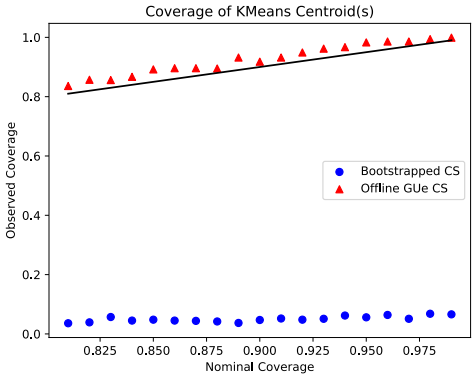

Figure 6: Coverage of μ_3 for the bootstrapped confidence set versus joint coverage of (μ_1, μ_2, μ_3) for the offline GUe confidence set, $\omega = 30$ fixed.

$k = 1, \ldots, K$, that minimizes

$$\ell(\theta; Z^n) = \sum_{k=1}^{K} |\theta_k| \ \mathrm{Var}(\{Z_i : i \in \theta_k\}),$$

where $|A|$ denotes the cardinality of the set A and the Var operator returns the sample variance of its data-set-valued argument; note that we can define $\mathrm{Var}(\varnothing)$ arbitrarily, here, since multiplying by the cardinality [of $\varnothing$] eliminates the dependence on this arbitrary choice. This partition θ implicitly defines the centroids $\mu_1, \ldots, \mu_K$, where $\mu_k = |\theta_k|^{-1} \sum_{z \in \theta_k} z$. These centroids are typically the quantities of interest.

We generate bivariate normal data from $K = 3$ populations, $N_2(\mu_k, \sigma^2 I)$, where $\sigma^2 = 0.01$ and $\mu_1 = (1, 0)^\top$, $\mu_2 = (-1/2, \sqrt{3}/2)^\top$, and $\mu_3 = (-1/2, -\sqrt{3}/2)^\top$. Then the true centroids for K-means with $K = 3$ are approximately the means of each population.

One commonly-used method to construct approximate confidence sets for these centroids is via bootstrapping (Hofmans et al., 2015). That is, one resamples from the observed data set that has the estimate $\widehat{\mu}$ for the centroid, performs K-means again on the bootstrapped

data, and creates an ellipse with major and minor axes based on the covariance matrix necessary for the ellipse to contain the $\widehat{\mu}$ with the nominal level of coverage over the bootstrap resamples. Here, we compare this procedure for uncertainty quantification about the K-means centroids to our proposed generalized universal inference framework.

If we draw 100 samples with equal probability from the three populations, then these bootstrapped confidence sets indeed attain approximately the correct nominal coverage. However, this changes when the populations are unbalanced. Figure 6 illustrates that when the populations are sampled from with probabilities 0.96, 0.03, and 0.01, respectively, bootstrapping leads to abysmal coverage for the least frequent population centroid (and thus would perform even worse if it were used to create a joint confidence set for all three centroids), whereas the offline GUe (joint) confidence set for (μ_1, μ_2, μ_3) with a choice of $\omega = 30$ has essentially the correct level of coverage. Note that because of the computational burden of the K-means algorithm, we elected to fix $\omega = 30$ to speed up computation rather than use bootstrapping on every iteration; this learning rate was chosen after inspecting the range of learning rates that the nonparametric bootstrap approach to choosing ω had proposed after a few iterations. We also do not examine the performance of the online GUe—again due to the computational burden of having to run K-means n times—though we would expect similar, if slightly more conservative, results.

7. Millikan's electron charge study, revisited

The first experiment done to measure the charge on an electron was by Millikan (1913). About this experiment, Feynman (1974) noted the following:

> "Millikan measured the charge on an electron by an experiment with falling oil drops and got an answer which we now know not to be quite right... It's interesting to look at the history of measurements of the charge of the electron, after Millikan. If you plot them as a function of time, you find that one is a little bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher. Why didn't they discover that the new number was higher right away? It's a thing that scientists are ashamed of—this history—because it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off..."

Indeed, the charge of an electron is now known to be exactly 160.2176634 zC, whereas Millikan's experiment yielded a point estimate of 159.2 zC with standard error 0.07 zC. Thus, Millikan's point estimate was roughly 14 standard errors below the true value—in part due to Millikan's cherry-picking of data to artificially exclude data points he deemed to be outliers, much like in our Example 3.

Follow-up papers that attempted to calculate the charge of an electron include Wadlund (1928) at 159.24 zC, Bäcklin (1929) at 159.88 zC, and Bearden (1931) at 160.31 zC. After

the estimate of Bearden (1931), the timeline of results reported by Hill (2021) suggests that later estimates all tended to fall quite close to the true value of about 160.2 zC.

To see how the GUe-value applies to Millikan's oil drop experiment, we use the non-parametric bootstrap to choose the learning rate for the GUe confidence set; the offline GUe-value chooses $\omega \approx 0.1$ at all relevant significance levels, whereas the online GUe-value chooses learning rates in the interval $[0.03, 0.12]$ that steadily decrease as α increases. We find that the offline GUe confidence set from Millikan's cherry-picked data covers the true value of the charge of an electron until $\alpha \approx 0.32$, and the online GUe confidence set covers the true value until $\alpha \approx 0.44$. Had the uncertainty in measurement of the charge of the electron been calculated via the GUe-value, it is possible that chemists would have converged to the correct value faster than the multiple decades it took in real life, due to no longer being constrained by the too-narrow confidence interval generated by Millikan's cherry-picked data.

8. Conclusion

In this paper we considered a context common in modern statistical learning problems concerned with risk minimization. For such problems, we have proposed a new *generalized universal inference* framework that leverages the theory of e-values and e-processes, and have shown that the corresponding GUe-value tests and confidence sets for the unknown risk minimizer are provably valid in finite samples. These validity conclusions do not come for free, as one might hope based on the developments in Wasserman et al. (2020), but they follow from a general and relatively mild condition called the strong central condition. Furthermore, under certain weak consistency conditions, the diameter of the GUe-value confidence sets shrinks at the same rates achieved by the driving ERM, suggesting the finite-sample validity guarantees do not come at the cost of severe inefficiency. The online GUe-value features an additional anytime-validity property that means the validity claims hold uniformly over all stopping rules used in the data collection process. In particular, we showed that the method's reliable performance is stable across a variety of common stopping rules believed to contribute to the replication crisis in science. Furthermore, the practitioner has agency in choosing how conservative they wish to be, as the methodology they use to choose the learning rate for the GUe-value can be influenced by the assumptions they are willing to make regarding the collected data.

