Risk Estimators for Choosing Regularization Parameters in Ill-Posed Problems - Properties and Limitations

Felix Lucka * Katharina Proksch† Christoph Brune‡ Nicolai Bissantz§
Martin Burger¶ Holger Dette‖ Frank Wübbeling**

January 19, 2017

Abstract

This paper discusses the properties of certain risk estimators recently proposed to choose regularization parameters in ill-posed problems. A simple approach is Stein's unbiased risk estimator (SURE), which estimates the risk in the data space, while a recent modification (GSURE) estimates the risk in the space of the unknown variable. It seems intuitive that the latter is more appropriate for ill-posed problems, since the properties in the data space do not tell much about the quality of the reconstruction. We provide theoretical studies of both estimators for linear Tikhonov regularization in a finite dimensional setting and estimate the quality of the risk estimators, which also leads to asymptotic convergence results as the dimension of the problem tends to infinity. Unlike previous papers, who studied image processing problems with a very low degree of ill-posedness, we are interested in the behavior of the risk estimators for increasing ill-posedness. Interestingly, our theoretical results indicate that the quality of the GSURE risk can deteriorate asymptotically for ill-posed problems, which is confirmed by a detailed numerical study. The latter shows that in many cases the GSURE estimator leads to extremely small regularization parameters, which obviously cannot stabilize the reconstruction. Similar but less severe issues with respect to robustness also appear for the SURE estimator, which in comparison to the rather conservative discrepancy principle leads to the conclusion that regularization parameter choice based on unbiased risk estimation is not a reliable procedure for ill-posed problems. A similar numerical study for sparsity regularization demonstrates that the same issue appears in nonlinear variational regularization approaches.

Keywords: Ill-posed problems, regularization parameter choice, risk estimators, Stein's method, discrepancy principle.

*Centre for Medical Image Computing, University College London, WC1E 6BT London, UK email: f.lucka@ucl.ac.uk

†Institut fÃ¼r Mathematische Stochastik, Georg-August-Universität Göttingen, Goldschmidtstrasse 7, 37077 Göttingen, Germany, e-mail: kproksc@uni-goettingen.de

‡Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, e-mail:c.brune@utwente.nl

§Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany, e-mail: nicolai.bissantz@ruhr-uni-bochum.de

¶Institut für Numerische und Angewandte Mathematik, Westfälische Wilhelms-Universität (WWU) Münster. Einsteinstr. 62, D 48149 Münster, Germany. e-mail: martin.burger@wwu.de

‖Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany, e-mail: holger.dette@ruhr-uni-bochum.de

**Institut für Numerische und Angewandte Mathematik, Westfälische Wilhelms-Universität (WWU) Münster. Einsteinstr. 62, D 48149 Münster, Germany. e-mail: frank.wuebbeling@wwu.de

1 Introduction

Choosing suitable regularization parameters is a problem as old as regularization theory, which has seen a variety of approaches both from deterministic (e.g. L-curve criteria, [18, 17]) or statistical perspectives (e.g. Lepskij principles, [2, 20]), respectively in between (e.g. discrepancy principles motivated by deterministic bounds or noise variance, cf. [28, 3]). Recently, another class of statistical parameter choice rules based on risk estimation, more precisely using Stein's unbiased risk estimation [27], was introduced in problems related to image processing ([9, 29, 30, 12, 33, 10, 22, 32, 31, 25]). In addition to a classical Stein unbiased risk estimator (SURE), several authors have considered a generalized version (GSURE, [30, 11, 15]), which measures risk in the space of the unknown rather than in the data space and hence seems more appropriate for ill-posed problems. Previous investigations show that the performance of such parameter choice rules is reasonable in many different settings (cf. [16, 34, 8, 1, 26, 23, 13]). However, the problems considered in these works are very mildly ill-posed and therefore, a first motivation of this paper is to further study the properties of parameter choice by SURE and GSURE in Tikhonov-type regularization methods more systematically in dependence of the ill-posedness of the problem and the degree of smoothness of the unknown exact solution. For this purpose we provide a theoretical analysis of the quality of unbiased risk estiamtors in the case of linear Tikhonov regularization. Additionally we carry out extensive numerical investigations on appropriate model problems. While in very mildly ill-posed settings the performances of the parameter choice rules under consideration are reasonable and comparable, our investigations yield various interesting results and insights in ill-posed settings. For instance, we demonstrate that GSURE shows a rather erratic behaviour as the degree of ill-posedness increases. The observed effects are so strong that the meaning of a parameter chosen according to this particular criterion is unclear.
A second motivation of this paper is to study the discrepancy principle as a reference method and as we shall see it can indeed be put in a very similar context and analyzed by the same techniques. Although the popularity of the discrepancy principle is decreasing recently in favor of choices using more statistical details, our findings show that it is still more robust for ill-posed problems than risk-based parameter choices. The conservative choice by the discrepancy principle is well-known to rather overestimate the optimal parameter, but on the other hand it avoids to choose too small regularization as risk-based methods often do. In the latter case the reconstruction results are completely deteriorated, while the discrepancy principle yields a reliable, though not optimal, reconstruction.
Throughout the paper we consider a (discrete) inverse problem of the form

$$y = Ax^* + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}^m$ is a vector of observations, $A \in \mathbb{R}^{m\times n}$ is some known but possibly ill-conditioned matrix, and $\varepsilon \in \mathbb{R}^m$ is a noise vector. We assume that ε consists of independent and identically distributed (*i.i.d.*) Gaussian errors, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_m)$. The vector $x^* \in \mathbb{R}^n$ denotes the (unknown) exact solution to be reconstructed from the observations. In order to find an estimate $\hat{x}(y)$ of x^*, we apply a variational regularization method:

$$\hat{x}_\alpha(y) = \operatorname*{argmin}_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - y\|_2^2 + \alpha R(x), \tag{2}$$

where R is assumed convex and such that the minimizer is unique for positive regularization

parameter α. In what follows the dependence of $\hat{x}_\alpha(y)$ on α and the data y may be dropped where it is clear without ambiguity that $\hat{x} = \hat{x}_\alpha(y)$.
In practice there are two choices to be made: First, a regularization functional R needs to be specified in order to appropriately represent a-priori knowledge about solutions and second, a regularization parameter α needs to be chosen in dependence of the data y. The ideal parameter choice would minimize a difference between $\hat{x}_\alpha(y)$ and x^* over all α, which obviously cannot be computed and is hence replaced by a parameter choice rule that tries to minimize a worst-case or average error to the unknown solution, which can be referred to as a risk minimization. In the practical case of having a single observation only, the risk based on average error needs to be replaced by an estimate as well, and unbiased risk estimators that will be detailed in the following are a natural choice.
For the sake of a clearer presentation of methods and results we first focus on linear Tikhonov regularization, i.e.,

$$R(x) = \frac{1}{2}\|x\|_2^2,$$

leading to the explicit Tikhonov estimator

$$\hat{x}_\alpha(y) = T_\alpha y := (A^*A + \alpha I)^{-1}A^*y. \tag{3}$$

In this setting, a natural distance for measuring the error of $\hat{x}_\alpha(y)$ is given by its ℓ_2-distance to x^*. Thus, we define

