

Confidence intervals for functionals in constrained inverse problems via data-adaptive sampling-based calibration

Michael Stanley*, Pau Batlle, Pratik Patil†, Houman Owhadi‡, and Mikael Kuusela§

Abstract. We address functional uncertainty quantification for ill-posed inverse problems with rank-deficient forward models, known observation noise distributions, and parameter constraints common in scientific applications such as remote sensing and high-energy physics. We propose four constraint-aware confidence interval methods built upon a theoretical test inversion framework in [1] and optimization-based endpoint computations, resulting in intervals that are both computationally feasible and less conservative. Our approach first shrinks the potentially unbounded constraint set in a data-adaptive way, obtains samples of the relevant test statistic inside this set to estimate a quantile function, and then uses these computed quantities to produce the constraint-aware confidence intervals. Our approach to bounding the constraint set in a data-adaptive way is based on the approach by [2] and involves defining a subset of the constraint set where the true parameter is guaranteed to exist with high probability. The probabilistic guarantee of this subset is then incorporated into the final coverage guarantee in the form of an uncertainty budget. Both theoretical insights and numerical experiments demonstrate that our intervals achieve nominal coverage for specified functionals, consistently outperforming the commonly used one-at-a-time strict bound intervals in terms of coverage accuracy and interval length. Numerical results include a realistic, severely ill-posed unfolding problem from high-energy physics involving a rank-deficient forward model.

Key words. uncertainty quantification, frequentist coverage, constrained hypothesis testing, likelihood ratio test

MSC codes. 68Q25, 68R10, 68U05

1. Introduction. We consider frequentist uncertainty quantification in constrained inverse problems with observation model $\mathbf{y} = f(\mathbf{x}^*) + \boldsymbol{\varepsilon}$, where $f : \mathbb{R}^p \rightarrow \mathbb{R}^n$ is a known parametric forward map, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ has a known distribution, and the unknown parameter satisfies $\mathbf{x}^* \in \mathcal{X}$. Throughout the paper, $\mathcal{X}$ is primarily assumed to be a polyhedral constraint set of the form $\mathbf{A}\mathbf{x} \leq \mathbf{b}$. We allow the parameter dimension p to exceed the observation dimension n , and we do not require the forward map to be injective; both features are common in ill-posed inverse problems. In this work, the uncertainty quantification target is a confidence interval $\mathcal{C}_\alpha(\mathbf{y})$ for a one-dimensional quantity of interest (QoI) $\varphi(\mathbf{x}^*)$, where $\varphi : \mathcal{X} \rightarrow \mathbb{R}$. For a prescribed miscoverage level $\alpha \in (0, 1)$, we seek finite-sample frequentist coverage uniformly over the constrained parameter space:

$$(1.1) \quad \mathbb{P}_{\mathbf{x}^*}(\varphi(\mathbf{x}^*) \in \mathcal{C}_\alpha(\mathbf{y})) \geq 1 - \alpha \quad \text{for all } \mathbf{x}^* \in \mathcal{X}.$$

Here the probability is over the noise in the observation $\mathbf{y} = f(\mathbf{x}^*) + \boldsymbol{\varepsilon}$ for fixed $\mathbf{x}^*$. Thus, the interval must contain the true functional value $\varphi(\mathbf{x}^*)$ with probability at least $1 - \alpha$, no matter which feasible parameter $\mathbf{x}^*$ generated the data.

*Analytical Mechanics Associates, USA

†Department of Statistics and Data Sciences, University of Texas at Austin, USA

‡Department of Computing and Mathematical Sciences, California Institute of Technology, USA

§Department of Statistics and Data Science, Carnegie Mellon University, USA

1.1. Context and related work. High-dimensional, ill-posed inverse problems have become more prevalent, especially in remote sensing and data assimilation fields. This setting includes a wide array of physical science applications, spanning Earth science [26], atmospheric science and remote sensing [17, 22], and high energy physics [14, 31], among many others. Providing guarantees for UQ in parameter inference is essential for assessing the precision of downstream scientific inferences. However, the inherent ill-posed nature of these problems often leads to inferences that are highly sensitive to noise, posing significant challenges for UQ. Namely, the ill-posedness leads to an identifiability issue in which statistical inference is impossible without providing some form of regularization, which usually takes either a deterministic (e.g., SVD truncation [12] and Tikhonov regularization [29]) or probabilistic (e.g., priors and Bayesian inference) form. Under some assumptions, these two approaches are mathematically equivalent and are therefore subject to the same pitfalls, such as a disruption of the coverage guarantees of downstream intervals as a result of the incurred regularization bias, as thoroughly discussed in [14]. Our method’s focus on one-dimensional QoI’s and incorporation of parameter constraints allow for implicit regularization. Therefore, it produces intervals with the promised coverage guarantee while avoiding the problem of regularization bias altogether.

Including parameter constraints shifts complexity to statistical inference under constraints, a non-trivial problem in even elementary settings [11, 36, 25, 30, 37, 20]. *Optimization-based UQ*, originating in the work of [3] and [27], offers one solution. [32] extended and generalized their approach, calling this method *strict bounds*, since it produces guaranteed simultaneous coverage interval estimators complying with known physical constraints on the model parameters. This collection of work defines endpoint optimization problems over the physically constrained parameter space to compute confidence intervals for one-dimensional QoI’s directly. Not only does the optimization form of the confidence interval computation shift statistical inference complexity to numerical optimization, but it also allows known physical constraints to be directly included in the endpoint optimizations. This practical advantage is dulled by the difficulty of proving the coverage properties of the intervals resulting from the defined endpoint optimizations in the one-at-a-time setting. Even under the relatively strong assumptions in [27] of a linear forward model, non-negativity parameter constraints, and linear QoI, the authors were only able to conjecture the coverage of their interval (known as the Burrus Conjecture). The interval coverage was unsuccessfully proven in [28] (the error pointed out in [34]) and finally generally refuted in [1]. These optimized-based intervals that have gained recent attention take the form:

$$(1.2) \quad \mathcal{I}(\psi_\alpha^2, \mathbf{y}) := \left[\varphi^l(\psi_\alpha^2, \mathbf{y}), \varphi^u(\psi_\alpha^2, \mathbf{y}) \right] = \left[\min_{\mathbf{x} \in D(\psi_\alpha^2, \mathbf{y})} \varphi(\mathbf{x}), \max_{\mathbf{x} \in D(\psi_\alpha^2, \mathbf{y})} \varphi(\mathbf{x}) \right],$$

where

$$(1.3) \quad D(\psi_\alpha^2, \mathbf{y}) := \{ \mathbf{x} \in \mathcal{X} : \|\mathbf{y} - f(\mathbf{x})\|_2^2 \leq \psi_\alpha^2 \}.$$

The statistical challenge for these intervals is choosing ψ_α^2 such that $\mathcal{I}(\psi_\alpha^2, \mathbf{y})$ has the desired coverage guarantee. Here $D(\psi_\alpha^2, \mathbf{y})$ assumes standardized noise; i.e., the covariance matrix of the noise has been transformed to the identity matrix.

The literature proposes two settings to guarantee coverage. One approach is to set ψ_α^2 such that $D(\psi_\alpha^2, \mathbf{y})$ is itself a confidence set for $\mathbf{x}^*$ in the parameter space. Since this choice would

automatically guarantee coverage for $\mathcal{I}(\psi_\alpha^2, \mathbf{y})$ regardless of the chosen QoI, it has been called “simultaneous” in [28], and “Simultaneous Strict Bounds” (SSB) in [31, 1] since it aligns with the setting of the strict bounds construction in [32]. Under the Gaussian assumption on the additive noise term ε , a $1 - \alpha$ confidence set for $\mathbf{x}^*$ can be obtained by setting $\psi_{SSB, \alpha}^2 := \chi_{n, \alpha}^2$, where $\chi_{n, \alpha}^2$ is the upper α -quantile for a chi-squared distribution with n degrees of freedom. This approach is conservative since the guarantee simultaneously holds for all possible QoI choices. By contrast, the “one-at-a-time” setting as described in [27, 28] (“one-at-a-time strict bounds” (OSB) as called in [31, 1]) tailors the interval coverage to one particular QoI. Under the same Gaussian assumption this proposed setting is $\psi_{OSB, \alpha}^2 = \chi_{1, \alpha}^2 + s(\mathbf{y})^2$, where $s(\mathbf{y})^2 := \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y} - f(\mathbf{x})\|_2^2$. Unlike the simultaneous setting, $D(\psi_{OSB, \alpha}^2, \mathbf{y})$ is not a $1 - \alpha$ confidence set for all $\mathbf{x}^*$, which makes proving the coverage guarantee difficult. The validity of this claim was proposed by the Burrus conjecture [27] and generally disproven by [1]. To address the challenge of calibrating strict bounds and optimization-based confidence intervals, [1] approached these intervals as an inverted likelihood ratio test. The distribution of the log-likelihood ratio (LLR) statistic is non-standard due to the constraints, complicating Type I error control.

Viewing QoI confidence sets as inverted likelihood ratio tests means that confidence set coverage can be achieved by controlling the Type I error rate of the hypothesis test. A mathematical challenge following from this formulation is that the parameter space of the null hypothesis is composite (i.e., contains more than one parameter setting) and can be unbounded. Since controlling the Type I error rate of a test with a composite null hypothesis test requires controlling at the worst-case parameter setting, the unboundedness poses a challenge.

1.2. Our approach and contributions. We address this unboundedness challenge by introducing the *Berger–Boos* set, inspired by [2] for handling composite null tests involving nuisance parameters, allowing us to create a bounded version of the composite null hypothesis space. We then develop maximum quantiles over the bounded Berger–Boos set to control the hypothesis test Type I error rates and thus construct calibrated confidence sets. In the additive Gaussian setting used in the algorithms below, this data-adaptive set has the simple form

$$\begin{aligned}
 \Gamma_\eta(\mathbf{y}) &:= \{\mathbf{y}' \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{y}'\|_2^2 \leq \chi_{n, \eta}^2\}, \\
 \mathcal{B}_\eta(\mathbf{y}) &:= f^{-1}(\Gamma_\eta(\mathbf{y})) \cap \mathcal{X} \\
 &= \{\mathbf{x} \in \mathcal{X} : \|\mathbf{y} - f(\mathbf{x})\|_2^2 \leq \chi_{n, \eta}^2\}.
 \end{aligned}
 \tag{1.4}$$

The set $\mathcal{B}_\eta(\mathbf{y})$ contains the true parameter with probability at least $1 - \eta$; the remaining miscoverage budget is used by calibrating the LLR tests at level γ , with $\gamma \leq \alpha - \eta$. The formal construction and coverage argument are given in Subsection 3.1.

Since the aforementioned OSB intervals are the state-of-the-art confidence intervals in this constrained inference setting, we conduct a series of numerical experiments showing that our intervals are competitive with OSB in situations where OSB achieves nominal coverage. In situations where OSB does not achieve nominal coverage, we show that our intervals do, while also achieving smaller expected interval lengths in a high-dimensional experiment.

Methodologically, we propose new confidence interval constructions for the setting described at the beginning of the section. Although the use of the Berger–Boos set is inspired by

[2], they originally applied it in a hypothesis testing setting for handling nuisance parameters. [19] applied a similar idea to nuisance parameters in a classification setting, but to the best of our knowledge, our paper is the first to use the idea in the ill-posed inverse problem UQ setting. Inspired by the methods used in simulation-based inference [8, 9, 18], we apply the approach of estimating the quantile function of a test statistic using simulated data. This work differs in the underlying model assumptions and composite nature of the null hypotheses. In addition, the prior work assumed bounded and relatively low-dimensional parameter spaces. The novelty of our method is therefore not the Berger–Boos construction, specific sampling algorithms, or quantile regression in isolation, but their integration into a coverage-guaranteed test-inversion procedure for functionals in constrained, ill-posed inverse problems. To the best of our knowledge, this work is the first to combine test inversion, sampling, and quantile regression to produce constraint-aware confidence intervals for functionals in high-dimensional, potentially unbounded settings.

Computationally, we adapt sampling-based calibration ideas to draw design points from the data-adaptive Berger–Boos set and to estimate the relevant LLR quantiles from simulated data. The sampling problem can be challenging because this set is an observed-data-dependent pre-image intersected with parameter constraints, and its geometry reflects the ill-posedness of the forward map. To efficiently sample from this set, we consider a Voelker–Gosmann–Stewart (VGS) [35] sampler in low dimensions and a custom-built Polytope/Vaidya-walk sampler [5] in higher dimensions and demonstrate their effectiveness in concrete ill-posed uncertainty quantification problems.