The test consistency results in Theorems 4 and 5 above are related to a more fundamental question concerning the asymptotic growth rate of the proposed GUe-value, akin to the investigations in Grünwald et al. (2023) for the well-specified statistical model setting. For a given (possibly composite) null hypothesis $H_0 : \theta^* \in \Theta_0$, recall that $G_n^{(\omega)}(\Theta_0) = \inf_{\theta \in \Theta_0} G_n^{(\omega)}(\theta)$. Following Theorem 2 in Dixit and Martin (2023), our claim is that, under certain conditions (e.g., the learning rate ω is sufficiently small), the asymptotic growth rate of our GUe-value is

$$\log G_n^{(\omega)}(\Theta_0) = n \times \omega \Big\{ \inf_{\theta \in \Theta_0} R(\theta) - \inf_{\theta \notin \Theta_0} R(\theta) \Big\} + o(n), \quad \text{almost surely.}$$

Note that, if the hypothesis is true in the sense that $\Theta_0 \ni \theta^*$, then the GUe-value vanishes as $n \to \infty$, as expected. Similarly, if the hypothesis is false in the sense that $\Theta_0 \not\ni \theta^*$, then

the GUe-value diverges to ∞ as $n \to \infty$, again as expected. Moreover, the (exponential) rate at which these limits are approached corresponds, e.g., in the latter case, to the degree of separation between θ^* and Θ_0 determined by the risk function: the further θ^* is from Θ_0, as measured by $\omega \inf_{\theta \in \Theta_0}\{R(\theta) - R(\theta^*)\}$, the faster the growth rate. Our further conjecture is that the asymptotic growth rate of the GUe-value above is the "optimal growth rate for e-values aimed at inference on a risk minimizer," but we leave a proper formulation and verification of these claims for follow-up work.

Future investigations will consider how this theory might extend to the non-i.i.d. setting to allow for inference on risk minimizers in longitudinal or spatial data. Another important open question is how best to choose the learning rate for the GUe-value, and under what conditions the proposed bootstrapping strategy offers GUe-value confidence sets with provable validity guarantees. The theory we have presented for learning rate selection is quite limited, even for the special case of the L^2 loss function, despite how critical the choice of learning rate is in providing finite-sample validity guarantees for the GUe-value; thus, further work in this direction is necessary. Finally, we hope to further investigate the utility of the GUe-value in more modern machine learning models (where model parameters need not be interpretable), through its connection to the Gibbs posterior and thus PAC-Bayes learning theory.

A. Omitted Proofs

The following lemma is of use in the proofs of Theorems 4 and 5:

Lemma 6. *For any n, we have that*

$$\sum_{i=1}^{n} \ell(\widehat{\theta}_i;\, Z_i) \leq \sum_{i=1}^{n} \ell(\widehat{\theta}_n;\, Z_i).$$

Proof We proceed by induction. When $n = 1$, the statement trivially holds. Thus, suppose

$$\sum_{i=1}^{n-1} \ell(\widehat{\theta}_i;\, Z_i) \leq \sum_{i=1}^{n-1} \ell(\widehat{\theta}_{n-1};\, Z_i).$$

Then

$$\begin{aligned}
\sum_{i=1}^{n} \ell(\widehat{\theta}_i;\, Z_i) &= \left\{ \sum_{i=1}^{n-1} \ell(\widehat{\theta}_i;\, Z_i) \right\} + \ell(\widehat{\theta}_n;\, Z_n) \\
&\leq \left\{ \sum_{i=1}^{n-1} \ell(\widehat{\theta}_{n-1};\, Z_i) \right\} + \ell(\widehat{\theta}_n;\, Z_n) \\
&\leq \left\{ \sum_{i=1}^{n-1} \ell(\widehat{\theta}_n;\, Z_i) \right\} + \ell(\widehat{\theta}_n;\, Z_n) \\
&= \sum_{i=1}^{n} \ell(\widehat{\theta}_n;\, Z_i)
\end{aligned}$$

where the first inequality uses the inductive hypothesis and the second inequality uses the fact that $\widehat{\theta}_{n-1}$ minimizes the sum within the braces. ■

A.1 Proof of Theorem 4

We divide the proof of this theorem in two parts—Part I proves the consistency result (4) for the online GUe-value, and Part II does so for the offline GUe-value.

Part I Online GUe

For convenience in notation, define $\sigma_n := \frac{1}{n}\sum_{i=1}^n \ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\widehat{\theta}_n;\, Z_i)$. Further, let θ be such that $R(\theta) > \inf_\vartheta R(\vartheta)$, and define $\Delta := (R(\theta) - \inf_\vartheta R(\vartheta))/2$. We then have that

$$\begin{aligned}
\Pr\left\{G^{(\omega)}_{n,\mathrm{on}}(\theta) \geq \frac{1}{\alpha}\right\} &= \Pr\left\{\sum_{i=1}^n \ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\theta;\, Z_i) \leq \frac{\log\alpha}{\omega}\right\} \\
&= \Pr\left\{\sum_{i=1}^n \ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\widehat{\theta}_n;\, Z_i) + \ell(\widehat{\theta}_n;\, Z_i) - \ell(\theta;\, Z_i) \leq \frac{\log\alpha}{\omega}\right\} \\
&= \Pr\left\{\sigma_n + \widehat{R}_n(\widehat{\theta}_n) - \widehat{R}_n(\theta) \leq \frac{\log\alpha}{n\omega}\right\} \\
&\geq \underbrace{\Pr\left\{\sigma_n \leq \frac{\Delta}{2}\right\}}_{\text{(A)}} + \underbrace{\Pr\left\{\widehat{R}_n(\widehat{\theta}_n) - \widehat{R}_n(\theta) \leq \frac{\log\alpha}{n\omega} - \frac{\Delta}{2}\right\}}_{\text{(B)}} - 1.
\end{aligned}$$