$$\alpha^* := \operatorname*{argmin}_{\alpha \geqslant 0} \ \|\hat{x}_\alpha(y) - x^*\|_2^2$$

as the optimal, but inaccessible, regularization parameter. Many different rules for the choice of the regularization parameter α are discussed in the literature. Here, we focus on strategies that rely on an accurate estimate of the noise variance σ^2. A classical example of such a rule is given by the *discrepancy principle*: The regularization parameter $\hat{\alpha}_{\mathrm{DP}}$ is given as the solution of the equation

$$\|A\hat{x}_\alpha(y) - y\|_2^2 = m\sigma^2. \tag{4}$$

The discrepancy principle is robust and easy-to-implement for many applications (cf. [4, 19, 24]) and is based on the heuristic argument, that $x_\alpha(y)$ should only explain the data y up to the noise level. Several other parameter choice rules are based on Stein's famous unbiased risk estimator (SURE).
The basic idea is to choose the α that minimizes the estimated quadratic risk function

$$\hat{\alpha}^*_{\mathrm{SURE}} \in \operatorname*{argmin}_{\alpha \geqslant 0} \ \mathrm{R}_{\mathrm{SURE}}(\alpha) := \operatorname*{argmin}_{\alpha \geqslant 0} \ \mathbb{E}\left[\|Ax^* - A\hat{x}_\alpha(y)\|_2^2\right] \tag{5}$$

Since $\mathrm{R}_{\mathrm{SURE}}$ depends on the unknown vector x^*, we replace it by an unbiased estimate:

$$\hat{\alpha}_{\mathrm{SURE}} \in \operatorname*{argmin}_{\alpha \geqslant 0} \ \mathrm{SURE}(\alpha, y) := \operatorname*{argmin}_{\alpha \geqslant 0} \ \|y - A\hat{x}_\alpha(y)\|_2^2 - m\sigma^2 + 2\sigma^2 \mathrm{df}_\alpha(y) \tag{6}$$

with

$$\mathrm{df}_\alpha(y) = \mathrm{tr}\left(\nabla_y \cdot A\hat{x}_\alpha(y)\right).$$

As an analogue of the SURE-criterion, a generalized version (GSURE) is often considered. In contrast to SURE, which aims at optimizing the MSE in the image of the operator, GSURE operates in the domain instead and considers the MSE of the reconstruction of x:

$$\hat{\alpha}^*_{\mathrm{GSURE}} \in \operatorname*{argmin}_{\alpha \geqslant 0} \ \mathrm{R}_{\mathrm{GSURE}}(\alpha) := \operatorname*{argmin}_{\alpha \geqslant 0} \ \mathbb{E}\left[\|\Pi(x^* - \hat{x}_\alpha(y))\|_2^2\right],$$

where $\Pi := A^+A$ denotes the orthogonal projector onto the range of A^*. Again, we replace $\mathrm{R}_{\mathrm{GSURE}}$ by an unbiased estimator to obtain

$$\hat{\alpha}_{\mathrm{GSURE}} \in \operatorname*{argmin}_{\alpha \geqslant 0} \mathrm{GSURE}(\alpha, y) := \operatorname*{argmin}_{\alpha \geqslant 0} \|x_{\mathrm{ML}}(y) - \hat{x}_\alpha(y)\|_2^2 - \sigma^2 \mathrm{tr}\left((AA^*)^+\right) + 2\sigma^2 \mathrm{gdf}_\alpha(y) \tag{7}$$

with

$$\mathrm{gdf}_\alpha(y) = \mathrm{tr}((AA^*)^+ \nabla_y A\hat{x}_\alpha(y)), \qquad x_{\mathrm{ML}} = A^+y = A^*(AA^*)^+y,$$

where M^+ denotes the Pseudoinverse of M.
Notice that all parameter choice rules depend on the data y and hence on the random errors $\varepsilon_1, \ldots, \varepsilon_m$. Therefore, $\hat{\alpha}_{\mathrm{DP}}$, $\hat{\alpha}_{\mathrm{SURE}}$ and $\hat{\alpha}_{\mathrm{GSURE}}$ are random variables, described in terms of their probability distributions. We first investigate these distributions by numerical simulation studies. The results point to several problems of the presented parameter choice rules, in particular of GSURE, and motivate our further theoretical investigation.
In the next section, we will describe a simple inverse problem scenario in terms of quadratic Tikhonov regularization and fix the setting and notations both for further numerical simulation as well as the analysis of the risk based estimators. The latter will be carried out in Section 3 and supplemented by an exhaustive numerical study in Section 4. Finally we extend the numerical investigation in Section 5 to a sparsity-promoting LASSO-type regularization, for which we find similar behaviour. Conclusions are given in Section 6.

2 Risk Estimators for Quadratic Regularization

In the following we discuss the setup in the case of a quadratic regularization functional $R(x) = \frac{1}{2}\|x\|^2$, i.e. we recover the well-known linear Tikhonov regularization scheme. The linearity can be used to simplify arguments and gain analytical insight in the next section.

2.1 Singular System and Risk Representations

Considering a quadratic regularization allows to analyze $\hat{x}_\alpha$ in a singular system of A in a convenient way. Let $r = \mathrm{rank}(A)$, $l = \min(n, m)$. Let

$$A = U\Sigma V^*, \ \Sigma = \mathrm{diag}\,(\gamma_1, \ldots, \gamma_l) \in \mathbb{R}^{m\times n}, \ \gamma_1 \geq \ldots \geq \gamma_r > 0, \ \gamma_{r+1} \ldots \gamma_m := 0$$

denote a singular value decomposition of A with

$$U = (u_1, \ldots, u_m) \in \mathbb{R}^{m\times m}, V = (v_1, \ldots, v_n) \in \mathbb{R}^{n\times n} \text{ unitary.}$$

Defining

$$y_i = \langle u_i, y\rangle, \qquad x_i^* = \langle v_i, x^*\rangle, \qquad \tilde{\epsilon}_i = \langle u_i, \epsilon\rangle \tag{8}$$

we can rewrite model (1) in its spectral form

$$y_i = \gamma_i x_i^* + \tilde{\varepsilon}_i, \ i = 1 \ldots l; \ y_i = \tilde{\varepsilon}_i, \ i = l+1 \ldots m,$$

where $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_m$ are still *i.i.d.* $\sim \mathcal{N}(0, \sigma^2)$. We will express some more terms in the singular system that are frequently used throughout this paper. In particular, we have for x_{ML}, the

regularized solution and its norm

$$x_{ML} = A^+ y = V\Sigma^+ U^* y, \text{ with } \Sigma^+ = \operatorname{diag}(\frac{1}{\gamma_1}, \dots, \frac{1}{\gamma_r}, 0 \dots 0) \in \mathbb{R}^{n\times m}$$
$$\hat{x}_\alpha(y) = (A^*A + \alpha I)^{-1} A^* y =: V\Sigma_\alpha^+ U^* y, \quad \text{with} \quad \Sigma_\alpha^+ = \operatorname{diag}\left(\frac{\gamma_i}{\gamma_i^2 + \alpha}\right) \in \mathbb{R}^{n\times m}$$
$$\|\hat{x}_\alpha\|_2^2 = \sum_{i=1}^m \frac{\gamma_i^2}{(\gamma_i^2+\alpha)^2} y_i^2$$

as well as the residual and distance to the maximum likelihood estimate

$$\|A\hat{x}_\alpha - y\|_2^2 = \sum_{i=1}^m \frac{\alpha^2}{(\gamma_i^2+\alpha)^2} y_i^2. \tag{9}$$
$$\|x_{\mathrm{ML}} - \hat{x}_\alpha\|_2^2 = \|A^*(AA^*)^+ y - (A^*A + \alpha I)^{-1} A^* y\|_2^2 = \|V\left(\Sigma^+ - \Sigma_\alpha^+\right) U^* y\|_2^2$$
$$= \sum_{i=1}^r \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{(\gamma_i^2+\alpha)}\right)^2 y_i^2.$$

