A Parameter Choice Rule for Tikhonov Regularization Based on Predictive Risk

Federico Benvenuto* Bangti Jin†

December 30, 2019

Abstract

In this work, we propose a new criterion for choosing the regularization parameter in Tikhonov regularization when the noise is white Gaussian. The criterion minimizes a lower bound of the predictive risk, when both data norm and noise variance are known, and the parameter choice involves minimizing a function whose solution depends only on the signal-to-noise ratio. Moreover, when neither noise variance nor data norm is given, we propose an iterative algorithm which alternates between a minimization step of finding the regularization parameter and an estimation step of estimating signal-to-noise ratio. Simulation studies on both small- and large-scale datasets suggest that the approach can provide very accurate and stable regularized inverse solutions and, for small sized samples, it outperforms discrepancy principle, balancing principle, unbiased predictive risk estimator, L-curve method generalized cross validation, and quasi-optimality criterion, and achieves excellent stability hitherto unavailable.

Keywords: Tikhonov regularization, regularization parameter, predictive risk optimization

1 Introduction

In this work, we study discrete linear inverse problems of recovering an unknown object from noisy indirect measurements. When the matrix is ill-conditioned, the solution given by the generalized inverse is generally unsatisfactory, especially in the presence of data noise. Then one often employs a regularization method, which provides a one-parameter family of candidate solutions and a criterion for selecting the optimal parameter (and a corresponding solution) [43, 7, 19]. In the case of white Gaussian noise, popular linear regularization methods include truncated SVD [13], Tikhonov regularization [43] and Landweber method [26], and they all require specifying one scalar parameter, i.e., regularization parameter, which is notoriously challenging. A large number of choice rules have been proposed in the literature. Most popular criteria include discrepancy principle (DP) [35], Unbiased Predictive Risk Estimator (UPRE) [40] and balancing principle (BP) / Lepskii's principle [27, 34, 2, 33] (see also [11] and references therein), when the noise level is given, and Generalized Cross Validation (GCV) [9, 46], L-Curve (LC) method [14, 16] (see also [39, 21] for an algebraic variant), quasi-optimality criterion (QOC) [44, 42] and Hanke-Raus rule [12], when the noise level is unknown. The heuristic rules (without requiring a knowledge of the noise level) can categorically grouped into the general context of choice rules based on functional minimization. See the review [10] for detailed experimental comparison and in-depth discussions for various linear regularization strategies; see also the work [23]. However, none of the aforementioned methods can consistently find a near-optimal regularization parameter for all test problems and noise realizations in simulation studies.

DP, UPRE and GCV have a common drawback in that they depend heavily on the sample realization and thus can lead to instable solutions. In particular, DP selects a regularized solution with a predictive error not smaller than the given noise level [35]. Thus, when the predictive error is larger (smaller) than

*Dipartimento di Matematica, Università degli Studi di Genova, Via Dodecaneso 35, 16146, Genova, Italy. benvenuto@dima.unige.it

†Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK. bangti.jin@gmail.com,b.jin@ucl.ac.uk

the true noise level or the noise level is very small (large), it leads to under-regularized (over-regularized) solutions, and the error of the approximate solutions may be arbitrarily large. Indeed, in practice, DP often includes an additional tuning parameter [7]. BP is theoretically better than DP [34, 10]: the convergence holds if the ratio of the actual and given noise is bounded (which may be larger than one), thereby avoiding the convergence issue of DP. GCV provides an asymptotically unbiased estimator of the predictive risk, but for small sized samples it can give unstable solutions [46, 31], due to the flatness of the GCV curve. Finally, UPRE can fail to approximate the predictive error when the data size is not large enough or the data norm is too small [5]. Further, the LC method still defies a complete convergence analysis even though it performs extremely well in many applications [45]. QOC is very prominent among heuristic rules: it is easy to implement, gives excellent empirical performance, and further, both convergence and convergence rates have also been established for a wide variety of noise models, including the infinite-dimensional setting [24, 3, 22, 23]. However, it fails almost surely for severely ill-posed problems with Gaussian white noise [25].

In this work we propose a novel method for choosing the crucial Tikhonov regularization parameter, which falls within the realm of rules based on functional minimization. The proposed strategy is inspired by the predictive risk, an idea that has gained increasing attention in recent years (see the works [30, 29] and references therein). However, instead of minimizing the predictive risk or its unbiased estimators directly, it minimizes a lower bound of the predictive risk, when both data norm and noise level are known. The lower bound is constructed only for the approximation error (a.k.a. bias) and does not change the noise amplification (a.k.a. variance). This construction ensures that the method chooses a regularization parameter that is an upper bound of the optimal parameter (with respect to the predictive risk). The minimizer of the lower bound depends only on the ratio between the data norm and noise level, i.e. signal-to-noise ratio (SNR), and the procedure is termed as predictive risk optimization (`PRO`). In practice, this lower bound is quite tight, and thus the choice rule enjoys excellent accuracy and stability. We discuss its utility under three different information availability scenarios: (i) SNR is known, (ii) the data noise level is known, but the true data norm is unknown and (iii) both noise level and data norm are unknown. In case (i), `PRO` can be applied directly. In case (ii), we employ an unbiased estimator of the data norm to efficiently approximate `PRO`. In case (iii), we propose an iterative algorithm, each iteration of which is composed of two steps: given estimated data norm and noise variance, one uses `PRO` to find the regularization parameter, and given the estimated regularization parameter, one re-estimates the data norm and noise variance. The procedure is termed as iterative `PRO` (or `I-PRO`). Theoretically, we prove a number of properties of the rule for Tikhonov regularization, e.g., convergence to zero as the noise level tends to zero, upper bound property and the monotone convergence of the `I-PRO` procedure. Further, to demonstrate their performance, we carry out extensive simulation studies on benchmark problems from two public software packages, i.e., Regularization tools [15] and AIR tools [17], which involve small-scale 1d applications and large-scale 2d tomography problems, respectively, and give a detailed comparative study with several popular existing choice rules.

The rest of the paper is organized as follows. In Section 2, we describe the setup of a discrete inverse problem in a Gaussian framework and introduce the `PRO` criterion. In Section 3, we describe optimization strategies for three different information availability scenarios. In Section 4, we investigate the `PRO` criterion for Tikhonov regularization, establish the well-posedness of the rule and bounds on the chosen parameter, analyze the convergence of the `I-PRO` scheme, and describe their efficient implementation for large-scale problems. In Section 5, we report simulation results and discuss the pros and cons of `PRO` and `I-PRO`, when compared with existing choice rules. Section 6 contains a summary of the work.

2 Background and motivation

Consider the linear inverse problem

$$\mathbf{g}^\dagger = A\mathbf{f}^\dagger, \tag{1}$$

where $\mathbf{f}^\dagger \in \mathbb{R}^m$, $\mathbf{g}^\dagger \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times m}$. The goal is to recover the true signal $\mathbf{f}^\dagger$ from a given noisy measurement of $\mathbf{g}^\dagger$, denoted by $\mathbf{g}^\eta$. Throughout, $\mathbf{g}^\eta := \mathbf{g}^\dagger + \boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is a Gaussian distributed random vector with $\mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{\eta}] = 0$, $\mathbb{E}_{\boldsymbol{\eta}}[\eta_i\eta_j] = \sigma^2\delta_{ij}$, with δ_{ij} denote the standard Kronecker symbol, where $\mathbb{E}_{\boldsymbol{\eta}}[\cdot]$ denotes taking expectation with respect to the distribution of $\boldsymbol{\eta}$. A linear regularization method R_α

provides a family of estimates of the solution

$$\mathbf{f}_\alpha^\eta = R_\alpha \mathbf{g}^\eta, \tag{2}$$

for any realization $\mathbf{g}^\eta$. Upon ignoring the usual scaling factor $1/n$, the predictive risk $p_\alpha(\mathbf{g}^\eta)$ is defined as

$$p_\alpha(\mathbf{g}^\eta) = \mathbb{E}_{\boldsymbol{\eta}}[\|\mathbf{g}_\alpha^\eta - \mathbf{g}^\dagger\|^2],$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{g}_\alpha^\eta = X_\alpha \mathbf{g}^\eta$ is the predictive data, where

$$X_\alpha = AR_\alpha \tag{3}$$

is the so-called influence matrix. Since the predicted data $\mathbf{g}_\alpha^\eta$ can be split into

$$\mathbf{g}_\alpha^\eta = X_\alpha \mathbf{g}^\dagger + X_\alpha \boldsymbol{\eta},$$

by the standard bias-variance dcomposition, the predictive risk $p_\alpha(\mathbf{g}^\eta)$ is given by

$$p_\alpha(\mathbf{g}^\eta) = \|(X_\alpha - I)\mathbf{g}^\dagger\|^2 + \sigma^2 \|X_\alpha\|_F^2, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. Equation (4) indicates the predictive risk $p_\alpha(\mathbf{g}^\eta)$ consists of two terms: the bias term $\|(X_\alpha - I)\mathbf{g}^\dagger\|^2$ due to the approximation error, which depends only the regularity of the exact solution $\mathbf{f}^\dagger$, and the variance term $\sigma^2\|X_\alpha\|_F^2$ due to noise amplification.

Note that the expression in (4) is not directly computable, since the exact data $\mathbf{g}^\dagger$ is unknown. However, this quantity may be estimated using sample data $\mathbf{g}^\eta$ to obtain various predictive risk estimators for parameter choice. This idea has long been pursued in the literature; see, e.g., [32, 41, 8, 38, 6, 28, 30, 29] and references therein for various practical applications and theoretical developments. For example, with $\mathrm{tr}(\cdot)$ denoting the trace of a matrix, the following two estimators are frequently adopted:

$$\hat{p}_\alpha(\mathbf{g}^\eta) = \|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\|^2 - 2\sigma^2 \mathrm{tr}(I - X_\alpha)$$

for UPRE and

$$\tilde{p}_\alpha(\mathbf{g}^\eta) = \frac{\|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\|^2}{\mathrm{tr}(I - X_\alpha)^2}$$

for GCV, when the noise level σ^2 is known and unknown, respectively. We refer to [28] for further approximations of the predictive risk in parameter choice. However, it is known that they tend to suffer from the notorious overfitting issue, i.e., the selected parameter tends to be too small [30], which is also confirmed by the simulation study in Section 5. See also the recent work [29] for the order optimality of risk estimators for a class of (ordered) filter based regularization methods, where the numerical challenges are also highlighted [29, Fig. 1].

GCV and UPRE estimate the minimum value of the predictive risk $p_\alpha(\mathbf{g}^\eta)$ using the noisy sample $\mathbf{g}^\eta$. The accuracy of these estimators depends strongly on the sample size, and when the size of $\mathbf{g}^\eta$ is small, the estimation can be unsatisfactory. To illustrate this point, in Fig. 1(a), we show the predictive risk, its bias-variance decomposition and the optimal α (with respect to predictive risk), for *shaw* from Regularization tools. In Fig. 1(b), we show the GCV and UPRE estimators (including their global minimizers). Clearly, both curves do not have a unique minimizer, and more importantly, the global minimizers are smaller than the optimal one by several orders of magnitude. Consequently, the corresponding reconstructions by GCV and UPRE are hugely corrupted by noise amplification and completely useless. This kind of behavior can be observed in the majority of test problems available in the package Regularization tools and actually for each problem therein, its presence has a non-negligible significant percentage of randomly generated samples $\mathbf{g}^\eta$.