We need to show that both terms (A) and (B) go to 1 as $n \to \infty$. For the former, we have by Theorem 6 that

$$\Pr\left\{\sigma_n \leq \frac{\Delta}{2}\right\} \geq \Pr\left\{\frac{1}{n}\sum_{i=1}^n \ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\widehat{\theta}_i;\, Z_i) \leq \frac{\Delta}{2}\right\}.$$

Next, we have by the stability hypothesis that each $\ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\widehat{\theta}_i;\, Z_i) \leq \beta_i$ for some positive scalars β_i that satisfy $\lim_{n\to\infty} \beta_n = 0$. This implies that $\frac{1}{n}\sum \beta_i \to 0$, and so

$$\lim_{n\to\infty} \Pr\left\{\sigma_n \leq \frac{\Delta}{2}\right\} \geq \lim_{n\to\infty} \Pr\left\{\frac{1}{n}\sum_{i=1}^n \beta_i \leq \frac{\Delta}{2}\right\} = 1$$

as desired.

To show that (B) has limit 1, we note that

$$\begin{aligned}
&\Pr\left\{\widehat{R}_n(\widehat{\theta}_n) - \widehat{R}_n(\theta) \leq \frac{\log(\alpha)}{n\omega} - \frac{\Delta}{2}\right\} \\
&= \Pr\left\{\widehat{R}_n(\widehat{\theta}_n) - R(\widehat{\theta}_n) + R(\widehat{\theta}_n) - \widehat{R}_n(\theta) \leq \frac{\log(\alpha)}{n\omega} - \frac{\Delta}{2}\right\} \\
&\geq \underbrace{\Pr\left\{\widehat{R}_n(\widehat{\theta}_n) - R(\widehat{\theta}_n) \leq \frac{\Delta}{2}\right\}}_{\text{(C)}} + \underbrace{\Pr\left\{R(\widehat{\theta}_n) - \widehat{R}_n(\theta) \leq \frac{\log\alpha}{n\omega} - \Delta\right\}}_{\text{(D)}} - 1.
\end{aligned}$$

It now suffices to show that each of (C) and (D) also has limit 1. To do so for (C), suppose $\widetilde{\theta}$ is such that $R(\widetilde{\theta}) \leq \inf_\vartheta R(\vartheta) + \Delta/4$. Then

$$\begin{aligned}
&\Pr\left\{\widehat{R}_n(\widehat{\theta}_n) - R(\widehat{\theta}_n) \leq \frac{\Delta}{2}\right\} \\
&\geq \Pr\left\{\widehat{R}_n(\widetilde{\theta}) - R(\widehat{\theta}_n) \leq \frac{\Delta}{2}\right\} \\
&\geq \Pr\left\{\widehat{R}_n(\widetilde{\theta}) - R(\widetilde{\theta}) + \inf_\vartheta R(\vartheta) - R(\widehat{\theta}_n) \leq \frac{\Delta}{4}\right\} \\
&\geq \Pr\left\{\widehat{R}_n(\widetilde{\theta}) - R(\widetilde{\theta}) \leq \frac{\Delta}{8}\right\} + \Pr\left\{\inf_\vartheta R(\vartheta) - R(\widehat{\theta}_n) \leq \frac{\Delta}{8}\right\} - 1.
\end{aligned}$$

As $n \to \infty$, the first term goes to 1 by the weak law of large numbers, and it follows from the uniform convergence of $\widehat{R}_n$ to R that the second term goes to 1 as well.

To show that (D) has limit 1, we see that

$$\begin{aligned}
&\Pr\left\{R(\widehat{\theta}_n) - \widehat{R}_n(\theta) \leq \frac{\log\alpha}{n\omega} - \Delta\right\} \\
&= \Pr\left\{R(\widehat{\theta}_n) - \inf_\vartheta R(\vartheta) + \inf_\vartheta R(\vartheta) - \widehat{R}_n(\theta) \leq \frac{\log\alpha}{n\omega} - \Delta\right\} \\
&\geq \Pr\left\{R(\widehat{\theta}_n) - \inf_\vartheta R(\vartheta) \leq \frac{\log\alpha}{n\omega} + \Delta\right\}.
\end{aligned}$$

Now if $n \geq 2\log(1/\alpha)/(\omega\Delta)$, the above is lower bounded by

$$\Pr\left\{R(\widehat{\theta}_n) - \inf_\vartheta R(\vartheta) \leq \frac{\Delta}{2}\right\},$$

which we again know to have limit 1 as $n \to \infty$ by uniform convergence.

Part II Offline Gue

For convenience in notation, define the function $\Phi(\vartheta) := \widehat{R}_{S_2}(\vartheta) - R(\vartheta)$. Further, let θ be such that $R(\theta) > \inf_\vartheta R(\vartheta)$, and define $\Delta := (R(\theta) - \inf_\vartheta R(\vartheta))/3$. We then have that

$$\begin{aligned}
&\Pr\left\{G^{(\omega)}_{S,\text{off}}(\theta) \geq \frac{1}{\alpha}\right\} \\
&= \Pr\left\{\widehat{R}_{S_2}(\widehat{\theta}_{S_1}) - \widehat{R}_{S_2}(\theta) \leq \frac{\log\alpha}{\omega n_2}\right\} \\
&= \Pr\left\{[\widehat{R}_{S_2}(\widehat{\theta}_{S_1}) - R(\widehat{\theta}_{S_1})] + [R(\widehat{\theta}_{S_1}) - R(\theta)] + [R(\theta) - \widehat{R}_{S_2}(\theta)] \leq \frac{\log\alpha}{\omega n_2}\right\} \\
&\geq \underbrace{\Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta\right\}}_{\text{(A)}} + \underbrace{\Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta\right\}}_{\text{(B)}} + \underbrace{\Pr\left\{-\Phi(\theta) \leq \Delta\right\}}_{\text{(C)}} - 2.
\end{aligned}$$

It suffices to show that each of (A) (B), and (C) has limit 1 as $(n_1, n_2) \to (\infty, \infty)$. For (A), we have from uniform convergence in probability of $\widehat{R}_S$ to R that for any $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that if $n_2 \geq N$,

$$1 - \Pr\left\{\sup_{\vartheta \in \Theta} |\Phi(\vartheta)| \leq \Delta\right\} < \varepsilon.$$