Based on the generalized inverse we compute

$$(AA^*)^+ = U(\Sigma\Sigma^*)^+ U^* = U\operatorname{diag}\left(\frac{1}{\gamma_1^2}, \dots, \frac{1}{\gamma_r^2}, 0, \dots, 0\right) U^*$$
$$A^*(AA^*)^+ A = V\operatorname{diag}(\underbrace{1, \dots, 1}_{r}, \underbrace{0, \dots, 0}_{n-r}) V^*,$$

which yields the degrees of freedom and the generalized degrees of freedom

$$\mathrm{df}_\alpha := \nabla_y \cdot A\hat{x} = \operatorname{tr}\left(A(A^*A + \alpha I)^{-1} A^*\right) = \sum_{i=1}^r \frac{\gamma_i^2}{\gamma_i^2 + \alpha}$$
$$\mathrm{gdf}_\alpha := \operatorname{tr}((AA^*)^+ \nabla_y \cdot A\hat{x}) = \operatorname{tr}\left((AA^*)^+ A(A^*A + \alpha I)^{-1} A^*\right)$$
$$= \operatorname{tr}((\Sigma\Sigma^*)^+ \Sigma\Sigma_\alpha^{-1}) = \sum_{i=1}^r \frac{1}{\gamma_i^2} \gamma_i \frac{\gamma_i}{\gamma_i^2+\alpha} = \sum_{i=1}^r \frac{1}{\gamma_i^2+\alpha}.$$

Next, we derive the spectral representations of the parameter choice rules. For the discrepancy principle, we use (9) to define

$$\mathrm{DP}(\alpha, y) := \sum_{i=1}^m \frac{\alpha^2}{(\gamma_i^2+\alpha)^2} y_i^2 - m\sigma^2, \tag{10}$$

and now, (4) can be restated as $\mathrm{DP}(\hat{\alpha}_{\mathrm{DP}}, y) = 0$. For (6) and (7), we find

$$\mathrm{SURE}(\alpha, y) = \sum_{i=1}^m \frac{\alpha^2}{(\gamma_i^2+\alpha)^2} y_i^2 - m\sigma^2 + 2\sigma^2 \sum_{i=1}^m \frac{\gamma_i^2}{\gamma_i^2+\alpha} \tag{11}$$
$$\mathrm{GSURE}(\alpha, y) = \sum_{i=1}^r \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{\gamma_i^2+\alpha}\right)^2 y_i^2 - \sigma^2 \sum_{i=1}^r \frac{1}{\gamma_i^2} + 2\sigma^2 \sum_{i=1}^r \frac{1}{\gamma_i^2+\alpha}. \tag{12}$$

2.2 An Illustrative Example

We consider a simple imaging scenario which exhibits typical properties of inverse problems. The unknown function $x^*_\infty : [-1/2, 1/2] \to \mathbb{R}$ is mapped to a function $y_\infty : [-1/2, 1/2] \to \mathbb{R}$ by a periodic convolution with a compactly supported kernel of width $l \leq 1/2$:

$$y_\infty(s) = A_{\infty,l} x^*_\infty := \int_{-\frac{1}{2}}^{\frac{1}{2}} k_l(s-t)\, x^*_\infty(t)\, \mathrm{d}t, \quad s \in [-1/2, 1/2],$$

where the 1-periodic $C^\infty_0(\mathbb{R})$ function $k_l(t)$ is defined for $|t| \leq 1/2$ by

$$k_l(t) := \frac{1}{N_l} \begin{cases} \exp\left(-\frac{1}{1-t^2/l^2}\right) & \text{if} \quad |t| < l \\ 0 & l \leq |t| \leq 1/2 \end{cases}, \qquad N_l = \int_{-l}^{l} \exp\left(-\frac{1}{1-t^2/l^2}\right) \mathrm{d}t,$$

and continued periodically for $|t| > 1/2$. Examples of $k_l(t)$ are plotted in Figure 1(a). The normalization ensures that $A_{\infty,l}$ and suitable discretizations thereof have the spectral radius $\gamma_1 = 1$ which simplifies our derivations and the corresponding illustrations. The x^*_∞ used in the numerical examples is the sum of four delta distributions:

$$x^*_\infty(t) := \sum_{i=1}^{4} a_i \delta\left(b_i - \frac{1}{2}\right), \text{ with } a = [0.5, 1, 0.8, 0.5], \quad b = \left[\frac{1}{\sqrt{26}}, \frac{1}{\sqrt{11}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3/2}}\right].$$

The locations of the delta distributions approximate $[-0.3, -0.2, 0.1, 0.3]$ by irrational numbers which simplifies the discretization.

Discretization For a given number of degrees of freedom n, let

$$E^n_i := \left[\frac{i-1}{n} - \frac{1}{2}, \frac{i}{n} - \frac{1}{2}\right], \qquad i = 1, \ldots, n$$

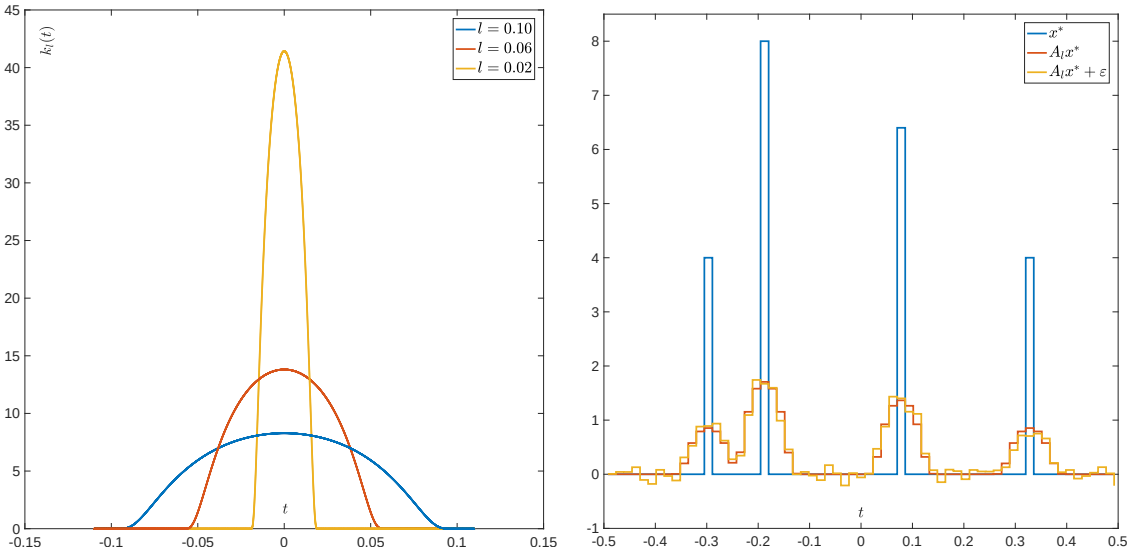
denote the equidistant partition of $[-1/2, 1/2]$ and $\psi^n_i(t) = \sqrt{n}\, \mathbb{1}_{E^n_i}(t)$ an ONB of piecewise constant functions over that partition. If we use m and n degrees of freedom to discretize range and domain of $A_{\infty,l}$, respectively, we arrive at the discrete inverse problem (1) with