In this work we present a novel parameter choice strategy inspired by predictive risk, which aims at overcoming the aforementioned drawback of lacking robustness. It gives an upper bound of the optimal parameter, and thus can provide solutions sufficiently accurate yet very stable with respect to sample variation. The rule is based on the following simple observation:

$$s_n(I - X_\alpha)^2 \|\mathbf{g}^\dagger\|^2 \le \|(X_\alpha - I)\mathbf{g}^\dagger\|^2 \le s_1(I - X_\alpha)^2 \|\mathbf{g}^\dagger\|^2,$$

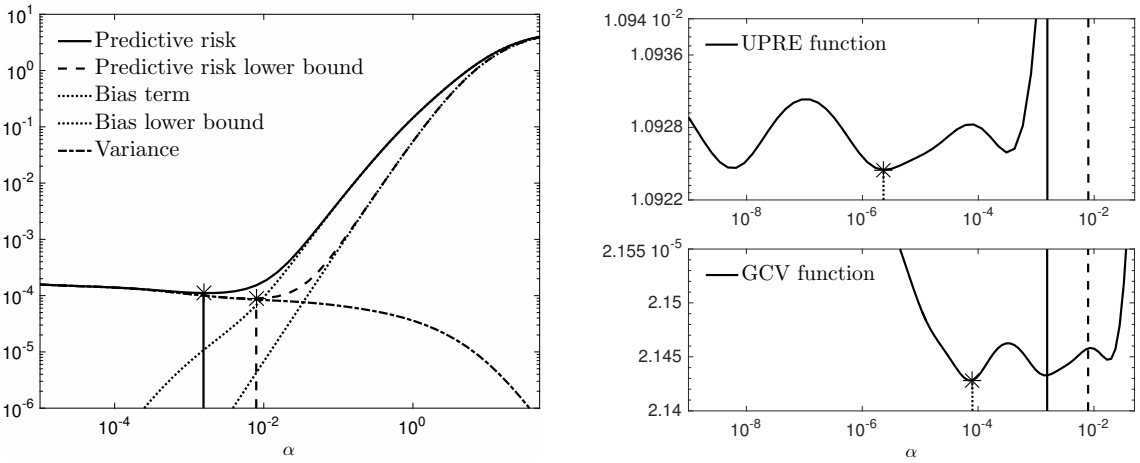


Figure 1: Behavior of the predictive risk and the lower bound for *shaw*.

where $s_n(I-X_\alpha)$ and $s_1(I-X_\alpha)$ denote the minimum and maximum singular value of the matrix $I-X_\alpha$, respectively. Then given the norm of the true data $\mathbf{g}^\dagger$, i.e., $\|\mathbf{g}^\dagger\| = \rho$, we can bound the predictive risk $p_\alpha(\mathbf{g}^\eta)$ both from above and below accordingly using the preceding bounds. It turns out that the upper bound is not useful since it can trivially reduce to a constant. Surprisingly, graphically, the lower bound has the same shape as the predictive risk and possesses a minimizer close to the optimal one. In Fig. 1(a), we show the minimizers of the predictive risk (left one) and lower bound (right one), and the lower bound of the bias term, which empirically motivates the use of lower bound for parameter choice. Thus, under the hypothesis $\|\mathbf{g}^\dagger\| = \rho$, we take the minimum of the predictive risk

$$T_\rho(\alpha) = \min_{\|\mathbf{g}^\dagger\|=\rho} p_\alpha(\mathbf{g}^\eta), \tag{5}$$

and then seek the minimizer of $T_\rho(\alpha)$ along α :

$$\alpha^* = \arg\min_\alpha T_\rho(\alpha), \tag{6}$$

which approximates the optimal parameter and thus can be used as a choice rule. Note that the function $T_\rho(\alpha)$ depends on the variance σ^2 of the noise η and the (squared) data norm ρ^2. We can rewrite the lower bound of the predictive risk as

$$T_{(\rho^2,\sigma^2)}(\alpha) = \rho^2 s_n(X_\alpha - I)^2 + \sigma^2 \|X_\alpha\|_F^2. \tag{7}$$

In the construction of the approximation (7), the variance term remains intact, but the bias term is subsumed by a lower bound. Since the bias term is increasing and the variance one is decreasing, α^* is an upper bound of the "optimal" regularization parameter, which ensures the non-overfitting of the choice rule; see Proposition 2 for a proof in the case of Tikhonov regularization. It is worth noting that problem (7) has a nice variational structure similar to the standard Tikhonov regularization itself: the first term can be viewed as fidelity (with respect to the identity matrix I), and the second term is a penalty, controlling the size). This variational structure lends itself to rigorous analysis.

In (7), we give the explicit dependency of $T_{(\rho^2,\sigma^2)}(\alpha)$ on the noise variance σ^2. Clearly, its minimizer(s) α^* depends only on the SNR ρ^2/σ^2. Thus, the choice rule determines a proper level of regularization according to the SNR. Nevertheless, the latter information is not always available in practice. In Section 3, we discuss three different strategies for minimizing the function $T_{(\rho^2,\sigma^2)}(\alpha)$ according to the available SNR information.

3 Optimization strategies

Now we apply the lower bound $T_{(\rho^2,\sigma^2)}(\alpha)$ given in (7) for parameter choice. The minimization of $T_{(\rho^2,\sigma^2)}(\alpha)$ is possible only if both ρ and σ are known. When some of this information is unavailable, as

is often the case in practice, we propose to approximate $T_{(\rho^2,\sigma^2)}(\alpha)$ (and accordingly the minimizer α^*). We describe three optimization approaches according to the level of available information. The second and third cases are more realistic in practical applications.

3.1 ρ and σ known

First, we consider the case of known ρ and σ. This is the least likely case in practice, since usually the true data norm $\|\mathbf{g}^\dagger\|$ is not known *a priori*, but theoretically it is the easiest case. Then the function $T_{(\rho^2,\sigma^2)}$ applies directly. Note that minimizers of $T_{(\rho^2,\sigma^2)}$ only depend on the SNR ρ^2/σ^2 or equivalently

$$\xi := 10\log_{10}\left(\frac{\rho^2}{n\sigma^2}\right), \tag{8}$$

which is generally known as SNR in decibel (dB), according to ISO 15739:2017, and will also be called SNR below, by slightly abusing the terminology. Note that given SNR ξ, the minimizer does not depend on noisy data $\mathbf{g}^\eta$ and the choice rule is of *a priori* nature [7]. For Tikhonov regularization, it can be shown that the rule is indeed well defined, and the function $T_{(\rho^2,\sigma^2)}$ is relatively well behaved and numerically tractable; see Section 4 below for further details.

3.2 ρ unknown and σ known

In many applications, the SNR ξ is unknown and only an estimate of σ^2 is available. This happens when the measurement is acquired by an industrial device, and by collecting multiple repetitive measurements of a zero excitation signal, the read-out noise allows estimating the noise level σ^2 (e.g., via maximum likelihood). This is a standard calibration procedure for providing the noise level (measurement precision) σ^2 of a device in industry. Using the following unbiased estimator of ρ^2

$$\hat{\rho^2} = \|\mathbf{g}^\eta\|^2 - n\sigma^2, \tag{9}$$

we can define an M-estimator of α^*. Specifically, we take the minimizer

$$\alpha^* := \arg\min_\alpha \{U^\eta(\alpha) := T_{(\hat{\rho^2},\sigma^2)}(\alpha)\} \tag{10}$$

where $U^\eta(\alpha)$ is an unbiased estimator of $T_{(\rho^2,\sigma^2)}(\alpha)$, i.e. $\mathbb{E}_{\boldsymbol{\eta}}[U^\eta(\alpha)] = T_{(\rho^2,\sigma^2)}(\alpha)$. Upon slightly abusing terminology, we also refer it to as `PRO`.

3.3 ρ and σ unknown

When both ρ and σ are unknown, `PRO` cannot be applied directly, and we propose an alternating estimating/minimization strategy. The procedure is summarized in the `I-PRO` algorithm below.

Formally, the `I-PRO` algorithm alternates between estimating the SNR (given α) and estimating α (given the SNR), in a manner similar to the classical EM algorithm or coordinate ascent in variational Bayes. We term it as iterative predictive risk optimization (`I-PRO`). Surprisingly, this simple procedure can provide excellent estimates of the SNR for ill-posed problems, and the obtained solutions are often very close to the ones corresponding to known ρ and σ, as confirmed by extensive simulation results in Section 5.

Computationally, `PRO` is based on two invariants, i.e., the smallest singular value $s_n(X_\alpha - I)$ of $X_\alpha - I$ and trace of $X_\alpha^T X_\alpha$. They can be both computed by means of SVD of the matrix A, whereas X_α can be written in terms of SVD of A, which is very convenient when the problem size is small. Otherwise, they can be computed without the SVD of A, using randomized algorithms; see Section 4.3 for further details.

4 Tikhonov regularization

Now we consider `PRO` and `I-PRO` for standard Tikhonov regularization. The corresponding influence matrix $X_\alpha \in \mathbb{R}^{n\times n}$ is given by

$$X_\alpha = A(A^*A + \alpha I)^{-1}A^*.$$

Algorithm 1 `I-PRO`: Iterative Predictive Risk Optimization.

1: Fix $\epsilon = 10^{-16}$, $\alpha_0 \neq 0$, $\alpha_1 = 0$, $k = 1$
2: **while** $|\alpha_k - \alpha_{k-1}| > \epsilon\alpha_k$ **do**
3: Update the variance according to formula

$$\sigma_k^2(\alpha) = \frac{1}{n}\|\mathbf{r}_{\alpha_k}^{\eta}\|^2 \tag{11}$$

where $\mathbf{r}_{\alpha}^{\eta} := A\mathbf{f}_{\alpha}^{\eta} - \mathbf{g}^{\eta}$ is the residual.
4: Update the signal norm using

$$\rho_k^2(\alpha_k) = \|\mathbf{g}^{\eta}\|^2 - \|\mathbf{r}_{a_k}^{\eta}\|^2. \tag{12}$$

5: Update α_k by computing minimizer of equation (7) given ρ_k^2 and σ_k^2, i.e.,

$$\alpha_{k+1} := \arg\min_{\alpha} T_{(\rho_k^2,\sigma_k^2)}(\alpha). \tag{13}$$

6: **end while**
7: **return** α_k

Clearly, X_α is symmetric and positive semidefinite, with its singular values $s_n(X_\alpha) \leq \ldots s_1(X_\alpha) < 1$. Let $s_1 \geq ... \geq s_r > 0 = s_{r+1} = \ldots = s_{\min(m,n)}$ be the singular values of the matrix A, with r being the rank of A. Then the singular values $s_i(X_\alpha)$ of X_α are given by

$$s_i(X_\alpha) = \begin{cases} \dfrac{s_i^2}{s_i^2+\alpha}, & i = 1, \ldots, r, \\ 0, & i = r+1, \ldots, n. \end{cases} \tag{14}$$

Since $s_n(X_\alpha - I) = 1 - s_1(X_\alpha) > 0$, a lower bound of the predictive risk $p_\alpha(\mathbf{g}^\eta)$ is given by

$$T_{(\rho^2,\sigma^2)}(\alpha) = \rho^2 \frac{\alpha^2}{(s_1^2+\alpha)^2} + \sigma^2 \sum_{i=1}^{r} \frac{s_i^4}{(s_i^2+\alpha)^2}. \tag{15}$$

The explicit form of $T_{(\rho^2,\sigma^2)}$ facilitates the analytical study of `PRO`. It is convenient to introduce two auxiliary functions. We denote the first and the second terms (involving α) of equation (15) by $f_1(\alpha)$ and $f_2(\alpha)$, respectively, i.e.,

$$f_1(\alpha) = \frac{\alpha^2}{(\alpha+s_1^2)^2} \quad \text{and} \quad f_2(\alpha) = \sum_{i=1}^{r} \frac{s_i^4}{(\alpha+s_i^2)^2}.$$

Next, for any fixed $h = \sigma^2/\rho^2$, we define a parameterized function

$$T_h(\alpha) = f_1(\alpha) + hf_2(\alpha) \quad \text{and} \quad \alpha^* = \arg\min_{\alpha\in[0,s_1^2/2]} T_h(\alpha).$$

The re-parameterization does not change the minimizer α^*, but is more convenient for the analysis. The following identities hold for the function $T_h(\alpha)$:

$$\lim_{\alpha\to 0^+} T_h(\alpha) = rh \quad \text{and} \quad \lim_{\alpha\to\infty} T_h(\alpha) = 1.$$

4.1 The `PRO` function $T_h(\alpha)$

The next lemma gives the convexity and monotonicity of $f_1(\alpha)$ and $f_2(\alpha)$.