The rest of the paper is organized as follows. Section 2 provides the necessary statistical preliminaries for our method, including the hypothesis test we invert, the test statistic we use, and the test statistic quantile functions necessary for confidence set calibration. Section 3 defines the Berger–Boos set and downstream critical values used to calibrate several different confidence set varieties. It also provides two algorithms for computing the confidence sets. Section 4 provides a necessary theoretical result to justify that the algorithms described in Section 3 produce the correct confidence sets. Section 5 describes two core computational components necessary for the construction of our confidence sets in practice. Section 6 presents simple numerical experiments to illustrate the method’s operation and a high-dimensional particle physics-inspired experiment to demonstrate our method’s performance relative to the state-of-the-art. Finally, Section 7 summarizes our results and presents next steps.

2. Statistical preliminaries. A key part of the approach in [1] was to view the intervals in Equation (1.2) as inverted hypothesis tests, more specifically, a particular inverted likelihood ratio test. This change in perspective allowed for interval analysis using the properties of a particular log-likelihood ratio (LLR) test. We summarize this connection here, as it is critical to our method development carried out in Section 3.

Suppose $\mathbf{y} \sim P_{\mathbf{x}^*}$, where $P_{\mathbf{x}^*}$ is a distribution parameterized by a fixed but unknown $\mathbf{x}^* \in \mathbb{R}^p$. Let $\ell_{\mathbf{x}}(\mathbf{y})$ denote the log-likelihood of $\mathbf{x}$ evaluated at $\mathbf{y}$. We furthermore suppose we know a set $\mathcal{X} \subseteq \mathbb{R}^p$ such that $\mathbf{x}^* \in \mathcal{X}$. The case $P_{\mathbf{x}} = \mathcal{N}(\mathbf{K}\mathbf{x}, \mathbf{I})$, where the mean is linear in $\mathbf{x}$ and the additive noise is Gaussian with standardized covariance, is a scenario of particular interest; we henceforth refer to it as the *linear-Gaussian* case. In this case, the log-likelihood takes the following form, $\ell_{\mathbf{x}}(\mathbf{y}) = -\frac{1}{2}\|\mathbf{y} - \mathbf{K}\mathbf{x}\|_2^2 - \frac{n}{2}\log 2\pi$. Although our numerical

experiments in Section 6 focus on the linear-Gaussian case, we present the framework using the generic likelihood function to show that it theoretically holds more generally.

The duality between hypothesis tests and confidence sets is well known in statistics (see, e.g., Chapter 7 of [4] or Chapter 5 of [21]). We invert the following hypothesis test:

$$(2.1) \quad H_0 : \mathbf{x}^* \in \Phi_\mu \cap \mathcal{X} \quad \text{versus} \quad H_1 : \mathbf{x}^* \in \mathcal{X} - \Phi_\mu,$$

where $\Phi_\mu := \{\mathbf{x} : \varphi(\mathbf{x}) = \mu\}$. Since we seek a confidence interval on the real line, each hypothesis test is defined by $\mu \in \mathbb{R}$. This null hypothesis is *composite*: fixing the value of the QoI, $\varphi(\mathbf{x}) = \mu$, generally does not identify a single parameter vector, but instead leaves all parameter settings in the level set $\Phi_\mu \cap \mathcal{X}$.

We define the following test statistic to evaluate Test (2.1):

$$(2.2) \quad \lambda(\mu, \mathbf{y}) := \begin{cases} \inf_{\mathbf{x} \in \Phi_\mu \cap \mathcal{X}} (-2\ell_{\mathbf{x}}(\mathbf{y})) - \inf_{\mathbf{x} \in \mathcal{X}} (-2\ell_{\mathbf{x}}(\mathbf{y})), & \text{if } \Phi_\mu \cap \mathcal{X} \neq \emptyset, \\ \infty, & \text{otherwise.} \end{cases}$$

where μ denotes the level set of the null hypothesis, $\mathbf{y}$ is the observed data, and $\mathcal{X}$ is the constraint set known to contain the parameter. The test is rejected if the test statistic is large, and hence we automatically reject the null if $\Phi_\mu \cap \mathcal{X} = \emptyset$ independently of the observed data. Alternatively, we can consider the test (2.1) only for $\mu \in \varphi(\mathcal{X})$ to perform test inversion. Under the linear-Gaussian case, the LLR test statistic is,

$$(2.3) \quad \lambda(\mu, \mathbf{y}) = \min_{\mathbf{x} \in \Phi_\mu \cap \mathcal{X}} \|\mathbf{y} - \mathbf{K}\mathbf{x}\|_2^2 - \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y} - \mathbf{K}\mathbf{x}\|_2^2.$$

We control the behavior of this test statistic, and therefore the test, by bounding from above the probability of erroneously rejecting the null hypothesis (i.e., Type I error). As such, we consider the distribution of $\lambda(\mu, \mathbf{y})$ under the null. For each $\mathbf{x} \in \mathcal{X}$, let $\mu = \varphi(\mathbf{x})$. Define $Q_{\mathbf{x}} : (0, 1) \rightarrow \mathbb{R}$ such that, for all $\alpha \in (0, 1)$,

$$(2.4) \quad \mathbb{P}(\lambda(\mu, \mathbf{y}) \leq Q_{\mathbf{x}}(1 - \alpha)) = 1 - \alpha,$$

where the probability is over $\mathbf{y} \sim P_{\mathbf{x}}$. We refer to $Q_{\mathbf{x}}$ as the *quantile function* of the LLR under the null hypothesis at $(\mathbf{x}, \varphi(\mathbf{x}) = \mu)$. Since the null hypothesis is composite, using this quantile function to define a cutoff is not enough to control Type I error. Thus, we use the *sliced* maximum quantile function over the level set under consideration:

$$(2.5) \quad Q_{\mu, 1-\alpha}^{\max} := \sup_{\mathbf{x} \in \Phi_\mu \cap \mathcal{X}} Q_{\mathbf{x}}(1 - \alpha).$$

Note that we refer to this version as “sliced” since it considers the maximum of the quantile function defined by the level set of the functional, defining a slice through the constrained parameter space. $Q_{\mu, 1-\alpha}^{\max}$ as defined above controls the Type I error for a specific value of μ . We can define a more conservative confidence set using the following *global* maximum quantile:

$$(2.6) \quad Q_{1-\alpha}^{\max} := \sup_{\mu \in \varphi(\mathcal{X})} Q_{\mu, 1-\alpha}^{\max} = \sup_{\mathbf{x} \in \mathcal{X}} Q_{\mathbf{x}}(1 - \alpha).$$

By Lemma 2.1 in [1], the set

$$(2.7) \quad \mathcal{C}_\alpha^\mu(\mathbf{y}; \mathcal{X}) := \{\mu : \lambda(\mu, \mathbf{y}) \leq Q_{\mu, 1-\alpha}^{\max}\} \subset \mathbb{R}$$

defines a $1 - \alpha$ confidence set for the true functional, $\mu^* = \varphi(\mathbf{x}^*)$. The global quantile (2.6) is more conservative than the sliced (2.5) since it holds for all null hypotheses, and thus,

$$(2.8) \quad \mathcal{C}_\alpha(\mathbf{y}; \mathcal{X}) := \{\mu : \lambda(\mu, \mathbf{y}) \leq Q_{1-\alpha}^{\max}\} \subset \mathbb{R}$$

is also a $1 - \alpha$ confidence set.

To connect this framework back to Interval (1.2), under the linear-Gaussian assumption producing the LLR in Equation (2.3), [1] (Theorem 2.4) proved that

$$(2.9) \quad \left[\inf_{\mu: \lambda(\mu, \mathbf{y}) \leq Q_{1-\alpha}^{\max}} \mu, \sup_{\mu: \lambda(\mu, \mathbf{y}) \leq Q_{1-\alpha}^{\max}} \mu \right] = \left[\inf_{\mathbf{x} \in D(Q_{1-\alpha}^{\max} + s(\mathbf{y})^2, \mathbf{y})} \varphi(\mathbf{x}), \sup_{\mathbf{x} \in D(Q_{1-\alpha}^{\max} + s(\mathbf{y})^2, \mathbf{y})} \varphi(\mathbf{x}) \right],$$

where $D(\cdot, \mathbf{y})$ is defined by Equation (1.3). This equivalence asserts that by setting $\psi_\alpha^2 := Q_{1-\alpha}^{\max} + s(\mathbf{y})^2$, where $s(\mathbf{y})^2 := \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y} - \mathbf{K}\mathbf{x}\|_2^2$, we guarantee coverage for OSB Interval (1.2). It asserts that the original OSB interval formulation is valid if and only if $\chi_{1,\alpha}^2 \geq Q_{1-\alpha}^{\max}$ (see Lemma 2.2 and Corollary 2.3). [1] showed that this inequality generally does not hold.

As explored in [1] and reiterated above, Interval (1.2) can be calibrated by computing $Q_{\mu, 1-\alpha}^{\max}$ or $Q_{1-\alpha}^{\max}$. However, pursuing calibration in this way is computationally challenging and statistically conservative. Both of these values require the ability to evaluate $Q_{\mathbf{x}}$. Without parameter constraints, this quantile function can be constant under linear-Gaussian assumptions [1], i.e., the test statistic is pivotal. But even with a relatively simple two-dimensional Gaussian noise model with non-negativity constraints, this quantile function becomes non-trivial (e.g., see Figure 5.3 in [1]). Beyond the practical difficulty of dealing with the underlying quantile function, both (2.5) and (2.6) can be expressed as chance-constrained optimization problems, which are known to be NP-hard in general and would need to be solved over an unbounded constraint set [10, 24, 1]. Statistically, since both (2.5) and (2.6) are the largest quantile function values subject to their respective constraints, they are, by definition, conservative, especially in scenarios where the true $\mathbf{x}^*$ is far from these most conservative points. Since we do not know $\mathbf{x}^*$, this conservatism is necessary under the framework of [1].

As we see in Section 3, our method addresses both challenges. In scenarios where evaluating $Q_{\mathbf{x}}$ is difficult, we sample a collection of design points in a bounded subset of the constraint set, sample the LLR under the null at each design point, and use quantile regression to estimate the quantile surface. The most conservative points can potentially be removed from consideration by only considering parameter values that are *not unlikely* given the observed data.

3. Interval construction methodology. This section presents four related interval constructions building on the theory of [1] by addressing the aforementioned key challenges. The first implementation challenge is to handle the potentially unbounded constraint set, for which we define and apply the Berger–Boos set to create a data-dependent subset of the original constraints. This set leads to our four interval definitions, which follow a two-stage taxonomy: Global versus Sliced and Inverted versus Optimized. The second implementation challenge

is computing these interval constructions, which we achieve using a combination of novel sampling algorithms and quantile regression. We will cover this in a later section (Section 5).

We rewrite the data-generating process articulated in the introduction:

$$(3.1) \quad \mathbf{y} = f(\mathbf{x}^*) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{x}^* \in \mathcal{X}$$

where $f : \mathbb{R}^p \rightarrow \mathbb{R}^n$ is a known forward model and $\text{Cov}(\varepsilon) = \mathbf{I}$, without loss of generality. Note, in the linear-Gaussian case, we have $f(\mathbf{x}^*) = \mathbf{K}\mathbf{x}^*$. Let $f^{-1}(A) := \{\mathbf{x} : f(\mathbf{x}) \in A\}$ be the pre-image of a set $A \subset \mathbb{R}^n$ under the forward map f .

Although the assumed data-generating process in Equation (3.1) is less general than $\mathbf{y} \sim P_{\mathbf{x}^*}$ assumed in Section 2, it is sufficiently general to contain the application areas mentioned in Subsection 1.1. Slightly generalizing the form of the additive noise distribution would make the construction of the Berger–Boos set more complicated, but it is possible in principle.