We then note that for any $n_1 \in \mathbb{N}$, if $\sup_{\vartheta \in \Theta} |\Phi(\vartheta)| \leq \Delta$, it is certainly also the case that $\Phi(\widehat{\theta}_{S_1}) \leq \Delta$. Hence, we have for any $\varepsilon > 0$ that there exists an $N \in \mathbb{N}$ such that for any $n_1 \in \mathbb{N}$, if $n_2 \geq N$,

$$1 - \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta\right\} < \varepsilon.$$

That is to say that as $n_2 \to \infty$, $\Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta\right\} \to 1$ uniformly in n_1. Since this uniform limit does not depend on the value of n_1, we have that the double limit exists and is equal to the single limit:

$$\lim_{(n_1,n_2)\to(\infty,\infty)} \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta\right\} = \lim_{n_2 \to \infty} \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta\right\} = 1.$$

We now examine term (B):

$$\begin{aligned}
&\Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log \alpha}{\omega n_2} - 2\Delta\right\} \\
&= \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log \alpha}{\omega n_2} - 2\Delta + R(\theta) - \inf_{\vartheta} R(\vartheta)\right\} \\
&= \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log \alpha}{\omega n_2} + \Delta\right\}
\end{aligned}$$

where the final equality comes from our choice for Δ. We now show that the double limit of the above expression exists and is equal to 1. To this end, let $\varepsilon > 0$ be arbitrary. Since $R(\widehat{\theta}_{S_1}) \xrightarrow{p} \inf_{\vartheta} R(\vartheta)$, we have that there exists $N \in \mathbb{Z}^+$ such that if $n_1 \geq N$,

$$\Pr\left\{|R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta)| \leq \frac{\Delta}{2}\right\} > 1 - \varepsilon.$$

Thus, if $n_1, n_2 \geq \max(-\frac{2\log\alpha}{\omega\Delta}, N)$, we have that

$$\begin{aligned}
&\Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log \alpha}{\omega n_2} + \Delta\right\} \\
&\geq \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq -\frac{\log \alpha}{2\omega \cdot \log(\alpha)/(\omega\Delta)} + \Delta\right\} \\
&= \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\Delta}{2}\right\} \\
&\geq \Pr\left\{|R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta)| \leq \frac{\Delta}{2}\right\} \\
&> 1 - \varepsilon
\end{aligned}$$

and thus our double limit is 1:

$$\lim_{(n_1,n_2)\to(\infty,\infty)} \Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta\right\} = 1.$$

Finally, we note that (C) has limit 1 by the law of large numbers, as $\mathrm{E}[\Phi(\theta)] = 0$.

A.2 Proof of Theorem 5

We again divide the proof of the theorem in three parts, with each part proving the corresponding numbered result in Theorem 5.

Part I Online GUe

Define $\sigma_n := \frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\widehat{\theta}_n;\, Z_i)$ and $\Phi_n(\theta) := \widehat{R}_n(\theta) - R(\theta)$. Further define $\Delta_n = (R(\theta_n) - \inf_\vartheta R(\vartheta))/2$; note that there exists $c > 0$ such that $\Delta_n \geq c \cdot n^{-\beta}/2$ due to the definition of $(\theta_n)_{n\in\mathbb{N}}$. Then using the same arguments as in Part I of the proof of Theorem 4, we have that

$$\begin{aligned}
&\Pr\left\{G^{(\omega)}_{n,\mathrm{on}}(\theta_n) \geq \frac{1}{\alpha}\right\} \\
&\geq \Pr\left\{\sigma_n \leq \frac{\Delta_n}{2}\right\} + \Pr\left\{\Phi_n(\widehat{\theta}_n) \leq \frac{\Delta_n}{2}\right\} + \Pr\left\{R(\widehat{\theta}_n) - \widehat{R}_n(\theta_n) \leq \frac{\log\alpha}{n\omega} - \Delta_n\right\} - 2.
\end{aligned}$$

We must show that each term has limit 1. We have that second addend immediately has limit 1 since we've assumed that $\sup_\theta |\Phi_n(\theta)| = o_p(n^{-\beta})$. That the first addend converges in probability to 1 follows by essentially the same argument as in the previous theorem: We have by Theorem 6 that

$$\Pr\left\{\sigma_n \leq \frac{\Delta_n}{2}\right\} \geq \Pr\left\{\frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{\theta}_{i-1};\, Z_i) - \ell(\widehat{\theta}_i;\, Z_i) \leq \frac{\Delta_n}{2}\right\}$$

and by stability, there exists a sequence $\varepsilon_n = o(n^{-\beta})$ such that for each $n \in \mathbb{N}$, $\ell(\widehat{\theta}_{n-1};\, Z_n) - \ell(\widehat{\theta}_n;\, Z_n) \leq \varepsilon_n$. Hence, since $\frac{1}{n}\sum \varepsilon_i = o(n^{-\beta})$ and $\Delta_n \geq c \cdot n^{-\beta}/2$, we have that

$$\lim_{n\to\infty} \Pr\left\{\sigma_n \leq \frac{\Delta_n}{2}\right\} \geq \lim_{n\to\infty} \Pr\left\{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_n \leq \frac{\Delta_n}{2}\right\} = 1$$

as required.