Lemma 1. *The function $f_1(\alpha)$ is convex over $[0, s_1^2/2]$ and monotonically increasing, and $f_2(\alpha)$ is convex and monotonically decreasing over $[0, \infty)$.*

Proof. Direct computation gives

$$\begin{aligned} f_1'(\alpha) &= \tfrac{2\alpha}{(\alpha+s_1^2)^2} - \tfrac{2\alpha^2}{(\alpha+s_1^2)^3} = \tfrac{2\alpha s_1^2}{(\alpha+s_1^2)^3} > 0, \\ f_1''(\alpha) &= \tfrac{2s_1^2(\alpha+s_1^2)-6\alpha s_1^2}{(\alpha+s_1^2)^4} = \tfrac{2s_1^2(s_1^2-2\alpha)}{(\alpha+s_1^2)^4}. \end{aligned}$$

Thus, f_1 is convex over the interval $(0, s_1^2/2]$. Further,

$$f_2'(\alpha) = -\sum_{i=1}^{r} \frac{2s_i^4}{(\alpha+s_i^2)^3} < 0, \quad f_2''(\alpha) = \sum_{i=1}^{r} \frac{6s_i^4}{(\alpha+s_i^2)^4}.$$

This completes the proof of the lemma. □

Lemma 2. *For any minimizer $\alpha^*(h) \in [0, s_1^2/2]$ to the function $T_h(\alpha)$, it is strictly monotonically increasing in h.*

Proof. We prove the assertion by the implicit function theorem. The existence of a unique $\alpha^* \equiv \alpha^*(h) \in [0, s_1^2/2]$ follows from the strict convexity of T_h over the interval, cf Lemma 1. First, we claim that $\alpha^*(h)$ is not $\alpha = 0$. Indeed, it follows from straightforward computation that

$$T_h'(\alpha) = \frac{2\alpha s_1^2}{(\alpha+s_1^2)^3} - h\sum_{i=1}^{r} \frac{2s_i^4}{(\alpha+s_i^2)^3} \tag{16}$$

and thus

$$\lim_{\alpha\to 0^+} T_h'(\alpha) = -2h\sum_{i=1}^{r} s_i^{-2} < 0.$$

Thus, T_h is differentiable at $\alpha^*(h)$, and the optimal $\alpha^* \equiv \alpha^*(h)$ satisfies the optimality condition

$$f_1'(\alpha^*) + hf_2'(\alpha^*) = 0.$$

Then by the chain rule, we have

$$f_1''(\alpha^*)\frac{d\alpha^*}{dh} + f_2'(\alpha^*) + hf_2''(\alpha^*)\frac{d\alpha^*}{dh} = 0,$$

i.e.,

$$\frac{d\alpha^*}{dh} = \frac{-f_2'(\alpha^*)}{f_1''(\alpha^*) + hf_2''(\alpha^*)}$$

Then in view of Lemma 1, the denominator is strictly positive for any $\alpha^* > 0$, from which the desired assertion follows directly. □

The next proposition summarizes some properties of the minimizer to $T_h(\alpha)$.

Proposition 1. *Let r be the rank of the matrix A. Then the following statements hold for the function $T_h(\alpha)$.*

(i) Over the interval $[0, s_1^2/2]$, the function T_h has a unique minimizer α^*.

(ii) If $s_1 > 0$ and $h \le (27r)^{-1}$, the function $T_h(\alpha)$ admits a unique global minimizer α^* in $(0, s_1^2/2]$.

(iii) For $h \le \zeta := s_1^2/\mathrm{tr}(A^*A)$, the minimizer $\alpha^* \equiv \alpha^*(h)$ satisfies

$$s_1^2 h \le \alpha^* \le (1-(h/\zeta)^{\frac{1}{3}})^{-1} s_1^2 (h/\zeta)^{\frac{1}{3}}.$$

Proof. Part (i) is already shown in Lemma 2.

Part (ii). By Lemma 1, the function $T_h(\alpha)$ is differentiable and strictly convex in the interval $\alpha \in [0, s_1^2/2]$. Furhter, T_h is increasing in the interval $\alpha \in (s_1^2/2, \infty)$ if $h < (27r)^{-1}$. Indeed, in view of (16),

$$\begin{aligned} T_h'(\alpha) &\geq \tfrac{2\alpha s_1^2}{(\alpha+s_1^2)^3} - h\sum_{i=1}^r \tfrac{2s_i^2}{(\alpha+s_i^2)^2} \\ &\geq \tfrac{2\alpha s_1^2}{(\alpha+s_1^2)^3} - \tfrac{2hrs_1^2}{\alpha^2} = \tfrac{2s_1^2}{\alpha^2}\Big(\tfrac{\alpha^3}{(\alpha+s_1^2)^3} - hr\Big). \end{aligned}$$

Clearly $\frac{\alpha^3}{(\alpha+s_1^2)^3}$ is an increasing in α, and its minimum over $[s_1^2/2, +\infty)$ is achieved at $\alpha = s_1^2/2$ with a minimum value $1/27$. Therefore under the condition $h \leq (27r)^{-1}$, there exists a unique global minimizer to $T_h(\alpha)$ in $[0, +\infty)$ and it is located within the interval $[0, s_1^2/2]$. Moreover, as argued in Lemma 2 that if $h > 0$, $\alpha = 0$ cannot be a minimizer. This shows part (ii).

Part (iii). Clearly, by Lemma 2, as h tends to zero monotonically, the minimizer $\alpha^*(h)$ also decreases monotonically. Further, the optimality condition $\frac{d}{d\alpha}T_h(\alpha^*) = 0$ implies

$$\frac{\alpha^* s_1^2}{(s_1^2+\alpha^*)^3}\left(\sum_{i=1}^r \frac{s_i^4}{(s_i^2+\alpha^*)^3}\right)^{-1} = h. \tag{17}$$

Next we bound the quantity on the left hand side. Since $\sum_{i=1}^r \frac{s_i^4}{(s_i^2+\alpha^*)^3} \geq \frac{s_1^4}{(\alpha^*+s_1^2)^3}$, we deduce

$$h \leq \frac{\alpha^* s_1^2}{(s_1^2+\alpha^*)^3}\frac{(\alpha^*+s_1^2)^3}{s_1^4} = \frac{\alpha^*}{s_1^2}$$

This shows the first inequality. Meanwhile, from the inequality

$$\sum_{i=1}^r \frac{s_i^4}{(s_i^2+\alpha^*)^3} \leq \sum_{i=1}^r \frac{s_i^2}{(\alpha^*)^2} = \frac{\mathrm{tr}(A^*A)}{(\alpha^*)^2},$$

we obtain

$$h \geq \frac{\alpha^* s_1^2}{(s_1^2+\alpha^*)^3}\frac{(\alpha^*)^2}{\mathrm{tr}(A^*A)} = \frac{s_1^2}{\mathrm{tr}(A^*A)}\frac{(\alpha^*)^3}{(s_1^2+\alpha^*)^3},$$

i.e., $\frac{\alpha^*}{\alpha^*+s_1^2} \leq (\mathrm{tr}(A^*A)h/s_1^2)^{\frac{1}{3}}$. Solving for the inequality gives the assertion in part (iii). □

Remark 1. *According to Theorem 1, the function T_h always has a finite positive and unique minimizer within the interval $[0, s_1^2/2]$, which is also the unique global minimizer over $[0, \infty)$ if the SNR is sufficiently large. In practice, minimizing over $[0, s_1^2/2]$ is sufficient for Tikhonov regularization. Since the function $T_h(\alpha)$ is strictly convex in $\alpha \in [0, s_1^2/2]$, in principle, its minimization is numerically tractable. For example, one may apply Newton type methods, which is guaranteed to converge globally [37]. Thus it is numerically more amenable than choice rules based on predictive risk, e.g., UPRE and GCV, which are known to suffer from (bad) local minima as well as flatness near global minima; see Fig. 1.*

Next we turn to the upper bound property of α^*.

Lemma 3. *For any $\mathbf{g}^\dagger$, there exists an $\alpha_0 > 0$, such that there holds for all $\alpha \leq \alpha_0$*

$$\sum_{i=1}^r \frac{2\alpha s_i^2}{(s_i^2+\alpha)^3}(\mathbf{g}^\dagger, \mathbf{u}_i)^2 \geq \frac{2\alpha s_1^2}{(s_1^2+\alpha)^3}\|\mathbf{g}^\dagger\|^2,$$

where $\mathbf{u}_i$ denotes the ith left singular vector of the matrix A.