$$(A_l)_{i,j} = \left\langle \psi^m_i, A_{\infty,l} \psi^n_j \right\rangle = \sqrt{mn} \int_{E^m_i} \int_{E^n_j} k_l(s-t)\, \mathrm{d}t\, \mathrm{d}s \tag{13}$$

$$x^*_j = \left\langle \psi^n_j, x^*_\infty \right\rangle = \sqrt{n} \int_{E^n_j} x^*_\infty(t)\, \mathrm{d}t = \sqrt{n} \sum_{i}^{4} a_i \mathbb{1}_{E^n_i} \delta\left(b_i - \frac{1}{2}\right)$$

The two dimensional integration in (13) is computed by the trapezoidal rule with equidistant spacing, employing 100×100 points to partition $E^m_i \times E^n_i$. Note that we drop the subscript l from A_l whenever the dependence on this parameter is not of importance for the argument being carried out.

As the convolution kernel k_l has mass 1 and the discretization was designed to be mass-preserving, we have $\gamma_1 = 1$ and the condition of A is given by $\mathrm{cond}(A) = 1/\gamma_r$, where $r = \mathrm{rank}(A)$. Figure 2 shows the decay of the singular values for various parameter settings and Table 1 lists the corresponding condition numbers.



(a) (b)

Figure 1: (a) The convolution kernel $k_l(t)$ for different values of l. (b) True solution x^*, clean data $A_l x^*$ and noisy data $A_l x^* + \varepsilon$ for $m = n = 64$, $l = 0.06$, $\sigma = 0.1$.

Table 1: Condition of A_l computed different values of $m = n$ and l.

	$l = 0.02$	$l = 0.04$	$l = 0.06$	$l = 0.08$	$l = 0.1$
$m = 16$	1.27e+0	1.75e+0	2.79e+0	6.77e+0	2.31e+2
$m = 32$	1.75e+0	6.77e+0	6.94e+1	6.88e+2	2.30e+2
$m = 64$	6.77e+0	6.88e+2	6.42e+2	1.51e+3	4.22e+3
$m = 128$	6.88e+2	1.51e+3	1.51e+4	4.29e+3	4.29e+4
$m = 256$	1.70e+3	4.70e+4	1.87e+6	4.07e+6	1.79e+6
$m = 512$	4.70e+4	1.11e+7	1.22e+7	2.12e+7	3.70e+7

Empirical Distributions Using the above formulas and $m = n = 64$, $l = 0.06$, $\sigma = 0.1$, we computed the empirical distibutions of the different parameter choice rules by evaluating (10), (11) and (12) on a fine logarithmical α-grid, i.e., $\log_{10}(\alpha_i)$ was increased linearly in between -40 and 40 with a step size of 0.01. We draw $N_\varepsilon = 10^6$ samples of ε. The results are displayed in Figures 3 and 4: In both figures, we use a logarithmic scaling of the empirical probabilities wherein empirical probabilities of 0 have been set to $1/(2N_\varepsilon)$. While this presentation complicates the comparison of the distributions as the probability mass is deformed, it facilitates the examination of small values and tails.

First, we observe in Figure 3(a) that $\hat{\alpha}_{\mathrm{DP}}$ typically overestimates the optimal α^*. However, it performs robustly and does not cause large ℓ_2-errors as can be seen in Figure 3(b). For $\hat{\alpha}_{\mathrm{SURE}}$ and $\hat{\alpha}_{\mathrm{GSURE}}$, the latter is not true: While being closer to α^* than $\hat{\alpha}_{\mathrm{DP}}$ most often, and, as can be seen from the joint error histograms in Figure 4, producing smaller ℓ_2-errors most often, both distributions show outliers, i.e., occasionally, very small values of $\hat{\alpha}$ are estimated that cause large ℓ_2-errors. In the case of $\hat{\alpha}_{\mathrm{GSURE}}$, we even observe two clearly separated modes in the distributions. These findings motivate the theoretical examinations carried out in the

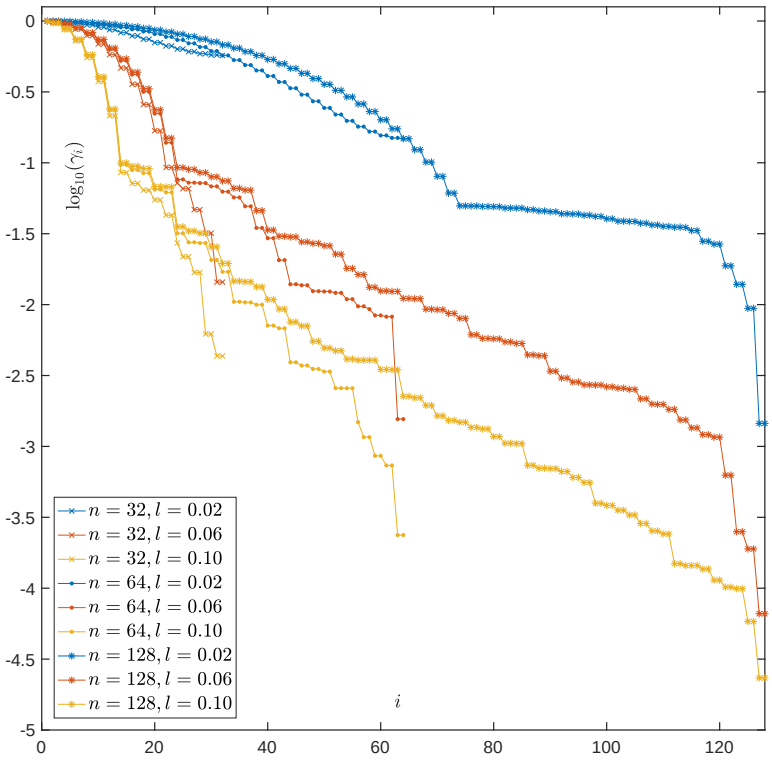
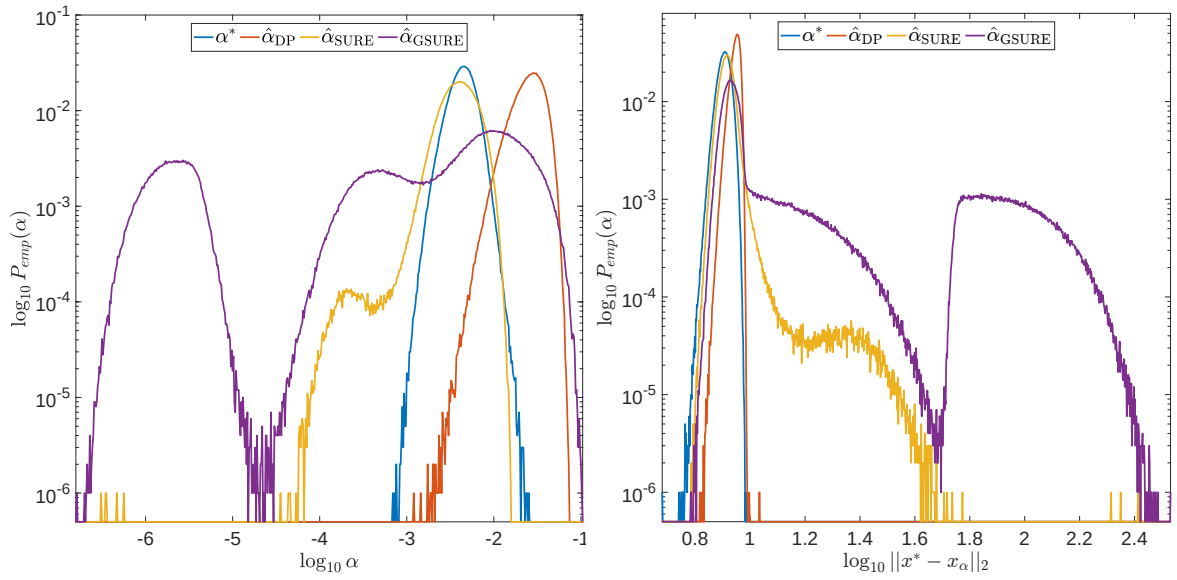