To show that the third addend has limit 1, define $\varepsilon_n = \frac{\log\alpha}{\omega} n^{\beta-1} + \frac{c}{2}$. We note that

$$\begin{aligned}
&\Pr\left\{R(\widehat{\theta}_n) - \widehat{R}_n(\theta_n) \leq \frac{\log\alpha}{n\omega} - \Delta_n\right\} \\
&= \Pr\left\{R(\widehat{\theta}_n) - \inf_\vartheta R(\vartheta) \leq \frac{\log\alpha}{n\omega} + \Delta_n\right\} \\
&\geq \Pr\left\{R(\widehat{\theta}_n) - \inf_\vartheta R(\vartheta) \leq \frac{\log\alpha}{n\omega} + \frac{c}{2n^\beta}\right\} \\
&= \Pr\left\{R(\widehat{\theta}_n) - \inf_\vartheta R(\vartheta) \leq \varepsilon_n n^{-\beta}\right\}.
\end{aligned}$$

Since $\beta \in (0,1)$, whenever $n > (\frac{2\log(1/\alpha)}{\omega c})^{1/(1-\beta)}$ we have that $\varepsilon_n > 0$. Thus, for large enough n, there exists $\varepsilon > 0$ such that

$$\Pr\left\{R(\widehat{\theta}_n) - \widehat{R}_n(\theta_n) \leq \frac{\log\alpha}{n\omega} - \Delta_n\right\} \geq \Pr\left\{R(\widehat{\theta}_n) - \inf_{\vartheta} R(\vartheta) \leq 2\varepsilon n^{-\beta}\right\}. \tag{7}$$

Next, note that if $\sup_\vartheta |\Phi(\vartheta)| < \varepsilon n^{-\beta}$, then

$$R(\widehat{\theta}_n) \leq \widehat{R}_n(\widehat{\theta}_n) + \varepsilon n^{-\beta} \leq \widehat{R}_n(\theta) + \varepsilon n^{-\beta} \leq R(\theta) + 2\varepsilon n^{-\beta}$$

for any $\theta \in \Theta$. Taking the infimum over θ, we arrive at the implication

$$\sup_{\vartheta} |\Phi(\vartheta)| < \varepsilon n^{-\beta} \implies R(\widehat{\theta}_n) \leq \inf_{\vartheta} R(\vartheta) + 2\varepsilon n^{-\beta},$$

and so

$$\Pr\left\{R(\widehat{\theta}_n) \leq \inf_{\vartheta} R(\vartheta) + 2\varepsilon n^{-\beta}\right\} \geq \Pr\left\{\sup_{\vartheta} |\Phi(\vartheta)| \leq \varepsilon n^{-\beta}\right\}. \tag{8}$$

Since the right hand side of (8) has limit 1 by hypothesis, combining it with inequality (7) yields the desired result.

Part II Offline GUe (Fixed Validation Set)

For convenience in notation, we define $\Phi(\theta) := \widehat{R}_{S_2}(\theta) - R(\theta)$; we also define $\Delta_{n_2} = [R(\theta_{n_2}) - \inf_\vartheta R(\vartheta)]/3$, so that there exists some $c > 0$ such that $\Delta_{n_2} \geq cn^{-\beta}/3$. Then in the same manner as in Part II of the proof of Theorem 4, we have that

$$\begin{aligned}
&\Pr\left\{G^{(\omega)}_{S,\text{off}}(\theta_n) \geq 1/\alpha\right\} \\
&\geq \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta_{n_2}\right\} + \Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta_n) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta_{n_2}\right\} + \Pr\left\{-\Phi(\theta_n) \leq \Delta_{n_2}\right\} - 2.
\end{aligned}$$

As usual, we show that each term limit 1.

For the first term, we have that

$$\lim_{n_2\to\infty} \Pr\left\{\widehat{R}_{S_2}(\widehat{\theta}_{S_1}) - R(\widehat{\theta}_{S_1}) \leq \Delta_{n_2}\right\} \geq \lim_{n_2\to\infty} \Pr\left\{\widehat{R}_{S_2}(\widehat{\theta}_{S_1}) - R(\widehat{\theta}_{S_1}) \leq \frac{c}{3n_2^\beta}\right\} = 1$$

where the convergence is uniform by essentially the same arguments as in in the proof of Theorem 4, but we now use the fact that $\sup_\vartheta |\widehat{R}_S(\vartheta) - R(\vartheta)|$ is $o_p(n^{-\beta})$ rather than simply $o_p(1)$.

For the second term, we note that

$$\begin{aligned}
\Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta_{n_2}\right\} &= \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log\alpha}{\omega n_2} + \Delta_{n_2}\right\} \\
&\geq \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log\alpha}{\omega n_2} + \frac{c}{3n_2^\beta}\right\}.
\end{aligned}$$

Now we notice that $\log(\alpha)/(\omega n_2) + c/(3n_2^\beta) > 0$ if and only if $n_2^{1-\beta} > \log(1/\alpha)/(c\omega)$; since $\beta \in (0,1)$ there exists some $N_2 \in \mathbb{Z}^+$ such that the right hand side is positive for all $n_2 \geq N_2$. Next, since $R(\widehat{\theta}_{S_1}) \xrightarrow{p} \inf_\vartheta R(\vartheta)$, we know that there exists $N_1 \in \mathbb{Z}^+$ such that for any $\varepsilon > 0$,

$$\Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_\vartheta R(\vartheta) \leq \frac{\log\alpha}{\omega N_2} + \frac{c}{3N_2^\beta}\right\} > 1-\varepsilon$$

for all $n_1 \geq N_1$. We thus have that for any $\varepsilon > 0$, if $n_1, n_2 \geq \max(N_1, N_2)$, then

$$\Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta_{n_2}\right\} > 1-\varepsilon$$

so the second addend has double limit 1 again.

For the third addend, we simply apply our uniform convergence in probability at rate $n_2^{-\beta}$ since $\Delta_{n_2} \geq cn_2^{-\beta}/3$.

Part III Offline GUe (Growing Validation Set)

We define Φ as in the previous part, and we also define $\Delta_n := [R(\theta_n) - \inf_\vartheta R(\vartheta)]/3$ so that there exists $c > 0$ such that $\Delta_n > cn^{-\beta}/3$. Then as in the previous parts,

$$\begin{aligned}
&\Pr\left\{G_S^{(\omega)}(\theta_n) \geq 1/\alpha\right\} \\
&\geq \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta_n\right\} + \Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta_n) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta_n\right\} + \Pr\left\{-\Phi(\theta_n) \leq \Delta_n\right\} - 2.
\end{aligned}$$

and we again show that each term has limit 1.