3.1. Global and sliced confidence sets using the Berger–Boos set. Both $Q_{\mu, 1-\alpha}^{\max}$ and $Q_{1-\alpha}^{\max}$ in Section 2 suffer from the same theoretical and practical concerns. Theoretically, they are *conservative* in the sense that they control Type I error under the worst case truth (i.e., the parameter setting with the largest quantile). Practically, not only can $Q_{\mathbf{x}}$ be difficult to compute, but if the constraint set $\mathcal{X}$ is unbounded, computing quantiles becomes even more difficult. Both challenges stem from the composite nature of the null hypothesis (H_0 in Equation (2.1)). Namely, for a given μ , $\Phi_\mu \cap \mathcal{X}$ usually contains more than one feasible parameter value, so Type I error control must hold uniformly over all of these possible truths. [2] introduce a compelling solution in the context of hypothesis testing with nuisance parameters (a special case of hypothesis testing on $\varphi(\mathbf{x})$), which is to control Type I error only over a data-informed region of the parameter space. Following their construction, we build a confidence interval that, instead of using the maximum $1 - \alpha$ quantile over $\mathcal{X}$ in (2.6) (or $\mathcal{X} \cap \Phi_\mu$ in (2.5)), uses a larger quantile ($1 - \gamma$ with $\gamma < \alpha$) but that is maximized over a smaller set. The construction follows a three-step process:

(1) Choose $\eta \in (0, \alpha)$ and build a $1 - \eta$ confidence set for $\mathbf{x}^*$, $\mathcal{B}_\eta$. Under the additive Gaussian noise assumption in Equation (3.1), letting $\Gamma_\eta(\mathbf{y}) := \{\mathbf{y}' \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{y}'\|_2^2 \leq \chi_{n,\eta}^2\}$ we have $\mathbb{P}(f(\mathbf{x}^*) \in \Gamma_\eta(\mathbf{y})) = 1 - \eta$, hence

$$(3.2) \quad \mathcal{B}_\eta := f^{-1}(\Gamma_\eta(\mathbf{y})) \cap \mathcal{X} = \{\mathbf{x} \in \mathcal{X} : \|\mathbf{y} - f(\mathbf{x})\|_2^2 \leq \chi_{n,\eta}^2\}$$

is a $1 - \eta$ confidence set for $\mathbf{x}^{*1}$. This pre-image confidence set, $\mathcal{B}_\eta$, is the Berger–Boos set.

(2) Optimize the $1 - \gamma$ quantile of the test statistic over $\mathcal{B}_\eta$ (or $\mathcal{B}_\eta \cap \Phi_\mu$), instead of $\mathcal{X}$ (or $\mathcal{X} \cap \Phi_\mu$), where $\gamma < \alpha$ is chosen to ensure calibration. Lemma 3.1 proves that $\gamma = \alpha - \eta$ is valid.

(3) Use the obtained quantiles, which we define as

$$(3.3) \quad \bar{q}_{\gamma,\eta}(\mu) := \sup_{\mathbf{x} \in \mathcal{B}_\eta \cap \Phi_\mu} Q_{\mathbf{x}}(1 - \gamma),$$

$$(3.4) \quad \bar{q}_{\gamma,\eta} := \sup_{\mathbf{x} \in \mathcal{B}_\eta} Q_{\mathbf{x}}(1 - \gamma),$$

¹Intersecting the pre-image $f^{-1}(\Gamma_\eta(\mathbf{y}))$ with the constraint set $\mathcal{X}$ does not change the coverage probability since we know $\mathbf{x}^* \in \mathcal{X}$.

and construct the following sliced (sl) and global (gl) confidence sets:

$$(3.5) \quad C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta) := \{\mu : \lambda(\mu, \mathbf{y}) \leq \bar{q}_{\gamma, \eta}(\mu)\},$$

$$(3.6) \quad C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta) := \{\mu : \lambda(\mu, \mathbf{y}) \leq \bar{q}_{\gamma, \eta}\}.$$

Figure 1 provides a visual roadmap of the calibration procedure after the relevant quantities have been defined. The left panel illustrates how the LLR quantile function $Q_{\mathbf{x}}$ can vary over the constrained parameter space. The right panels show how the data-adaptive Berger–Boos set restricts the region over which the quantiles are maximized and how the resulting global or sliced cutoffs are used to invert the LLR test.

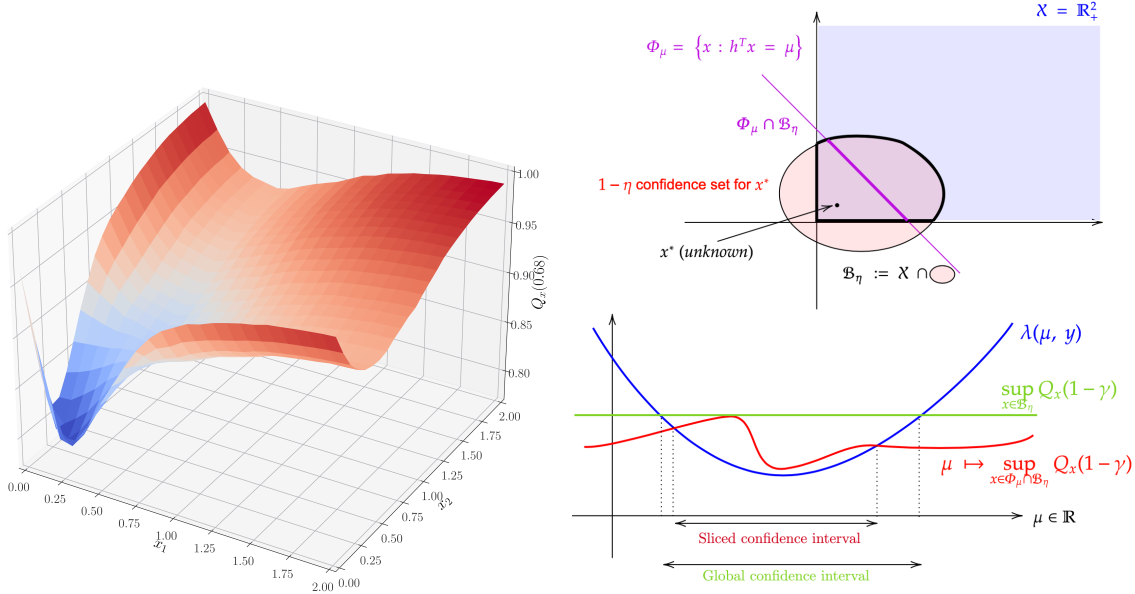


Figure 1: Visual roadmap of the data-adaptive calibration procedure. Left: an illustrative LLR quantile surface, $Q_{\mathbf{x}}(1 - \alpha)$, in a two-dimensional constrained-Gaussian example, showing how the relevant quantile can vary over the constrained parameter space. Right-top: the Berger–Boos set $\mathcal{B}_\eta$ restricts the constrained parameter space to a data-adaptive region. Right-bottom: global or sliced LLR cutoffs, based on $\bar{q}_{\gamma, \eta}$ or $\bar{q}_{\gamma, \eta}(\mu)$, yield different inverted intervals for the QoI.

Analogous to the presentation in [1], we define *sliced* and *global* max-quantiles that include the Berger–Boos set to control the Type I error of Test (2.1). Both max-quantiles maximize a larger γ -quantile over the set of interest instead of the α -quantile, as shown in (2.5) and (2.6).

The following Lemma (analogous to Lemma 2.1 in [1]) gives a sufficient condition to select the values γ and η to ensure a $1 - \alpha$ coverage.

Lemma 3.1 (Setting η and γ to guarantee $1 - \alpha$ coverage). *Let $\alpha \in (0, 1)$ and define $\bar{q}_{\gamma, \eta}(\mu)$ according (3.3) and $C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta)$ according to (3.5). For $\eta \in (0, \alpha)$, $C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta)$ is a $1 - \alpha$ confidence set for $\mu^* = \varphi(\mathbf{x}^*)$ if $\gamma \leq \alpha - \eta$. The length of the obtained interval is a non-increasing function of γ , so the tightest interval will be obtained when equality is satisfied.*

The coverage guarantee implied by [Lemma 3.1](#) also implies coverage for $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$, as shown in the following Corollary. Proofs of both statements are provided in [Section SM1](#) (in the Supplementary Material).

Corollary 3.2. *In the setting of [Lemma 3.1](#), $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$ is a $1 - \alpha$ confidence set for $\mu^* = \varphi(\mathbf{x}^*)$ if $\gamma \leq \alpha - \eta$.*

Remark 3.3. Following the generality of the original construction, [Lemma 3.1](#) generalizes to any $1 - \eta$ confidence set of x^* and test statistic calibrated with the $1 - \gamma$ quantile. In our particular case, since both the confidence set $\mathcal{B}_\eta$ and the test statistic $\lambda(\mu, \mathbf{y})$ depend on the Gaussian log-likelihood $\|\mathbf{y} - f(\mathbf{x})\|_2^2$, there is a stronger interpretation of the Berger-Boos construction. Whenever $\mathcal{B}_\eta \cap \Phi_\mu \neq \emptyset$, the Berger-Boos construction is equivalent to a data-dependent reduction of the constraint set, replacing $\mathcal{X}$ for $\mathcal{B}_\eta$ both in the test statistic and the quantile optimization problems. Since this is done after seeing the data, the optimized quantile needs to increase to $1 - \alpha + \eta$ to maintain $1 - \alpha$ coverage.

Navigating the Berger-Boos parameter choices and trade-offs. Setting $\eta = 0$ in $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$ we maximize $Q_{\mathbf{x}}(1 - \gamma) = Q_{\mathbf{x}}(1 - \alpha)$ over $\mathcal{X}$, which returns $Q_{1-\alpha}^{\text{max}}$ from Equation (2.6). For $\eta > 0$, the Berger-Boos construction restricts the quantile maximization to a smaller subset of the parameter space while maximizing a larger $1 - \gamma$ quantile to maintain the desired $1 - \alpha$ confidence level of the final confidence set. Fundamentally, the η choice is a trade-off between these two opposing effects: As we increase η from 0, the set over which the quantile function is optimized shrinks, but the quantile level is increased. In [Subsection 6.2](#), we perform a numerical experiment comparing average interval length for different values of η , showing that a small $\eta > 0$ can be beneficial both numerically (since in many instances it allows sampling over a bounded set) and in terms of the average interval length. In practice, we treat η as a small pre-specified calibration parameter rather than choosing it after inspecting coverage outcomes. The original Berger-Boos recommendation and our experiments both support using a small positive value and, when computationally affordable, checking the sensitivity of interval length to a small grid of candidate values. A fully automated rule for choosing an expected-length-optimal η while preserving the finite-sample coverage guarantee is an important direction for future work. Related calibration-parameter tuning ideas have been explored in other UQ settings, including objective frequentist uncertainty quantification for atmospheric CO₂ retrievals in [\[22\]](#); adapting such ideas to the present Berger-Boos/test-inversion setting while preserving finite-sample coverage is left for future work.

3.2. Algorithmic implementations. There are possibly many ways to compute $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$ and $C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta)$ in practice. Obtaining either set reduces to the computation of $\bar{q}_{\gamma, \eta}(\mu)$ and $\bar{q}_{\gamma, \eta}$. If the quantile function $Q_{\mathbf{x}}$ and its gradient $\nabla_{\mathbf{x}} Q_{\mathbf{x}}$ could be evaluated, computation could potentially be achieved using a first-order numerical optimizer. However, since the sliced max-quantile, $\bar{q}_{\gamma, \eta}(\mu)$, is a function of the level-set parameter, μ , such an optimization would have to be done for each possible functional value. Since easy function and gradient evaluations rarely exist (see [\[1\]](#) for some examples where such evaluations are possible), this paper develops a *sampling*-based approach to estimate these quantities to construct Confidence Sets (3.5) and (3.6). Once we estimate $\bar{q}_{\gamma, \eta}(\mu)$ and $\bar{q}_{\gamma, \eta}$, we can either use the output from the sampling algorithm to compute the Global or Sliced interval via classical test inversion, or we can use

the estimated max-quantiles in optimizations similar to those in Interval (2.9). As such, we introduce four interval constructions: Global Inverted, Global Optimized, Sliced Inverted, and Sliced Optimized. These four options are summarized in Table 1.

We use our ability to sample $\lambda(\mu, \mathbf{y})$ via $\mathbf{y} \sim \mathcal{N}(f(\mathbf{x}), \mathbf{I})$ to estimate the desired max-quantiles. We present two algorithms, both of which first generate a random set of design points from the Berger–Boos set. If we can directly *efficiently* compute the desired quantile at each design point, we use Algorithm 3.1. Absent such an ability, Algorithm 3.2 presents an alternative, which first samples a realization of the LLR test statistic at each design point and then performs quantile regression with the generated pairs of design points and LLR values to estimate the underlying quantile surface. The sampled design points and either the exact or estimated quantiles at the design points are then used to produce the final intervals.

