Non-Asymptotic Uncertainty Quantification in High-Dimensional Learning

Frederik Hoppe[*]
RWTH Aachen University
hoppe@mathc.rwth-aachen.de

Claudio Mayrink Verdun[*]
Harvard University
claudioverdun@seas.harvard.edu

Hannah Laus[*]
TU Munich & MCML
hannah.laus@tum.de

Felix Krahmer
TU Munich & MCML
felix.krahmer@tum.de

Holger Rauhut
LMU Munich & MCML
rauhut@math.lmu.de

Abstract

Uncertainty quantification (UQ) is a crucial but challenging task in many high-dimensional learning problems to increase the confidence of a given predictor. We develop a new data-driven approach for UQ in regression that applies both to classical optimization approaches such as the LASSO as well as to neural networks. One of the most notable UQ techniques is the debiased LASSO, which modifies the LASSO to allow for the construction of asymptotic confidence intervals by decomposing the estimation error into a Gaussian and an asymptotically vanishing bias component. However, in real-world problems with finite-dimensional data, the bias term is often too significant to disregard, resulting in overly narrow confidence intervals. Our work rigorously addresses this issue and derives a data-driven adjustment that corrects the confidence intervals for a large class of predictors by estimating the means and variances of the bias terms from training data, exploiting high-dimensional concentration phenomena. This gives rise to non-asymptotic confidence intervals, which can help avoid overestimating certainty in critical applications such as MRI diagnosis. Importantly, our analysis extends beyond sparse regression to data-driven predictors like neural networks, enhancing the reliability of model-based deep learning. Our findings bridge the gap between established theory and the practical applicability of such methods.

1 Introduction

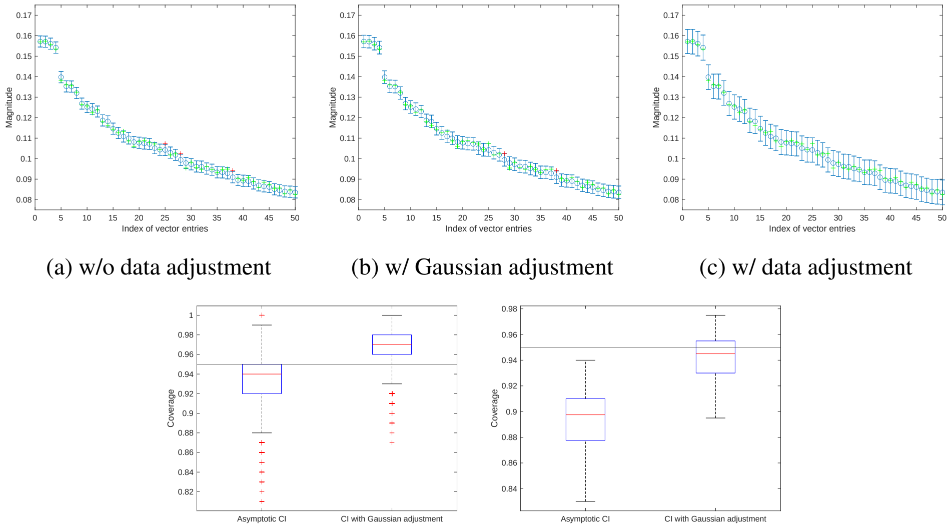
The past few years have witnessed remarkable advances in high-dimensional statistical models, inverse problems, and learning methods for solving them. In particular, we have seen a surge of new methodologies and algorithms that have revolutionized our ability to extract insights from complex, high-dimensional data [1–3]. Also, the theoretical underpinnings of the techniques in these fields have achieved tremendous success. However, the development of rigorous methods for quantifying the uncertainty associated with their estimates and underlying parameters, such as constructing confidence intervals for a given solution, has lagged behind, with much of the underlying theory remaining elusive.

In high-dimensional statistics, for example, even for classical regularized estimators such as the LASSO [4–6], it was shown that a closed-form characterization of the probability distribution of the estimator in simple terms is not possible, e.g., [7, Theorem 5.1]. This, in turn, implies that it is very challenging to establish rigorous confidence intervals that would quantify the uncertainty of

[*] : Equal contribution. Correspondence to hoppe@mathc.rwth-aachen.de.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3894



(a) w/o data adjustment

(b) w/ Gaussian adjustment

(c) w/ data adjustment

(d) coverage over all components

(e) coverage on the support

Figure 1: Illustration of the confidence interval correction. Figs. 1a, 1b, 1c show the construction of CIs with standard debiased techniques (w/o data adjustment) and with our proposed method (w/ Gaussian adjustment - Thm. 3 - in Fig. 1b and data adjustment - Thm. 2 - in Fig. 1c), respectively. The red points represent the entries that are not captured by the CIs. Additionally, Fig. 1d shows box plots of coverage over all components, and Fig. 1e shows them on the support, non-zero pixels. In the last two plots, the left box refers to the asymptotic and the right to the non-asymptotic CIs based on Gaussian adjustment of 500 feature vectors. We solve a sparse regression problem $y = Ax + \varepsilon$ via the LASSO, where $A \in \mathbb{C}^{4000\times 10000}$, $x \in \mathbb{C}^N$ is 200-sparse, and the noise level is $\approx 10\%$. The averaged coverage over 250 vectors with significance level $\alpha = 0.05$ of the asymptotic confidence intervals is $h^W(0.05) = 0.9353$ and on the support $h^W_S(0.05) = 0.8941$. Confidence intervals built with our proposed method yield for Gaussian adjustment $h^G(0.05) = 0.9684$ and on the support $h^G_S(0.05) = 0.9421$, and for data-driven adjustment $h(0.05) = h_S(0.05) = 1$. For more details, cf. Section 5.1 and Appendix D.

such estimated parameters. To overcome this, a series of papers [8–10] proposed and analyzed the *debiased LASSO*, also known as the desparsified LASSO, a procedure to fix the bias introduced by the ℓ_1 penalty in the LASSO; see [9, Corollary 11] and [11] for a discussion on the bias induced by the ℓ_1 regularizer. The debiased estimator derived in the aforementioned works has established a principled framework for obtaining sharp confidence intervals for the LASSO, initiating a statistical inference approach with UQ guarantees for high-dimensional regression problems where the number of predictors significantly exceeds the number of observations. Recently, this estimator was also extended in several directions beyond ℓ_1-minimization which include, for example, deep unrolled algorithms [12, 13] and it has been applied to many fields like magnetic resonance imaging both with classical high-dimensional regression techniques as well as recent learning ones [12, 14]; see the paragraph *related works* in Section 2 below.

The idea of the debiased LASSO is that its estimation error, i.e., the difference between the debiased estimator and the ground truth, can be decomposed into a **Gaussian** and a **remainder/bias** component. It has been shown in certain cases that the ℓ_∞ norm of the remainder component vanishes with high probability, assuming an *asymptotic* setting, i.e., when the dimensions of the underlying model grow within a specific rate, see [10, 15] for details. In this case, the estimator is proven to be *approximately Gaussian* from which the confidence intervals are derived. However, in practice, one needs to be in a very high-dimensional regime with enough data for these assumptions to kick in. In many applications with a finite set of observations, the *remainder term does not vanish*; it can rather be substantially large, and the confidence intervals constructed solely based on the Gaussian component *fail to account for the entire estimation error*. Consequently, the derived confidence intervals are narrower, resulting in an overestimation of certainty. This issue is particularly problematic in applications where it is crucial to estimate the magnitude of a vector coefficient with a high degree of confidence, such as in medical imaging applications.

Moreover, according to the standard theory of debiased estimators, the estimation of how small the **remainer term** is depends on how well one can quantify the ℓ_2 and ℓ_1 bounds for the corresponding *biased estimator*, e.g., the LASSO [10, 15]. Although sharp oracle inequalities exist for such classical regression estimators, cf. *related works*, the same cannot be said about when unrolled algorithms are employed. For the latter, generalization bounds are usually not sharp or do not exist.

In this paper, we tackle the challenge of constructing valid confidence intervals around debiased estimators for the parameters in high-dimensional learning. The key difficulty lies in accounting for the remainder term in the estimation error decomposition – see Equation 3 – which hinders the development of finite-sample confidence intervals. We propose a **novel non-asymptotic theory that explicitly characterizes the remainder term**, enabling us to construct reliable confidence intervals in the finite-sample regime. Furthermore, we extend our framework to quantify uncertainty for the output of model-based neural networks, which, in turn, are used to solve inverse problems. This paves the way for a rigorous theory of data-driven UQ for modern deep learning techniques. We state an informal version of our main result, discussed in detail in Section 3.

Theorem 1 (Informal Version)**.** *Let $x^{(1)}, \dots, x^{(l)} \in \mathbb{C}^N$ be i.i.d. data. Let $b^{(i)} = Ax^{(i)} + \varepsilon^{(i)}$ be a high-dimensional regression model with noise $\varepsilon^{(i)} \sim \mathcal{CN}(0, \sigma^2 I_{N\times N})$. With the data, derive, for a significance level α, a confidence radius $r_j(\alpha)$ for a new sample's component $x_j^{(l+1)}$. Let $(\hat{x}^u)_j^{(l+1)}$ be the debiased estimator based on a (learned) high-dimensional regression estimator $\hat{x}_j^{(i)}$. Then, it holds that*

$$\mathbb{P}\left(\left|(\hat{x}^u)_j^{(l+1)} - x_j^{(l+1)}\right| \leq r_j(\alpha)\right) \geq 1-\alpha.$$

Theorem 1 has far-reaching implications that transcend the classical regularized high-dimensional regression setting. For example, it enables the establishment of rigorous confidence intervals for learning algorithms such as unrolled networks [16]. To our knowledge, obtaining rigorous UQ results for neural networks without relying on non-scalable Monte Carlo methods remains a challenging problem [17]. To address this and quantify uncertainty, our approach combines model-based prior knowledge with data-driven statistical techniques. The model-based component harnesses the Gaussian distribution of the noise to quantify the uncertainty arising from the noisy data itself. We note that the Gaussian assumption for the noise is not a limitation, and extensions to non-Gaussian distributions are also possible via, e.g., a Central Limit Theorem-type argument, as clarified by [10]. We make a Gaussian noise assumption here for the sake of clarity. Complementing this, the data-driven component is imperative for quantifying the uncertainty inherent in the estimator's performance. Moreover, *our approach does not require any assumptions regarding the convergence or quality properties of the estimator*. This flexibility enables the debiased method to apply to a wide range of estimators.

Contributions. The key contributions in this work are threefold [1]

1. We solve the problem illustrated in Fig. 1 by **developing a non-asymptotic theory for constructing confidence intervals around the debiased LASSO estimator**. Unlike existing approaches that rely on asymptotic arguments and ignore the remainder term, *our finite-sample analysis explicitly accounts for the remainder*, clarifying an important theoretical gap and providing rigorous guarantees without appealing to asymptotic regimes.
2. We establish a general framework that **extends the debiasing techniques to model-based deep learning approaches for high-dimensional regression**. Our results enable the principled measurement of uncertainty for estimators learned by neural networks, a capability crucial for reliable decision-making in safety-critical applications. We test our approach with state-of-the-art unrolled networks such as the It-Net [18].
3. For real-world medical imaging tasks, we demonstrate that the remainder term in the debiased LASSO estimation error can be accurately modeled as a Gaussian distribution. Leveraging this finding, we derive **Gaussian adjusted CIs that provide sharper uncertainty estimates than previous methods**, enhancing the practical utility of debiased estimators in high-stakes medical domains.