For the first addend, since $n_1 \lesssim n_2$, there exist $k > 0$ and $N_1 \in \mathbb{Z}^+$ such that if $n_1 \geq N_1$, then $n_1 \leq k \cdot n_2$. Furthermore, we have from uniform convergence of $\widehat{R}_{S_2}$ to R at rate $o_p(n_2^{-\beta})$ that for every $\varepsilon > 0$, there exists N_2 such that if $n_2 \geq N_2$,

$$1 - \Pr\left\{\sup_{\vartheta\in\Theta} |\Phi(\vartheta)| \leq \frac{c}{3((k+1)n_2)^\beta}\right\} < \varepsilon$$

for some $c > 0$. Similarly to Part II in the proof of Theorem 4, we then have that for any $\varepsilon > 0$, there exists N_2 such that for any $n_1 \geq N_1$, if $n_2 \geq N_2$

$$1 - \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \frac{c}{3((k+1)n_2)^\beta}\right\} < \varepsilon. \tag{9}$$

But when $n_1 \geq N_1$, we have that

$$\Delta_n \geq \frac{c}{3n^\beta} = \frac{c}{3(n_1+n_2)^\beta} \geq \frac{c}{3(k+1)n_2)^\beta}$$

and so equation (9) reduces to

$$1 - \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta_n\right\} < \varepsilon.$$

We hence have that

$$\lim_{n_2\to\infty} \Pr\left\{\Phi(\widehat{\theta}_{S_1}) \leq \Delta_n\right\} = 1$$

and the limit is uniform in n_1, as necessary for the double limit to exist and equal 1.

For the second addend, we note that

$$\begin{aligned}\Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta_n\right\} &= \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log\alpha}{\omega n_2} + \Delta_n\right\} \\ &\geq \Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log\alpha}{\omega n_2} + \frac{c}{3(n_1+n_2)^\beta}\right\}.\end{aligned}$$

Similarly to the first addend, there exist $k > 0$ and $N_1 \in \mathbb{Z}^+$ such that if $n_1 \geq N_1$, $n_1 \leq k{\cdot}n_2$. So for all $n_1 \geq N_1$, the above is at least

$$\Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log\alpha}{\omega n_2} + \frac{c}{3((k+1)n_2)^\beta}\right\}.$$

Next, we notice that since $\beta \in (0,1)$ there exists some $N_2 \in \mathbb{Z}^+$ such that the right hand side is positive if $n_2 \geq N_2$. Then since $R(\widehat{\theta}_{S_1}) \xrightarrow{p} \inf_\vartheta R(\vartheta)$, we know that there exists $M \in \mathbb{Z}^+$ such that for any $\varepsilon > 0$,

$$\Pr\left\{R(\widehat{\theta}_{S_1}) - \inf_{\vartheta} R(\vartheta) \leq \frac{\log\alpha}{\omega N_2} + \frac{c}{3(k+1)^\beta N_2^\beta}\right\} > 1-\varepsilon$$

for all $n_1 \geq M$. We thus have that for any $\varepsilon > 0$, if $n_1, n_2 \geq \max(N_1, N_2, M)$, then

$$\Pr\left\{R(\widehat{\theta}_{S_1}) - R(\theta) \leq \frac{\log\alpha}{\omega n_2} - 2\Delta_n\right\} > 1-\varepsilon$$

so the second addend has double limit 1 again.

For the third addend, we simply apply our uniform convergence in probability at rate $n^{-\beta}$ since $\Delta_n \geq cn^{-\beta}/3$.

B. Learning Rates: Analytical Results

As mentioned in Section 4, closed form-learning rates for the L^2 loss function can be derived. For example, we have the following proposition for normally distributed data:

Proposition 7. *Suppose* $X_1, \ldots, X_{2n} \overset{iid}{\sim} N(\theta^*, \sigma^2)$*, and define*

$$b^{(\omega)}_{\alpha,\sigma^2}(z) := \frac{\log(1/\alpha)}{2\omega\sigma^2 z} + \frac{z}{2}.$$

Then the learning rate ω *for the offline GUe-value that obtains exactly* $(1-\alpha)$*-level coverage for* θ^* *under the* L^2 *loss is given by the solution to the equation*

$$\int_0^\infty \int_{b^{(\omega)}_{\alpha,\sigma^2}(z_2)}^\infty \frac{\exp\left(-\frac{z_1^2+z_2^2}{2}\right)}{2\pi}\,\mathrm{d}z_1\,\mathrm{d}z_2 + \int_{-\infty}^0 \int_{-\infty}^{b^{(\omega)}_{\alpha,\sigma^2}(z_2)} \frac{\exp\left(-\frac{z_1^2+z_2^2}{2}\right)}{2\pi}\,\mathrm{d}z_1\,\mathrm{d}z_2 = \alpha \qquad (10)$$

Proof Let $\overline{X}$ denote the sample mean of $X_1, \ldots, X_n$ and $\widehat{\theta}$ denote the sample mean of $X_{n+1}, \ldots, X_{2n}$. Then by expanding the definition of the GUe-value and using the law of total probability, we have that

$$\begin{aligned}
&\Pr\left\{G_S^{(\omega)}(\theta^*) \geq 1/\alpha\right\} \\
&= \Pr\left\{\overline{X}(\widehat{\theta}-\theta^*) - \frac{(\widehat{\theta}-\theta^*)(\widehat{\theta}+\theta^*)}{2} \geq \frac{\log(1/\alpha)}{2n\omega}\right\} \\
&= \frac{\Pr\left\{\overline{X} \geq \frac{\log(1/\alpha)}{2n\omega(\widehat{\theta}-\theta^*)} + \frac{\widehat{\theta}+\theta^*}{2} \mid \widehat{\theta} > \theta^*\right\}}{2} + \frac{\Pr\left\{\overline{X} \leq \frac{\log(1/\alpha)}{2n\omega(\widehat{\theta}-\theta^*)} + \frac{\widehat{\theta}+\theta^*}{2} \mid \widehat{\theta} \leq \theta^*\right\}}{2} \\
&= \frac{1}{2}\Pr\left\{Z_1 \geq \frac{\log(1/\alpha)}{2\omega\sigma^2 Z_2} + \frac{Z_2}{2} \mid Z_2 > 0\right\} + \frac{1}{2}\Pr\left\{Z_1 \leq \frac{\log(1/\alpha)}{2\omega\sigma^2 Z_2} + \frac{Z_2}{2} \mid Z_2 \leq 0\right\} \\
&= \frac{1}{2}\Pr\left\{Z_1 \geq b^{(\omega)}_{\alpha,\sigma^2}(Z_2) \mid Z_2 > 0\right\} + \frac{1}{2}\Pr\left\{Z_1 \leq b^{(\omega)}_{\alpha,\sigma^2}(z_2) \mid Z_2 \leq 0\right\}
\end{aligned}$$

where $Z_1 = \sqrt{n}(\overline{X}-\theta^*)/\sigma$ and $Z_2 = \sqrt{n}(\widehat{\theta}-\theta^*)/\sigma$ are independent standard normal random variables, whence the result follows by substituting integrals of standard normal densities for the probability statements. ■