Both algorithms start by sampling $\mathcal{B}_\eta$ uniformly at random to generate a collection of M design points, i.e., $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_M \sim \mathcal{U}(\mathcal{B}_\eta)$. Section 5.1 provides the practical sampler choices used for this step: a VGS-based sampler in low-dimensional full-rank settings and a Polytope sampler based on the Vaidya walk in higher-dimensional or rank-deficient settings. These geometric difficulties are well known in convex-body and polytope sampling. Our contribution is not a new generic sampler for such sets, but rather the use of appropriate samplers to generate design points over the data-adaptive Berger–Boos set and across the range of QoI values needed for calibrated test inversion. If computing $\lambda(\mu, \mathbf{y})$ for a given $(\mu, \mathbf{y})$ is inexpensive, Algorithm 3.1 directly estimates the quantile function at each design point. This computation is most likely inexpensive when there are closed-form solutions to the LLR’s subordinate optimizations. As shown in Algorithm 3.1, an easy way to estimate each design point’s quantile is to sample its test statistic N times and take the appropriate percentile. More often, computing $\lambda(\mu, \mathbf{y})$ is expensive since it involves two constrained optimizations, which are possibly non-convex due to either the constraints or the forward model, or just numerically challenging due to the ill-posedness of the problem. In this scenario, Algorithm 3.2 samples a realization of the test statistic at each design point and estimates the quantile function over the Berger–Boos set using quantile regression. While Algorithm 3.1 relies upon computational strength to compute the LLR $N \times M$ times, Algorithm 3.2 shifts complexity to the quantile regression and relieves the computational burden by assuming a smooth quantile surface. We discuss some considerations to this end in Section 5.2. We note that although Algorithm 3.1 and Algorithm 3.2 make use of the additive Gaussian noise assumption, they are not necessarily limited to this assumption given the proper adjustments to the Berger–Boos set.

The approach of Algorithm 3.2 is inspired by recent UQ approaches in likelihood-free inference, where one can sample from a likelihood but cannot easily compute it. Specifically, we draw inspiration from the use of quantile regression in [8, 9, 18, 19]. Although these approaches differ in detail and implementation from our approach (e.g., they typically focus on settings with low-dimensional $\mathbf{x}$), they overlap in the sampling and quantile regression perspectives, which effectively allow machine learning to supplement computationally intensive or intractable quantities. Here, rather than assuming a purely stochastic forward model and therefore only being able to sample from the likelihood, we assume f is a deterministic function involved in $\lambda(\mu, \mathbf{y})$, which is a random quantity due to the additive noise. Since Algorithm 3.2 involves training a quantile regressor, it includes separate training and testing sets of design points over the Berger–Boos set and samples from their respective test statistics.

Algorithm 3.1 Direct estimation of quantiles**Input:** $\alpha, \gamma, \eta \in (0, 1)$ such that $\gamma = \alpha - \eta$, $M, N \in \mathbb{N}$.

- 1: **Construct the Berger–Boos set:** Define $\Gamma_\eta(\mathbf{y}) := \{\mathbf{y}' \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{y}'\|_2^2 \leq \chi_{n,\eta}^2\}$, so that $\mathbb{P}(f(\mathbf{x}^*) \in \Gamma_\eta(\mathbf{y})) \geq 1 - \eta$, and set $\mathcal{B}_\eta$ as in Equation (3.2).
- 2: **Sample from the Berger–Boos confidence set, $\mathcal{B}_\eta$:** Sample $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_M \sim \mathcal{U}(\mathcal{B}_\eta)$.
- 3: **for** $k = 1, 2, \dots, M$ **do**
- 4: **Sample noise realizations:** Sample N noise realizations: $\varepsilon_1, \dots, \varepsilon_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- 5: **Generate LLR draws:** For $i = 1, \dots, N$, set $\mathbf{y}_i := f(\bar{\mathbf{x}}_k) + \varepsilon_i$ and compute $\lambda_i := \lambda(\varphi(\bar{\mathbf{x}}_k), \mathbf{y}_i; \mathcal{X})$.
- 6: **Compute percentile estimate of the γ -quantile:** Compute the $(1 - \gamma) \times 100$ percentile of the LLR samples for the data-generating process under $\bar{\mathbf{x}}_k$, i.e., $\hat{q}_\gamma^k := \lambda_{(\{(1-\gamma)N\})}$, where $\{\cdot\}$ denotes the nearest whole number and $\lambda_{(i)}$ denotes the i -th order statistic.
- 7: **end for**

Output: Pairs of sampled design points and their respective γ -quantiles, i.e., $\{(\bar{\mathbf{x}}_k, \hat{q}_\gamma^k)\}_{k=1}^M$.**Algorithm 3.2** Quantile regression estimate of quantile surface**Input:** $\alpha, \gamma, \eta \in (0, 1)$ such that $\gamma = \alpha - \eta$; $M_{\text{tr}}, M \in \mathbb{N}$.

- 1: **Construct the Berger–Boos set:** Define $\Gamma_\eta(\mathbf{y}) := \{\mathbf{y}' \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{y}'\|_2^2 \leq \chi_{n,\eta}^2\}$, so that $\mathbb{P}(f(\mathbf{x}^*) \in \Gamma_\eta(\mathbf{y})) \geq 1 - \eta$, and set $\mathcal{B}_\eta$ as in Equation (3.2).
- 2: **Sample from the Berger–Boos confidence set $\mathcal{B}_\eta$:** Sample $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_{M_{\text{tr}}} \sim \mathcal{U}(\mathcal{B}_\eta)$ design points to train the quantile regressor and $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_M \sim \mathcal{U}(\mathcal{B}_\eta)$ test points to invert the interval, generating $M_{\text{tr}} + M$ total samples. Since the test points are used for the interval inversion, they are used as out-of-sample points for the quantile regressor.
- 3: **for** $k = 1, 2, \dots, M_{\text{tr}}$ **do**
- 4: **Sample a noise realization:** Sample a noise realization: $\varepsilon_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- 5: **Generate one LLR draw:** Set $\mathbf{y}_k := f(\bar{\mathbf{x}}_k) + \varepsilon_k$ and compute $\lambda_k := \lambda(\varphi(\bar{\mathbf{x}}_k), \mathbf{y}_k; \mathcal{X})$.
- 6: **end for**
- 7: **Estimate the quantile function using quantile regression:** Using the generated pairs $\{(\bar{\mathbf{x}}_k, \lambda_k)\}_{k=1}^{M_{\text{tr}}}$, estimate the upper γ -conditional quantile function, $\hat{q}_\gamma(\mathbf{x})$, using quantile regression.

Output: Generate γ -quantile predictions at out-of-sample test points, $\{(\tilde{\mathbf{x}}_k, \hat{q}_\gamma(\tilde{\mathbf{x}}_k))\}_{k=1}^M$.

With the generated pairs from Algorithms (3.1) and (3.2), we present two strategies to estimate each of $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$ and $C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta)$. To streamline notation, let $(\mathbf{x}_k, q_\gamma^k)$ denote the k -th pair from either algorithm. This is notationally helpful since Algorithm 3.1 only generates one set of parameter samples, whereas Algorithm 3.2 generates two. That is, using Algorithm 3.1, $\mathbf{x}_k := \bar{\mathbf{x}}_k$ and $q_\gamma^k := \hat{q}_\gamma^k$ and using Algorithm 3.2, $\mathbf{x}_k := \tilde{\mathbf{x}}_k$ and $q_\gamma^k := \hat{q}_\gamma(\tilde{\mathbf{x}}_k)$.

To estimate the global confidence set, we estimate $\bar{q}_{\gamma,\eta}$ using the empirical maximum,

$\hat{q} := \max_k q_\gamma^k$. This results in the following two interval constructions:

$$(3.7) \quad C_{\text{opt}}^{\text{gl}}(\mathbf{y}) := \min / \max \{ \varphi(\mathbf{x}) : \mathbf{x} \in D(\hat{q} + s(\mathbf{y})^2, \mathbf{y}) \}$$

$$(3.8) \quad C_{\text{inv}}^{\text{gl}}(\mathbf{y}) := \min / \max \{ \varphi(\mathbf{x}_k) : k = 1, \dots, M \text{ and } \lambda(\varphi(\mathbf{x}_k), \mathbf{y}; \mathcal{X}) \leq \hat{q} \},$$

We refer to Interval (3.8) as the Global Inverted interval construction since its endpoints are defined by only those sampled parameter values that comport with the maximum estimated quantile LLR cutoff. We refer to Interval (3.7) as the Global Optimized interval since its endpoints are defined by the extreme functional values of a feasible region defined by $\hat{q}$. Although, $C_{\text{inv}}^{\text{gl}}(\mathbf{y}) \neq C_{\text{opt}}^{\text{gl}}(\mathbf{y})$ due to finite sample, they are asymptotically equal and in practice show similar performance in terms of coverage and expected length, as shown in Section 6. We prove the consistency of the interval constructed via inversion and via optimization in Theorem 4.1. Since the endpoints of both the inverted and optimized intervals converge in probability to the endpoints of $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$, the two interval constructions are asymptotically equivalent.

To estimate the sliced set, we estimate $\bar{q}_{\gamma, \eta}(\mu)$ using a rolling maximum of the sampled q_γ^k values, as ordered by the sampled functional values, and directly accept or reject functional values based on their estimated quantile. These approaches result in two interval constructions:

$$(3.9) \quad C_{\text{opt}}^{\text{sl}}(\mathbf{y}) := \min / \max \{ \mu \in \mathbb{R} : \lambda(\mu, \mathbf{y}; \mathcal{X}) \leq \hat{m}_\gamma(\mu) \},$$

$$(3.10) \quad C_{\text{inv}}^{\text{sl}}(\mathbf{y}) := \min / \max \{ \varphi(\mathbf{x}_k) : k = 1, \dots, M \text{ and } \lambda(\varphi(\mathbf{x}_k), \mathbf{y}; \mathcal{X}) \leq q_\gamma^k \},$$

where $\hat{m}_\gamma(\mu)$ denotes rolling estimate of $\bar{q}_{\gamma, \eta}(\mu)$ defined as follows. We refer to $I^{\text{sl}}(\mathbf{y})$ as the “sliced” index set, i.e., those indices for which the LLR at a particular functional value is less than the estimated quantile at a design point generating that functional value.

The rolling maximum quantile is defined using estimated quantiles ordered by the sampled functional values. Choose “rolling” parameter, $T \in \mathbb{N}$, and let $\sigma(1), \sigma(2), \dots, \sigma(M)$ define an ordering such that $\mu_{\sigma(k)} \leq \mu_{\sigma(k+1)}$ for all $k = 1, \dots, M-1$. Define

$$Q_k := \{q_\gamma^{\sigma(k)}, q_\gamma^{\sigma(k-1)}, \dots, q_\gamma^{\sigma(k-T)}\}.$$

Then, for a given $\mu \in [\min_{\mathbf{x} \in \mathcal{B}_\eta} \varphi(\mathbf{x}), \max_{\mathbf{x} \in \mathcal{B}_\eta} \varphi(\mathbf{x})]$, define $k^*(\mu) := \arg\min_k |\mu - \mu_{\sigma(k)}|$ and define $\hat{m}_\gamma(\mu) := \max\{q \in Q_{k^*(\mu)}\}$. Using the rolling maximum quantile is one way to estimate $\bar{q}_{\gamma, \eta}(\mu)$. One could also bin the quantiles by functional value, compute the maximum predicted quantile in each bin, and then fit a nonparametric regression to fit the maximum binned quantiles to the functional values.

This estimator choice for $\hat{m}_\gamma(\mu)$ affects how one computes $C_{\text{opt}}^{\text{sl}}(\mathbf{y})$. To see why, note that we can re-express $C_{\text{opt}}^{\text{sl}}(\mathbf{y})$ as follows:

$$(3.11) \quad C_{\text{opt}}^{\text{sl}}(\mathbf{y}) := \min / \max \{ \varphi(\mathbf{x}) : \mathbf{x} \in D(\hat{m}_\gamma(\varphi(\mathbf{x})) + s(\mathbf{y})^2, \mathbf{y}) \}.$$

As such, the estimated $\hat{m}_\gamma(\cdot)$ could be substituted in each endpoint optimization. However, this estimated curve is likely not convex in $\mathbf{x}$ and therefore complicates optimizations. One could also pursue a root-finding approach to find the set of μ such that $\lambda(\mu, \mathbf{y}) = \hat{m}_\gamma(\mu)$ since these intersection points define a set of accepted functional values. This approach can

Table 1: Summary of methods based on global and sliced max quantile, and whether they are optimization-based or inversion-based.