[1] The code for our findings is available on GitHub : `https://github.com/frederikhoppe/UQ_high_dim_learning`

https://doi.org/10.52202/079017-3894

2 Background and Problem Formulation

In numerous real-world applications, we encounter high-dimensional regression problems where the number of features far exceeds the number of observations. This scenario, known as high-dimensional regression, arises when we aim to estimate N features, described by $x^0 \in \mathbb{C}^N$ from only a few m target measurements $b \in \mathbb{C}^m$, where $m \ll N$. Mathematically, this can be expressed as a linear model $b = Ax^0 + \varepsilon$, where $A \in \mathbb{C}^{m \times N}$ is the measurement matrix and $\varepsilon \sim \mathcal{CN}(0, \sigma^2 I_{N\times N})$ is additive Gaussian noise with variance σ^2. In the presence of sparsity, where the feature vector x^0 has only s non-zero entries ($s \ll N$), a popular approach is to solve the LASSO, which gives an estimator $\hat{x}$ obtained by solving the following ℓ_1-regularized optimization problem:

$$\min_{x\in\mathbb{C}^N} \frac{1}{2m}\|Ax - b\|_2^2 + \lambda\|x\|_1. \tag{1}$$

However, the LASSO estimator is known to exhibit a systematic bias, and its distribution is intractable, posing challenges for uncertainty quantification [7]. To address this limitation, debiasing techniques have been developed in recent years [8–10]. The debiased LASSO estimator, $\hat{x}^u$, is defined as:

$$\hat{x}^u = \hat{x} + \frac{1}{m}MA^*(A\hat{x} - b), \tag{2}$$

where M is a correction matrix that could be chosen such that $\max_{i,j\in\{1,\ldots,N\}} |(M\hat{\Sigma} - I_{N\times N})_{ij}|$ is small. Here, $\hat{\Sigma} = \frac{A^*A}{m}$. We refer to [15] for a more detailed description of how to choose M. Remarkably, the estimation error

$$\hat{x}^u - x^0 = \underbrace{MA^*\varepsilon/m}_{=:W} + \underbrace{(M\hat{\Sigma} - I_{N\times N})(x^0 - \hat{x})}_{=:R}, \tag{3}$$

can be decomposed into a Gaussian component $W \sim \mathcal{CN}(0, \frac{\sigma^2}{m}\hat{\Sigma})$ and a remainder term R that vanishes asymptotically with high probability [15, Theorem 3.8], assuming a Gaussian measurement matrix A. Such a result was extended to matrices associated to a bounded orthonormal system like a subsampled Fourier matrix, allowing for extending the debiased LASSO to MRI [19]. The decomposition (3) and the asymptotic behavior of R enable the construction of asymptotically valid CIs for the debiased LASSO estimate, providing principled UQ for high-dimensional sparse regression problems.

However, in real-world applications involving finite data regimes, the remainder term can be significant, rendering the asymptotic confidence intervals imprecise or even misleading, as illustrated in Fig. 1. This issue is particularly pronounced in high-stakes domains like medical imaging, where reliable UQ is crucial for accurate diagnosis and treatment planning. Second, the debiasing techniques have thus far been restricted to estimators whose error is well quantifiable, leaving the **challenge of how they would behave for deep learning architectures open**. In such cases, the behavior of the remainder term is largely unknown, precluding the direct application of existing debiasing methods and hindering the deployment of these methods in risk-sensitive applications.

A prominent example for solving the LASSO problem in (1) with an unrolled algorithm is the ISTA [20, 21]:

$$x^{k+1} = \mathcal{S}_\lambda\left((I_{N\times N} - \frac{1}{\mu}A^TA)x^k + \frac{1}{\mu}A^Tb\right), \qquad k \geq 0.$$

Here, $\mu > 0$ is a step-size parameter, and $\mathcal{S}_\lambda(x)$ is the soft-thresholding operator. The work [22] interpreted each ISTA iteration as a layer of a recurrent neural network (RNN). The Learned ISTA (LISTA) approach learns the parameters W_1^k, W_2^k, λ^k instead of using the fixed ISTA updates:

$$x^{k+1} = \mathcal{S}_{\lambda^k}(W_2^k x^k + W_1^k b).$$

In this formulation, LISTA unrolls K iterations into K layers, with learnable parameters (W^k, λ^k) per layer. The parameters are learned by minimizing the reconstruction error $\min_{\lambda,W} \frac{1}{l}\sum_{i=1}^{l}\|x_i^k(\lambda, W, b^{(i)}, x^{(i)}) - x^{(i)}\|_2^2$ on training data $(x^{(i)}, b^{(i)})$. Unrolled neural networks like LISTA have shown promise as model-based deep learning solutions for inverse problems, leveraging domain knowledge for improved performance. Such iterative end-to-end network schemes provide state-of-the-art reconstructions for inverse problems [18]. Recently, the work [12] proposes a

https://doi.org/10.52202/079017-3894

framework based on the debiasing step to derive confidence intervals specifically for the unrolled LISTA estimator. However, similar to the previously mentioned debiased LASSO literature, *it only handles the asymptotic setting*. One of the main goals of this paper is to overcome such limitation of the current theory.

Related Works.

High-dimensional regression. High-dimensional regression and sparse recovery is now a well-established theory, see [1, 3, 23] and references therein. In this context, several extensions of the LASSO have been proposed such as the elastic net [24], the group LASSO [25], the LASSO with a nuclear norm penalization [26], the Sorted L-One Penalized Estimation (SLOPE) [27] which adapts the ℓ_1-norm to control the false discovery rate. In addition to convex penalty functions, concave penalties have been explored to address some limitations of the LASSO, e.g., the Smoothly Clipped Absolute Deviation (SCAD) penalty [28] and the Minimax Concave Penalty (MCP) [29]. Non-convex variants of the LASSO for ℓ_p-norm ($p < 1$) minimization were also studied [30, 31] as well as noise-blind variants such as the square-root LASSO [32, 33]. Scalable and fast algorithms for solving the LASSO and its variants include semi-smooth Newton methods [34] and IRLS [35].

*LASSO theory.*Given how ubiquitous and studied such an estimator is, it is difficult to do justice to all the papers that have contributed to such a theory. Several works have established oracle inequalities for the LASSO [36–40]. Another key theoretical result is the consistency of the LASSO in terms of variable selection. [41] and [42] established the consistency of the LASSO while [43] analyzed the sparsity behavior of the LASSO when the design matrices satisfy the Restricted Isometry Property.

Debiased estimators. After the first papers about the debiased LASSO [8–10], some works have focused on improving its finite-sample performance and computational efficiency [15, 44]. The size of the confidence intervals derived for the debiased LASSO has been proven to be sharp in the minimax sense [45]. Debiased estimators have been extended in several directions, e.g., [19, 44, 46, 47]. Recently, [48] established asymptotic normality results for a debiased estimator of convex regularizers beyond the ℓ_1-norm. In the context of MR images, [49] explored a debiased estimator for inverse problems with a total variation regularizer. Debiased estimators have also been recently extended to unrolled estimators – see discussion in the next paragraph – in [12, 13].

Algorithm unrolling and model-based deep learning for inverse problems. The idea of unfolding the iterative steps of classical algorithms into a deep neural network architecture dates back to [22], which proposed the Learned ISTA (LISTA) to fast approximate the solution of sparse coding problems. Several works have extended and improved upon the original LISTA framework [50–55]. [56] proposed the Learned Primal-Dual algorithm, unrolling the primal-dual hybrid gradient method for tomographic reconstruction. [57] proposed the Deep Cascade of Convolutional Neural Networks (DC-CNN) for dynamic MRI reconstruction. [58] unfolded proximal gradient descent solvers to learn their parameters for 1D TV regularized problems. [16] introduced a general framework for algorithm unrolling. [59] developed MoDL, a model-based deep learning approach for MRI reconstruction that unrolls the ADMM algorithm. [60] proposed a proximal alternating direction network (PADNet) to unroll nonconvex optimization. See also the surveys for more information about unrolling and also the connection with physics-inspired methods [61, 62]. [18, 63] developed the It-Net, an unrolled proximal gradient descent scheme where the proximal operator is replaced by a U-Net. This scheme won the AAPM Challenge 2021 [64] whose goal was *to identify the state-of-the-art in solving the CT inverse problem with data-driven techniques*. A generalization of the previous paradigm is the *learning to optimize* framework that develops an optimization method by training, i.e., learning from its performance on sample problem [65, 66].

Uncertainty Quantification. There have been a few attempts to quantify uncertainty on a pixel level for unrolled networks used in imaging processing, e.g., [67]. However, such approaches are based on Bayesian networks and MC dropout [68], which requires significant inference time paired with a loss of reconstruction performance since the dropout for UQ is a strong regularizer in the neural network. Unlike prior work, our contribution focuses on a scalable data-driven method that is easily implementable in the data reconstruction pipeline.

Another method that became popular in the last couple of years is conformal prediction [69, 70], which addresses the problem of constructing prediction bands $\hat{C}_m : X \to \{\text{subsets of } B\}$ for a given level α with the property that for a new i.i.d. pair (a_i, b_i), we get $\mathbb{P}(b_{m+1} \in \hat{C}_n(a_{m+1})) \geq 1 - \alpha$, where the pairs $(a_i, b_i) \sim P, i = 1, \ldots, m$ are i.i.d. feature and response pairs from a distribution

https://doi.org/10.52202/079017-3894

P on $A \times B$ and the probability is over all of available data $(a_i, b_i), i = 1, \ldots, m+1$. Such a method assumes that the regression coefficient x, in the model $b_i = \langle a_i, x \rangle + \epsilon$, is fixed. In contrast to that, the debiased LASSO produces confidence intervals for individual pixels $x_j^{(l+1)}$, where $j = 1, \ldots, N$ when a new set of data points $b_1^{(l+1)}, \ldots, b_m^{(l+1)}$ is given for a fixed set of measurement vectors $a_1, \ldots, a_m$. Recently, a few papers developed conformal prediction-based uncertainty masks for imaging tasks [71, 72]. Unlike the former, our method provides computationally inexpensive confidence intervals for a new given test image. Unlike the latter, which constructs an interval-valued function for each pixel of a new image that holds in expectation (see Equation 2 in [72]), our method comes with pixel-wise coverage guarantees for each new test image.

3 Data-Driven Confidence Intervals

We now introduce our data-driven approach to correct the CIs. Instead of deriving asymptotic CIs from the decomposition $\hat{x}^u - x^0 = W + R$, by assuming that R asymptotically vanishes, we utilize data $\left(b^{(i)}, x^{(i)}\right)_{i=1}^{l}$ along with concentration techniques to estimate the size of the bias component R. We continue to leverage the prior knowledge of the Gaussian component W while extending the CIs' applicability to a broad class of estimators, including neural networks. Our method is summarized in Algorithm 1, where the data is used to estimate the radii of the CIs, and in Algorithm 2, which constructs the estimator around which the CIs are centered. The following main result proves the validity of our method.

Theorem 2. *Let $x^{(1)}, \ldots, x^{(l)} \in \mathbb{C}^N$ be i.i.d. complex random variables representing ground truth data drawn from an unknown distribution $\mathbb{Q}$. Suppose, that $\varepsilon^{(i)} \sim \mathcal{CN}(0, \sigma^2 I_{m\times m})$ is noise in the high-dimensional models $b^{(i)} = Ax^{(i)} + \varepsilon^{(i)}$, where $A \in \mathbb{C}^{m\times N}$, and independent of the $x^{(i)}$'s. Let $\hat{X} : \mathbb{C}^m \to \mathbb{C}^N$ be a (learned) estimation function that maps the data $b^{(i)}$ to $\hat{x}^{(i)}$, which is an estimate for $x^{(i)}$. Set $|R_j^{(i)}| = |e_j^T (M\hat{\Sigma} - I_{N\times N})(\hat{x}^{(i)} - x^{(i)})|$ for fixed A and M. For $j = 1, \ldots, N$, we denote the true but unknown mean with $\mu_j = \mathbb{E}[|R_j^{(1)}|]$ and the unknown variance with $(\sigma_R^2)_j := \mathbb{E}[(|R_j^{(1)}| - \mu_j)^2]$, respectively. Let $\hat{S}_j = \frac{1}{l}\sum_{i=1}^{l} |\ R_j^{(i)}|$ be the unbiased sample mean estimator and $(\hat{\sigma}_R^2)_j = \frac{1}{l-1}\sum_{i=1}^{l}(|R_j^{(i)}| - \hat{S}_j)^2$ the unbiased variance estimator. Let $\alpha \in (0,1)$ and $\gamma_j \in \left(0, 1 - \frac{1}{l\alpha}\right)$. Furthermore, set the confidence regions for the sample $x^{(l+1)} \sim \mathbb{Q}$ in the model $b^{(l+1)} = Ax^{(l+1)} + \varepsilon^{(l+1)}$ as $C_j(\alpha) = \{z \in \mathbb{C} : |(\hat{x}^u)_j^{(l+1)} - z| \leq r_j(\alpha)\}$ with radius*

$$r_j(\alpha) = \frac{\sigma(M\hat{\Sigma}M^*)_{jj}^{1/2}}{\sqrt{m}} \sqrt{\log\left(\frac{1}{\gamma_j \alpha}\right)} + c_l\,(\alpha) \cdot (\hat{\sigma}_R)_j + \hat{S}_j, \qquad c_l(\alpha) := \sqrt{\frac{l^2 - 1}{l^2(1-\gamma_j)\alpha - l}}. \tag{4}$$

Then, it holds that

$$\mathbb{P}\left(x_j^{(l+1)} \in C_j(\alpha)\right) \geq 1 - \alpha. \tag{5}$$

Theorem 2 presents a way to achieve conservative confidence intervals that are proven to be valid, i.e., are proven to contain the true parameter with a probability of $1 - \alpha$. Its main advantage is that there are *no assumptions* on the distribution $\mathbb{Q}$ (except that σ_R^2 exists), making it widely applicable. Hence, Theorem 2 includes the worst-case distribution showing a way to quantify uncertainty even in such an ill-posed setting. Especially in medical imaging, such certainty guarantees are crucial for accurate diagnosis. The proof exploits the Gaussianity of the component W as well as an empirical version of Chebyshev's inequality, which is tight when there is no information on the underlying distribution. The detailed proof can be found in Appendix B. For a thorough discussion on Theorem 2 including practical simplifications and the dependence of γ_j on the confidence interval length, we refer to Appendix A.