We note that equation (10) can be solved numerically for ω quite quickly, so it is a convenient choice for learning rate whenever Theorem 7 is applicable. The proof of the proposition illustrates that in order to use equation (10) for non-normal random variables, we need the sample size to be large enough for sample means to be reasonably approximated as normal and for the sample variance to act as a good estimator for σ^2. If safety at smaller sample sizes is a concern, however, one could simply solve for ω from equation (10) and divide the learning rate by two (for example) to be confident that the learning rate is small enough to be a safe choice.

To ensure safety for non-normal data, we may use the following proposition:

Proposition 8. *Let* $X_1, \ldots, X_{2n}$ *be i.i.d. from any distribution with mean* θ^**, variance* σ^2*, and third absolute moment* ρ*. Let* $c_B \approx 0.4748$ *be the Berry-Esseen constant and* Φ *denote*

the standard normal CDF, and define

$$l_\alpha^{(\omega)}(z) := \max\left(\Phi(b_{\alpha,\sigma^2}^{(\omega)}(z)) - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0\right)$$
$$u_\alpha^{(\omega)}(z) := \min\left(\Phi(b_{\alpha,\sigma^2}^{(\omega)}(z)) + \frac{c_B\rho}{\sigma^3\sqrt{n}}, 1\right).$$

Furthermore, let $l_\alpha^{(\omega)}(z)$ be maximized on $[0,\infty)$ at $z=\beta$ and $u_\alpha^{(\omega)}(z)$ be maximized on $(-\infty,0]$ at $z=\gamma$. Then any ω that satisfies

$$\begin{aligned}\alpha \geq 1 &- \max\left(\frac{1}{2} - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0\right) - \max\left(1 - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0\right) \\ &+ \left[\max(1 - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0) + \min(\frac{c_B\rho}{\sigma^3\sqrt{n}}, 1)\right] \cdot \min\left(\frac{1}{2} + \frac{c_B\rho}{\sigma^3\sqrt{n}}, 1\right) \\ &+ \int_0^\beta l_\alpha^{(\omega)}(z)\,\mathrm{d}l_\alpha^{(\omega)}(z) + \int_\beta^\infty u_\alpha^{(\omega)}(z)\,\mathrm{d}l_\alpha^{(\omega)}(z) \\ &+ \int_{-\infty}^\gamma u_\alpha^{(\omega)}(z)\,\mathrm{d}u_\alpha^{(\omega)}(z) + \int_\gamma^0 l_\alpha^{(\omega)}(z)\,\mathrm{d}u_\alpha^{(\omega)}(z)\end{aligned}$$

obtains at least $(1-\alpha)$-level coverage for θ^* under the L^2 loss using the offline GUe-value.

Proof We can follow the proof of Theorem 7 up until the point where we assume Z_1 and Z_2 are standard normal; supposing they instead have CDF F, we arrive at

$$\Pr\left\{G_S^{(\omega)}(\theta^*) \geq 1/\alpha\right\} = \int_0^\infty 1 - F(b_{\alpha,\sigma^2}^{(\omega)}(z_2))\,\mathrm{d}F(z_2) + \int_{-\infty}^0 F(b_{\alpha,\sigma^2}^{(\omega)}(z_2))\,\mathrm{d}F(z_2)\,.$$

Although we do not know F, we do know by the Berry-Esseen inequality that

$$\sup_{x\in\mathbb{R}} |F(x) - \Phi(x)| \leq \frac{c_B\rho}{\sigma^3\sqrt{n}}$$

where Φ denotes the standard normal CDF. For convenience, we suppress all unnecessary parameters so that we denote $b(z) = b_{\alpha,\sigma^2}^{(\omega)}(z)$, $l(z) = l_\alpha^{(\omega)}(z)$, and $u(z) = u_\alpha^{(\omega)}(z)$. Then we can obtain a safe learning rate by repeatedly integrating by parts and applying the

Berry-Esseen bounds to upper bound this expression:

$$
\begin{aligned}
&\int_0^\infty 1 - F(b(z))\,\mathrm{d}F(z) + \int_{-\infty}^0 F(b(z))\,\mathrm{d}F(z) \\
&= (1 - F(0)) - \int_0^\infty F(b(z))\,\mathrm{d}F(z) + \int_{-\infty}^0 F(b(z))\,\mathrm{d}F(z) \\
&\leq 1 - F(0) - \int_0^\infty l(z)\,\mathrm{d}F(z) + \int_{-\infty}^0 u(z)\,\mathrm{d}F(z) \\
&= 1 - F(0) - \left[l(\infty) - l(0)F(0) - \int_0^\infty F(z)\,\mathrm{d}l(z)\right] + \left[u(0)F(0) + \int_{-\infty}^0 F(z)\,\mathrm{d}u(z)\right] \\
&= 1 - F(0) - l(\infty) + l(0)F(0) + u(0)F(0) \\
&\qquad + \int_0^\beta F(z)\,\mathrm{d}l(z) + \int_\beta^\infty F(z)\,\mathrm{d}l(z) + \int_{-\infty}^\gamma F(z)\,\mathrm{d}u(z) + \int_\gamma^0 F(z)\,\mathrm{d}u(z) \\
&\leq 1 - F(0) - l(\infty) + [l(0) + u(0)] \cdot F(0) \\
&\qquad + \int_0^\beta l(z)\,\mathrm{d}l(z) + \int_\beta^\infty u(z)\,\mathrm{d}l(z) + \int_{-\infty}^\gamma u(z)\,\mathrm{d}u(z) + \int_\gamma^0 l(z)\,\mathrm{d}u(z) \\
&\leq 1 - \max\left(\frac{1}{2} - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0\right) - \max\left(1 - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0\right) \\
&\qquad + \left[\max\left(1 - \frac{c_B\rho}{\sigma^3\sqrt{n}}, 0\right) + \min\left(\frac{c_B\rho}{\sigma^3\sqrt{n}}, 1\right)\right] \cdot \min\left(\frac{1}{2} + \frac{c_B\rho}{\sigma^3\sqrt{n}}, 1\right) \\
&\qquad + \int_0^\beta l(z)\,\mathrm{d}l(z) + \int_\beta^\infty u(z)\,\mathrm{d}l(z) + \int_{-\infty}^\gamma u(z)\,\mathrm{d}u(z) + \int_\gamma^0 l(z)\,\mathrm{d}u(z)
\end{aligned}
$$