Category	Global max quantile	Sliced max quantile
Inversion-based methods	Interval (3.8)	Interval (3.10)
Optimization-based methods	Interval (3.7)	Interval (3.9)

be as complex as the complexity of the estimated curve. In our view, the most pragmatic approach is to determine all μ_k such that $\lambda(\mu_k, \mathbf{y}; \mathcal{X}) \leq \hat{m}_\gamma(\mu_k)$, and then define the endpoints of $C_{\text{opt}}^{\text{sl}}(\mathbf{y})$ as the minimum and maximum of the accepted sample values. The consistency of the computed endpoints for both inverted and optimized sliced intervals is proven in [Theorem 4.1](#). Similarly to the global interval constructions, the consistency of both constructions to $C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta)$ establishes the asymptotic equivalence of the inverted and optimized approaches.

Although both [Algorithms 3.1](#) and [3.2](#) can produce a quantile estimate, it is worth noting a few key differences between the two. First, [Algorithm 3.1](#) involves a nested sampling loop. Thus, any statistical guarantee regarding the validity of its output has the two moving parts of the accuracy of $\hat{q}_\gamma^k$ as N gets large and the proximity of the approximated quantile to the true maximum quantile function value over $\mathcal{B}_\eta$ as M gets large. By contrast, [Algorithm 3.2](#) only has one sampling loop and estimates the full quantile function surface over $\mathcal{B}_\eta$, producing an estimate of the maximum as a consequence, but is reliant on smoothness to do an accurate regression with finitely many data points. In [Section 4](#), we provide a rigorous theoretical justification for the consistency of each algorithm estimate and prove that the consistency propagates to all defined interval constructions.

4. Theoretical justification. In this section, we provide convergence results for the interval constructions of the previous section. We prove that, under certain assumptions, whenever [Algorithm 3.1](#) or [Algorithm 3.2](#) are used to estimate the maximum quantiles, the global interval constructions (3.7) and (3.8) converge in probability to the true global interval $C_\alpha^{\text{gl}}(\mathbf{y}; \mathcal{B}_\eta)$ in (3.6), and the sliced interval constructions (3.9) and (3.10) converge in probability to the true sliced interval $C_\alpha^{\text{sl}}(\mathbf{y}; \mathcal{B}_\eta)$ in (3.5). Although our approach involving quantile regression is inspired by the approaches in [\[8, 9, 18\]](#) and therefore requires similar theoretical results to connect the consistency of the quantile estimation with the interval validity, our approach is sufficiently different so that it requires novel theoretical insights. First, the inversion of a hypothesis test with a composite null adds a layer of complexity to the proofs. [\[9\]](#) do address composite null hypotheses, but not in the functional setting, as we do here. Second, since we use the sampled points to invert the interval, it is insufficient to show that the cutoff is consistent as the results in [\[8, 9, 18\]](#) show. Instead, our inverted intervals require proof that the samplers can get arbitrarily close to the true endpoint boundaries.

Throughout the following results, we assume that for fixed $\gamma \in (0, 1)$, $Q_{\mathbf{x}}(1 - \gamma)$ is a continuous function of $\mathbf{x}$. We refer the reader to [\[13\]](#) for a complete analysis of the properties of parametrized quantile functions. We furthermore assume that $\mathcal{B}_\eta$ is a compact set without isolated points so that the sampling methods eventually sample close to every point.

Theorem 4.1. Assume that $Q_{\mathbf{x}}(1 - \gamma)$ is a continuous function of $\mathbf{x}$, and let $\mathcal{B}_\eta$ be compact and without isolated points. Let the quantile regression in [Algorithm 3.2](#) be consistent for all $\mathbf{x}$, i.e., such that $\mathbb{P}(|\hat{q}_\gamma(\mathbf{x}) - Q_{\mathbf{x}}(1 - \gamma)| > \varepsilon) \rightarrow 0$ as $M_{tr} \rightarrow \infty$ is satisfied $\forall \varepsilon > 0$. We will write $\xrightarrow{P}$ for convergence in probability, understood as $N, M \rightarrow \infty$ if [Algorithm 3.1](#) is used and as $M_{tr}, M \rightarrow \infty$ if [Algorithm 3.2](#) is used. For either algorithm, we have, for a given observation $\mathbf{y}$: $C_{inv}^{gl}(\mathbf{y}) \xrightarrow{P} C_\alpha^{gl}(\mathbf{y})$, $C_{inv}^{sl}(\mathbf{y}) \xrightarrow{P} C_\alpha^{sl}(\mathbf{y})$. Further assume that there exists a point $\bar{\mu} \in \varphi(\mathcal{B}_\eta)$ satisfying $\lambda(\bar{\mu}, \mathbf{y}; \mathcal{X}) < \bar{q}_{\gamma, \eta}$, then $C_{opt}^{gl}(\mathbf{y}) \xrightarrow{P} C_\alpha^{gl}(\mathbf{y})$. Finally, further assuming technical conditions discussed in [Section SM2](#), then $C_{opt}^{sl}(\mathbf{y}) \xrightarrow{P} C_\alpha^{sl}(\mathbf{y})$.

Note that if $\lambda(\mu, \mathbf{y})$ is convex in μ , as it is for linear forward models and quantities of interest [[1](#), Proposition 2.5], the condition of the existence of $\bar{\mu}$ satisfying $\lambda(\bar{\mu}, \mathbf{y}; \mathcal{X}) < \bar{q}_{\gamma, \eta}$ is equivalent to the interval not being empty. For nonlinear forward models, it is a slightly stronger condition. For the convergence of the sliced optimized version, one needs to show that $\inf_{\mu: \lambda(\mu, \mathbf{y}) \leq \hat{m}_\gamma(\mu)} \mu$ converges to $\inf_{\mu: \lambda(\mu, \mathbf{y}) \leq m_\gamma(\mu)} \mu$ as $\hat{m}_\gamma(\mu)$ converges to $m_\gamma(\mu)$. In order to do so, we study the convergence of optimization problems of the form $\inf_{\mu: \hat{f}(\mu) \geq 0} \mu$ to $\inf_{\mu: f(\mu) \geq 0} \mu$ as $\hat{f}$ converges to f . Although the result is not true in general, we provide sufficient technical conditions on f and the uniformity of the convergence of $\hat{f}$ to f for the result to hold. The technical details and proofs of these results are provided in [Section SM2](#) (in the Supplement).

5. Implementation methodology.

5.1. Sampling the pre-image Berger–Boos set. The viability of this method directly relies upon our ability to sample from the Berger–Boos set, $\mathcal{B}_\eta = f^{-1}(\Gamma_\eta(\mathbf{y})) \cap \mathcal{X}$. Under the assumed data-generating process in Equation (3.1), the Berger–Boos set is defined as follows:

$$(5.1) \quad \mathcal{B}_\eta = f^{-1}(\Gamma_\eta(\mathbf{y})) \cap \mathcal{X} = \{\mathbf{x} \in \mathcal{X} : (\mathbf{y} - f(\mathbf{x}))^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - f(\mathbf{x})) \leq \chi_{n, \eta}^2\},$$

where we have generalized to a non-identity covariance matrix, $\boldsymbol{\Sigma}$, to make the following exposition more general. This set is equivalent to the set over which the strict bounds intervals are optimized in [[33](#)], also called “SSB” intervals in [[31](#), [1](#)]. Note, however, that one would use $\chi_{n, \alpha}^2$ instead of $\chi_{n, \eta}^2$ for $1 - \alpha$ interval computation in that scenario.

We focus on sampling strategies under the linear-Gaussian setting ($f(\mathbf{x}) = \mathbf{K}\mathbf{x}$) corresponding to the numerical experiments described in [Section 6](#). Sampling from constrained ellipsoids, polytopes, and convex bodies is a well-developed topic; the purpose of this subsection is therefore to explain which existing sampling ideas are useful for the particular Berger–Boos sets arising in our calibrated interval construction, and what trade-offs they introduce. In low-dimensional cases with full column rank, $\mathcal{B}_\eta$ becomes an ellipsoid intersected with the parameter constraints $\mathcal{X}$. In such settings, we apply the efficient Voelker–Gosmann–Stewart (VGS) algorithm from [[35](#)], which samples uniformly from the ellipsoid by using a suitable linear transformation of uniform samples from a ball, followed by a constraint-based accept-reject step.

However, the VGS algorithm becomes inefficient or infeasible as dimensionality and ill-conditioning of the forward model increase, due to rapidly declining accept-reject probabilities. To address these challenging scenarios, we introduce an MCMC-based sampling approach called the Polytope sampler, inspired by the Vaidya walk algorithm from [[5](#)]. This method

constructs a bounding polytope around $\mathcal{B}_\eta$ defined by linear constraints (e.g., derived from the parameter constraints and eigenvectors of the forward model) and performs parallel random walks starting from strategically chosen points. These starting points span the space between the extremes of the functional of interest and the polytope’s Chebyshev center. Such parallelization helps to ensure exploration of the Berger–Boos set without relying solely on a single chain reaching stationarity, which may be challenging due to the geometry induced by parameter constraints. Additional technical details, including exact algorithms and further illustrations of both samplers, are provided in [Section SM3](#) (in the Supplementary Material).

5.2. Quantile regression. [Algorithm 3.2](#) explained in [Subsection 3.2](#) uses quantile regression to learn a quantile surface from a collection of pairs of design points and LLR test statistic samples. As mentioned previously, similar approaches have been taken in [\[8, 9, 18, 19\]](#), and since quantile regression is a technique that facilitates our interval constructions, we give a brief overview of quantile regression and different ways to implement it in [Section SM4](#).

Using the design points in the Berger–Boos set as sampled via the VGS or Polytope samplers, [Algorithm 3.2](#) shows how we sample from the test statistic distribution defined at each design point to define a data set to fit a quantile regressor. We use the Gradient Boosting Regressor implemented in scikit-learn [\[23\]](#) with the “quantile” loss function (i.e., pinball loss defined in [\(SM4.2\)](#)) for our numerical examples in [Section 6](#). This algorithm involves a collection of hyperparameters (i.e., the minimum number of samples required to split an internal node, the minimum number of samples needed to be a leaf node, the maximum depth of any individual estimator, the learning rate, and the number of estimators) which we determine using 10-fold cross validation in a pilot study ahead of our simulation experiments in [Section 6](#). Although one may use any quantile regressor to estimate the quantile surface, we emphasize the importance of choosing a strategy that can accommodate a nonlinear surface in the parameters (such as the Gradient Boosting Regressor) as the parameter constraints are known to produce nonlinear quantile surfaces in even simple examples as seen in [\[1\]](#). Practically, the quantile-regression step trades repeated LLR simulation at every design point for a modeling assumption about the smoothness of the conditional quantile surface. This trade-off is beneficial when LLR evaluations require expensive constrained optimizations, but it also makes diagnostic checks important: in our experiments we use a pilot cross-validation study to choose hyperparameters and reserve out-of-sample design points for interval construction. For substantially higher-dimensional problems, the same pipeline can be used in principle, but its performance will depend on obtaining enough design points to cover the relevant functional range and on using a regressor with sufficient inductive bias for the quantile surface.

6. Numerical experiments. For scenarios within the case of the linear-Gaussian data-generating process of [\(3.1\)](#), we regard the OSB interval as the previous state-of-the-art option for computing constraint-aware confidence intervals. Although these intervals have empirically achieved nominal coverage in applications [\[22, 31\]](#), they generally do not guarantee coverage [\[1\]](#). As such, we use the OSB interval in the following numerical experiments as a main comparison point to the intervals defined in this paper. In scenarios where the OSB interval achieves at least nominal coverage, we show that our intervals are either competitive or better in terms of expected interval length. In scenarios where the OSB interval does not achieve nominal coverage, our intervals do achieve nominal coverage and can have shorter expected length.