More certainty comes with the price of larger confidence intervals. If there is additional information on the distribution of R, like the ability to be approximated by a Gaussian distribution, then the confidence intervals become tighter. This case, which includes relevant settings such as MRI, is discussed in Section 4.

https://doi.org/10.52202/079017-3894

Algorithm 1 Estimation of Confidence Radius

```
Input: Estimation function $\hat{X}$, dictionary matrix $A$, correction matrix $M$, data points $\left(b^{(i)}, x^{(i)}\right)_{i=1}^{l}$, significance level $\alpha$
for $i = 1, \ldots, l$ do
    Compute $\hat{x}^{(i)}, R^{(i)} \in \mathbb{C}^N$ via $\hat{x}^{(i)} = \hat{X}(b^{(i)})$ and $R_j^{(i)} = e_j^T (M\hat{\Sigma} - I_{N\times N})(\hat{x}^{(i)} - x^{(i)})$.
end for
for $j = 1, \ldots, N$ do
    Estimate $\hat{S}_j = \frac{1}{l}\sum_{i=1}^{l} |R_j^{(i)}|$ and $(\hat{\sigma}_R^2)_j = \frac{1}{l-1}\sum_{i=1}^{l}(|R_j^{(i)}| - \hat{S}_j)^2$
    Solve $r_j(\alpha) = \min_{\gamma \in \left(0, 1-\frac{1}{l\alpha}\right)} \frac{\sigma(M\hat{\Sigma}M^*)_{jj}^{1/2}}{\sqrt{m}} \sqrt{\log\left(\frac{1}{\gamma\alpha}\right)} + c_l\left((1-\gamma)\alpha\right) \cdot (\hat{\sigma}_R)_j + \hat{S}_j$
end for
Output: Radii of confidence regions $(r_j(\alpha))_{j=1}^{N}$
```

Algorithm 2 Construction of Confidence Regions

```
Input: Estimation function $\hat{X}$, dictionary matrix $A$, correction matrix $M$, measurement $b$, radii $(r_j(\alpha))_{j=1}^{N}$ (derived from Algorithm 1 or Theorem 3)
Compute estimator $\hat{x} = \hat{X}(b)$.
Construct debiased estimator via $\hat{x}^u = \hat{x} + \frac{1}{m} M A^*(b - A\hat{x})$
for $j = 1, \ldots, N$ do
    Construct confidence region $C_j(\alpha) = \{z \in \mathbb{C} \mid |\hat{x}_j^u - z| \leq r_j(\alpha)\}$
end for
Output: Debiased estimator $x^u$ and confidence regions $(C_j(\alpha))_{j=1}^{N}$
```

4 Confidence Intervals for Gaussian Remainders

Valid confidence intervals, i.e., those with correct coverage probability, can be derived most straightforwardly when the distribution of the remainder term is known and easily characterized. In such cases, more informative distributional assumptions lead to potentially tighter confidence intervals compared to Theorem 2, which makes no assumptions about the remainder component. In this section, we derive non-asymptotic confidence intervals assuming the remainder term to be approximated by a Gaussian distribution.

Theorem 3. *Let $\hat{x}^u \in \mathbb{C}^N$ be a debiased estimator for $x \in \mathbb{C}^N$ with a remainder term $R \sim \mathcal{CN}(0, \Sigma_R/m)$. Then, $C_j(\alpha) = \{z \in \mathbb{C} \mid |z - \hat{x}_j^u| \leq r_j(\alpha)\}$ with radius*

$$r_j^G(\alpha) = \frac{(\sigma^2 (M\hat{\Sigma}M^*)_{jj} + (\Sigma_R)_{jj})^{1/2}}{\sqrt{m}} \sqrt{\log\left(\frac{1}{\alpha}\right)}. \tag{6}$$

is valid, i.e. $\mathbb{P}\left(x_j \in C_j(\alpha)\right) \geq 1 - \alpha$.

For the proof, we refer to Appendix B. We note, however, that the theorem can be generalized beyond the Gaussian case. In particular, we present in Appendix C a proof of this theorem for heavy-tailed distributions. In Appendix E, we demonstrate empirically that the Gaussian assumption for the remainder term holds in a wide range of relevant practical settings. This validation enables the application of the proposed confidence intervals derived under this assumption. These confidence intervals strike a careful balance between non-asymptotic reliability, ensuring valid coverage even in finite-sample regimes, and tightness, providing informative and precise uncertainty estimates. By leveraging the Gaussian approximation, which becomes increasingly accurate in higher dimensions as illustrated in Figure 7, our framework offers a principled and computationally efficient approach to quantifying uncertainty in high-dimensional prediction problems. The variance of R can be estimated with the given data using, e.g., the unbiased estimator for the variance as in Theorem 2.

5 Numerical Experiments

We evaluate the performance of our non-asymptotic confidence intervals through extensive numerical experiments across two settings: (i.) the *classical debiased LASSO framework* to contrast our non-asymptotic confidence intervals against the asymptotic ones. (ii.) the *learned framework* where we employ learned estimators, specifically the U-net [73] as well as the It-Net [18], to reconstruct real-world MR images and quantify uncertainty. Our experiments demonstrate the importance of properly accounting for the remainder term in practical, non-asymptotic regimes. Each experiment follows the same structure:

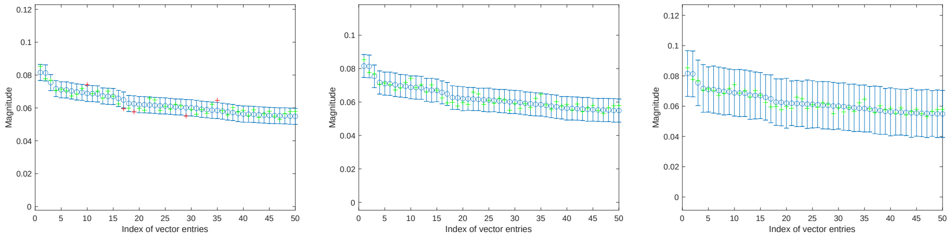
1. Data Generation and Management: We fix the forward operator A and generate $n > 2$ feature vectors ${x^{(i)}}_{i=1}^{n}$ and noise vectors ${\varepsilon^{(i)}}_{i=1}^{n}$ with $\varepsilon^{(i)} \sim \mathcal{CN}(0, \sigma^2 I_{m\times m})$. We obtain observations ${b^{(i)}}_{i=1}^{n}$ via $b^{(i)} = Ax^{(i)} + \varepsilon^{(i)}$. We split the data $(b^{(i)}, x^{(i)})_{i=1}^{n}$ into an *estimation* dataset of size l and a test dataset of size k ($l + k = n$). If we learn an estimator, we further split the data into training, estimation, and test sets.
2. Reconstruction: Depending on the experiment, we obtain a reconstruction function $\hat{X}$ in one of the following ways: for the classical LASSO setting, we use the LASSO; for the learned estimator experiment, we train a U-Net [73] or It-net [18] on the training data to serve as the reconstruction function $\hat{X}$.
3. Estimation of Confidence Radii: We run Algorithm 1 with $A, \hat{X}, M$ (that is chosen according to [15]), the estimation data $(b^{(i)}, x^{(i)})_{i=1}^{l}$, and a predefined significance level $\alpha \in (0, 1)$ to obtain radii ${r_j(\alpha)}_{j=1}^{N}$. We optimize over γ, therefore the γ we use is the optimal one which leads to the smallest confidence intervals. To construct the final confidence intervals, the radii need to be centered according to the debiased estimator. For every new measurement b, we run Algorithm 2 to obtain tailored confidence intervals for the feature vector x corresponding to b. In addition, we compute the CI for the Gaussian adjustment based on Theorem 3 using the estimation set to quantify the variance of R with the unbiased estimator for the variance as before.
4. Evaluation: We use the test dataset $(b^{(i)}, x^{(i)})_{i=l+1}^{n}$ to evaluate our adjustments. For each $b^{(i)}$, we run Algorithm 2 to obtain confidence intervals ${C_j^{(i)}(\alpha)}_{j=1}^{N}$ for $x^{(i)}$. We estimate $\mathbb{P}(x_j^{(i)} \in C_j(\alpha))$ by $h_j(\alpha) = \frac{1}{k}\sum_{i=l+1}^{n} \mathbb{1}_{\{x_j^{(i)} \in C_j(\alpha)\}}$ and average over all components $h(\alpha) = \frac{1}{N}\sum_{j=1}^{N} h_j(\alpha)$. Since performance on the support S, the non-zero pixels, is crucial, we define the hit rate on S as $h_S^{(i)} = \frac{1}{|S|}\sum_{j=1}^{N} \mathbb{1}_{\{x_j^{(i)} \in C_j(\alpha)\}}$ and average $h_S(\alpha) = \frac{1}{k}\sum_{i=l+1}^{n} h_S^{(i)}$. Note that the support may change with i. Moreover, we do the same for the CI based on the Gaussian-adjusted radii.

5.1 UQ for Sparse Model-Based Regression

We consider a setting aligned with existing debiased LASSO literature, e.g., [9] to demonstrate our approach's extension of current UQ methods. The forward operator is a complex Gaussian matrix $A \in \mathbb{C}^{m\times N}$ with dimensions $N = 10000$, $m = 0.6N$, and $A_{ij} \sim \mathcal{CN}(0, 1)$. We generate $n = 750$ $s = 0.1N$-sparse features $x^{(i)}$ by randomly selecting m distinct indices from $1, \dots, N$ and drawing magnitudes from $\mathcal{CN}(0, 1)$. With relative noise $\frac{\|\varepsilon^{(i)}\|}{\|Ax^{(i)}\|} \approx 0.2$, we split the data $(b^{(i)}, x^{(i)})_{i=1}^{n}$ into $l = 500$ estimation and $k = 250$ test data. For reconstruction, we solve the LASSO $\hat{X}(b) := \operatorname{argmin}_{x\in\mathbb{C}^N} \frac{1}{m}\|Ax - b\| + \lambda\|x\|_1$ with $\lambda = 10\frac{\sigma}{\sqrt{m}}(2 + \sqrt{12\log(N)})$ following [19].

With significance level $\alpha = 0.05$, we run Algorithm 1 to obtain confidence radii, choosing $M = I_{N\times N}$ [15] and exploiting the relaxation (9). Averaged over the l estimation data points, the ℓ_2 and ℓ_∞ norm ratios are: $\frac{\|R\|_2}{\|W\|_2} = 0.9993$ and $\frac{\|R\|_\infty}{\|W\|_\infty} = 1.1581$. In existing literature, the ℓ_∞ norm is typically measured when the remainder term vanishes, as it is relevant for pixel-wise confidence intervals. Here, the remainder term is of comparable order as the Gaussian term and hence, too significant to neglect in confidence intervals derivation.

https://doi.org/10.52202/079017-3894



(a) w/o adjustment

(b) w/ Gaussian adjustment

(c) w/ data adjustment

Figure 2: Confidence intervals of asymptotic type 2a, with Gaussian adjustment 2b and data-driven adjustment 2c for one evaluation feature vector in the sparse regression setting described in Section 5.1.