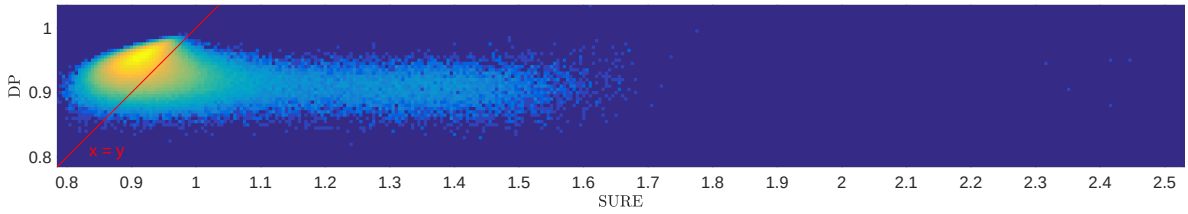
Figure 2: Decay of the singular values γ_i of A_l for different choices of m and l. As expected, increasing the width l of the convolution kernel leads to a faster decay. For a fixed l, increasing m corresponds to using a finer discretization and γ_i converges to the corresponding singular value of $A_{\infty,l}$, as can be seen for the largest γ_i, e.g., for $l = 0.02$.

following section.

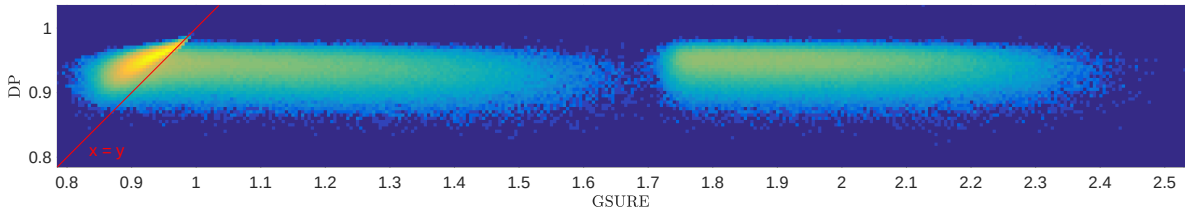


(a) (b)

Figure 3: Empirical probabilities of (a) $\hat{\alpha}$ and (b) the corresponding ℓ_2-error for different parameter choice rules using $m = n = 64$, $l = 0.06$, $\sigma = 0.1$ and $N_\varepsilon = 10^6$ samples of ε.

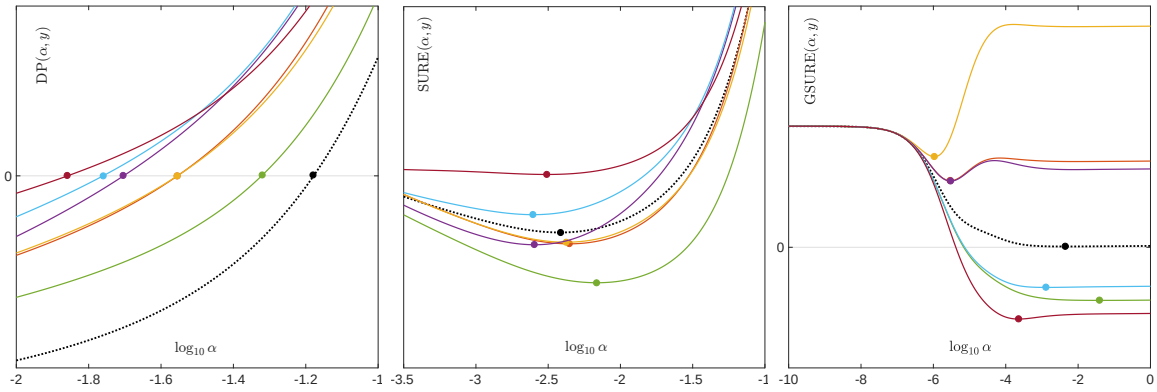


(a) Discrepancy principle vs SURE



(b) Discrepancy principle vs GSURE

Figure 4: Joint empirical probabilities of $\log_{10} \|x^* - x_{\hat{\alpha}}\|_2$ using $m = n = 64$, $l = 0.06$, $\sigma = 0.1$ and $N_\varepsilon = 10^6$ samples of ε (the histograms in Figure 3(b) are the marginal distributions thereof). As in Figure 3(b), the logarithms of the probabilities are displayed (here in form of a color-coding) to facilitate the identification of smaller modes and tails. The red line at $x = y$ divides the areas where one method performs better than the other: In (a), all samples falling into the area on the right of the red line correspond to a noise realization where the discrepancy principle leads to a smaller error than SURE.



(a) (b) (c)

Figure 5: True risk functions (black dotted line), their estimates for six different realizations y^k, $k = 1 \dots 6$ (solid lines), and their corresponding minima/roots (dots on the lines) in the setting described in Figure 1 using ℓ_2-regularization: (a) $\mathrm{DP}(\alpha, Ax^*)$ and $\mathrm{DP}(\alpha, y^k)$. (b) $\mathrm{R}_{\mathrm{SURE}}(\alpha)$ and $\mathrm{SURE}(\alpha, y^k)$. (c) $\mathrm{R}_{\mathrm{GSURE}}(\alpha)$ and $\mathrm{GSURE}(\alpha, y^k)$.