Proof. By the identity $\sum_{i=1}^r (\mathbf{g}^\dagger, \mathbf{u}_i)^2 = \|\mathbf{g}^\dagger\|^2 = \rho^2$, the assertion is equivalent to

$$\sum_{i=1}^r \frac{2\alpha s_i^2}{(s_i^2+\alpha)^3}(\mathbf{g}^\dagger, \mathbf{u}_i)^2 \geq \sum_{i=1}^r \frac{2\alpha s_1^2}{(s_1^2+\alpha)^3}(\mathbf{g}^\dagger, \mathbf{u}_i)^2.$$

Now we claim that for all sufficiently small α, there holds

$$\frac{s_i^2}{(s_i^2+\alpha)^3} \geq \frac{s_1^2}{(s_1^2+\alpha)^3}, \quad \text{i.e.} \quad \frac{s_1^2+\alpha}{s_i^2+\alpha} \leq \Big(\frac{s_1^2}{s_i^2}\Big)^{\frac{1}{3}},$$

which, with $\lambda = (\frac{s_1^2}{s_i^2})^{\frac{1}{3}}$, is equivalent to $\alpha \leq s_1^2 \frac{1-\lambda^{-2}}{\lambda-1} = s_1^2\lambda^{-1}(1+\lambda^{-1})$. Thus, it suffices to choose $\alpha_0 = s_1^{\frac{4}{3}} s_n^{\frac{2}{3}}$. □

Remark 2. *The bound α_0 given in Lemma 3 is very loose. In practice, the coefficients $(\mathbf{g}^\dagger, \mathbf{u}_i)^2$ can decay very rapidly to zero as i increases, especially for severely ill-posed problems, due to smoothing property of A (and regularity condition on $\mathbf{f}^\dagger$, e.g., source conditions [7, 19]). Then, the first few terms in the summation are dominating, and one expects a much larger bound. In practice, it is often $O(1)$.*

Proposition 2. *If the minimizer α^* to the function T_h is sufficiently small, then it is an upper bound of a local minimizer to the predictive risk $p_\alpha(\mathbf{g}^\eta)$.*

Proof. Straightforward computation shows that the predictive risk $p_\alpha(\mathbf{g}^\eta)$ is given by

$$p_\alpha(\mathbf{g}^\eta) = \sum_{i=1}^r \frac{\alpha^2}{(s_i^2+\alpha)^2}(\mathbf{g}^\dagger, \mathbf{u}_i)^2 + \sigma^2 \sum_{i=1}^r \frac{s_i^4}{(s_i^2+\alpha)^2},$$

It suffices to show that $\frac{d}{d\alpha} p_\alpha(\mathbf{g}^\eta)|_{\alpha=\alpha^*} > 0$. Note that $\frac{d}{d\alpha} p_\alpha(\mathbf{g}^\eta)|_{\alpha=\alpha^*}$ is given by

$$\sum_{i=1}^r \frac{2\alpha^* s_i^2}{(s_i^2+\alpha^*)^3}(\mathbf{g}^\dagger, \mathbf{u}_i)^2 - \sigma^2 \sum_{i=1}^r \frac{2s_i^4}{(s_i^2+\alpha^*)^3}.$$

Since the minimizer α^* of $T_{(\rho^2,\sigma^2)}$ is sufficiently small, by Lemma 3, $\frac{d}{d\alpha} p_\alpha(\mathbf{g}^\eta)|_{\alpha=\alpha^*} > 0$. This and the monotonicity of the sums in $p_\alpha(\mathbf{g}^\eta)$ complete the proof. □

4.2 Convergence of the iterative scheme

Now we analyze the convergence of the `I-PRO` algorithm for Tikhonov regularization. First, by viewing the estimates ρ and σ as functions of α, we have the following monotonicity result.

Lemma 4. *The functions $\rho^2(\alpha)$ and $\sigma^2(\alpha)$ are monotonically decreasing and increasing, respectively, in α.*

Proof. The variance estimate $\sigma^2(\alpha)$ is proportional to $\|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\|^2$. For any $\alpha_0, \alpha_1 > 0$, by the minimizing property of $\mathbf{f}_{\alpha_0}^\eta$ and $\mathbf{f}_{\alpha_1}^\eta$, i.e.,

$$\begin{aligned} \|A\mathbf{f}_{\alpha_0}^\eta - \mathbf{g}^\eta\|^2 + \alpha_0\|\mathbf{f}_{\alpha_0}^\eta\|^2 &\leq \|A\mathbf{f}_{\alpha_1}^\eta - \mathbf{g}^\eta\|^2 + \alpha_0\|\mathbf{f}_{\alpha_1}^\eta\|^2, \\ \|A\mathbf{f}_{\alpha_1}^\eta - \mathbf{g}^\eta\|^2 + \alpha_1\|\mathbf{f}_{\alpha_1}^\eta\|^2 &\leq \|A\mathbf{f}_{\alpha_0}^\eta - \mathbf{g}^\eta\|^2 + \alpha_1\|\mathbf{f}_{\alpha_0}^\eta\|^2, \end{aligned}$$

we deduce that the residual $\|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\|$ is monotonically increasing in α. Then the assertion follows immediately. □

Now we can prove that the sequence $\{\alpha_k\}_{k=1}^\infty$ of regularization parameters generated by the `I-PRO` algorithm is actually monotone. Once it is bounded, the result provides a constructive proof of the existence of a fixed point.

Theorem 1. *For any $\alpha_0 \in [0, s_1^2/2]$, the sequence $\{\alpha_k\}$ generated by the `I-PRO` algorithm is monotone.*

Proof. If $\alpha_0 < \alpha_1$, then by the monotonicity in Lemma 4, $h(\alpha) = \sigma^2(\alpha)/\rho^2(\alpha)$ is monotonically increasing, and thus $h(\alpha_0) < h(\alpha_1)$. Then by Lemma 2, $\alpha_{h(\alpha_0)}^* < \alpha_{h(\alpha_1)}^*$, i.e., $\alpha_1 < \alpha_2$. The case $\alpha_0 > \alpha_1$ follows similarly. This shows the monotonicity of the iterates $\{\alpha_k\}_{k=1}^\infty$. □

Note that by the optimality condition to T_h, at the fixed point α^* satisfies

$$\frac{\alpha s_1^2}{(\alpha+s_1)^3} - \frac{\sigma^2(\alpha)}{\rho^2(\alpha)} \sum_{i=1}^{r} \frac{s_i^4}{(\alpha+s_i^2)^3} = 0. \tag{18}$$

Note that this equation can be rewritten into a (high-degree) polynomial in α, and thus the existence of a root is ensured. Further, by Lemma 4, $\|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\|$ decreases monotonically to the least-squares residual as α tends to zero, and thus the limit $\lim_{\alpha\to 0^+} \|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\| > 0$ makes sense.

Proposition 3. *The following statements hold for the* `I-PRO` *algorithm.*

(i) If $\lim_{\alpha\to 0^+} \|A\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\| > 0$*, then* 0 *is not a fixed point of the* `I-PRO` *algorithm.*

(ii) If the estimate $\sigma^2(\alpha^*)/\rho(\alpha^*)$ *is smaller (greater) than the exact one, then* `I-PRO` *estimate* α^* *is smaller (greater) than the* `PRO` *estimate.*

Proof. Assertion (i) is direct from (18) and an argument by contradiction, and (ii) follows from Lemma 2. □

Proposition 3(ii) suggests that one may monitor the quantity $\sigma^2(\alpha_k)/\rho^2(\alpha_k)$ during the iteration as an *a posteriori* check: if it is deemed to be too small, then the iteration should be terminated.

4.3 SVD-free implementation

For large-scale problems, SVD is usually too costly to apply (randomized SVD may be applied instead, when the problem has low-rank structure [20]). Instead we implement `PRO` and `I-PRO` as follows. It follows from (14) that the first term in the function $T_{(\rho^2,\sigma^2)}$ depends on the smallest eigenvalue of the influence matrix $I - X_\alpha$, or the largest eigenvalue $\lambda_1(A^*A)$ of A^*A. Thus, we employ classical algorithms for computing $\lambda_1(A^*A)$, e.g., power method, and take

$$s_n(X_\alpha - I)^2 = \left(\frac{\alpha}{\lambda_1(A^*A)+\alpha}\right)^2. \tag{19}$$

To compute the term $\|X_\alpha\|_F^2$, we employ a randomized trace estimator [18]:

$$\begin{aligned} \|X_\alpha\|_F^2 &= \operatorname{tr}(X_\alpha^* X_\alpha) = \mathbb{E}_{\mathbf{z}}[\|X_\alpha \mathbf{z}\|^2] \\ &\simeq \textstyle\sum_{i=1}^{p} \|A(A^*A+\alpha I)^{-1} A^* \mathbf{z}_{(i)}\|^2, \end{aligned}$$

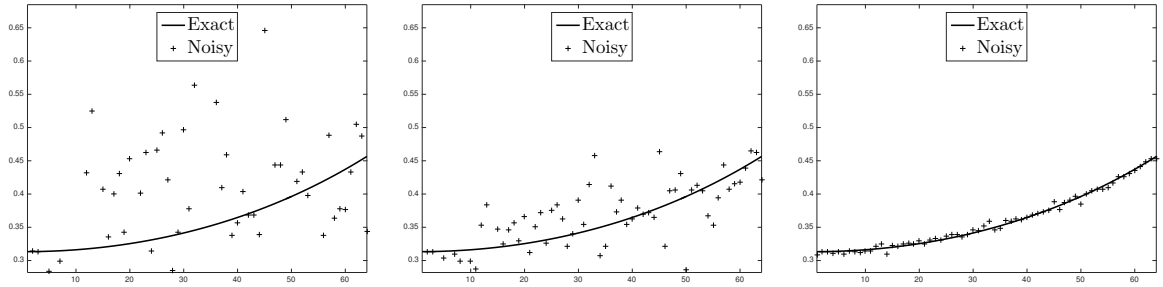
where $\mathbf{z}$ follows the standard Gaussian distribution and $\{\mathbf{z}_{(i)}\}_{i=1}^p$ are p i.i.d. samples of Z (other choices are also possible). Each summand involves solving one linear system: $(A^*A+\alpha I)\mathbf{w}_{(i)} = A^*\mathbf{z}_{(i)}$. Once $\mathbf{w}_{(i)}$ are obtained, we have

$$\|X_\alpha\|_F^2 \simeq \sum_{i=1}^{p} \|A\mathbf{w}_{(i)}\|^2. \tag{20}$$

In practice, only a few samples are required for an accurate estimation (see [1] for relevant error bounds on randomized trace estimators). The randomized trace estimation is widely used in implementing GCV. The difference lies in the fact that for `PRO`, it is applied to the matrix $X_\alpha^* X_\alpha$, instead of X_α in GCV.

5 Numerical simulations and discussions

Now we present two sets of simulation studies with synthetic data from popular public software packages, i.e., Regularization tools and AIR tools. The first aims to show the robustness of the proposed method for problems with small and moderately sized samples, and the second to show its performance on large-scale problems.



(a) $\xi = 10$ (b) $\xi = 20$ (c) $\xi = 40$

Figure 2: Synthetic data for *baart* with $n = 64$ and random seed 9, i.e. `rng(9)`.

5.1 Small and moderately sized samples

Below we compare the Tikhonov solutions by `PRO` and `I-PRO` with that by existing choice rules in two scenarios: (i) `PRO` versus DP, BP and UPRE, if σ is known; (ii) `I-PRO` versus LC, GCV and QOC, if σ is unknown. DP is implemented by solving for α in

$$\|\mathbf{A}\mathbf{f}_\alpha^\eta - \mathbf{g}^\eta\| = \sqrt{n}\sigma.$$

The implementation of BP follows [11, equation (1.7)] with the constants $\gamma = 1/4$ and c given in [11, Table 1, p. 961]. All the rules are computed on an equally distributed grid on a logarithmical scale. We evaluate them on eight inverse problems (i.e., *baart*, *deriv2*, *foxgood*, *gravity*, *heat*, *i_laplace*, *phillips*, *shaw*) from the public package Regularization tools (downloaded from `http://people.compute.dtu.dk/pcha/Regutools/index.html`, accessed on July 1, 2019). They are all one-dimensional linear integral equations of first kind, with different degree of ill-posedness. Each problem refers to one example, except *heat* and *i_laplace*. The *heat* family depends on one parameter controlling its conditioning, and we choose the values 1 and 5 to simulate mildly ill-posed and almost well-posed scenarios, respectively. The *i_laplace* family has four cases, of which the first three cases have smooth solutions and thus standard Tikhonov regularization is suitable. This gives rise to a total of eleven examples. For each example, we take 100 independent and identically distributed (i.i.d.) realizations $\{\boldsymbol{\eta}_i\}_{i=1}^{100}$ of the Gaussian noise with mean zero and covariance $\sigma^2 I$, seeded by MATLAB function `rng(i)` with $i = 1, ..., 100$, leading to 100 replicates of noisy data $\mathbf{g}^{\eta_i} \sim \mathcal{N}(\mathbf{g}, \sigma^2 I)$; see Fig. 2 for typical noisy data for *baart*.