Evaluating it on the remaining $k = 250$ data points, the data-driven and Gaussian-adjusted averaged hit rates are $h(0.05) = 1$, $h_S(0.05) = 1$ and $h^G(0.05) = 0.9691$, $h_S^G(0.05) = 0.8948$, respectively. Neglecting the remainder term yields $h^W(0.05) = 0.8692$ and $h_S^W(0.05) = 0.6783$, which is substantially lower and violates the specified 0.05 significance level. Fig. 2 presents confidence intervals of each type for one data point $x^{(i)}$. A detailed visualization of $h_j(0.05)$, $h_S^{(i)}(0.05)$, $h_j^G(0.05)$, and $(h_S^G)^{(i)}(0.05)$ is illustrated in Fig. 4c and 4i in the appendix. Further experiments with different sparse regression settings, including subsampled Fourier matrices, are also presented in Appendix D.

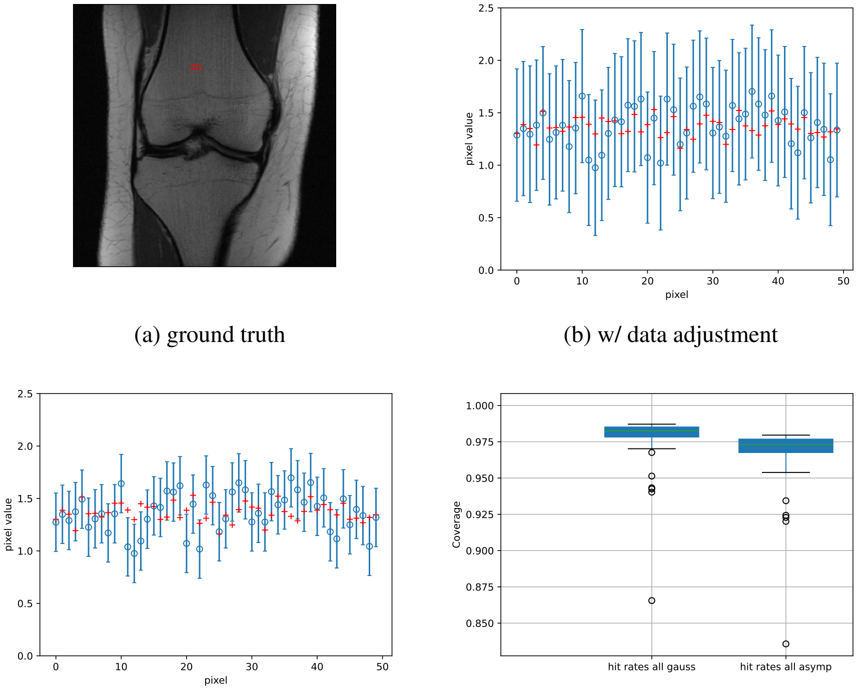
5.2 UQ for MRI Reconstruction with Unrolled Neural Networks

We extend the debiasing approach to model-based deep learning for MRI reconstruction using the U-Net and It-Net on single-coil knee images from the NYU fastMRI dataset [2] [74, 75]. Here, the forward operator is the undersampled Fourier operator $\mathcal{PF} \in \mathbb{C}^{m \times N}$ with $N = 320 \times 320$, $m = 0.6N$, the Fourier matrix $\mathcal{F}$ and a radial mask $\mathcal{P}$, see Figure 5b. The noise level σ is chosen such that the relative noise is approximately 0.1. The data is split into training (33370 slices), validation (5346 slices), estimation (1372 slices), and test (100 slices) datasets.

We then train an It-Net [18] with 8 layers, a combination of MS-SSIM [76] and ℓ_1-losses and Adam optimizer with learning rate $5e^{-5}$ for 15 epochs to obtain our reconstruction function $\hat{X}$.

With significance level $\alpha = 0.1$, we run Algorithm 1 to construct confidence radii, this time for the real part of the image instead of the magnitude, choosing $M = I_{N \times N}$ [15] and exploiting the relaxation (9). Averaged over the l estimation data points, we have $\frac{\|R\|_2}{\|W\|_2} = 0.38$ and $\frac{\|R\|_\infty}{\|W\|_\infty} = 0.49$, which indicates that the remainder term is significant and cannot be neglected. Evaluating the test data, the averages of the data-driven adjustment hit rates are $h(0.1) = 0.9998$, $h_S(0.1) = 0.9995$ and the averages of the Gaussian adjusted hit rates are $h^G(0.1) = 0.9485$, $h_S^G(0.1) = 0.9353$, where the support refers to the largest 10% of the image pixels as the background cannot be clearly delineated numerically. Neglecting the remainder term, the hit rates of the asymptotic CIs are $h^W(0.1) = 0.9306$ and $h_S^W(0.1) = 0.9152$. As in the sparse regression setting, they are significantly lower. Fig. 3 presents confidence intervals based on the data-driven adjustment and the asymptotic confidence intervals for a region in one image $x^{(i)}$. In addition, it contains a box plot showing the distribution of the hit rates based on the Gaussian adjustment and the asymptotic hit rates. More experiments for UQ for MRI reconstruction can be found in Appendix D and Tables 2 and 3.

[2] We obtained the data, which we used for conducting the experiments in this paper from the NYU fastMRI Initiative database (fastmri.med.nyu.edu) [74, 75]. The data was only obtained from the NYU fastMRI investigators, but they did not contribute any ideas, analysis, or writing to this paper. The list of the NYU fastMRI investigators, can be found at fastmri.med.nyu.edu, it is subject to updates.



(a) ground truth

(b) w/ data adjustment

(c) w/o data adjustment

(d) coverage over all components

Figure 3: Reconstruction obtained with the It-Net as described in 5.2. Data-driven adjustment 90% confidence intervals 3b and asymptotic 90% confidence intervals 3c for the region (50 pixels) in 320x320 knee image 3a; Boxplots of hit rates 3d for 95% confidence level for the Gaussian adjusted and asymptotic confidence intervals.

6 Final Remarks

In this work, we proposed a data-driven uncertainty quantification method that derives non-asymptotic confidence intervals based on debiased estimators. Our approach corrects asymptotic confidence intervals by incorporating an estimate of the remainder component and has solid theoretical foundations. While the correction can be based on prior knowledge, e.g., a Gaussian distribution of the remainder term, we also derive CI based on a data-driven adjustment without further information. This data-driven nature enhances its applicability to a wide range of estimators, including model-based deep-learning techniques. We conducted experiments that confirm our theoretical findings, demonstrating that even in classical sparse regression settings, the remainder term is too significant to be neglected. Furthermore, we applied the proposed method to MRI, achieving significantly better rates on the image support.

Limitations and Future Directions. While our method corrects for the remainder term, larger remainder terms necessitate greater corrections, resulting in wider confidence intervals. The goal is that those intervals should be narrow enough to be informative but wide enough to be realistic, given the sample size and variability in the data. Therefore, it is crucial to achieve a small remainder term to avoid excessively large confidence intervals. Additionally, the accuracy of our method depends on the quality of the estimates for the mean and variance of the remainder term, which improves with more available data. Additionally, the length of the intervals can be minimized over a larger parameter set, provided that more data is available. We leave as a future direction to study the sharpness of the proposed confidence intervals and radii for a given amount of data. Moreover, we would like to investigate how the length of the confidence intervals could be improved when estimating higher moments. We believe that our method is applicable to a wide variety of deep learning architectures, including vision transformers in MRI, e.g., [77]. Testing the generality of the method with state-of-the-art architectures for different problems would demonstrate its broad usefulness.

https://doi.org/10.52202/079017-3894

Acknowledgments

We gratefully acknowledge financial support with funds provided by the German Federal Ministry of Education and Research in the grant *"SparseMRI3D+: Compressive Sensing und Quantifizierung von Unsicherheiten für die beschleunigte multiparametrische quantitative Magnetresonanztomografie (FZK 05M20WOA)"*.

References

[1] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

[2] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[3] John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[4] Scott Chen and David L Donoho. Examples of basis pursuit. In *Wavelet Applications in Signal and Image Processing III*, volume 2569, pages 564–574. SPIE, 1995.

[5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[6] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[7] Keith Knight and Wenjiang Fu. Asymptotics for Lasso-Type Estimators. *The Annals of Statistics*, 28(5):1356–1378, 2000. ISSN 00905364.

[8] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. ISSN 13697412. doi: 10.1111/rssb.12026.

[9] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

[10] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1221.

[11] Cun Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[12] Frederik Hoppe, Claudio Mayrink Verdun, Hannah Laus, Felix Krahmer, and Holger Rauhut. Uncertainty quantification for learned ISTA. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2023.

[13] Pierre C Bellec and Kai Tan. Uncertainty quantification for iterative algorithms in linear models with application to early stopping. *arXiv preprint arXiv:2404.17856*, 2024.

[14] Frederik Hoppe, Felix Krahmer, Claudio Mayrink Verdun, Marion I Menzel, and Holger Rauhut. High-Dimensional Confidence Regions in Sparse MRI. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[15] A. Javanmard and A. Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A), 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1630.

[16] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[17] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1): 1513–1589, 2023.

[18] Martin Genzel, Ingo Gühring, Jan Macdonald, and Maximilian März. Near-exact recovery for tomographic inverse problems via deep learning. In *International Conference on Machine Learning*, pages 7368–7381. PMLR, 2022.

https://doi.org/10.52202/079017-3894

[19] Frederik Hoppe, Felix Krahmer, Claudio Mayrink Verdun, Marion I Menzel, and Holger Rauhut. Uncertainty quantification for sparse Fourier recovery. *arXiv preprint arXiv:2212.14864*, 2022.

[20] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11): 1413–1457, 2004.

[21] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

[22] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on machine learning*, pages 399–406, 2010.

[23] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer New York, New York, NY, 2013. ISBN 978-0-8176-4947-0. doi: 10.1007/978-0-8176-4948-7.

[24] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[25] Ming Yuan and Yi Lin. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 12 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00532.x.

[26] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302 – 2329, 2011. doi: 10.1214/11-AOS894.

[27] Malgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel Candès. SLOPE–Adaptive Variable Selection via Convex Optimization. *The Annals of Applied Statistics*, 9(65):1103–1140, 2015.

[28] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[29] Cun Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010.

[30] Le Zheng, Arian Maleki, Haolei Weng, Xiaodong Wang, and Teng Long. Does ℓ_p-Minimization Outperform ℓ_1-Minimization? *IEEE Transactions on Information Theory*, 63(11):6896–6935, 2017.

[31] Alain Rakotomamonjy, Rémi Flamary, Joseph Salmon, and Gilles Gasso. Convergent working set algorithm for lasso with non-convex sparse regularizers. In *International Conference on Artificial Intelligence and Statistics*, pages 5196–5211. PMLR, 2022.

[32] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[33] Claudio Mayrink Verdun, Oleh Melnyk, Felix Krahmer, and Peter Jung. Fast, blind, and accurate: Tuning-free sparse regression with global linear convergence. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3823–3872. PMLR, 2024.

[34] Xudong Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018.

[35] Christian Kümmerle, Claudio Mayrink Verdun, and Dominik Stöger. Iteratively reweighted least squares for basis pursuit with global linear convergence rate. *Advances in Neural Information Processing Systems*, 34:2873–2886, 2021.

[36] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[37] Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pages 7–57, 2009.

[38] Fei Ye and Cun-Hui Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

[39] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10): 6976–6994, 2011.

[40] Arnak Dalalyan, Mohamed Hebiri, and Johannes C Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

[41] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[42] Martin J Wainwright. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55 (5):2183–2202, 2009.

[43] Simon Foucart, Eitan Tadmor, and Ming Zhong. On the sparsity of LASSO minimizers in sparse data recovery. *Constructive Approximation*, 57(2):901–919, 2023.

[44] Sai Li. Debiasing the debiased Lasso with bootstrap. *Electronic Journal of Statistics*, 14: 2298–2337, 2020.

[45] T Tony Cai and Zijian Guo. Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity. *The Annals of Statistics*, pages 615–646, 2017.

[46] Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *Annals of statistics*, 50(3):1320, 2022.

[47] Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022.

[48] Pierre C Bellec and Cun-Hui Zhang. Debiasing convex regularized estimators and interval estimation in linear models. *The Annals of Statistics*, 51(2):391–436, 2023.

[49] Frederik Hoppe, Claudio Mayrink Verdun, Hannah Laus, Sebastian Endt, Marion I. Menzel, Felix Krahmer, and Holger Rauhut. Imaging with Confidence: Uncertainty Quantification for High-dimensional Undersampled MR Images. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2024.

[50] Daisuke Ito, Satoshi Takabe, and Tadashi Wadayama. Trainable ISTA for sparse signal recovery. *IEEE Transactions on Signal Processing*, 67(12):3113–3125, 2019.

[51] Kailun Wu, Yiwen Guo, Ziang Li, and Changshui Zhang. Sparse coding with gated learned ISTA. In *International conference on learning representations*, 2019.