The first two experiments use a constrained Gaussian noise model in two or three dimensions to make the above points while demonstrating parts of the interval construction. Given the straightforward mathematics of these problems, both the mechanics of our intervals and their performance improvements over the OSB interval can be clearly seen. The second two experiments (inspired by particle unfolding in high-energy physics [31]) consider a wide-bin deconvolution setup featuring a substantially more challenging 80-dimensional parameter space with a rank-deficient forward model. These examples demonstrate the superior performance of our intervals over OSB in terms of both coverage and expected length. The experiments are intentionally ordered from pedagogical to application-oriented. The two-dimensional example is a visualization and sanity-check case in which the geometry of the Berger–Boos set, sampled design points, and LLR cutoffs can be inspected directly. The three-dimensional example is a calibration stress test in which the OSB interval fails to achieve nominal coverage, while the proposed intervals retain the desired coverage. The wide-bin deconvolution experiments are the application-oriented inverse-problem UQ examples: they use a rank-deficient, high-dimensional forward model motivated by particle unfolding and test whether the full sampling and quantile-regression pipeline remains effective in the setting for which the method is designed.

In the following experiments, we form 68% confidence intervals, set $\eta = 0.01$, and compute γ according to Lemma 3.1. We draw 10^3 observations from each data-generating process to estimate both coverage and expected interval length for our four interval constructions and the OSB interval. We also provide 95% confidence intervals in the form of orange line segments to characterize the statistical error for both coverage and expected length estimates. The coverage confidence intervals are Clopper–Pearson intervals for the success probability parameter of a binomial distribution, while the expected length confidence intervals are the average length plus/minus the appropriately scaled standard error of the mean. The source code to reproduce our experiments is available at <https://github.com/mcstanle/adaptive-functional-ci>.

6.1. Constrained Gaussian in two dimensions. The two-dimensional Gaussian noise model is defined as follows:

$$(6.1) \quad \mathbf{y} = \mathbf{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2), \quad \mathbf{x}^* \in \mathbb{R}_+^2,$$

where $\varphi(\mathbf{x}) = x_1 - x_2$ and $\mathbf{x}^* = (0.5 \ 0.5)^\top$. The LLR is then given as follows:

$$(6.2) \quad \lambda(\mu, \mathbf{y}) = \min_{\substack{x_1 - x_2 = \mu \\ \mathbf{x} \in \mathbb{R}_+^2}} \|\mathbf{y} - \mathbf{x}\|_2^2 - \min_{\mathbf{x} \in \mathbb{R}_+^2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

This example first appeared in [34] as a case where the OSB interval allegedly fails to achieve nominal coverage when the true parameter $\mathbf{x}^*$ is such that $\varphi(\mathbf{x}^*) = 0$. Since [1] overturned this result by proving OSB validity in this case, this example is important to include for its historical context and OSB validity. The proof of OSB interval coverage in this case relies upon showing that $Q_{1-\alpha}^{\max} = \chi_{1,\alpha}^2$ for all $\alpha \in (0, 1)$, where $\chi_{1,\alpha}^2$ is the upper α -quantile of a chi-squared distribution with one degree of freedom. This result is shown in Lemma 4.4 of [1].

Estimated coverage and length results are shown in Figure 2. We note that all four of our interval constructions are competitive with OSB in terms of coverage, while all of our interval

constructions have higher estimated expected length, apart from the Sliced constructions, which are within statistical error of OSB. Since the OSB interval is defined using $Q_{1-\alpha}^{\max} = \chi_{1,\alpha}^2$ and the α -quantile surface of the LLR rapidly approaches this global max-quantile as one moves away from the origin (see Figure 5.3 in [1]), the OSB interval lengths are difficult to beat in practice with intervals based on the Berger–Boos sets since these sets likely contain parameter settings with quantiles near $\chi_{1,\alpha}^2$. The left panel of Figure 3 shows four realizations of the data-generating process with the observations shown as red points. For each observation, the blue points show uniformly distributed draws within its Berger–Boos set, sampled using the VGS sampler. Cross-referencing the spread of the Berger–Boos set samples in Figure 3 with Figure 5.3 in [1], it is clear that there are always samples in the parameter space where the quantile surface is nearly the same as the $\chi_{1,\alpha}^2$ quantile. Further, when including the Berger–Boos set in the interval construction, we instead construct our intervals using the γ -quantile, where $\gamma < \alpha$, as the LLR cutoff, resulting in a more relaxed constraint. This fact can be clearly observed in the central panel of Figure 3, showing the sampled γ -quantiles within the Berger–Boos set of one observation from the data-generating process. Since a non-trivial portion of this distribution is above $\chi_{1,\alpha}^2$, the longer average length of the Global intervals is explained. In the right panel of Figure 3, for the same observation, we show the estimated sliced max-quantile function, $\hat{m}_\gamma(\mu)$, in orange alongside $\chi_{1,\alpha}^2$. Since this estimated function is above $\chi_{1,\alpha}^2$ at their intersection points with the underlying LLR function shown in the solid blue line, it further makes sense that the Sliced interval constructions provide no additional length improvement compared to the OSB interval in this particular setting.

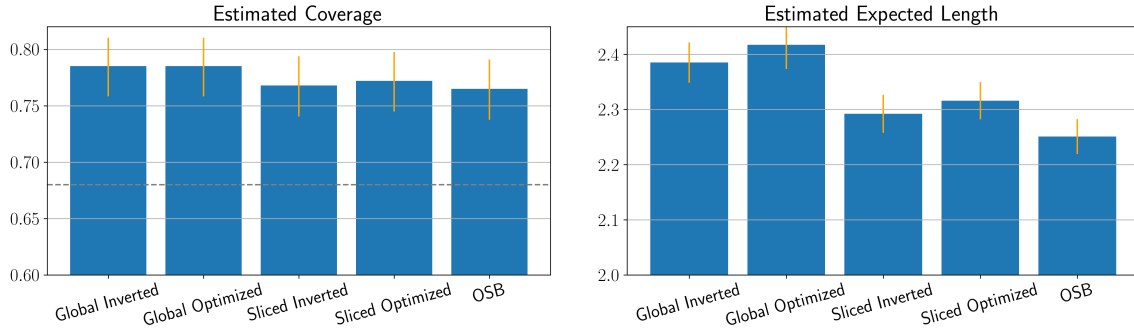


Figure 2: Estimated coverages and expected lengths across all four interval constructions and OSB for comparison at the 68% level for the two-dimensional constrained Gaussian setting. All four of our interval constructions are comparable to OSB with respect to coverage, but OSB shows better expected length performance, aside from our two sliced interval constructions. Although the OSB intervals are defined using the global max-quantile ($Q_{1-\alpha}^{\max}$) and therefore can potentially be improved upon by limiting the considered parameter space via the Berger–Boos set, due to the rapidity with which the α -quantile surface meets the $\chi_{1,\alpha}^2$ quantile (see Figure 5.3 in [1]), the OSB interval lengths are difficult to beat in practice.

6.2. Constrained Gaussian in three dimensions. In situations where the OSB coverage guarantee is difficult to prove, our intervals provide a clear theoretical advantage. We apply

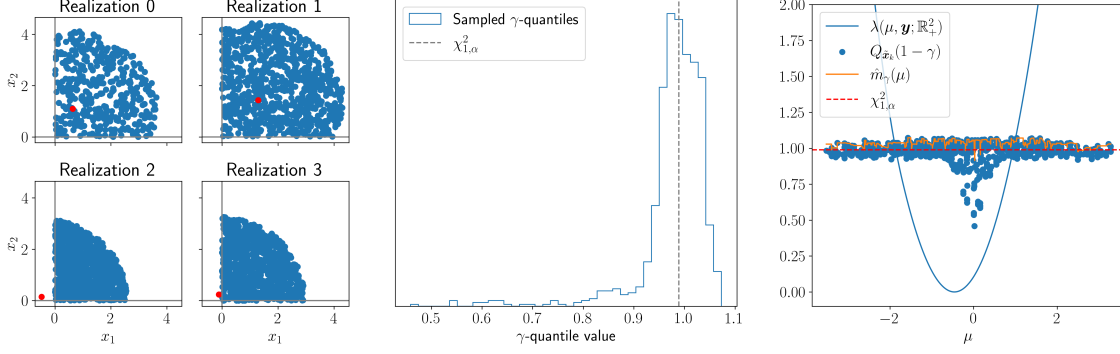


Figure 3: **(Left)** Four realizations of the data-generating process where the observations are shown in red. For each realization, the blue points are uniformly distributed samples from its Berger–Boos set, sampled using the VGS sampler. **(Center)** For a realization of the data-generating process, we plot the distribution of γ -quantiles for the points sampled by the VGS sampler. Notably, a non-trivial percent of these are above $\chi^2_{1,\alpha}$ defining the OSB interval. **(Right)** For the same realization, we plot the estimated sliced max-quantile function, $\hat{m}_\gamma(\mu)$ in orange alongside $\chi^2_{1,\alpha}$ in red. The blue points correspond to sampled parameter values, each of which has a functional and quantile value, while the solid blue line shows the LLR over the functional varies. All intervals can be read immediately from this image by inspecting where the blue LLR curve intersects the sampled points.

our four intervals to one such case involving a three-dimensional constrained Gaussian case as explored in [1]. The three-dimensional Gaussian noise model is defined as follows:

$$(6.3) \quad \mathbf{y} = \mathbf{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_3), \quad \mathbf{x}^* \in \mathbb{R}_+^3,$$

where $\varphi(\mathbf{x}) = x_1 + x_2 - x_3$ and $\mathbf{x}^* = (0.03 \ 0.03 \ 1)^\top$. The LLR is then defined as follows:

$$(6.4) \quad \lambda(\mu, \mathbf{y}) = \min_{\substack{x_1+x_2-x_3=\mu \\ \mathbf{x} \in \mathbb{R}_+^3}} \|\mathbf{y} - \mathbf{x}\|_2^2 - \min_{\mathbf{x} \in \mathbb{R}_+^3} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Figure 4 shows estimated coverage and expected length across all four interval constructions and the OSB interval. While the OSB interval fails to attain nominal coverage, all four of our interval constructions do, with the Sliced constructions providing the best calibration. The Sliced constructions best navigate the trade-off between coverage and expected length, paying for coverage with only slightly longer intervals compared to the OSB interval.

The setting of $\mathbf{x}^*$ used here is slightly different than that of [1], where $\mathbf{x}^* = (0 \ 0 \ 1)^\top$ was used. In [1], this setting was used as a counter-example for OSB coverage, since $Q_{\mathbf{x}^*}(1 - \alpha) > \chi^2_{1,\alpha}$ for at least some $\alpha \in (0, 1)$. However, as shown in Figure 5.5 of [1], when $\alpha = 0.05$, $Q_{\mathbf{x}^*}(1 - \alpha) \leq \chi^2_{1,\alpha}$ for $\mathbf{x}^* = (t \ t \ 1)^\top$ when t is approximately greater than $e^{-2} \approx 0.135$, which indicates that parameter settings violating stochastic dominance by χ^2_1 exist close to the parameter constraint boundary. The location of these key parameter settings presents a challenge for the Polytope sampler (Algorithm SM3.2). In Subsection SM5.1, Algorithm SM5.1 presents a modified version of Algorithm SM3.2 that better handles this situation.

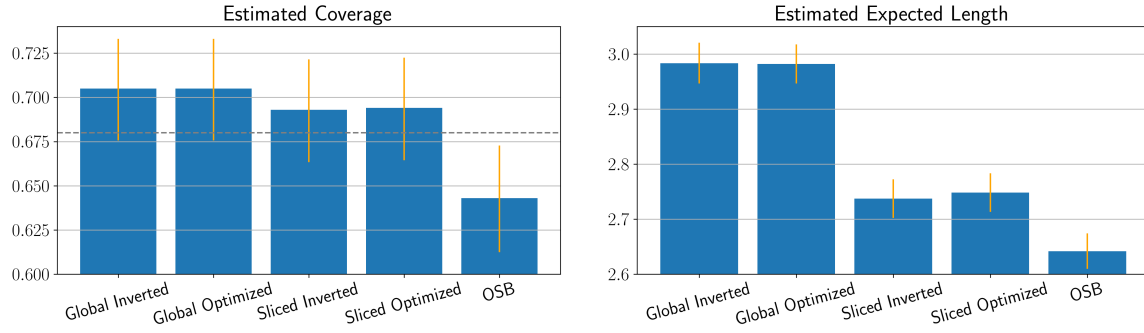


Figure 4: Estimated coverage and expected length across all four interval constructions and OSB for comparison at the 68% level for the three-dimensional constrained Gaussian example. All four of our interval constructions achieve nominal coverage, while the OSB interval does not. While the Global interval constructions pay a steep price in expected length compared to OSB, the Sliced constructions are only slightly longer than OSB.