To measure the accuracy of a reconstruction $\mathbf{f}_\alpha^\eta$, we use the relative ℓ_2 norm error. ratio between the reconstruction $\mathbf{f}_\alpha^\eta$ and the ground truth $\mathbf{f}^\dagger$

$$\varepsilon(\alpha, \eta) = \frac{\|\mathbf{f}_\alpha^\eta - \mathbf{f}^\dagger\|}{\|\mathbf{f}^\dagger\|}$$

Other quality measures can be used, and the observations below remain unchanged. Further, we define the 'oracle'/optimal ℓ_2 norm error, denoted by $\varepsilon_o(\eta)$, as the minimum value of ℓ_2 norm error achieved along the regularization path. Then we measure the efficiency of a parameter choice rule by computing the ratio between the oracle ℓ_2 norm error over the ℓ_2 norm error of the solution given by the rule

$$\text{eff}_*(\eta) = \frac{\varepsilon_o}{\varepsilon(\alpha_*, \eta)}, \tag{21}$$

where $* =$ DP, UPRE, BP, `PRO`, LC, GCV, QOC, `I-PRO`. Finally, to show the overall performance of a choice rule on a fixed problem setup, we take the median of its efficiency over the 100 replicates.

In Tables 1 and 2, we present quantitative results for all the eight choice rules under consideration at three typical SNR values (i.e., $\xi = 10, 20, 40$) and two sample sizes (i.e., $n = 64$ and 1024). The results are the median (columns from 3 to 10) of the empirical efficiency distributions computed from the 100 replicates. Below we examine the results more closely.

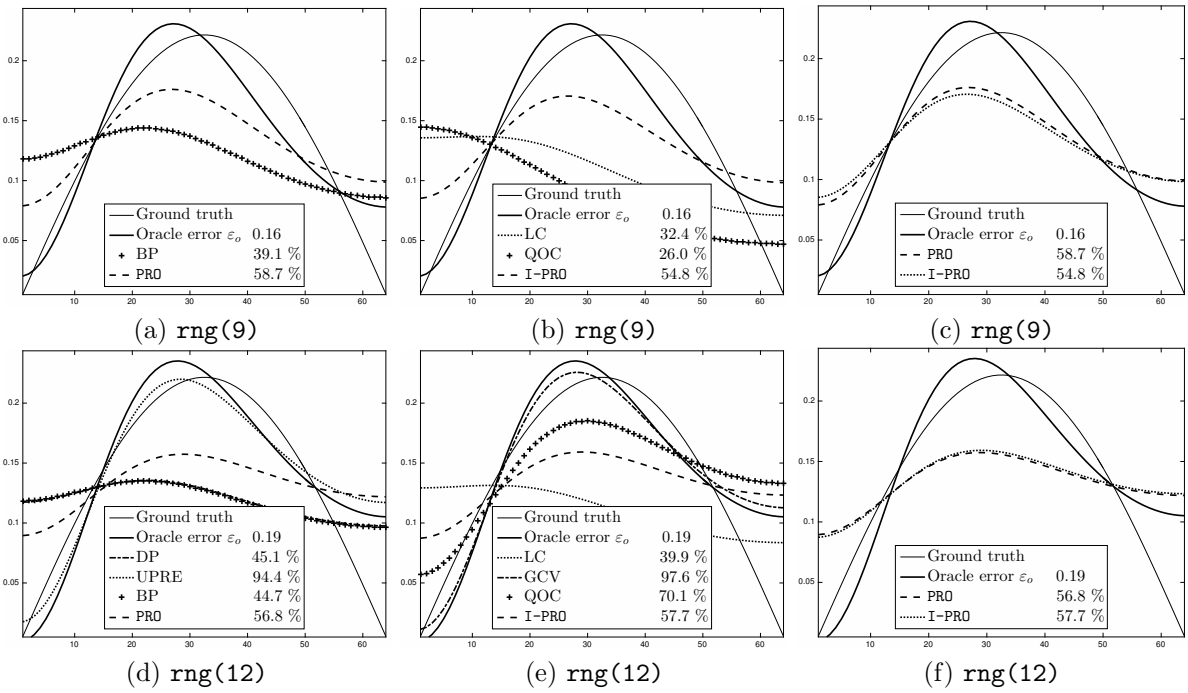
First, the oracle ℓ_2 norm error, ε_o, in the second column depends crucially on the problem type, condition number and ground truth. When $n = 64$ (see Table 1), except for one case, ε_o ranges from 0.11

	ℓ_2 error	σ know λ unknown				σ and λ unknowns			
$n = 64$	Oracle	DP	UPRE	BP	PRO	LC	GCV	QOC	I-PRO
		low statistic ($\xi = 10$)							
baart	0.34	43.1%	74.5%	61.1%	78.7%	71.1%	71.2%	79.3%	79.5%
deriv2	0.46	76.9%	90.2%	87.5%	99.0%	94.4%	90.2%	84.9%	98.8%
foxgood	0.11	12.7%	51.5%	38.5%	77.2%	56.0%	51.6%	60.5%	76.4%
gravity	0.19	73.8%	76.3%	86.9%	87.4%	93.3%	75.2%	96.9%	86.7%
heat(1)	0.50	87.0%	96.9%	91.9%	89.4%	0.1%	95.2%	82.4%	89.1%
heat(5)	0.38	98.5%	90.2%	80.2%	98.5%	88.0%	88.0%	55.0%	55.9%
i_laplace(1)	0.22	63.2%	87.4%	84.0%	96.1%	96.6%	87.4%	96.2%	96.4%
i_laplace(2)	0.86	97.0%	99.2%	98.3%	99.4%	99.4%	99.3%	99.3%	99.4%
i_laplace(3)	0.28	65.8%	89.8%	84.4%	95.9%	96.7%	87.7%	94.9%	96.3%
phillips	0.19	77.6%	85.4%	87.6%	95.5%	91.6%	82.8%	97.2%	94.4%
shaw	0.24	57.9%	88.7%	74.1%	95.9%	95.2%	86.3%	95.5%	96.2%
		medium statistic ($\xi = 20$)							
baart	0.21	45.7%	78.4%	35.3%	68.8%	72.8%	72.3%	68.4%	69.4%
deriv2	0.39	87.7%	94.5%	93.2%	97.5%	98.5%	92.8%	85.1%	97.4%
foxgood	0.06	14.5%	48.8%	32.4%	79.5%	83.7%	48.7%	86.9%	79.6%
gravity	0.11	84.4%	78.2%	85.5%	90.1%	76.0%	78.3%	97.3%	87.5%
heat(1)	0.33	90.4%	89.8%	96.1%	74.2%	0.3%	86.9%	58.6%	71.3%
heat(5)	0.18	98.4%	92.8%	98.5%	99.4%	94.5%	94.5%	80.9%	80.9%
i_laplace(1)	0.16	74.8%	88.1%	82.9%	95.1%	95.3%	88.2%	92.1%	95.2%
i_laplace(2)	0.84	97.9%	99.0%	98.3%	98.8%	99.4%	99.0%	99.0%	98.8%
i_laplace(3)	0.18	60.6%	86.2%	69.0%	89.4%	92.2%	85.8%	77.7%	89.6%
phillips	0.09	82.9%	75.2%	80.4%	90.3%	73.9%	73.5%	99.0%	88.0%
shaw	0.18	70.8%	84.4%	82.6%	97.6%	95.8%	84.5%	95.5%	97.5%
		high statistic ($\xi = 40$)							
baart	0.15	51.7%	76.9%	24.9%	72.2%	88.9%	71.2%	89.8%	70.6%
deriv2	0.26	95.0%	95.6%	97.7%	88.0%	91.6%	95.5%	88.9%	86.2%
foxgood	0.02	24.7%	56.1%	10.5%	84.9%	49.5%	41.8%	84.7%	84.6%
gravity	0.04	86.4%	84.8%	80.4%	95.9%	45.6%	84.9%	96.2%	95.5%
heat(1)	0.13	95.4%	82.5%	23.8%	59.7%	1.6%	78.9%	96.1%	36.2%
heat(5)	0.02	90.3%	99.8%	85.3%	99.8%	16.4%	16.4%	99.6%	99.6%
i_laplace(1)	0.10	73.2%	87.0%	83.4%	92.3%	85.5%	86.9%	87.8%	92.7%
i_laplace(2)	0.82	98.4%	99.1%	98.6%	97.9%	99.6%	99.1%	98.5%	97.7%
i_laplace(3)	0.06	60.6%	78.2%	68.0%	68.9%	69.2%	73.4%	94.1%	67.3%
phillips	0.03	93.5%	56.8%	90.6%	72.0%	22.3%	51.1%	98.0%	70.3%
shaw	0.11	64.4%	87.0%	68.4%	79.6%	91.8%	87.4%	69.1%	80.0%

Table 1: Performance comparison between choice rules DP, UPRE, BP, PRO, LC, GCV, QOC, I-PRO on examples from Regularization tools. For each row: column one gives the problem and column two shows the oracle ℓ_2-norm error. From the third to tenth column the median efficiency of the choice rule (see equation 21). The three blocks correspond to three different levels of statistic. The sample size n of each problem is 64.