[52] Jialin Liu and Xiaohan Chen. ALISTA: Analytic weights are as good as learned weights in LISTA. In *International Conference on Learning Representations (ICLR)*, 2019.

[53] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Hyperparameter tuning is all you need for LISTA. *Advances in Neural Information Processing Systems*, 34:11678–11689, 2021.

[54] Aviad Aberdam, Alona Golts, and Michael Elad. Ada-lista: Learned solvers adaptive to varying models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9222–9235, 2021.

[55] Ziyang Zheng, Wenrui Dai, Duoduo Xue, Chenglin Li, Junni Zou, and Hongkai Xiong. Hybrid ISTA: Unfolding ISTA with convergence guarantees using free-form deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3226–3244, 2022.

[56] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37(6):1322–1332, 2018.

[57] Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.

[58] Hamza Cherkaoui, Jeremias Sulam, and Thomas Moreau. Learning to solve TV regularised problems with unrolled algorithms. *Advances in Neural Information Processing Systems*, 33: 11513–11524, 2020.

[59] Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.

https://doi.org/10.52202/079017-3894

[60] Risheng Liu, Shichao Cheng, Long Ma, Xin Fan, and Zhongxuan Luo. Deep proximal unrolling: Algorithmic framework, convergence analysis and applications. *IEEE Transactions on Image Processing*, 28(10):5013–5026, 2019.

[61] Jian Zhang, Bin Chen, Ruiqin Xiong, and Yongbing Zhang. Physics-inspired compressive sensing: Beyond deep unrolling. *IEEE Signal Processing Magazine*, 40(1):58–72, 2023.

[62] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[63] Martin Genzel, Jan Macdonald, and Maximilian März. Solving inverse problems with deep neural networks–robustness included? *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1119–1134, 2022.

[64] Emil Y Sidky and Xiaochuan Pan. Report on the AAPM deep-learning sparse-view CT grand challenge. *Medical physics*, 49(8):4935–4943, 2022.

[65] Ke Li and Jitendra Malik. Learning to Optimize. In *International Conference on Learning Representations*, 2016.

[66] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

[67] Canberk Ekmekci and Mujdat Cetin. Uncertainty quantification for deep unrolling-based computational imaging. *IEEE Transactions on Computational Imaging*, 8:1195–1209, 2022.

[68] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[69] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[70] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[71] Gilad Kutiel, Regev Cohen, Michael Elad, Daniel Freedman, and Ehud Rivlin. Conformal prediction masks: Visualizing uncertainty in medical imaging. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 163–176. Springer, 2023.

[72] Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR, 2022.

[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[74] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI, 2019.

[75] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J. Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzalv, Adriana Romero, Michael Rabbat, Pascal Vincent, James Pinkerton, Duo Wang, Nafissa Yakubova, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: A Publicly Available Raw k-Space and DICOM Dataset of Knee Images for Accelerated MR Image Reconstruction Using Machine Learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020. doi: 10.1148/ryai.2020190007. PMID: 32076662.

[76] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[77] Kang Lin and Reinhard Heckel. Vision transformers enable fast and robust accelerated mri. In *International Conference on Medical Imaging with Deep Learning*, pages 774–795. PMLR, 2022.

[78] John G. Saw, Mark C. K. Yang, and Tse Chin Mo. Chebyshev Inequality with Estimated Mean and Variance. *The American Statistician*, 38(2):130–132, 1984. ISSN 00031305.

[79] Santiago Aja-Fernández and Gonzalo Vegas-Sánchez-Ferrero. Statistical analysis of noise in MRI. *Switzerland: Springer International Publishing*, 2016.

[80] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[81] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statist. Sinica*, pages 35–67, 2016.

[82] Christopher Kennedy and Rachel Ward. Greedy variance estimation for the LASSO. *Appl. Math. Optim.*, 82(3):1161–1182, 2020.

[83] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012.

[84] Lee H Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2): 269–284, 2014.

[85] X Liu, S Zheng, and X Feng. Estimation of error variance via ridge regression. *Biometrika*, 107 (2):481–488, 2020.

[86] Guo Yu and Jacob Bien. Estimating the error variance in a high-dimensional linear model. *Biometrika*, 106(3):533–546, 2019.

[87] Bartolomeo Stellato, Bart P. G. van Parys, and Paul J. Goulart. Multivariate Chebyshev Inequality With Estimated Mean and Variance. *The American Statistician*, 71(2):123–127, 2017. ISSN 0003-1305. doi: 10.1080/00031305.2016.1186559.

[88] Esa Ollila, David E. Tyler, Visa Koivunen, and H. Vincent Poor. Complex Elliptically Symmetric Distributions: Survey, New Results and Applications. *IEEE Transactions on Signal Processing*, 60(11):5597–5625, 2012. doi: 10.1109/TSP.2012.2212433.

[89] S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical programming computation*, 3(3):165–218, 2011.

Supplementary material to the paper *Non-Asymptotic Uncertainty Quantification in High-Dimensional Learning*.

In this supplement to the paper, we present in Section A a detailed discussion about aspects of the main result that are not mentioned in the main body of the paper. Moreover, Section B presents the proof Theorem 2 and Theorem 3. The former establishes data-driven confidence intervals, while the latter assumes the remainder component to be approximated by a Gaussian distribution. Appendix C constructs confidence intervals similar to Theorem 3 under the assumption of a heavy-tailed distribution. In Section D, we confirm our theoretical findings with several numerical experiments for classical high-dimensional regression as well as model-based neural networks. In Section E, we visualize the approximate Gaussian distribution of the remainder terms, demonstrating the applicability of Theorem 3 in relevant settings.

A Further Discussion of Main Result

In this appendix, we provide an in-depth discussion of various technical aspects and practical considerations regarding our main theoretical results, Theorem 2, including analysis of confidence interval radii, probabilistic interpretations, and implementation-oriented relaxations of the theoretical assumptions.

Length of radius: To minimize the length of the radius, $\gamma_j \in \left(0, 1 - \frac{1}{l\alpha}\right)$ should be chosen as the minimizer of the problem

$$\min_{\gamma_j \in \left(0, 1-\frac{1}{l\alpha}\right)} \frac{\sigma (M\hat{\Sigma}M^*)^{1/2}_{jj}}{\sqrt{m}} \sqrt{\log\left(\frac{1}{\gamma_j \alpha}\right)} + \sqrt{\frac{l^2-1}{l^2(1-\gamma_j)\alpha - l}} \cdot (\hat{\sigma}_R)_j, \tag{7}$$

In order to minimize over a large set for a given significance level α, a large number of data l is needed. For fixed estimates $\hat{S}_j$ and $(\hat{\sigma}^2_R)_j$, more data leads to a potentially smaller confidence interval length. If we assume $R_j = 0$, it follows that $\hat{S}_j = 0$ and $(\hat{\sigma}^2_R)_j = 0$. Then, $\gamma = 1$ is a valid choice, for which the function $\frac{\sigma(M\hat{\Sigma}M^*)^{1/2}_{jj}}{\sqrt{m}} \sqrt{\log\left(\frac{1}{\gamma_j\alpha}\right)}$ is well-defined and is minimized. In this case, the radius coincides with the asymptotic radius derived in [9, 10, 15] (except for that these works handle the real case) and the ones in [12] with $M = I_{N\times N}$. In this sense, the asymptotic confidence intervals can be seen as a special case of the proposed method.

The significance level α depends on γ_j and l to assure $c_l(\,\cdot\,)$ to be well-defined. For a large dataset $x^{(1)}, \dots, x^{(l)}$, i.e. if l is large, then it holds that $\lim_{l\to\infty} c_l(\alpha) = \lim_{l\to\infty} \sqrt{\frac{1-\frac{1}{l^2}}{(1-\gamma_j)\alpha - \frac{1}{l}}} = \frac{1}{\sqrt{(1-\gamma_j)\alpha}}$.

Probabilistic discussion: The probability in (5) is over the randomness of the noise as well as $\mathbb{Q}$. The confidence circles $C_j(\alpha)$ consist of two random variables, the debiased estimator $\hat{x}^u_j$ and the radius $r_j(\alpha)$. The former depends on the random noise and potentially on training data, while the latter depends on the estimators $\hat{S}_j$ and $\hat{\sigma}_R$, which in turn depend on both the noise and the data $x^{(1)}, \dots, x^{(l)}$.

A crucial requirement of applying the empirical version of Chebyshev's inequality [78] is the independence and identical distribution of the variables $|R^{(1)}_j|, \dots, |R^{(l)}_j|$. Therefore, it is essential that the estimator function $\hat{X}$ is independent of the data $x^{(1)}, \dots, x^{(l)}$. To achieve this, we train the estimator function $\hat{X}$ using a dataset that is independent of the data $x^{(1)}, \dots, x^{(l)}$, used for estimating $R^{(1)}, \dots, R^{(1)}$. However, the mean and variance of $|R^{(1)}|$ and hence of $|R^{(i)}|$ depend on the variance of the noise $\varepsilon^{(1)}$, i.e. σ^2. Thus, different noise levels σ require a new estimation of the mean and variance of $|R^{(1)}|$. Throughout this paper, we assume the noise level to be fixed and known. The latter assumption is motivated by two factors. First, the size of the confidence intervals relies on σ. Given that the primary focus of this paper is to determine the size of the confidence intervals based on the remainder term $R^{(1)}$, we seek to mitigate other influencing factors such as noise level estimation. Second, in domains like medical imaging, there is substantial knowledge about the noise level. For instance, the noise level in MRI can be directly measured from the scanner [79]. If the noise level is unknown, there are methods to estimate it. In the debiased LASSO literature, the most used method is the scaled LASSO [80]. Other methods for sparse regression, either in the LASSO context or more general for high-dimensional models, are [81–86].

Relaxation of assumptions in practice: In practice, it is often the case, that $|R_1^{(i)}|, \dots, |R_N^{(i)}|$ are identical distributed resulting in $\mu_1 = \dots = \mu_N$ and $(\sigma_R^2)_1 = \dots = (\sigma_R^2)_N$. Although the proof requires independence of the $|R_j^{(i)}|$, there are cases when it might suffice to relax this assumption by estimating the mean and variance pixel-wise uniformly, i.e.,

$$\hat{S} = \frac{1}{l \cdot N} \sum_{i=1}^{l} \sum_{j=1}^{N} R_j^{(i)} \qquad \text{and} \qquad \hat{\sigma}_R^2 = \frac{1}{l \cdot N - 1} \sum_{i=1}^{l} \sum_{j=1}^{N} (R_j^{(i)} - \hat{S})^2. \tag{8}$$

In addition to saving computational resources, accuracy improves due to the higher number of samples. Furthermore, instead of solving the optimization problem (7) for every $j \in \{1, \dots, N\}$, it might be a good idea to choose $\gamma_1 = \dots = \gamma_N$ as the minimizer of

$$\min_{\gamma_j \in \left(0, 1 - \frac{1}{l\alpha}\right)} \frac{\sigma \sum_{j=1}^{N} (M\hat{\Sigma}M^*)_{jj}^{1/2}}{\sqrt{m}N} \sqrt{\log\left(\frac{1}{\gamma_j \alpha}\right)} + c_l\left((1-\gamma_j)\alpha\right) \cdot \frac{1}{N} \sum_{j=1}^{N} (\hat{\sigma}_R)_j. \tag{9}$$

Then, one γ can be used for computing the potentially different radii $r_j(\alpha)$.

Type of uncertainty: The decomposition of the estimation error into a Gaussian term and a remainder term allows for handling both types of uncertainty, aleatoric and epistemic, almost separately. With the Gaussian term W, we quantify the aleatoric uncertainty from the inherent measurement noise. The remainder term R handles the epistemic uncertainty, which we quantify using a purely data-driven approach since for two different backward models (e.g., two different neural networks), it can be used to compare the estimation error of both models with respect to the ground truth. In this sense, our technique is rather a more general inferential uncertainty method.