3 Properties of the Parameter Choice Rules for Quadratic Regularization

In this section we consider the theoretical (risk) properties of SURE, GSURE and the discrepancy principle.

Assumption 1. *For the sake of simplicity we only consider $m = n$ in this first analysis. Furthermore, we assume*

$$1 = \gamma_1 \geq \ldots \geq \gamma_m > 0 \tag{14}$$

and that $\|x^*\|_2^2 = O(m)$. Note that all assumptions are fulfilled in the numerical example we described in the previous section.

We mention that we consider here a rather moderate size of the noise, which remains bounded in variances as $m \to \infty$. A scaling corresponding to white noise in the infinite dimensional limit is rather $\sigma^2 \sim m$ and an inspection of the estimates below shows that the risk estimate is potentially far from the expected values in such cases additionally.

3.1 SURE-Risk

We start with an investigation of the well-known SURE risk estimate. Based on (11) and Stein's result, the representation for the risk is given as

$$\begin{aligned}
\mathrm{R}_{\mathrm{SURE}}(\alpha) &= \mathbb{E}[\mathrm{SURE}(\alpha, y)] \\
&= \sum_i \frac{\alpha^2}{(\gamma_i^2 + \alpha)^2} \mathbb{E}[y_i^2] - \sigma^2 m + 2\sigma^2 \sum_i \frac{\gamma_i^2}{\gamma_i^2 + \alpha} \\
&= \sum_i \frac{\alpha^2}{(\gamma_i^2 + \alpha)^2} (\gamma_i^2 \cdot (x_i^*)^2 + \sigma^2) - \sigma^2 m + 2\sigma^2 \sum_i \frac{\gamma_i^2}{\gamma_i^2 + \alpha}.
\end{aligned} \tag{15}$$

Figure 5(b) illustrates the typical shape of $\mathrm{R}_{\mathrm{SURE}}(\alpha)$ and SURE estimates thereof. Following [35, 14], who investigated the performance of Stein's unbiased risk estimate in the different context of hierarchical modeling, we show that, with the definition of the loss l by

$$l(\alpha) := \frac{1}{m}\|Ax^* - A\hat{x}_\alpha(y)\|_2^2,$$

$1/m$ SURE(α, y) is close to l for large m. Note that SURE is an unbiased estimate of the expectation of l.

Theorem 1. *If Assumption 1 holds, then*

$$\sup_{\alpha\in[0,\infty)} \left|\frac{1}{m}\,\mathrm{SURE}(\alpha, y) - l(\alpha)\right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{m}}\right).$$

Proof: We find

$$\begin{aligned} l &= \frac{1}{m}\|A\hat{x} - y + \varepsilon\|_2^2 = \frac{1}{m}\|A\hat{x} - y\|_2^2 + \frac{1}{m}\|\varepsilon\|_2^2 + \frac{2}{m}\langle\varepsilon, A\hat{x} - y\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{\alpha^2}{(\gamma_i^2+\alpha)^2}y_i^2 - \frac{1}{m}\|U^*\varepsilon\|_2^2 + \frac{2}{m}\langle\varepsilon, A\hat{x} - Ax^*\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{\alpha^2}{(\gamma_i^2+\alpha)^2}y_i^2 - \frac{1}{m}\|\tilde{\varepsilon}\|_2^2 + \frac{2}{m}\langle\varepsilon, A\hat{x} - Ax^*\rangle. \end{aligned}$$

Note that

$$\begin{aligned} A\hat{x} - Ax^* &= U\Sigma\Sigma_\alpha^{-1}U^*(Ax^* + \varepsilon) - U\Sigma V^*x^* \\ &= U\{\Sigma\Sigma_\alpha^{-1} - I\}\Sigma V^*x^* + U\Sigma\Sigma_\alpha^{-1}U^*\varepsilon \end{aligned}$$

and recall from (8) that $x_i^* = \langle v_i, x^*\rangle$. Since $U^*U = UU^* = I$, $\mathrm{Var}[\tilde{\varepsilon}_i] = \sigma^2$, where $\tilde{\varepsilon} = U^*\varepsilon$, that is, $\tilde{\varepsilon}_i = \langle u_i, \varepsilon\rangle$. This yields

$$\frac{2}{m}\langle\varepsilon\,,\, A\hat{x} - Ax^*\rangle = \frac{2}{m}\sum_{i=1}^{n}\frac{\tilde{\varepsilon}_i^2\gamma_i^2}{\gamma_i^2+\alpha} - \frac{2}{m}\sum_{i=1}^{n}\frac{\alpha\tilde{\varepsilon}_i\gamma_i x_i^*}{\gamma_i^2+\alpha}.$$

We obtain the representation

$$\begin{aligned} \frac{1}{m}\,\mathrm{SURE}(\alpha, y) - l &= -\sigma^2 + \frac{2\sigma^2}{m}\sum_{i=1}^{m}\frac{\gamma_i^2}{\gamma_i^2+\alpha} + \frac{1}{m}\sum_{i=1}^{m}\tilde{\varepsilon}_i^2 - \frac{2}{m}\sum_{i=1}^{m}\frac{\tilde{\varepsilon}_i^2\gamma_i^2}{\gamma_i^2+\alpha} + \frac{2}{m}\sum_{i=1}^{m}\frac{\alpha\tilde{\varepsilon}_i\gamma_i x_i^*}{\gamma_i^2+\alpha} \\ &= \frac{1}{m}\sum_{i=1}^{m}(\tilde{\varepsilon}_i^2 - \sigma^2) - \frac{2}{m}\sum_{i=1}^{m}\frac{\gamma_i^2}{\gamma_i^2+\alpha}(\tilde{\varepsilon}_i^2 - \sigma^2) + \frac{2}{m}\sum_{i=1}^{m}\frac{\alpha\gamma_i}{\alpha+\gamma_i^2}x_i^*\tilde{\varepsilon}_i \\ &=: Sl_1(\alpha) + Sl_2(\alpha) + Sl_3(\alpha), \end{aligned}$$

where the terms $Sl_j(\alpha)$, $j \in \{1, 2, 3\}$ are defined in an obvious manner. Since $\tilde{\varepsilon}_1^2, \ldots, \tilde{\varepsilon}_n^2$ are independent and identically distributed with expectation σ^2 we immediately obtain that

$$\sqrt{m}Sl_1(\alpha) = O_{\mathbb{P}}(\sigma^2).$$

Note that $Sl_1(\alpha)$ is independent of α. Next, we consider the term $Sl_2(\alpha)$.