	ℓ_2 error	σ know λ unknown				σ and λ unknowns			
$n = 1024$	Oracle	DP	UPRE	BP	PRO	LC	GCV	QOC	I-PRO
		low statistic ($\xi = 10$)							
baart	0.21	30.1%	86.7%	41.9%	69.8%	41.6%	83.3%	74.5%	70.3%
deriv2	0.37	70.6%	96.4%	67.4%	95.6%	78.0%	96.3%	90.1%	95.9%
foxgood	0.06	4.3%	55.6%	20.0%	84.4%	29.7%	49.3%	88.8%	84.2%
gravity	0.09	40.0%	77.8%	42.3%	90.1%	95.5%	76.1%	97.4%	90.4%
heat(1)	0.30	67.0%	91.7%	50.8%	71.8%	56.8%	90.2%	51.3%	71.5%
heat(5)	0.20	97.4%	62.2%	99.9%	83.5%	0.1%	60.9%	0.1%	81.8%
i_laplace(1)	0.26	81.8%	96.4%	76.4%	93.8%	97.0%	96.4%	98.0%	93.8%
i_laplace(2)	0.81	96.6%	99.4%	96.4%	99.5%	98.2%	99.4%	99.0%	99.5%
i_laplace(3)	0.20	45.7%	90.2%	47.7%	78.1%	65.0%	90.7%	92.1%	78.2%
phillips	0.08	44.4%	64.4%	39.0%	86.3%	98.3%	65.5%	99.1%	86.8%
shaw	0.18	41.1%	87.1%	54.0%	97.7%	86.6%	83.2%	97.6%	97.9%
		medium statistic ($\xi = 20$)							
baart	0.17	29.4%	80.8%	28.7%	67.8%	58.8%	79.6%	81.5%	68.0%
deriv2	0.31	76.5%	95.8%	71.6%	92.0%	85.5%	95.4%	89.6%	91.9%
foxgood	0.04	6.3%	49.1%	16.0%	91.5%	52.1%	45.9%	88.2%	91.2%
gravity	0.05	45.7%	81.1%	46.0%	91.7%	97.5%	81.6%	96.8%	91.2%
heat(1)	0.19	74.0%	86.8%	59.5%	64.1%	84.7%	86.3%	95.5%	62.7%
heat(5)	0.10	91.1%	53.7%	98.8%	74.3%	0.2%	52.6%	0.2%	72.3%
i_laplace(1)	0.24	87.0%	97.5%	83.5%	93.0%	97.3%	97.6%	95.8%	93.0%
i_laplace(2)	0.79	97.1%	99.5%	96.8%	98.9%	98.4%	99.5%	99.0%	99.0%
i_laplace(3)	0.12	46.5%	91.3%	45.3%	72.2%	76.2%	91.3%	95.1%	71.8%
phillips	0.04	53.7%	54.4%	41.8%	72.8%	90.0%	52.8%	99.1%	73.2%
shaw	0.15	47.9%	87.6%	68.9%	95.1%	93.7%	85.9%	91.6%	95.1%
		high statistic ($\xi = 40$)							
baart	0.12	30.1%	78.6%	19.3%	68.6%	78.1%	76.1%	76.2%	68.4%
deriv2	0.21	88.4%	95.4%	78.3%	82.1%	98.7%	95.1%	89.6%	81.6%
foxgood	0.02	11.8%	56.9%	7.1%	88.1%	90.1%	55.8%	72.3%	88.7%
gravity	0.02	54.6%	82.3%	49.4%	96.9%	79.3%	80.9%	94.0%	96.8%
heat(1)	0.07	87.7%	77.3%	12.4%	53.1%	85.6%	77.1%	99.0%	48.5%
heat(5)	0.02	71.8%	39.0%	92.2%	54.8%	0.5%	37.7%	0.5%	51.7%
i_laplace(1)	0.21	90.8%	98.3%	88.3%	91.2%	98.5%	98.3%	95.6%	91.0%
i_laplace(2)	0.77	98.3%	99.7%	97.9%	98.4%	99.2%	99.7%	98.9%	98.4%
i_laplace(3)	0.05	47.6%	86.8%	43.9%	59.0%	97.1%	87.4%	95.9%	58.6%
phillips	0.02	76.2%	65.5%	61.0%	83.7%	50.9%	63.6%	93.9%	83.1%
shaw	0.06	38.2%	87.9%	36.6%	56.7%	92.3%	87.1%	80.6%	56.2%

Table 2: See caption of Table 1, but with the sample size n of each problem being 1024.



(a) `rng(9)` (b) `rng(9)` (c) `rng(9)`

(d) `rng(12)` (e) `rng(12)` (f) `rng(12)`

Figure 3: Reconstructions for *baart* with $n = 64$ and $\xi = 10$ and two different noise realizations with `rng(9)` and `rng(12)`.

to 0.50, from 0.06 to 0.39, and from 0.02 to 0.26, for low, medium and high statistic, respectively. In the exceptional case, i.e., *i_laplace(2)*, the ground truth object $\mathbf{f}^\dagger$ to be recovered is actually almost invisible within the meaning of the inverse problems theory [4], and so ε_o fails to reach barely 0.82. For all choice rules with σ^2 known, the efficiency of any parameter choice rule is almost always larger than 60%. The efficiency depends on both the choice rule and specific problem. For example, *baart* and *foxgood* are fairly challenging for all eight choice rules. In particular, on these two problems, DP does not perform as well as the others. This may be related to the well known fact that DP is sensitive to the estimation accuracy of the noise variance σ^2. Despite a strong dependency on the problem, the percentage generally increases with the SNR, indicating the convergence of these rules (at least on average). It is worth noting that LC is very effective in most cases but it fails spectacularly on *heat*, for which all other methods work reasonably well. Moreover, the efficiency distributions for DP, UPRE, GCV (and occasionally also LC) have very low median, and thus in practice, these choice rules can fail to choose a suitable α value. The observations on GCV is consistent with prior empirical study in [30]. Surprisingly, `PRO` and `I-PRO` (and BP and QOC) can consistently provide satisfactory solutions for all noise realizations. This property is highly desirable in practice, and it is attributed to the fact that `PRO` and `I-PRO` are designed to achieve the minimum error in the mean squared sense (at least approximately!), and thus are fairly robust with respect to noise realization. This contrasts sharply with other existing choice rules, which always suffer from a significant number of failures. One example of failure is given in Fig. 3: whereas in (d) and (e) all choice rules yield good estimates of the ground truth, in (a) and (b) existing choice rules, e.g., UPRE or GCV, are highly inefficient and the corresponding solutions are omitted since they extend over several orders of magnitude above the ground truth.

In practice, it is also important to check the variance of the efficiency distribution. By considering each problem separately (given the SNR ξ and the size n), the simulation study indicates that the empirical efficiency distribution by `PRO` has a very small variance. This is due to the fact that given σ^2, the estimation $\hat{\rho^2}$ of the data norm ρ^2 given in (9) is relatively insensitive to the noise realization, especially when the sample size n is not too small. In contrast, the empirical efficiency distributions by DP, UPRE and GCV (and sometimes also LC) spread out, and thus these methods do not enjoy uniform performance with respect to noise realization. This behavior is clearly visible in Fig. 4: `PRO` and `I-PRO`

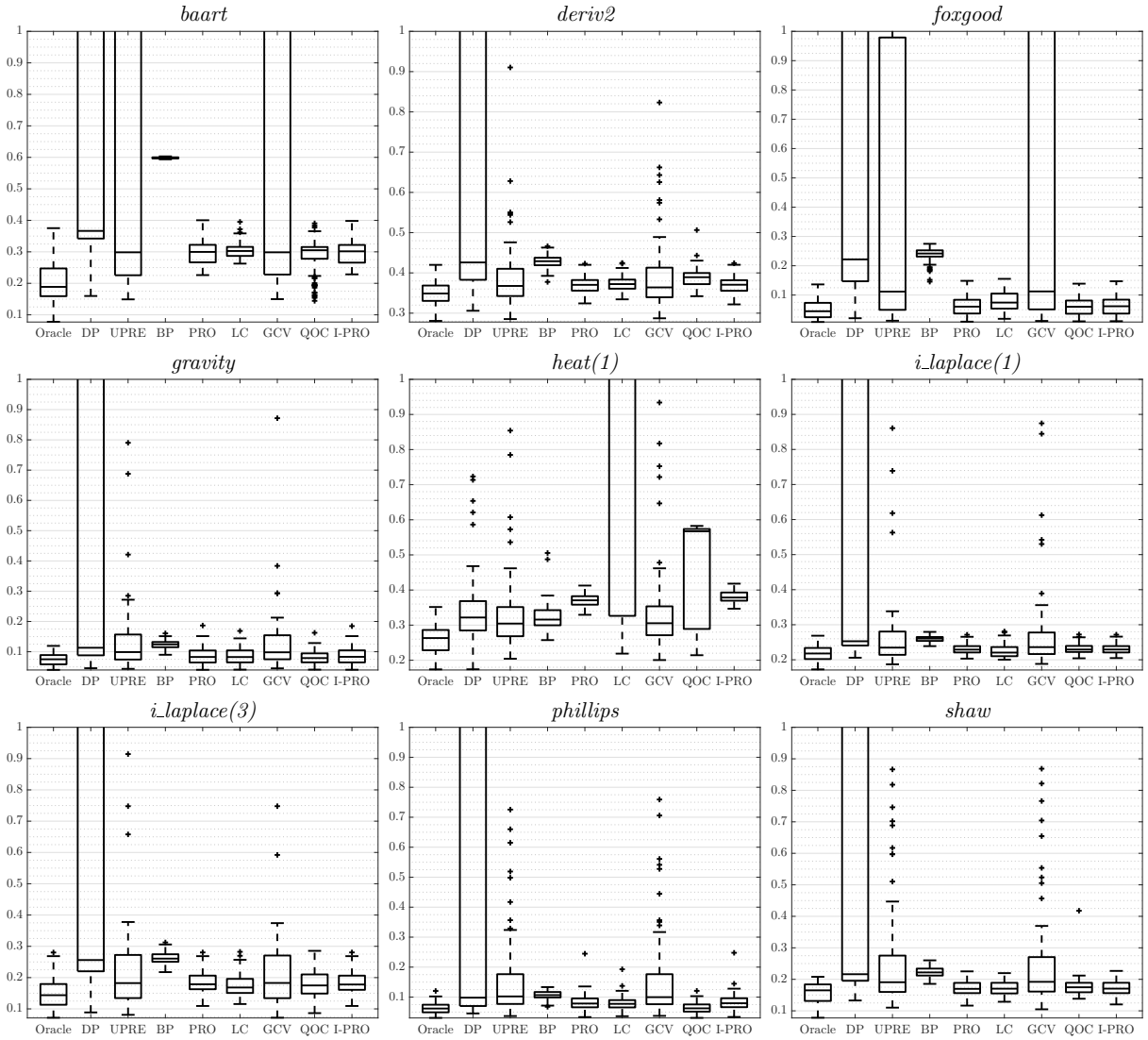
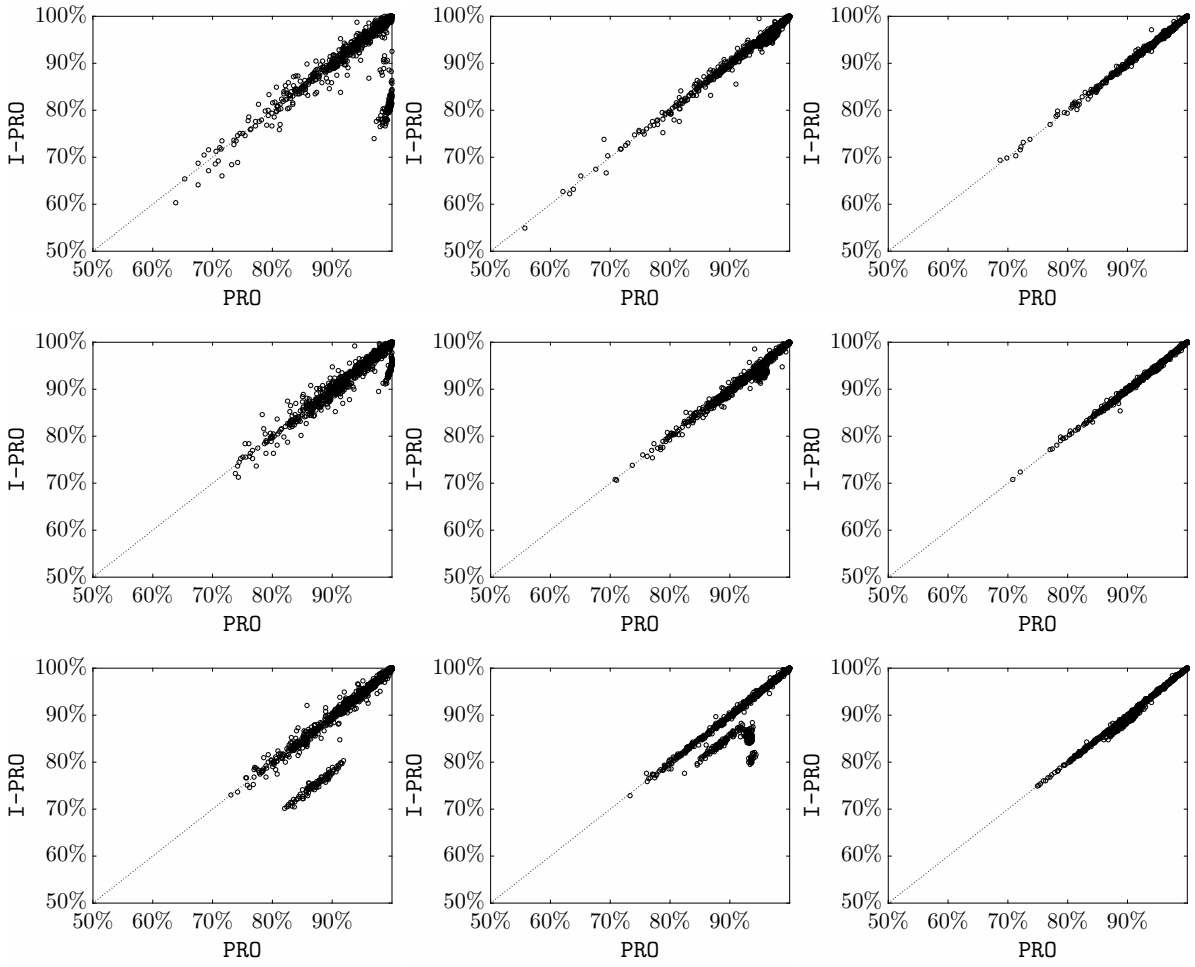