B Proofs

Proof of Theorem 2. The statement $x_j^{(l+1)} \in C_j(\alpha)$ is equivalent to $|(\hat{x}^u)_j^{(l+1)} - x_j^{(l+1)}| \leq r_j(\alpha)$. To prove (5), we show that

$$\mathbb{P}\left(|(\hat{x}^u)_j^{(l+1)} - x_j^{(l+1)}| \geq r_j(\alpha)\right) \leq \alpha$$

In the next step, we write the radius $r(\alpha)$ as the sum $r(\alpha) = r^W(\alpha) + r^R(\alpha)$. According to the decomposition $(\hat{x}^u)_j^{(l+1)} - x_j^{(l+1)} = W_j + R_j$ we obtain for fixed $j \in \{1, \dots, N\}$

$$\begin{aligned} &\mathbb{P}\left(|(\hat{x}_j^u)^{(l+1)} - x_j^{(l+1)}| \geq r_j^W(\alpha) + r_j^R(\alpha)\right) = \mathbb{P}\left(|W_j + R_j| \geq r_j^W(\alpha) + r_j^R(\alpha)\right) \\ \leq &\mathbb{P}\left(|W_j| + |R_j| \geq r_j^W(\alpha) + r_j^R(\alpha)\right) \leq \mathbb{P}\left(|W_j| \geq r_j^W(\alpha)\right) + \mathbb{P}\left(|R_j| \geq r_j^R(\alpha)\right) \end{aligned}$$

where the last step follows from the pigeonhole principle. To estimate the first summand, we set $r_j^W(\alpha) := \frac{\sigma(M\hat{\Sigma}M^*)_{jj}^{1/2}}{\sqrt{m}} \sqrt{\log\left(\frac{1}{\gamma_j\alpha}\right)}$. Since $|W_j| \sim \text{Rice}\left(0, \frac{\sigma(M\hat{\Sigma}M^*)_{jj}^{1/2}}{\sqrt{2m}}\right)$ we obtain

$$\begin{aligned} \mathbb{P}\left(|W_j| \geq r_j^W(\alpha)\right) &= \frac{2m}{\sigma^2 \hat{\Sigma}_{jj}} \int_{r_j^W(\alpha)}^{\infty} x \exp\left(-\frac{x^2 m}{\sigma^2 (M\hat{\Sigma}M^*)_{jj}}\right) dx = \int_{\frac{(r_j^W(\alpha))^2 m}{\sigma^2(M\hat{\Sigma}M^*)_{jj}}} \exp(-u) du \\ &= \exp\left(-\frac{(r_j^W(\alpha))^2 m}{\sigma^2 (M\hat{\Sigma}M^*)_{jj}}\right) = \exp\left(-\log(1/\gamma_j\alpha)\right) = \gamma_j \alpha. \end{aligned}$$

For estimating the term $\mathbb{P}\left(|R_j| \geq r_j^R(\alpha)\right)$, we set $r_j^R(\alpha) = c_l(\alpha) \cdot (\hat{\sigma}_R)_j + \hat{S}_j$. This choice leads to

$$\begin{aligned} \mathbb{P}\left(|R_j| \geq r_j^R(\alpha)\right) &= \mathbb{P}\left(|R_j| - \hat{S}_j \geq r_j^R(\alpha) - \hat{S}_j\right) \leq \mathbb{P}\left(\left||R_j| - \hat{S}_j\right| \geq r_j^R(\alpha) - \hat{S}_j\right) \\ &= \mathbb{P}\left(\frac{\left||R_j| - \hat{S}_j\right|}{(\hat{\sigma}_R)_j} \geq \frac{r_j^R(\alpha) - \hat{S}_j}{(\hat{\sigma}_R)_j}\right) = \mathbb{P}\left(\frac{\left||R_j| - \hat{S}_j\right|}{(\hat{\sigma}_R)_j} \geq c_l(\alpha)\right). \end{aligned}$$

Now, we apply an empirical version of Chebyshev's inequality [78, 87]. This leads to

$$\begin{aligned}\mathbb{P}\left(\frac{\left||R_j| - \hat{S}_j\right|}{(\hat{\sigma}_R)_j} \geq c_l(\alpha)\right) &\leq \min\left\{1, \frac{1}{l+1}\left\lfloor\frac{(l+1)(l^2-1+lc_l(\alpha)^2)}{l^2c_l(\alpha)^2}\right\rfloor\right\} \\ &\leq \min\left\{1, \frac{l^2-1+lc_l(\alpha)^2}{l^2c_l(\alpha)^2}\right\} = \min\left\{1, \frac{l^2-1+\frac{l^2-1}{l(1-\gamma_j)\alpha-1}}{\frac{l(l^2-1)}{l(1-\gamma_j)\alpha-1}}\right\} \\ &= \min\left\{1, \frac{1+\frac{1}{l(1-\gamma_j)\alpha-1}}{\frac{l}{l(1-\gamma_j)\alpha-1}}\right\} = \min\left\{1, (1-\gamma_j)\alpha\right\} = (1-\gamma_j)\alpha,\end{aligned}$$

where we used in the last step, that $(1-\gamma_j)\alpha < \alpha < 1$. To summarize,

$$\begin{aligned}\mathbb{P}\left(|(\hat{x}^u)_j^{(l+1)} - x_j^{(l+1)}| \geq r_j(\alpha)\right) &\leq \mathbb{P}\left(|W_j| \geq r_j^W(\alpha)\right) + \mathbb{P}\left(|R_j| \geq r_j^R(\alpha)\right) \\ &\leq \gamma_j\alpha + (1-\gamma_j)\alpha = \alpha.\end{aligned}$$

□

Proof of Theorem 3. Since $W \sim \mathcal{CN}(0, \frac{\sigma^2}{m}M\hat{\Sigma}M^*)$ and $R \sim \mathcal{CN}(0, \frac{1}{m}\Sigma_R)$ the estimation error $\hat{x}^u - x^0 = W + R$ follows again a multivariate normal distribution with zero mean and covariance matrix $\frac{1}{m}(\sigma^2M\hat{\Sigma}M^* + \Sigma_R)$. By exploiting the Gaussian distribution, we obtain

$$\begin{aligned}\mathbb{P}\left(|W_j + R_j| > r_j^G(\alpha)\right) &= \frac{2m}{\sigma^2(M\hat{\Sigma}M^*)_{jj} + (\Sigma_R)_{jj}} \int_{r_j^G(\alpha)}^{\infty} x \exp\left(-\frac{x^2m}{\sigma^2(M\hat{\Sigma}M^*)_{jj} + (\Sigma_R)_{jj}}\right) dx \\ &= \int_{\frac{r_j^G(\alpha)^2m}{\sigma^2(M\hat{\Sigma}M^*)_{jj}+(\Sigma_R)_{jj}}} \exp(-u)du = \exp\left(-\frac{r_j^G(\alpha)^2m}{\sigma^2(M\hat{\Sigma}M^*)_{jj} + (\Sigma_R)_{jj}}\right)\end{aligned}$$

Thus, we have

$$\mathbb{P}(|\hat{x}_j^u - x_j^*| > r_j^G(\alpha)) \leq \exp\left(-\frac{r_j^G(\alpha)^2m}{\sigma^2(M\hat{\Sigma}M^*)_{jj} + (\Sigma_R)_{jj}}\right),$$

which needs to be equal to $\alpha > 0$. Therefore,

$$r_j^G(\alpha) = \frac{(\sigma(M\hat{\Sigma}M_{jj} + (\Sigma_R)_{jj})^{1/2}}{\sqrt{m}}\sqrt{\log\left(\frac{1}{\alpha}\right)}.$$

□

C Heavy-Tailed Version of Theorem 3

The following version of Theorem 3 gives an example of how to derive the radii of the confidence intervals in the case of a known, heavy-tailed distribution of the remainder term.

Theorem 4. *Let $\hat{x}^u \in \mathbb{C}^N$ be a debiased estimator for $x \in \mathbb{C}^N$ with a remainder term following a complex t-distribution with a degree of freedom $\nu > 2$, i.e. $R \sim \mathcal{C}t_\nu(0, \Sigma_R)$. Then, $C_j(\alpha) = \{z \in \mathbb{C} \mid |z - \hat{x}_j^u| \leq r_j(\alpha)\}$ with radius*

$$r_j(\alpha) = \frac{\sigma(M\hat{\Sigma}M^*)_{jj}^{1/2}}{\sqrt{m}}\sqrt{\log\left(\frac{1}{\gamma_j\alpha}\right)} + \sqrt{\frac{(\Sigma_R)_{jj}\nu}{2}}\sqrt{(1-\gamma_j)^{-2/\nu}\alpha^{-2\nu} - 1}.$$

is valid, i.e. $\mathbb{P}\left(x_j \in C_j(\alpha)\right) \geq 1 - \alpha$.

Proof. We can bound the estimation error $|\hat{x}^u_j - x_j|$ analogously to the proof of Theorem 2 by

$$\mathbb{P}(|\hat{x}^u_j - x_j| \geq r_j(\alpha)) = \mathbb{P}(|W_j + R_j| > r_j(\alpha)) \leq \mathbb{P}(|W_j| > r^W_j(\alpha)) + \mathbb{P}(|R_j| > r^R_j(\alpha))$$

The distribution of $\mathbb{P}(|W_j| > r^W_j(\alpha))$ is determined by the Gaussian noise and can be computed as

$$r^W_j(\alpha) = \frac{\sigma(M\hat{\Sigma}M^*)^{1/2}_{jj}}{\sqrt{m}}\sqrt{\log\left(\frac{1}{\gamma_j\alpha}\right)}$$

similarly to Theorem 2. If $R \sim \mathcal{C}t_\nu(0, \Sigma_R)$, then the marginal distribution R_j is also complex t-distributed, i.e. $R_j \sim (0, (\Sigma_R)_{jj})$ (see [88]). Moreover, the probability density function of $|R_j|$ is

$$f(r) = \frac{2r}{(\Sigma_R)_{jj}}\left(1 + \frac{2r^2}{\nu(\Sigma_R)_{jj}}\right)^{-(\nu/2+1)}.$$

Hence,

$$\mathbb{P}(|R_j| > r^R_j(\alpha)) = \int\limits_{r_j(\alpha)}^{\infty} f(r)dr = \left[-\frac{1}{\left(\frac{2r^2}{(\Sigma_R)_{jj}\nu} + 1\right)^{\nu/2}}\right]^{\infty}_{r_j(\alpha)} = \frac{1}{\left(\frac{2r^2}{(\Sigma_R)_{jj}\nu} + 1\right)^{\nu/2}}.$$

Setting this probability equal to $(1-\gamma_j)\alpha$ requires

$$r^R_j(\alpha) = \sqrt{\frac{(\Sigma_R)_{jj}\nu}{2}}\sqrt{(1-\gamma_j)^{-2/\nu}\alpha^{-2\nu} - 1}.$$

Since we assumed $\nu > 2$ the inequality $(1-\gamma_j)^{-2/\nu}\alpha^{-2\nu} > 1$ holds. Combining $r_j(\alpha) = r^W_j(\alpha) + r^R_j(\alpha)$ concludes the proof. □