Berger–Boos set experiment. For this model, we investigate the effect of changing the parameter η that controls the Berger–Boos construction in the global interval. We compute intervals for different $\mathbf{y}$ and fixed $\mathbf{x}^* = (2, 2, 0)$, $\mathbf{x}^* = (3, 3, 0)$ and $\mathbf{x}^* = (5, 5, 0)$ as η ranges between 0 and $\alpha = 0.32$. The lengths of such intervals, averaged over the data $\mathbf{y}$, are shown in Figure 5. In this example the maximum quantile is achieved at $\mathbf{x} = (0, 0, t)$ for large t , and, as expected, the benefit of using a small $\eta > 0$ becomes more pronounced as the true $\mathbf{x}^*$ becomes farther from the point that achieves the maximum quantile. However, as η grows too close to α the downside of optimizing the $1 - \alpha + \eta$ quantile instead of $1 - \alpha$ outweighs the benefit of optimizing it over a smaller set, resulting in larger intervals. This suggests that a small $\eta > 0$ is a reasonable default, as proposed originally by [2].

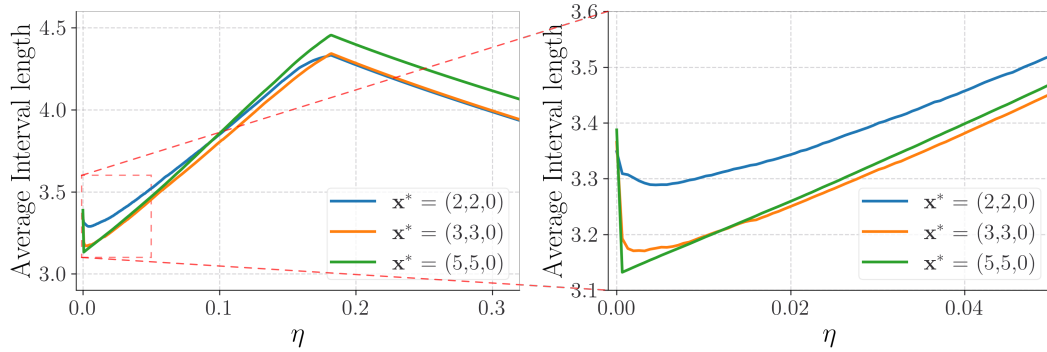


Figure 5: Confidence interval lengths in the Berger–Boos setting, averaged over values of $\mathbf{y}$, for varying η and $\mathbf{x}^*$. The minimum average length occurs at a small $\eta > 0$, showing that the construction is beneficial if η is tuned correctly. This occurs because even for moderately small η , the Berger–Boos set, which is a three-dimensional sphere intersected with the non-negative orthant, avoids the point with the highest $1 - \alpha$ quantile.

6.3. Wide-bin deconvolution. This section shows the superior performance of our interval constructions relative to OSB in a more complex high-dimensional setting by considering the problem of computing a confidence interval for the sum of adjacent bins of a deconvolved histogram as described in [31]. We deliberately use this previously studied model as a benchmark: the forward model is not the novelty of the present paper, but keeping the application model fixed allows the effect of the new data-adaptive calibration and sampling-based interval construction to be isolated and compared directly with OSB and related optimization-based intervals. This problem is a core statistical problem of particle unfolding in high-energy physics. We refer the readers to the following references for more detailed information [15, 14, 16, 6, 7].

The data-generating process is linear with Gaussian noise,

$$(6.5) \quad \mathbf{y} = \mathbf{K}\mathbf{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{x}^* \geq \mathbf{0},$$

where $\mathbf{K} \in \mathbb{R}^{40 \times 80}$ and $\varphi(\mathbf{x}) = \mathbf{h}^\top \mathbf{x}$. The vector $\mathbf{h}$ defines the bin-adjacent aggregation. In particle unfolding, the vectors $\mathbf{x}^*$ and $\mathbf{y}$ represent particle counts within discretized bins. We note that the dimension of these discretized bins is directly related to the resolution of the scientific inferences that can be made from the system, and it is advantageous to set the discretization so that the dimension of $\mathbf{x}^*$ is significantly larger than the dimension of $\mathbf{y}$. Other approaches typically cannot use this high-dimensional discretization and still achieve calibrated UQ on $\varphi(\mathbf{x}^*)$, since they require regularization and thus incur estimator bias. The LLR is then defined as follows:

$$(6.6) \quad \lambda(\mu, \mathbf{y}) = \min_{\substack{\mathbf{h}^\top \mathbf{x} = \mu \\ \mathbf{x} \in \mathbb{R}_+^{80}}} \|\mathbf{y} - \mathbf{K}\mathbf{x}\|_2^2 - \min_{\mathbf{x} \in \mathbb{R}_+^{80}} \|\mathbf{y} - \mathbf{K}\mathbf{x}\|_2^2.$$

The forward model, $\mathbf{K}$, has a non-trivial null space and a large condition number, making inverse problem point estimation and inference challenging (see [14] for more details). Although [1] and this paper propose a theoretical framework to perform inference with constraints, implementation is challenging in practice due to the high-dimensional parameter space of this scenario. As we show in the following sections, the Polytope sampler (Algorithm SM3.2) and quantile regression do an adequate job producing samples and fitting quantile surfaces in this high-dimensional space to ensure the desired coverage of the interval constructions.

As extensively discussed in [31], the OSB interval (i.e., using $\chi_{1,\alpha}^2$ in the optimization-based interval construction) produces empirically valid (though typically conservative) confidence intervals in all tested scenarios. Since the over-coverage in [31] is likely the result of relatively smooth mean vector ($\mathbf{x}^* \in \mathbb{R}_+^{80}$), we present two true parameter settings for $\mathbf{x}^*$ in (6.5) to highlight two advantages of our interval constructions over the OSB interval. First, we use the original smooth parameter setting from [31] to show improved over-coverage relative to the OSB interval via a reduction in the expected interval length. Second, we present an “adversarial” setting where our interval constructions achieve nominal coverage while the OSB interval does not. We constructed the adversarial setting (see Figure 6) by first computing our interval constructions on the smooth setting and then looking at the maximum out-of-sample predicted quantile for a generated observation with a large predicted quantile. For each observation drawn within both settings, we draw 2.1×10^4 samples using the Polytope sampler as described by Algorithm SM3.2.

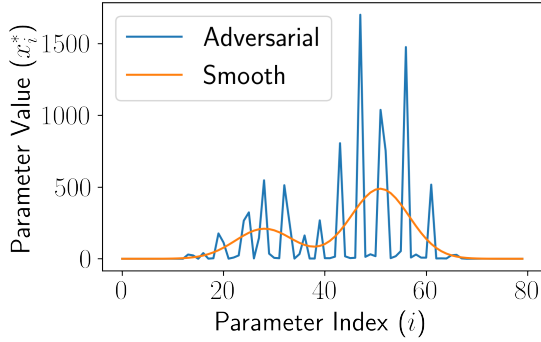


Figure 6: Parameter values for the smooth and adversarial settings for $\mathbf{x}^*$ used to illustrate our interval construction versus the OSB interval. The adversarial setting is made more difficult by the sharp jumps in parameter values.

6.3.1. Smooth setting. We show that our interval constructions achieve nominal coverage and the sliced constructions reduce over-coverage compared to OSB by producing shorter intervals on average. Estimated coverage and expected interval lengths are shown in Figure 7.

Both Global interval constructions and OSB dramatically over-cover, which highlights the conservatism of the Global constructions. These estimated coverage values indicate that within each observation’s Berger–Boos set, there is a parameter setting against which the method has to protect that is substantially more difficult to cover than the true realistic parameter setting. The Sliced intervals show lower over-coverage and significantly smaller average lengths, and both Sliced constructions are shorter on average compared to the OSB interval, with the Sliced Inverted interval showing an 18.7% reduction in average length and the Sliced Optimized showing a 11.1% reduction in average length.

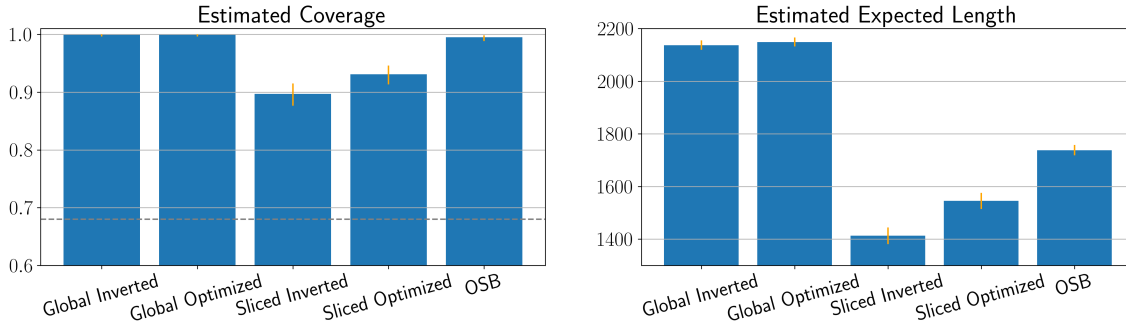


Figure 7: Estimated coverage and expected length across all four interval constructions and OSB at the 68% level for the **smooth** wide-bin deconvolution experiment. While the Global interval constructions over-cover like the OSB interval, the Sliced interval constructions reduce both over-coverage and expected interval length.

6.3.2. Adversarial setting. We show that for the adversarial parameter setting, the OSB interval does not achieve nominal coverage, whereas all four of our interval constructions do while still reducing the expected interval length in the case of the Sliced interval constructions compared to the OSB interval (see Figure 8 for primary results).

Both Global interval constructions and the Sliced Optimized interval over-cover, while the Sliced Inverted intervals achieve nominal coverage within statistical uncertainty. The estimated expected lengths are similar to those of the smooth example, with the Global intervals showing the longest average interval lengths, the Sliced intervals showing the shortest, and the OSB interval being between the two. Importantly, the Sliced intervals are again significantly shorter than the OSB interval, even though the OSB interval does not achieve nominal coverage. The Sliced Inverted interval shows a 18.9% average interval length reduction over OSB while the Sliced Optimized interval shows an 11.4% average interval length reduction.

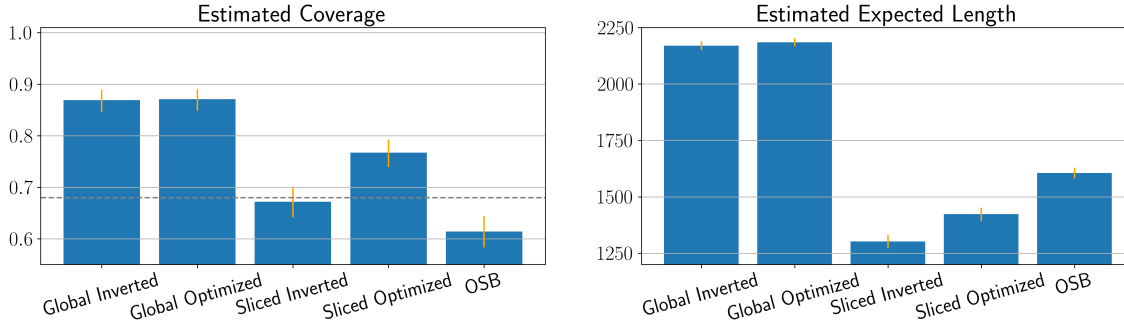


Figure 8: Estimated coverage and expected length across all four interval constructions and OSB at the 68% level for the **adversarial** wide-bin deconvolution experiment. While the OSB interval fails to achieve nominal coverage, all four of our interval constructions do. Interestingly, the Sliced interval constructions are meaningfully shorter than the OSB interval while also providing coverage.

For both experiments, we note the differences in coverage and expected length between the two Sliced intervals. Although both approaches theoretically compute the same interval, significant differences can arise from their construction. The Sliced Inverted intervals are constructed by accepting individual functional values, which makes each point’s acceptance dependent upon the quality of the quantile regressor at that point. By contrast, since the Sliced Optimized intervals essentially smooth over the max quantiles as a function of the functional space, the intervals are less sensitive to quantile regressor performance at any individual point. In the above experiments, there were realizations of the data for which the Sliced Inverted construction only accepted a single functional value sample, thereby making the confidence interval a single point. We found the Sliced Optimized construction to be more robust in these settings as the max quantile values were shared in a sliding window over the functional space.