Figure 4: Box plots of the ℓ_2-norm error distributions over 100 samples for nine 1d problems, with SNR $\xi = 20$ and $n = 256$.

plots show distributions with small variances, whereas that of the other methods have broad tails; see also Fig. 3 for the illustration of the excellent stability of the `PRO` and `I-PRO` solutions. Overall, BP works fairly well, but it is below par on the example *baart*. Fig. 4 indicates that among all eight rules, `PRO`, `I-PRO` and QOC emerge as the strongest contenders.

Now we provide numerical insights into the fixed point iteration in the `I-PRO` algorithm, when σ is unknown. Numerically, it always converges rapidly, and the corresponding fixed point equation has few fixed points, if not unique globally. (It was proved to be unique within the interval $[0, s_1^2/2]$ in Proposition 1.) However, the convergence of the algorithm is necessary but not sufficient for ensuring the good performance of `I-PRO`: the converged solution can potentially be very bad! Remarkably, `I-PRO` estimates are always very close to that by `PRO`, which is consistently observed on all benchmark examples. This is evident by comparing the 7-th column with the 11-th column of Tables 1 and 2: the reconstruction errors of `PRO` and `I-PRO` are nearly identical, except for *heat(5)*. In the exceptional case, i.e., *heat(5)* with $\xi = 10$ and $n = 64$, the condition number is very small (around 3) so the problem is actually well posed. In this case, the efficiency of `PRO` is about 99%, while the efficiency of `I-PRO` is only about 56%. This difference originates from the fixed point iteration: a close inspection shows that the fixed point



(a) $n = 64$ (b) $n = 256$ (c) $n = 1024$

Figure 5: Scatter plots of the efficiency of `PRO` and `I-PRO`. The top, middle and bottom rows are for $\xi = 10$, $\xi = 20$ and $\xi = 40$, respectively.

iteration converges to an α value very close to zero. Practically, this seems not a serious restriction, since reconstructions actually do not change much when varying α near zero, due to the well-conditioned nature of the specific problem. Although not proved, the fixed point iteration is observed to work well for all ill-conditioned problems.

In Fig. 5, we show the scattered plots of the efficiency of `PRO` and `I-PRO`. The `I-PRO` estimates are mostly close to that of `PRO`, except outliers for `I-PRO` on the well-conditioned problem, i.e., *heat(5)*, which concurs with the observation from Tables 1 and 2. `I-PRO` can also provide an estimation on SNR, which is generally very close to the true values. The precise mechanism is unclear but it essentially underpins the reliability of `I-PRO` for parameter choice when σ^2 is unknown.

5.2 Large-scale problems

Now we illustrate the proposed `PRO` and `I-PRO` on three tomography examples from public package AIR tools, i.e., *paralleltomo* (parallel beam tomography), *fanbeamtomo* (fan beam tomography) and *seismictomo* (seismic tomography) (available at `http://people.compute.dtu.dk/pcha/AIRtoolsII/index.html`, accessed on July 1, 2019). The problem parameters are taken as follows: for *paralleltomo* and *fanbeamtomo*, the number of discretization intervals in both dimensions is fixed at $\ell = 128$ (i.e., the domain consists of ℓ^2 cells), 175 parallel rays for each angle $\theta \in \{0, 1, ..., 179\}$; For *seismictomo*, we

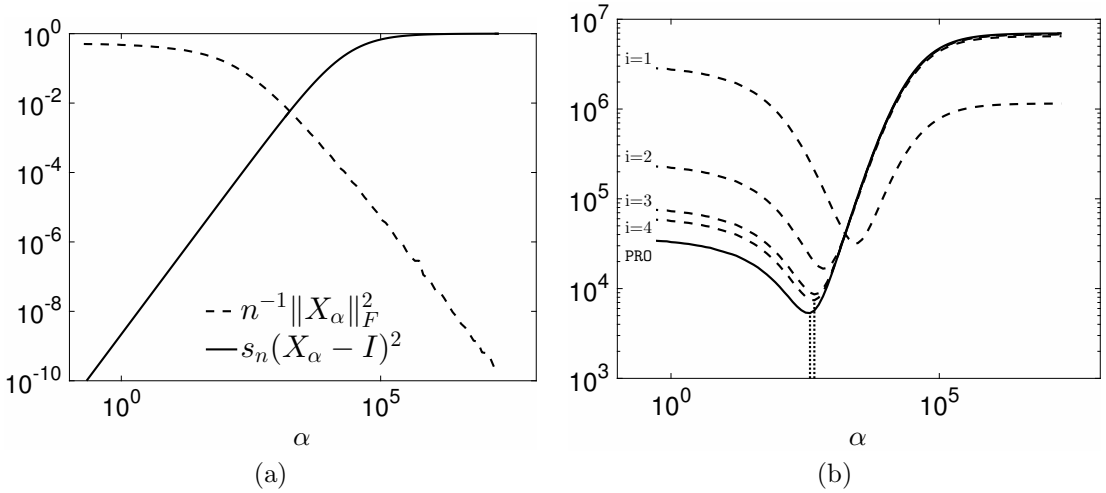
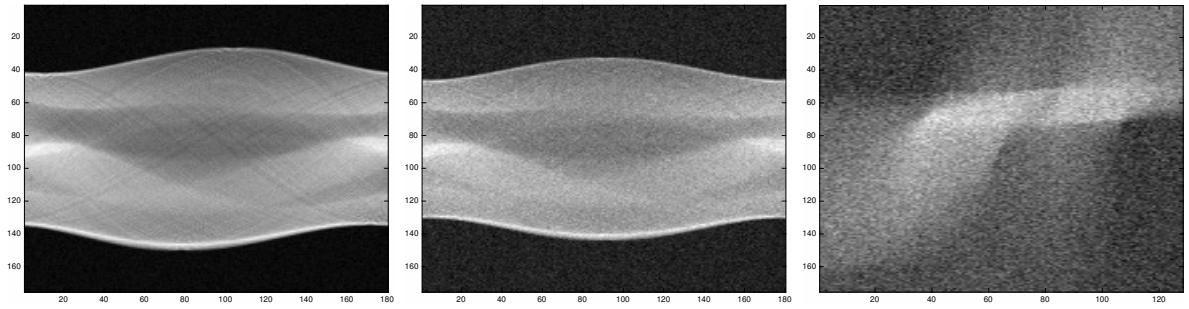


Figure 7: (a) The noise amplification error $n^{-1}\|X_\alpha\|_F^2$ and approximation error $s_n(I-X_\alpha)^2$ (lower bound) versus α for *paralleltomo*. (b): $T_{(\rho^2,\sigma^2)}(\alpha)$ (solid line) and its successive approximations $T_{(\hat{\rho^2},\hat{\sigma^2})}(\alpha)$ with `I-PRO`. `I-PRO` converges after four iterations (i.e., $i = 4$). The minimizers of these two functions provide the `PRO` and `I-PRO` solutions showed in the left panels of Fig. 8.

consider the same discretization $\ell = 128$ with 128 sources and 175 receivers (seismographs). See [17] for a complete description of these problems. For each example, we consider three noise levels (with SNR $\xi = 10, 20, 30$), each with 50 realizations η_i, $i = 1, \ldots, 50$, i.i.d. Gaussian noise with mean zero and variance σ^2. See Fig. 6 for exemplary noisy sinograms. In the implementation, for `PRO`, we take 100 α values equidistributed in a logarithmic scale over the interval $(10^{-8}, 10^{-2})\|A\|^2/2$ (as prescribed by Proposition 1) to find the minimizer of the function $T_{(\rho^2,\sigma^2)}(\alpha)$ (over the sample points), and finally compute the reconstructions using LSQR [36]. Meanwhile, `I-PRO` implements the fixed point iteration given in the `I-PRO` algorithm, initialized at the value $\alpha_1 := \alpha^{(50)}$ (from the samples). In all experiments below, the algorithm converges in less than 10 iterations.



(a) *fanbeamtomo*, $\xi = 30$ (b) *paralleltomo*, $\xi = 20$ (c) *seismictomo*, $\xi = 10$

Figure 6: Sinograms of tomography examples at various SNRs.

Since both terms of $T_{(\rho^2,\sigma^2)}(\alpha)$ depend only on the matrix A, they can be precomputed. In Fig. 7(a), we show the two terms $s_n(I - X_\alpha)^2$ and $\|X_\alpha\|_F^2$ as a function of the regularization parameter α for *paralleltomo*, and in Fig. 7(b), the sequence of four intermediate objective functions $T_{\hat{\rho}^2,\hat{\sigma}^2}(\alpha)$ by the `I-PRO` algorithm. The iterative process returns an α value close to the `PRO` value.