Due to the ordering of the singular values the vectors $\gamma_i^2/(\gamma_i^2+\alpha)$, which have entries in $(0,1]$ for $\alpha \in [0,\infty)$, and are monotonically decreasing. Thus, we find

$$\sup_{\alpha\in[0,\infty)} |Sl_2(\alpha)| = \sup_{\alpha\in[0,\infty)} \frac{1}{m}\left|\sum_{i=1}^{m} \frac{\gamma_i^2}{\gamma_i^2+\alpha}(\tilde{\varepsilon}_i^2-\sigma^2)\right| \leq \sup_{1\geq c_1\geq\ldots\geq c_m\geq 0} \frac{1}{m}\left|\sum_{i=1}^{m} c_i(\tilde{\varepsilon}_i^2-\sigma^2)\right|.$$

It follows from [21], Lemma 7.2.:

$$\sup_{1\geq c_1\geq\ldots\geq c_m\geq 0} \frac{1}{m}\left|\sum_{i=1}^{m} c_i(\tilde{\varepsilon}_i^2-\sigma^2)\right| = \sup_{1\leq j\leq m} \frac{1}{m}\left|\sum_{i=1}^{j} (\tilde{\varepsilon}_i^2-\sigma^2)\right|,$$

and an application of Kolmogorov's maximal inequality yields:

$$\sup_{\alpha\in[0,\infty)} |Sl_2(\alpha)| = O_{\mathbb{P}}\big(\sigma^2/\sqrt{m}\big),$$

where we also used that $\mathrm{Var}(\tilde{\varepsilon}_i^2-\sigma^2) = 2\sigma^4$, which follows because $\tilde{\varepsilon}_i \sim \mathcal{N}(0,\sigma^2)$.
Finally, we estimate $Sl_3(\alpha)$. The functions $\alpha \mapsto \alpha\gamma_i/(\gamma_i^2+\alpha)$ are monotonically increasing, which implies that $\alpha\gamma_i/(\gamma_i^2+\alpha) \subset [0,1]$, by condition (14). A further application of Kolmogorov's maximal inequality finally yields

$$\begin{aligned}\sup_{\alpha\in[0,\infty)} |Sl_3(\alpha)| &= \sup_{\alpha\in[0,\infty)} \frac{1}{m}\left|\sum_{i=1}^{m} \frac{\alpha\gamma_i}{\gamma_i^2+\alpha} x_i^*\tilde{\varepsilon}_i\right| \leq \sup_{1\geq c_1\geq\ldots\geq c_m\geq 0} \frac{1}{m}\left|\sum_{i=1}^{m} c_i x_i^*\tilde{\varepsilon}_i\right| \\ &= \sup_{1\leq j\leq m} \frac{1}{m}\left|\sum_{i=1}^{j} x_i^*\tilde{\varepsilon}_i\right| = O_{\mathbb{P}}\Big(\sigma\|x^*\|_2/m\Big) = O_{\mathbb{P}}\Big(\sigma/\sqrt{m}\Big).\end{aligned}$$

□

The latter result can be used to show that, in an asymptotic sense, if the loss l is considered, the estimator $\hat{\alpha}_{\mathrm{SURE}}$ does not have a larger risk than any other choice of regularization parameter. This statement is made precise in the following corollary.

Corollary 1. *Under Assumption 1 it holds that for all $\varepsilon > 0$ and any sequence of positive real numbers $(\alpha_m)_{m\in\mathbb{N}}$ we have*

$$\mathbb{P}(l(\hat{\alpha}_{\mathrm{SURE}}) \geq l(\alpha_m) + \varepsilon) \to 0.$$

Proof: By definition $\mathrm{SURE}(\hat{\alpha}_{\mathrm{SURE}}, y) \leq \mathrm{SURE}(\alpha_m, y)$. This yields

$$\mathbb{P}(l(\hat{\alpha}_{\mathrm{SURE}}) \geq l(\alpha_m) + \varepsilon) \leq \mathbb{P}\Big(l(\hat{\alpha}_{\mathrm{SURE}}) - \frac{1}{m}\mathrm{SURE}(\hat{\alpha}_{\mathrm{SURE}}, y) \geq l(\alpha_m) - \frac{1}{m}\mathrm{SURE}(\alpha_m, y) + \varepsilon\Big)$$

and the claim follows by an application of Theorem 1.

□

The following corollary is an extension of Corollary 1.

Corollary 2. *The claim of Corollary 1 remains true if the arbitrary but fixed positive constant $\varepsilon > 0$ is replaced by a sequence ε_m such that $1/\varepsilon_m = o(\sqrt{m})$.*

We finally mention that our estimates are rather conservative, in particular with respect to the quantity $Sl_3(\alpha)$, since we do not assume particular smoothness of x^*. With an additional source condition, i.e., certain decay speed of the x_i^*, it is possible to derive improved rates, which are however beyond the scope of our paper. We instead turn our attention to the convergence of the risk estimate as $m \to \infty$ as well as the convergence of the estimated regularization parameters.

Theorem 2. *If Assumption 1 holds, then, as* $m \to \infty$

$$\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m}\big(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha, y)\big)\Big| = O_{\mathbb{P}}\Big(\frac{1}{\sqrt{m}}\Big)$$

and

$$\mathbb{E}\bigg(\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m}\big(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha, y)\big)\Big|\bigg)^2 = O\Big(\frac{1}{m}\Big). \tag{16}$$

Proof: Observing (11) and (15) we find

$$\frac{1}{m}\big(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha)\big) = \frac{1}{m}\sum_{i=1}^{m} \frac{\alpha^2}{(\gamma_i^2+\alpha)^2}\check{\varepsilon}_i,$$

where $\check{\varepsilon}_i := y_i^2 - \mathbb{E}[y_i^2]$. The random variables $\check{\varepsilon}_1, \dots, \check{\varepsilon}_n$ are independent and centered. Notice that

$$\mathrm{Var}[\check{\varepsilon}_i] = \mathrm{Var}[y_i^2] = \mathbb{E}[y_i^4] - (\mathbb{E}[y_i^2])^2 = 4\gamma_i^2 {x_i^*}^2\sigma^2 + 2\sigma^4,$$

since $y_i \sim \mathcal{N}(\gamma_i x_i^*, \sigma^2)$. Consider the monotonically increasing function $\alpha \mapsto \frac{\alpha^2}{(\gamma_i^2+\alpha)^2} \subset [0,1]$ for $\alpha \in [0,\infty)$. With the same arguments as in the proof of Theorem 1, using Kolmogorov's maximal inequality, we estimate

$$\begin{aligned}
&\sup_{\alpha\in[0,\infty)} \Big|\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha)\Big| = \sup_{\alpha\in[0,\infty)} \Big|\sum_{i=1}^{m} \frac{\alpha^2}{(\gamma_i^2+\alpha)^2}\check{\varepsilon}_i\Big| = \sup_{1\le j\le m} \Big|\sum_{i=1}^{j} \check{\varepsilon}_i\Big| \\
&= O_{\mathbb{P}}\bigg(\Big(\sum_{j=1}^{m}(4\gamma_i^2(x_i^*)^2 + 2\sigma^4)\Big)^{\frac{1}{2}}\bigg)
\end{aligned}$$

It remains to show the L^2-convergence (16). To this end define the j-th partial sum

$$S_j := \sum_{i=1}^{j} \check{\varepsilon}_i$$

and observe that $\{S_j \,|\, j \in \mathbb{N}\}$ forms a martingale. The L^p-maximal inequality for martingales yields

$$\begin{aligned}
&\mathbb{E}\bigg(\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m}\big(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha)\big)\Big|\bigg)^2 = \mathbb{E}\bigg(\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m}\big(\mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha)\big)\Big|^2\bigg) \\
&= \frac{1}{m^2}\mathbb{E}\Big(\sup_{1\le j\le m} |S_j|^2\Big) \le \frac{4}{m^2}\mathbb{E}\bigg(\sum_{i=1}^{m} \check{\varepsilon}_i\bigg)^2 = O\bigg(\frac{1}{m^2}\sum_{j=1}^{m}(4\gamma_i^2(x_i^*)^2 + 2\sigma^4)\bigg)
\end{aligned}$$

as above.