D Further Numerical Evaluation

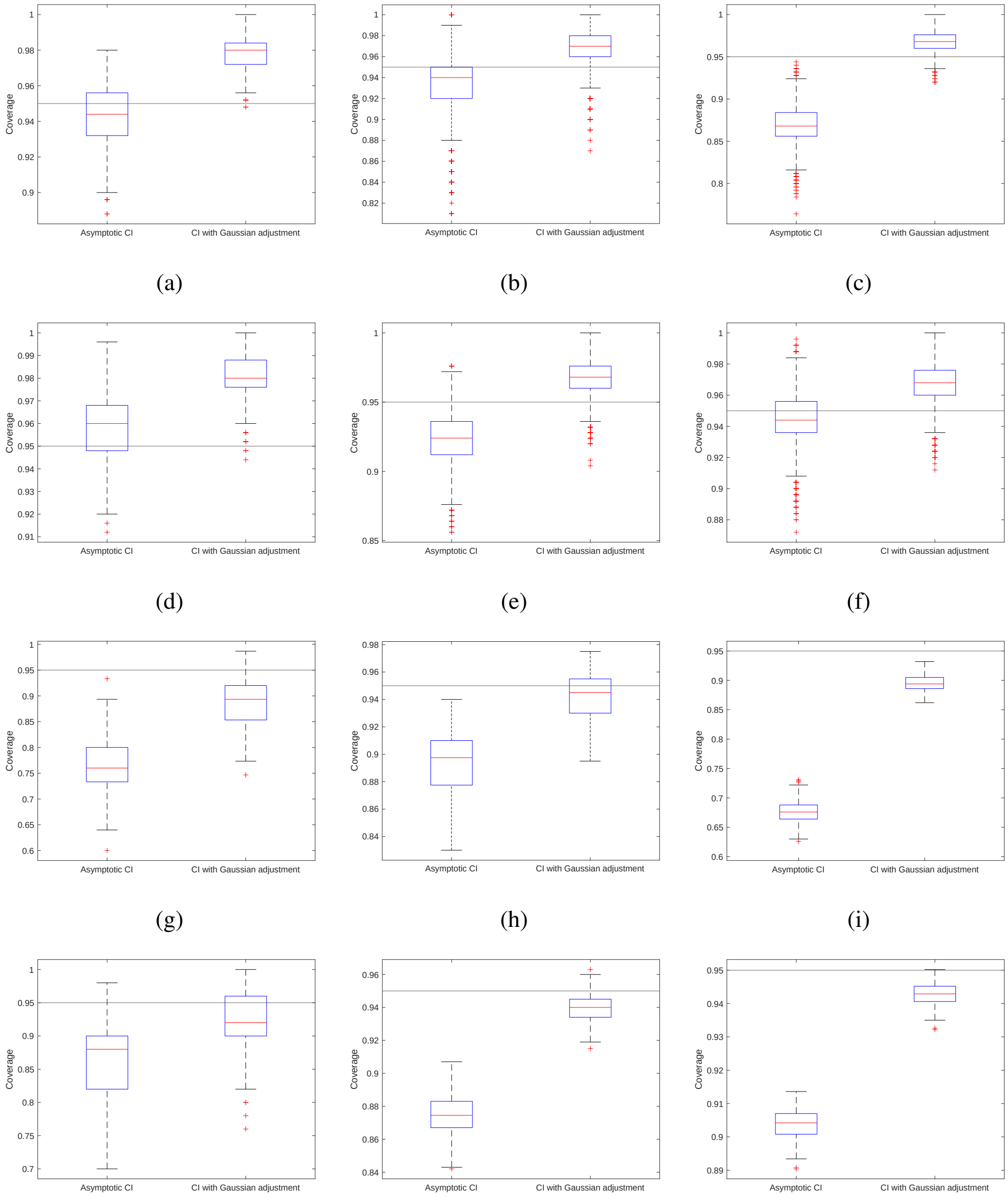
To confirm our theoretical findings claiming that the incorporation of the bias component renders the confidence intervals more robust, we present additional numerical experiments here.

UQ for Classical Model-Based Regression For the experiments described here, we use Tfocs [89]. Analogous to the experiment described in Section 5.1, we run further experiments in the classical sparse regression setting when the measurement matrix is a Gaussian and subsampled Fourier matrix. The different settings including the results can be found in Table 1. The results show that the Gaussian adjustment of our proposed method significantly increases the hit rates, especially on the support, while moderately increasing the confidence interval length. Our data-driven adjustment achieves even better hit rates, but the confidence intervals are larger. Although in well-posed settings, like the second column of Table 1, the hit rates $h^W(0.05)$ based on asymptotic confidence intervals lead *overall* to almost 95%, however on the *support*, which are the crucial features, the asymptotic hit rates fail. In particular, our corrections are essential in ill-posed regression problems as the third Gaussian column. The hit rates for the asymptotic CIs and the corrected ones with Gaussian adjustment are visualized in more detail in Figure 4.

UQ for MRI Reconstruction with Neural Networks In this section, we present more experiments for UQ for MRI reconstruction with neural networks. Our experimental settings, as well as our code for this experiment, are based on the paper and code [3] by [18]. The dataset used for conducting the experiments is the fastMRI single-coil knee dataset. For documentation, see [74, 75].

Table 2 represents the results obtained by learning the reconstruction function $\hat{X}$ using the It-net with 8 layers, with $60\%, 40\%$ and 30% radial undersampling and for noise levels obtained by adding complex gaussian noise with standard deviation $\sigma = 60$ and $\sigma = 84$, respectively. Similarly, Table 3 shows the results obtained by the U-Net. In Figure 6, the asymptotic hit rates and the Gaussian adjusted ones for the 95% confidence level are compared in a box plot for each experiment. Note that for these experiments, we construct only the confidence intervals for the real part of the image, as the

[3] https://github.com/jmaces/robust-nets



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

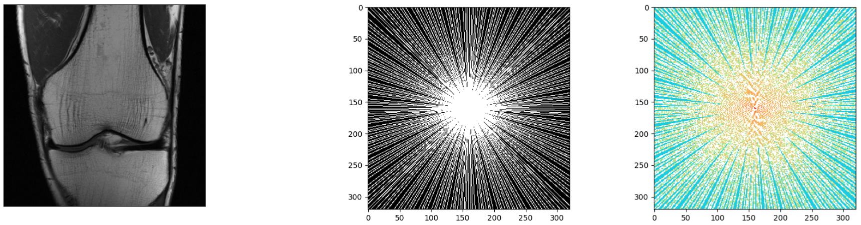
Figure 4: Box plots for hit rates of sparse regression experiments. The settings are those described in Table 1. The first row presents the hit rates over all components and the second the hit rates of the support, e.g., 4a and 4g correspond to the first column of the table, 4b and 4h to the second one and so forth. In each plot the left box represents the asymptotic hit rates, the right one the Gaussian adjusted hit rates. The horizontal line marks the desired 95%.

fastMRI dataset provides only real-valued data. Due to this artifact, the investigation of the imaginary part's distribution is not possible, and the Gaussian-adjusted confidence intervals are not applicable to the magnitude. For this reason, and to ensure a valid comparison of the three confidence interval types, we decided to investigate the confidence intervals for the real part, even in the asymptotic and data-driven setting.

All It-Net and U-Nets are trained with a combination of the MS-SSIM-loss [76], the ℓ_1-loss and the Adam optimizer with a learning rate of $5e^{-5}$, epsilon of $1e^{-4}$ and weight decay parameter $1e^{-5}$.

type	Gaussian			Fourier		
feature dimension	1000	10000		1000	10000	100000
undersampling	50%	40%	60%	40%	60%	50%
sparsity	7.5%	2%	10%	5%	10%	5%
relative noise	15%	10%	20%	15%	5%	10%
R/W ℓ_2-ratio	0.7062	0.5117	0.9993	0.5446	0.5924	0.4701
R/W ℓ_∞-ratio	0.8191	0.5161	1.1581	0.5752	0.6088	0.4794
average asympt. radius $r^W(0.05)$	0.0116	0.0027	0.0049	0.0130	0.0011	0.0008
av. radius $r^G(0.05)$ (Thm.3)	0.0142	0.0031	0.0069	0.0148	0.0013	0.0009
av. radius $r(0.05)$ (Thm.2)	0.0304	0.0060	0.0155	0.0295	0.0026	0.0016
$h^W(0.05)$	0.9437	0.9353	0.8692	0.9582	0.9246	0.9444
$h^G(0.05)$	0.9787	0.9684	0.9691	0.9799	0.9665	0.9687
$h(0.05)$	1	1	1	1	1	0.9999
$h_S^W(0.05)$	0.7661	0.8941	0.6783	0.8629	0.8745	0.9041
$h_S^G(0.05)$	0.8852	0.9421	0.8948	0.9226	0.9396	0.9425
$h_S(0.05)$	0.9999	1	1	0.9999	1	0.9998

Table 1: Experiments for sparse regression for Gaussian and Fourier matrix. Every experiment uses 500 estimation and 250 evaluation data.



(a) (b) (c)

Figure 5: Knee MRI groundtruth image from the fastMRI dataset 5a [74, 75], radial sampling mask 5b and undersampled k-space data 5c.

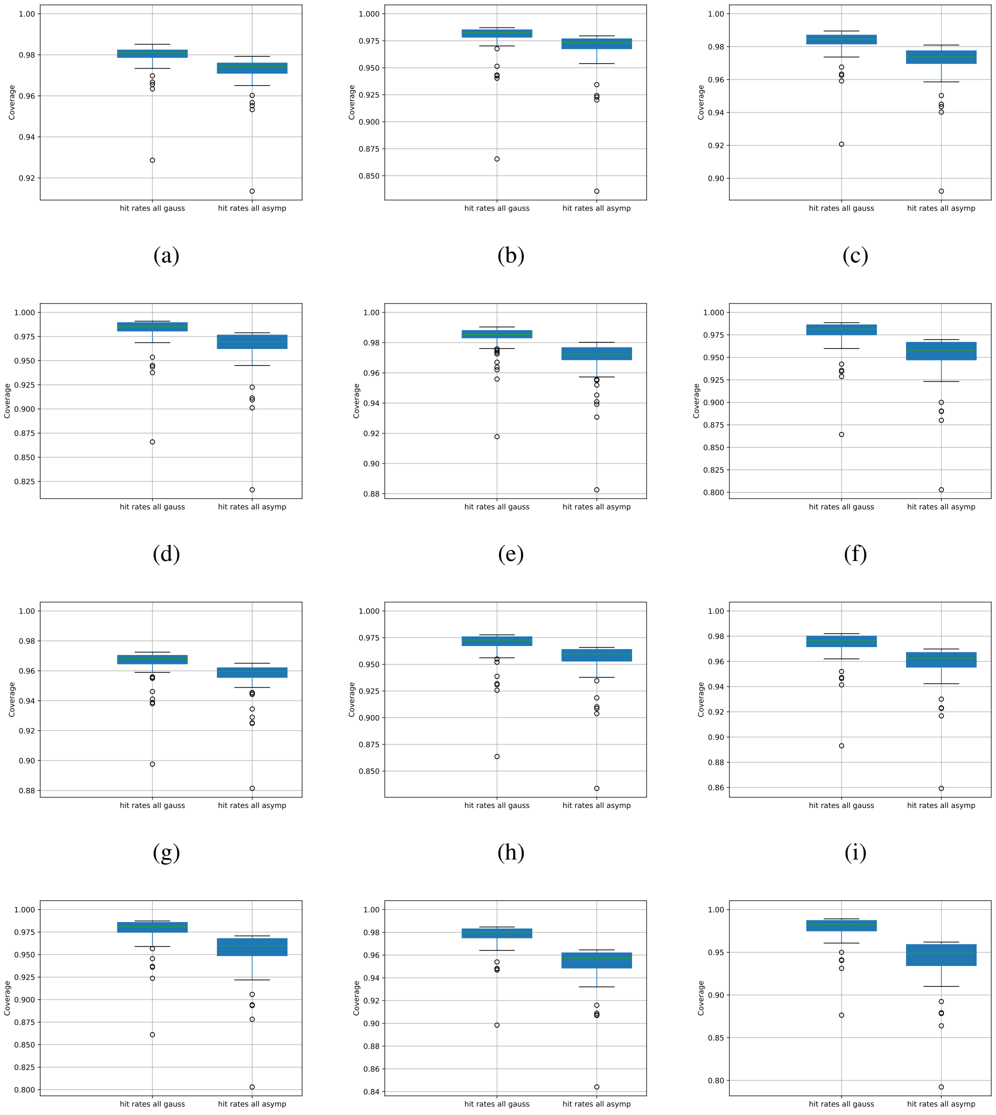
The It-Nets were trained for 15 epochs, and the U-Nets were trained for 20 epochs, both with batch size 40. Every U-Net has 2 input and output channels, 24 base channels, and encodes the image to a size of 20×20 and at most 384 channels. The It-Net employs the U-Net in each layer as a residual network and has a data consistency part around each U-Net in every layer.

Comparing the tables, the It-Net has, in general, better hit rates as well as a better R/W ratio than the U-Net due to its more accurate reconstruction. Further, the hit rates for all the pixels are higher than those obtained only for the support. For achieving reliable results for safety-critical applications, obtaining hit rates higher than the confidence level is crucial, especially on the support. Otherwise, one might achieve a certain confidence level overall but cannot trust the pixels of interest.

The experiments were conducted using Pytorch 1.9 on a desktop with AMD EPYC 7F52 16-Core CPUs and NVIDIA A100 PCIe030 030 40GB GPUs. The code for the experiments can be found in the supplementary material. The execution time of the code is around 5 hours for each It-Net, 2 hours for each U-Net, and around 30 minutes for the rest of each experiment. So, in total, this gives us a time of 48 hours for the MRI reconstruction experiments. The execution time for the classical model-based regression experiments takes 5 to 30 minutes each; therefore, in total it is less than 3 hours.