Due to the large number of cells ($\ell^2 \simeq 1.6 \times 10^4$), the variance of the reconstructions with respect to the noise realization is very small. In Table 3, we show the efficiency of the methods. For all three examples, all eight rules can choose a suitable regularization parameter, and in terms of efficiency, `PRO` and `I-PRO` perform better than existing methods when SNR is low and slightly worse when SNR is high, for *paralleltomo* and *fanbeamtomo*. This is attributed to the fact that by construction, `PRO` and `I-PRO`

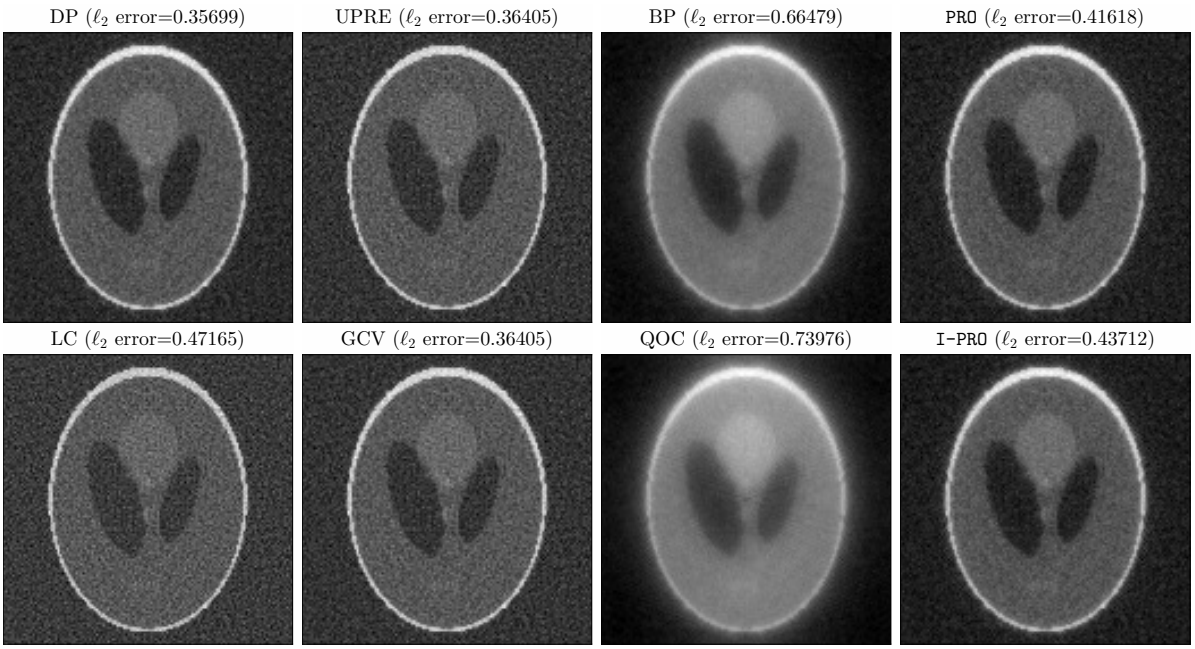


Figure 8: Reconstructions for *paralleltomo*, for the noisy sinogram showed in Fig. 6(b).

provide conservative estimates of the optimal parameter (with respect to predictive risk). In Fig. 8, we show exemplary reconstructions for *paralleltomo*.

6 Conclusion

In this work, we have proposed a new criterion, termed as PRO, to choose the crucial Tikhonov regularization parameter for discrete linear inverse problems. It is based on minimizing a lower bound of the predictive risk, and can handle effectively both known and unknown noise levels. In the latter case, we proposed an iterative scheme, i.e., I-PRO, which alternates between estimating the noise level and minimizing the predictive risk. Extensive numerical simulations show that both PRO and I-PRO are not only competitive with six existing choice rules including discrepancy principle, unbiased predictive risk estimator, balancing principle, generalized cross validation, L-curve criterion and quasi-optimality criterion) in terms of accuracy, but also can yield more stable results in terms of reliability for small-sized samples. Moreover, the methods apply also to large-scale inverse problems, and can produce solutions with improved accuracy for data with low signal-to-noise ratio. One preliminary theoretical analysis was provided to show several interesting properties of PRO and I-PRO, and the compelling empirical results promote further analysis, especially regularizing property and convergence rates, as well as developing extensions to other linear regularization techniques.

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement no. 640216. Federico Benvenuto thanks the National Group of Scientific Computing (GNCS-INDAM) that supported this research. The work of B. Jin is partially supported by UK EPSRC EP/T000864/1.

problem	stat	ℓ_2 error	σ known, ρ unknown				σ and ρ unknowns			
	ξ	Oracle	DP	UPRE	BP	PRO	LC	GCV	QOC	I-PRO
seismic	10	0.35	99.8%,	74.5%,	41.2%,	93.5%,	87.9%,	74.5%,	93.2%,	93.5%
	20	0.24	99.4%,	80.0%,	74.8%,	99.4%,	48.5%,	80.0%,	99.1%,	99.4%
	30	0.17	100.0%,	90.9%,	93.9%,	99.6%,	43.5%,	84.1%,	99.6%,	99.6%
fanbeam	10	0.55	99.8%,	88.6%,	89.2%,	98.7%,	99.0%,	88.6%,	72.3%,	98.7%
	20	0.37	98.8%,	94.9%,	55.9%,	87.2%,	75.2%,	94.9%,	51.6%,	83.5%
	30	0.24	96.6%,	97.8%,	33.9%,	76.9%,	82.9%,	97.8%,	96.5%,	61.5%
parallel	10	0.54	99.4%,	90.0%,	90.2%,	97.9%,	99.4%,	90.0%,	70.5%,	97.9%
	20	0.34	97.4%,	95.8%,	51.9%,	83.3%,	73.8%,	95.8%,	46.6%,	79.2%
	30	0.19	94.1%,	93.9%,	27.0%,	68.2%,	71.3%,	93.9%,	98.2%,	52.1%

Table 3: Performance comparison between the choice rules on the AIR tools dataset. The first column indicates the problem, the second gives the SNR of the data, the third shows the oracle ℓ_2-norm error, ε_o, and the 4th–11th columns: the median efficiency of the methods.

References

[1] H. Avron and S. Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *J. ACM*, 58(2):Art. 8, 17, 2011.

[2] F. Bauer and T. Hohage. A Lepskij-type stopping rule for regularized Newton methods. *Inverse Problems*, 21(6):1975–1991, 2005.

[3] F. Bauer and S. Kindermann. The quasi-optimality criterion for classical inverse problems. *Inverse Problems*, 24(3):035002, 20 pp., 2008.

[4] M. Bertero and P. Boccacci. *Introduction to Inverse Problems in Imaging*. IOP Publishing, Bristol, 1998.

[5] P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403, 1978/79.

[6] C.-A. Deledalle, S. Vaiter, J. Fadili, and G. Peyré. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM J. Imaging Sci.*, 7(4):2448–2487, 2014.

[7] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer, Dordrecht, 1996.

[8] N. P. Galatsanos and A. K. Katsaggelos. Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation. *IEEE Trans. Imag. Proc.*, 1(3):322–336, 1992.

[9] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[10] U. Hämarik, R. Palm, and T. Raus. On minimization strategies for choice of the regularization parameter in ill-posed problems. *Numer. Funct. Anal. Optim.*, 30(9-10):924–950, 2009.

[11] U. Hämarik and T. Raus. About the balancing principle for choice of the regularization parameter. *Numer. Funct. Anal. Optim.*, 30(9-10):951–970, 2009.

[12] M. Hanke and T. Raus. A general heuristic for choosing the regularization parameter in ill-posed problems. *SIAM J. Sci. Comput.*, 17(4):956–972, 1996.

[13] P. C. Hansen. The truncated SVD as a method for regularization. *BIT*, 27(4):534–553, 1987.

[14] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.*, 34(4):561–580, 1992.

[15] P. C. Hansen. Regularization tools version 3.0 for Matlab 5.2. *Numerical Algor.*, 20(2):195–196, 1999.

[16] P. C. Hansen and D. P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503, 1993.

[17] P. C. Hansen and M. Saxild-Hansen. AIR-tools—a MATLAB package of algebraic iterative reconstruction methods. *J. Comput. Appl. Math.*, 236(8):2167–2178, 2012.

[18] M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Comm. Statist. Simul. Comput.*, 19(2):433–450, 1990.

[19] K. Ito and B. Jin. *Inverse Problems: Tikhonov Theory and Algorithms*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2015.

[20] K. Ito and B. Jin. Regularized linear inversion with randomized singular value decomposition. Preprint, arXiv:1909.01947, 2019.

[21] K. Ito, B. Jin, and T. Takeuchi. A regularization parameter for nonsmooth Tikhonov regularization. *SIAM J. Sci. Comput.*, 33(3):1415–1438, 2011.

[22] B. Jin and D. A. Lorenz. Heuristic parameter-choice rules for convex variational regularization based on error estimates. *SIAM J. Numer. Anal.*, 48(3):1208–1229, 2010.

[23] S. Kindermann. Convergence analysis of minimization-based noise level-free parameter choice rules for linear ill-posed problems. *Electron. Trans. Numer. Anal.*, 38:233–257, 2011.

[24] S. Kindermann and A. Neubauer. On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. *Inverse Probl. Imaging*, 2(2):291–299, 2008.

[25] S. Kindermann, S. Pereverzyev, Jr., and A. Pilipenko. The quasi-optimality criterion in the linear functional strategy. *Inverse Problems*, 34(7):075001, 24, 2018.

[26] L. Landweber. An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.*, 73:615–624, 1951.

[27] O. V. Lepskiĭ. A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatnost. i Primenen.*, 35(3):459–470, 1990.

[28] E. Levin and A. Y. Meltzer. Estimation of the regularization parameter in linear discrete ill-posed problems using the Picard parameter. *SIAM J. Sci. Comput.*, 39(6):A2741–A2762, 2017.

[29] H. Li and F. Werner. Empirical risk minimization as parameter choice rule for general linear regularization methods. *Ann.l'Institut Henri Poincaré Prob. Stat.*, page in press. arXiv: 1703.07809, 2018.

[30] F. Lucka, K. Proksch, C. Brune, N. Bissantz, M. Burger, H. Dette, and F. Wübbeling. Risk estimators for choosing regularization parameters in ill-posed problems—properties and limitations. *Inverse Probl. Imaging*, 12(5):1121–1155, 2018.

[31] M. A. Lukas, F. R. De Hoog, and R. S. Anderssen. Performance of robust GCV and modified GCV for spline smoothing. *Scand. J. Stat.*, 39(1):97–115, 2012.

[32] C. L. Mallows. Some comments on c_p. *Technometrics*, 15(4):661–675, 1973.

[33] P. Mathé. The Lepskiĭ principle revisited. *Inverse Problems*, 22(3):L11–L15, 2006.

[34] P. Mathé and S. V. Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(3):789–803, 2003.

[35] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer-Verlag, New York, 1984.

[36] C. C. Paige and M. A. Saunders. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8(1):43–71, 1982.

[37] R. A. Polyak. Complexity of the regularized Newton's method. *Pure Appl. Funct. Anal.*, 3(2):327–347, 2018.

[38] S. Ramani, T. Blu, and M. Unser. Monte-Carlo sure: a black-box optimization of regularization parameters for general denoising algorithms. *IEEE Trans. Image Process.*, 17(9):1540–1554, 2008.

[39] T. Regińska. A regularization parameter in discrete ill-posed problems. *SIAM J. Sci. Comput.*, 17(3):740–749, 1996.

[40] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[41] A. M. Thompson, J. C. Brown, J. W. Kay, and D. M. Titterington. A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):326–339, 1991.

[42] A. N. Tihonov, V. B. Glasko, and J. A. Kriksin. On the question of quasi-optimal choice of a regularized approximation. *Dokl. Akad. Nauk SSSR*, 248(3):531–535, 1979.

[43] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York-Toronto, Ont.-London, 1977.

[44] A. N. Tikhonov and V. B. Glasko. Use of the regularization method in nonlinear problems. *USSR Comput. Math. Math. Phys.*, 5(3):93–107, 1965.

[45] C. R. Vogel. Non-convergence of the L-curve regularization parameter selection method. *Inverse Problems*, 12(4):535–547, 1996.

[46] G. Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, PA, 1990.