as desired. ■

This no longer depends on the distribution of the data (other than the second and third moments, which can often be well-approximated by the sample moments); thus, this proposition yields a provably safe choice of learning rate. We note that it is atypical for there to exist an ω that obtains exactly $(1-\alpha)$-level coverage from the above proposition—in general, all learning rates satisfying the proposition are more conservative than necessary.

References

E. Bäcklin. Eddington's hypothesis and the electronic charge. *Nature*, 123, 1929.

J. A. Bearden. Absolute wave-lengths of the copper and chromium k-series. *The Physical Review*, 37:1210–1220, 1931. doi: 10.1103/PhysRev.37.1210.

P. G. Bissiri, C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016. doi: 10.1111/rssb.12158.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference*, volume 120 of *Springer Texts in Statistics*. Springer, New York, NY, 2018.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

L. Cella and R. Martin. Direct and approximately valid probabilistic inference on a class of statistical functionals. *International Journal of Approximate Reasoning*, 2022. doi: 10.1016/j.ijar.2022.09.011.

P. De Blasi and S. G. Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23(1):169–187, 2013.

R. de Heide, A. Kirichenko, P. Grunwald, and N. Mehta. Safe-bayesian generalized linear regression. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2623–2633. PMLR, 26–28 Aug 2020.

V. Dixit and R. Martin. Anytime valid and asymptotically optimal statistical inference driven by predictive recursion. `arXiv:2309.13441`, 2023.

R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019.

R. P. Feynman. The cargo cult science, 1974. Commencement address given at the California Institute of Technology.

A. Gangrade, A. Rinaldo, and A. Ramdas. A sequential test for log-concavity. `arXiv:2301.03542`, 2023.

P. Grünwald and T. van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069 – 1103, 2017. doi: 10.1214/17-BA1085.

P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, to appear, 2023.

P. D. Grünwald and N. A. Mehta. Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, 2011. ISBN 978-1-118-15068-9.

S. Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

C. Hill. Measurements of the electron charge over time, March 2021. URL `https://scipython.com/blog/measurements-of-the-electron-charge-over-time/`.

J. Hofmans, E. Ceulemans, D. Steinley, and I. V. Mechelen. On the added value of bootstrap analysis for k-means clustering. *Journal of Classification*, 32:268–284, 2015. doi: 10.1007/s00357-015-9178-y.

C. C. Holmes and S. G. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 03 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx010.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

P. J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, 1981.

A. Hudson, M. Carone, and A. Shojaie. Inference on function-valued parameters using a restricted score test, 2021. URL `https://arxiv.org/abs/2105.06646`.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2):837–877, 2006. doi: 10.1214/009053606000000029.

S. P. Lyddon, C. C. Holmes, and S. G. Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 03 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz006.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, Ltd, 1st edition, 2006.

R. Martin and N. Syring. Direct Gibbs posterior inference on risk minimizers: Construction, concentration, and calibration. In A. S. Srinivasa Rao, G. A. Young, and C. Rao, editors, *Advancements in Bayesian Methods and Implementation*, volume 47 of *Handbook of Statistics*, pages 1–41. Elsevier, 2022. doi: 10.1016/bs.host.2022.06.004.

R. A. Millikan. On the elementary charge and the Avogadro constant. *Physical Review*, 2 (2):109–143, 1913. doi: 10.1103/PhysRev.2.109.

B. Park, S. Balakrishnan, and L. Wasserman. Robust universal inference, 2023.

L. Perrotta. Practical calibration of the temperature parameter in Gibbs posteriors, 2020.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(4):397–417, 2005.

R. V. Ramamoorthi, K. Sriram, and R. Martin. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015. doi: 10.1214/15-BA941.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

J. Ruf, M. Larsson, W. M. Koolen, and A. Ramdas. A composite generalization of Ville's martingale theorem using e-processes. *Electronic Journal of Probability*, 28:1–21, 2023. doi: 10.1214/23-EJP1019.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010.

N. Syring and R. Martin. Calibrating general posterior credible regions. *Biometrika*, 106 (2):479–486, 2019. doi: 10.1093/biomet/asy054.

N. Syring and R. Martin. Gibbs posterior concentration rates under sub-exponential type losses. *Bernoulli*, 29(2):1080–1108, 2023. doi: 10.3150/22-BEJ1491.

T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54): 1793–1861, 2015.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020.

A. P. R. Wadlund. Absolute x-ray wave-length measurements. *The Physical Review*, 14(7): 588–591, 1928. doi: 10.1103/PhysRev.32.841.

A. Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

A. Wald. *Sequential Analysis*. John Wiley & Sons, Inc., New York, 1947.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 01 2022. ISSN 1369-7412. doi: 10.1111/rssb.12489. URL `https://doi.org/10.1111/rssb.12489`.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023. doi: 10.1093/jrsssb/qkad009.

P.-S. Wu and R. Martin. A comparison of learning rate selection methods in generalized Bayesian inference. *Bayesian Analysis*, 18(1):105 – 132, 2023. doi: 10.1214/21-BA1302.

Z. Xu, R. Wang, and A. Ramdas. A unified framework for bandit multiple testing. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16833–16845. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/8c460674cd61bf189e62b4da4bd9d7c1-Paper.pdf`.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006. doi: 10.1109/TIT.2005.864439.

T. Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. ISBN 9781009093057. doi: 10.1017/9781009093057.