7. Conclusion and future work. This paper proposes several confidence interval constructions for functionals in constrained ill-posed inverse problems. Our approach is based on two key ideas: data-adaptive constraints using a Berger–Boos construction and sampling-based inversion. Two independent decisions when constructing intervals provide four different valid intervals: Global versus Sliced, using the quantile function either over the entire or along level-set slices of the Berger–Boos set and Inverted versus Optimized, constructing the interval either by individually accepted functional values via the estimated quantile function or using the estimated quantile function in endpoint optimizations. All of the constructions are

built upon the preliminary constraint by the data-informed Berger–Boos set, followed by a sampling procedure to estimate a quantile function that can be used to invert or optimize the interval endpoints. We have validated the method (including all four aforementioned interval constructions) through several numerical examples, demonstrating its ability to provide correct coverage, better calibration, and comparable or shorter interval length compared to the OSB interval baseline. Overall, our approach offers a flexible framework that can incorporate constraints directly and can be tailored to various types of inverse problems. The main take-away is that data-adaptive constraining helps improve the length of the resulting confidence intervals and enables sampling, which makes it feasible to carry out the test inversion needed to construct confidence intervals with a desired nominal coverage.

There are several promising directions for future work. One immediate direction is principled automated guidance for choosing the Berger–Boos parameter η . The finite-sample guarantee holds for any pre-specified $\eta \in (0, \alpha)$ with $\gamma \leq \alpha - \eta$, but the expected-length-optimal value is problem-dependent because it depends on both the geometry of $\mathcal{B}_\eta$ and the shape of the LLR quantile surface. Our experiments support using a small positive value and, when feasible, checking sensitivity over a small grid of pre-specified candidate values. In this paper, for example, the 68% experiments use $\alpha = 0.32$ and $\eta = 0.01$, so η uses only a small fraction of the total miscoverage budget while still making the data-adaptive set useful computationally. We do not claim that this value is universally optimal. Developing a fully automated, data-adaptive choice of η that optimizes interval length while preserving the coverage guarantee is left for future work. Another direction is to find ways to extend our method to even higher-dimensional problems, which are more challenging. This would involve developing improved techniques to handle the curse of dimensionality and exploring the trade-off between accuracy and computational complexity. For the approaches in this paper in particular, this extension would require a more tailored sampling approach. Another direction is to leverage more sophisticated machine learning algorithms (deep learning models or ensemble methods) to improve the estimate of the quantile function and, thus, improve the accuracy and efficiency of our confidence intervals. Additionally, applying our approach to other applications involving ill-posed inverse problems, such as medical imaging or geophysics, would provide further validation of the effectiveness of our strategy. Finally, it is of interest to conduct a further theoretical analysis of our approach under different constraints and noise conditions to better understand its limitations and strengths. This would involve studying the statistical properties of our confidence intervals and investigating the impact of various assumptions on their performance. Of particular interest are relaxations of the linear forward model and Gaussian noise assumptions to extend our method’s application domain. Although the original theoretical foundation developed in [1] does not make these assumptions, our implementation relies upon them for the tractability of the sampling algorithms. Overall, these future research directions have the potential to demonstrate the applicability and robustness of our approach in a wide range of domains.

Acknowledgment. We thank members of the STAMPS (Statistical Methods for the Physical Sciences) Research Center at Carnegie Mellon University for fruitful discussions on this work. MS and MK were partially supported by NSF grants DMS-2053804 and PHY-2020295, JPL RSAs No. 1670375, 1689177 & 1704914 and a grant from the C3.AI Digital Transforma-

tion Institute. PB and HO acknowledge support from the Air Force Office of Scientific Research under MURI awards number FA9550-20-1-0358, FOA-AFRL-AFOSR-2023-0004 and by the Department of Energy under award number DE-SC0023163. Additionally HO acknowledge support from the DoD Vannevar Bush Faculty Fellowship Program.

Supplementary material. To keep the main article focused while retaining reproducibility, the Supplementary Material collects the extended proofs, sampler algorithms, quantile-regression details, and numerical diagnostics supporting the main construction. The Supplementary Material includes: proofs from Sections 3 and 4 in Sections SM1 and SM2, additional algorithmic details and illustrations for sampling the Berger–Boos set (VGS and Polytope samplers) in Section SM3, quantile regression overview in Section SM4, and extended numerical details and illustrations, including an importance-like sampler used in numerical experiments in Section SM5.

REFERENCES

- [1] P. BATLLE, M. STANLEY, P. PATIL, M. KUUSELA, AND H. OWHADI, *Optimization-based frequentist confidence intervals for functionals in constrained inverse problems: Resolving the burrus conjecture*, arXiv preprint arXiv:2310.02461, (2023).
- [2] R. L. BERGER AND D. D. BOOS, *P values maximized over a confidence set for the nuisance parameter*, Journal of the American Statistical Association, 89 (1994), pp. 1012–1016.
- [3] W. BURRUS, U. A. E. COMMISSION, AND O. R. N. LABORATORY, *Utilization of a Priori Information by Means of Mathematical Programming in the Statistical Interpretation of Measured Distributions*, Oak Ridge National Laboratory, 1965, <https://books.google.com/books?id=ubUPYiYzRcC>.
- [4] G. CASELLA AND R. L. BERGER, *Statistical Inference*, Duxbery, 2 ed., 2002.
- [5] Y. CHEN, R. DWIVEDI, M. J. WAINWRIGHT, AND B. YU, *Fast mcmc sampling algorithms on polytopes*, Journal of Machine Learning Research, 19 (2018), pp. 1–86.
- [6] CMS COLLABORATION, *Measurement of differential cross sections for Higgs boson production in the diphoton decay channel in pp collisions at $\sqrt{s} = 8$ TeV*, The European Physical Journal C, 76 (2016), <https://doi.org/https://doi.org/10.1140/epjc/s10052-015-3853-3>.
- [7] CMS COLLABORATION, *Measurement of inclusive and differential Higgs boson production cross sections in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV*, Journal of High Energy Physics, 2019 (2019), [https://doi.org/https://doi.org/10.1007/JHEP01\(2019\)183](https://doi.org/https://doi.org/10.1007/JHEP01(2019)183).
- [8] N. DALMASSO, R. IZBICKI, AND A. LEE, *Confidence sets and hypothesis testing in a likelihood-free inference setting*, in International Conference on Machine Learning, PMLR, 2020, pp. 2323–2334.
- [9] N. DALMASSO, L. MASSERANO, D. ZHAO, R. IZBICKI, AND A. B. LEE, *Likelihood-free frequentist inference: Bridging classical statistics and machine learning for reliable simulator-based inference*, Electronic Journal of Statistics, (2024), pp. 5045–5090.
- [10] X. GENG AND L. XIE, *Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization*, Annual reviews in control, 47 (2019), pp. 341–363.
- [11] C. GOURIÉROUX, A. HOLLY, AND A. MONFORT, *Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters*, Econometrica, 50 (1982), pp. 63–80.
- [12] A. HÖCKER AND V. KARTVELISHVILI, *Svd approach to data unfolding*, Nuclear Instruments and Methods in Physics Research A, 372 (1996), pp. 469–481.
- [13] A. I. KIBZUN AND Y. S. KAN, *Stochastic programming problems with probability and quantile functions*, Journal of the Operational Research Society, 48 (1997), p. 849–849, <https://doi.org/10.1057/palgrave.jors.2600833>, <http://dx.doi.org/10.1057/palgrave.jors.2600833>.
- [14] M. KUUSELA, *Uncertainty quantification in unfolding elementary particle spectra at the Large Hadron Collider*, PhD thesis, École Polytechnique Fédérale de Lausanne, 7 2016.

- [15] M. KUUSELA AND V. M. PANARETOS, *Statistical unfolding of elementary particle spectra: Empirical bayes estimation and bias-corrected uncertainty quantification*, The Annals of Applied Statistics, 9 (2015), pp. 1671–1705.
- [16] M. KUUSELA AND P. B. STARK, *Shape-constrained uncertainty quantification in unfolding steeply falling elementary particle spectra*, The Annals of Applied Statistics, 11 (2017), pp. 1671–1710.
- [17] J. LIU, K. BOWMAN, L. MEEMONG, ET AL., *Carbon monitoring system flux estimation and attribution: impact of ACOS-GOSAT X_{CO_2} sampling on the inference of terrestrial biospheric sources and sinks*, Tellus B: Chemical and Physical Meteorology, 66 (2016), <https://doi.org/https://doi.org/10.3402/tellusb.v66.22486>.
- [18] L. MASSERANO, T. DORIGO, R. IZBICKI, M. KUUSELA, AND A. LEE, *Simulation-based inference with waldo: Confidence regions by leveraging prediction algorithms or posterior estimators for inverse problems*, AISTATS, (2023).
- [19] L. MASSERANO, A. SHEN, M. DORO, T. DORIGO, R. IZBIKI, AND A. LEE, *Classification under nuisance parameters and generalized label shift in likelihood-free inference*, arXiv preprint arXiv:2402.05330, (2024).
- [20] G. MOLENBERGHS AND G. VERBEKE, *Likelihood ratio, score, and wald tests in a constrained parameter space*, The American Statistician, 61 (2007), pp. 22–27.
- [21] V. M. PANARETOS, *Statistics for Mathematicians*, Springer International Publishing, 2016, <https://doi.org/10.1007/978-3-319-28341-8>, <https://doi.org/10.1007/978-3-319-28341-8>.
- [22] P. PATIL, M. KUUSELA, AND J. HOBBS, *Objective frequentist uncertainty quantification for atmospheric CO_2 retrievals*, SIAM/ASA Journal on Uncertainty Quantification, 10 (2022), <https://doi.org/https://doi.org/10.1137/20M1356403>.
- [23] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT, AND E. DUCHESNAY, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.
- [24] A. PENA-ORDIERES, J. R. LUEDTKE, AND A. WACHTER, *Solving chance-constrained problems via a smooth sample-based nonlinear approximation*, SIAM Journal on Optimization, 30 (2020), pp. 2221–2250.
- [25] T. ROBERTSON, F. T. WRIGHT, AND R. DYKSTRA, *Order Restricted Statistical Inference*, Wiley, 1988.
- [26] C. D. RODGERS, *Inverse Methods for Atmospheric Sounding*, World Scientific Publishing, 2000.
- [27] B. W. RUST AND W. R. BURRUS, *Mathematical Programming and the Numerical Solution of Linear Equations*, Elsevier, 1972.
- [28] B. W. RUST AND D. P. O’LEARY, *Confidence intervals for discrete approximations to ill-posed problems*, Journal of Computational and Graphical Statistics, 3 (1994), pp. 67–96.
- [29] S. SCHMITT, *Tunfold, an algorithm for correcting migration effects in high energy physics*, Journal of Instrumentation, 7 (2012).
- [30] A. SHAPIRO, *Towards a unified theory of inequality constrained testing in multivariate analysis*, International Statistical Review/Revue Internationale de Statistique, (1988), pp. 49–62.
- [31] M. STANLEY, P. PATIL, AND M. KUUSELA, *Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals*, Journal of Instrumentation, 17 (2022), <https://doi.org/https://doi.org/10.1088/1748-0221/17/10/P10013>.
- [32] P. B. STARK, *Inference in infinite-dimensional inverse problems: discretization and duality*, Journal of Geophysical Research: Solid Earth, 97 (1992), pp. 14055–14082.
- [33] P. B. STARK, *Inference in infinite-dimensional inverse problems: Discretization and duality*, Journal of Geophysical Research, 97 (1992), pp. 14055–14082.
- [34] L. TENORIO, A. FLECK, AND K. MOSES, *Confidence intervals for linear discrete inverse problems with a non-negativity constraint*, Inverse problems, 23 (2007), p. 669.
- [35] A. R. VOELKER, J. GOSMANN, AND T. C. STEWART, *Efficiently sampling vectors and coordinates from the n -sphere and n -ball*, tech. report, Centre for Theoretical Neuroscience, 2017.
- [36] F. A. WOLAK, *An exact test for multiple inequality and equality constraints in the linear regression model*, Journal of the American Statistical Association, 82 (1987), pp. 782–793.
- [37] F. A. WOLAK, *Testing inequality constraints in linear econometric models*, Journal of Econometrics, 41 (1989), pp. 205–235.