E Distribution Visualization of Remainder Term

In Figure 7, we present a series of histograms illustrating the empirical distribution of the remainder term's real part across all experimental settings in sparse regression, conducted in this



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 6: Box plots for hit rates of neural network experiments. The settings are those described in Table 2 and Table 3. The first row presents the hit rates for the different It-nets for 95% confident intervals and the second the hit rates for the U-Nets for 95% confident intervals, e.g., 6a and 6g correspond to the first columns of the tables. In each plot the left box represents the asymptotic hit rates, the right the Gaussian adjusted ones.

paper. These histograms provide evidence that the remainder term can be approximated by a Gaussian distribution, with the approximation becoming increasingly precise as the dimensionality increases. Across low-dimensional scenarios, the empirical distributions exhibit some deviations from the Gaussian form, but these discrepancies diminish as the dimensionality grows larger. In high-dimensional regimes, the empirical distributions demonstrate an exceptional degree of convergence to the Gaussian approximation. This close alignment lends strong support to the validity of the key assumption of Theorem 3, allowing a Gaussian adjustment to the confidence intervals.

In Figure 8 we present a series of histograms representing the empirical distribution of the remainder term's real part for the six different experimental settings, for the U-Net, conducted in this paper.

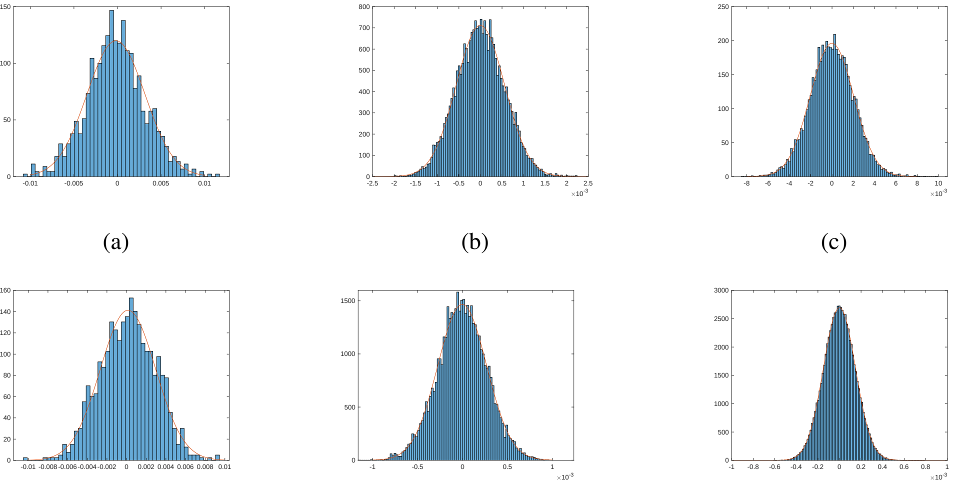
undersampling	60%		40%		30%	
noise level	15%	10%	12%	8%	10%	7%
R/W ℓ_2-ratio	0.3558	0.3800	0.6332	0.6922	0.8759	0.5759
R/W ℓ_∞-ratio	0.3842	0.4929	0.6268	0.6924	1.0614	0.6282
$h(0.05)$	1.0	1.0	1.0	1.0	1.0	1.0
$h^G(0.05)$	0.9797	0.9791	0.9826	0.9818	0.9831	0.9770
$h^W(0.05)$	0.9726	0.9689	0.9714	0.9649	0.9690	0.9518
$h(0.1)$	0.9998	0.9998	0.9999	0.9999	1.0	0.9999
$h^G(0.1)$	0.9486	0.9485	0.9544	0.9539	0.9553	0.9448
$h^W(0.1)$	0.9359	0.9306	0.9338	0.9241	0.9299	0.9033
$h_S(0.05)$	1.0	0.9999	1.0	1.0	1.0	1.0
$h_S^G(0.05)$	0.9739	0.9711	0.9789	0.9729	0.9798	0.9707
$h_S^W(0.05)$	0.9656	0.9589	0.9663	0.9520	0.9642	0.9421
$h_S(0.1)$	0.9997	0.9995	0.9999	0.9998	0.9999	0.9998
$h_S^G(0.1)$	0.9385	0.9353	0.9476	0.9392	0.9493	0.9347
$h_S^W(0.1)$	0.9246	0.9152	0.9257	0.9053	0.9225	0.8906

Table 2: Experiments for It-Net with 8 iterations. Results of hit rates averaged over $k = 100$ samples.

undersampling	60%		40%		30%	
noise level	15%	10%	12%	8%	10%	7%
R/W ℓ_2-ratio	0.2747	0.3976	0.6182	0.6671	1.2641	1.3399
R/W ℓ_∞-ratio	0.3923	0.4292	0.5681	0.7082	1.1668	1.2501
$h(0.05)$	0.9999	1.0	1.0	1.0	1.0	1.0
$h^G(0.05)$	0.9657	0.9688	0.9737	0.9767	0.9768	0.978
$h^W(0.05)$	0.9569	0.9552	0.9583	0.9535	0.9517	0.9418
$h(0.1)$	0.9993	0.9997	0.9998	0.9999	1.0	1.0
$h^G(0.1)$	0.9238	0.9296	0.9377	0.9443	0.9432	0.9465
$h^W(0.1)$	0.9098	0.9077	0.9125	0.9060	0.9022	0.8886
$h_S(0.05)$	0.9999	0.9999	1.0	1.0	1.0	1.0
$h_S^G(0.05)$	0.9679	0.9677	0.975	0.9727	0.9779	0.9757
$h_S^W(0.05)$	0.9594	0.9541	0.9602	0.9483	0.9542	0.9398
$h_S(0.1)$	0.9994	0.9996	0.9999	0.9998	1.0	0.9999
$h_S^G(0.1)$	0.9276	0.9282	0.9401	0.9388	0.946	0.9446
$h_S^W(0.1)$	0.9141	0.9064	0.9156	0.9002	0.9061	0.8876

Table 3: Experiments for U-Nets. Results of hit rates averaged over $k = 100$ samples.

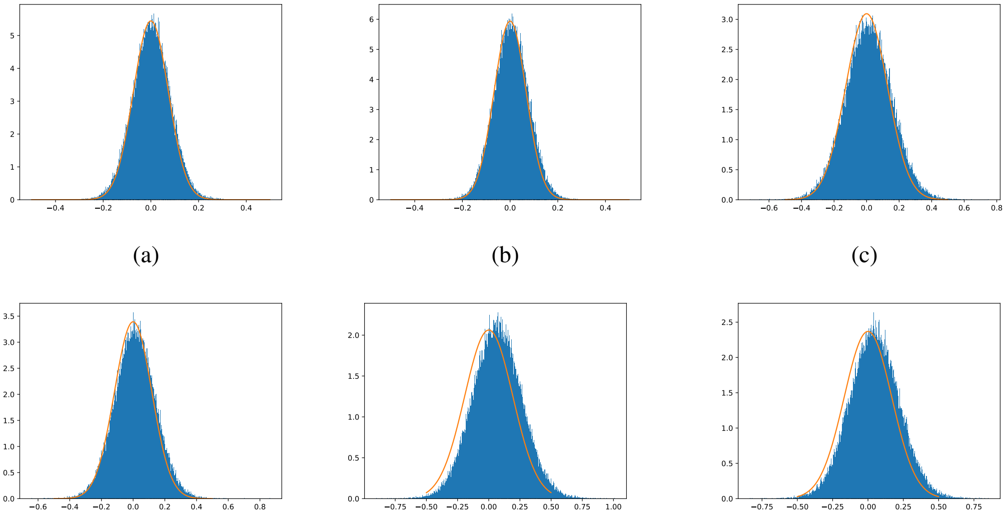
Figure 9 represents the histograms for the It-Net experiments. In most scenarios, the real part of the remainder term is Gaussian distributed with mean 0. The only exceptions are Figures 8e and 8f, which correspond to the U-Net experiments with 30% undersampling.



(a) (b) (c)

(d) (e) (f)

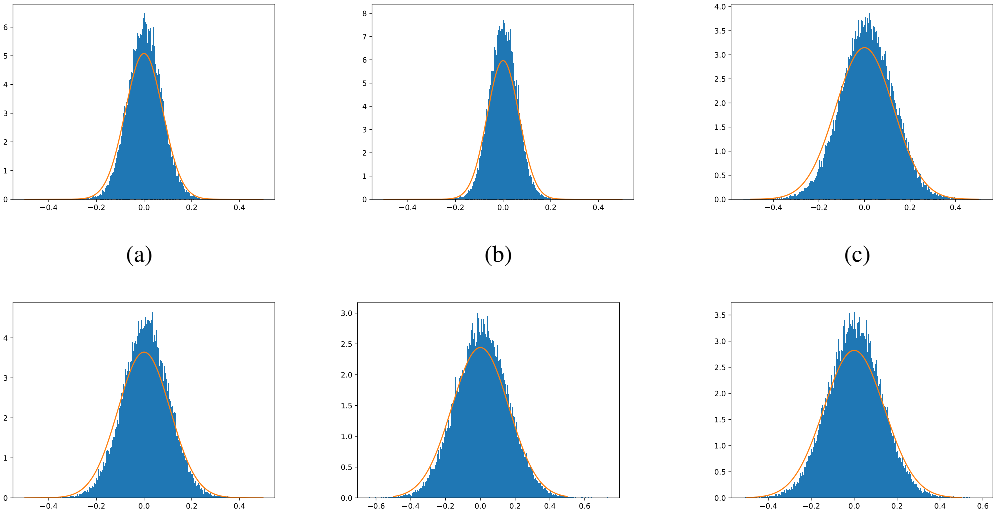
Figure 7: Histograms showing the empirical distribution of the remainder component. The real part of R is plotted as a histogram and the red line corresponds to a Gaussian fit. One realization of the remainder term is visualized for each of the experiments described in Table 1. (7a corresponds to the first column, 7b to the second column and so on. The plots for the imaginary part look similar.)



(a) (b) (c)

(d) (e) (f)

Figure 8: Histograms showing the empirical distribution of the remainder component. The real part of R is plotted as a histogram and the red line corresponds to a Gaussian fit. One realization of the remainder term is visualized for each of the experiments described in Table 3. (8a corresponds to the first column, 8b to the second column and so on.)



(a) (b) (c)

(d) (e) (f)

Figure 9: Histograms showing the empirical distribution of the remainder component. The real part of R is plotted as a histogram and the red line corresponds to a Gaussian fit. One realization of the remainder term is visualized for each of the experiments described in Table 2. (9a corresponds to the first column, 9b to the second column and so on.)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper will develop a theory for non-asymptotic uncertainty quantification in high-dimensional learning and, as described in the abstract, we also apply such theory to neural networks and learning algorithms, bridging theory and practice.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this paper are discussed in Section 6. In particular, we discussed that larger remainder terms necessitate greater corrections, resulting in wider confidence intervals. Also, the accuracy of our method depends on the quality of the estimates for the mean and variance of the remainder term, which improves with more available data. Lastly, we left the sharpness of the provided results as a future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

https://doi.org/10.52202/079017-3894

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results contain proofs, in detail Theorem 2 and Theorem 3 are proven in Section B and Theorem 1 is an informal version of Theorem 2. All assumptions are made in the statements of the theorems, and all results the proofs rely on are referenced.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main information about the experimental results and how they were obtained can be found in Section 5, further information can be found in the Appendix in Section D. We will also provide a github containing the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

https://doi.org/10.52202/079017-3894

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release Github containing all code and explanation how to reproduce our results once the paper is published, for now we attached the code as supplemetary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main experimental settings can be found in Section 5, more details about the experimental settings can be found in Appendix D and in the code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, the corresponding boxplots and confidence intervals can be found in Figures 2, 1, 3, 4 and 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All this information is provided in the Appendix D. Further, the paper did not require more computation and experiments than what is presented in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we checked the code of ethics and our paper conforms with every point in the code, also we made sure to preserve anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

https://doi.org/10.52202/079017-3894

Answer: [Yes]

Justification: This paper is developing a rigorous framework for confidence intervals for high-dimensional inverse problems. We believe that this will improve the diagnostics with medical images, as discussed in Section 6. We do not believe that such techniques could have a negative societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Since the paper develops confidence intervals for machine learning problems with solid theoretical foundations, there is no risk of misuse and no safeguards are necessary.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We used the NYU fastMRI dataset [74, 75], which we cited according to their homepage. We used publicly available code belonging to the papers [63] and [89] which we also cited at the according places (Section 5 or Section D).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is our code which is provided with this paper as supplementary material and is well documented on its own and also explained in Section 5 and in the Appendix in Section D.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The experiments in the paper are based on the NYU fastMRI dataset [74, 75] which is a publicily available dataset.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

https://doi.org/10.52202/079017-3894

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This is a theoretical paper (but with a potentially huge impact in the field of medical images) but all the images used for conducting our experiments are publicly available (and anonymized) and there is no need to get Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3894