□

In order to understand the behavior of the estimated regularization parameters we start with some bounds on $\hat{\alpha}^*_{\mathrm{SURE}}$, which recover a standard property of deterministic regularization methods, namely that $\frac{\sigma^2}{\alpha}$ does not diverge for suitable parameter choices.

Lemma 1. *A regularization parameter $\hat{\alpha}^*_{\mathrm{SURE}}$ obtained from $\mathrm{R}_{\mathrm{SURE}}$ satisfies*

$$\frac{\sigma^2}{\max_i |x_i^*|^2} \leq \hat{\alpha}^*_{\mathrm{SURE}} \leq \max\{1, 8\sigma^2 \frac{\sum \gamma_i^4}{\sum \gamma_i^4 (x_i^*)^2}\}$$

Proof: It is straightforward to see the differentiability of $\mathrm{R}_{\mathrm{SURE}}$ and to compute

$$\mathrm{R}_{\mathrm{SURE}}{}'(\alpha) = \sum_{i=1}^{m} \frac{2\gamma_i^4}{(\gamma_i^2 + \alpha)^3} (\alpha (x_i^*)^2 - \sigma^2).$$

Hence, for $\alpha < \frac{\sigma^2}{\max_i |x_i^*|^2}$, the risk $\mathrm{R}_{\mathrm{SURE}}$ is strictly decreasing, which implies the first inequality. Moreover, for $\alpha \geq 1$ we obtain

$$\begin{aligned} \alpha^3 \mathrm{R}_{\mathrm{SURE}}{}'(\alpha) &= 2 \sum_{i=1}^{m} \frac{\gamma_i^4}{(\gamma_i^2/\alpha + 1)^3} (\alpha (x_i^*)^2 - \sigma^2) \\ &> \frac{\alpha}{4} \sum_{i=1}^{m} \gamma_i^4 (x_i^*)^2 - 2\sigma^2 \sum_{i=1}^{m} \gamma_i^4 \end{aligned}$$

and we finally see that $\mathrm{R}_{\mathrm{SURE}}{}'$ is nonnegative if in addition $\alpha \geq 8\sigma^2 \frac{\sum \gamma_i^4}{\sum \gamma_i^4 (x_i^*)^2}$.

□

In order to make the convergence as $m \to \infty$ more clear we make the dependence on m explicit in the following by writing $\mathrm{R}_{\mathrm{SURE,m}}$ and $\alpha^*_{\mathrm{SURE},m}$ for the associated estimate of the regularization parameter. From a straight-forward estimate of the derivative of $\mathrm{R}_{\mathrm{SURE,m}}$ on sets where α is bounded away from zero we obtain the following result:

Lemma 2. *The sequence of functions $f_m := \frac{1}{m} \mathrm{R}_{\mathrm{SURE,m}}(\alpha)$ is equicontinuous on sets $[C_1, C_2]$ with $0 < C_1 < C_2$.*

As a consequence of the Arzela-Ascoli theorem we further derive a convergence result:

Proposition 1. *The sequence of functions $f_m := \frac{1}{m} \mathrm{R}_{\mathrm{SURE,m}}(\alpha)$ is equicontinuous on sets $[C_1, C_2]$ with $0 < C_1 < C_2$ and hence has a uniformly convergent subsequence f_{m_k} with continuous limit function f.*

In order to obtain convergence of minimizers it suffices to be able to choose uniform constants C_1 and C_2, which is possible if the bounds in Lemma 1 are uniform:

Theorem 3. *Let $\max_{i=1}^m |x_i^*|$ be uniformly bounded in m and $\frac{1}{m} \sum_{i=1}^m \gamma_i^4 (x_i^*)^2$ be uniformly bounded away from zero. Then there exists a subsequence $\hat{\alpha}_{SURE,m_k}$ that converges to a minimizer of the asymptotic risk f. Moreover $\hat{\alpha}_{\mathrm{SURE}}, m_k$ converges to to a minimizer of the asymptotic risk f in probability.*

Proof: From the uniform convergence of the sequence f_{m_k} in Proposition 1 we obtain the convergence of the minimizers $\hat{\alpha}^*_{\mathrm{SURE},m_k}$. Combined with Theorem 2 we obtain an analogous argument for $\hat{\alpha}_{\mathrm{SURE},m_k}$.

□

3.2 Discrepancy Principle

We now turn our attention to the discrepancy principle, which we can formulate in a similar setting as the SURE approach above. With a slight abuse of notation, in analogy to the other methods, we denote the expectation of $\mathrm{DP}(\alpha, y)$ by $\mathrm{R}_{\mathrm{DP}}(\alpha)$ and define $\hat{\alpha}_{\mathrm{DP}}$ as the solution of the equation

$$\mathrm{R}_{\mathrm{DP}}(\alpha) = \sum_{i=1}^{m} \frac{\alpha^2}{(\gamma_i^2 + \alpha)^2} \mathbb{E}[y_i^2] - m\sigma^2 = 0.$$

Figure 5(a) illustrates the typical shape of $\mathrm{R}_{\mathrm{DP}}(\alpha)$ and its DP estimates. Observing that

$$\mathrm{DP}(\alpha, y) - \mathrm{R}_{\mathrm{DP}}(\alpha) = \mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\beta)$$

we immediately obtain the following result:

Theorem 4. *If Assumption 1 holds, we have*

$$\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m} \big(\mathrm{DP}(\alpha, y) - \mathrm{R}_{\mathrm{DP}}(\alpha) \big) \right| = O_{\mathbb{P}}\Big(\frac{1}{\sqrt{m}} \Big)$$

and

$$\mathbb{E}\left(\sup_{\alpha \in [0,\infty)} \left| \frac{1}{m} \big(\mathrm{DP}(\alpha, y) - \mathrm{R}_{\mathrm{DP}}(\alpha) \big) \right| \right)^2 = O\Big(\frac{1}{m} \Big).$$

3.3 GSURE-Risk

Now we consider the GSURE-risk estimation procedure. Figure 5(c) illustrates the typical shape of $\mathrm{R}_{\mathrm{GSURE}}(\alpha)$ and GSURE estimates thereof. Based on (12), the risk can be written as

$$\mathrm{R}_{\mathrm{GSURE}}(\alpha, y) = \sum_{i=1}^{m} \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{(\gamma_i^2 + \alpha)} \right)^2 \left(\gamma_i^2 (x_i^*)^2 + \sigma^2 \right) - \sigma^2 \sum_{i=1}^{m} \frac{1}{\gamma_i^2} + 2\sigma^2 \sum_{i=1}^{m} \frac{1}{\gamma_i^2 + \alpha}.$$

For the SURE criterion we showed in Theorem 1 that $\mathrm{SURE}(\alpha, y)$ is close to the loss l in an asymptotic sense with the standard $\sqrt{m}$-rate of convergence. An analogous result can be shown for GSURE and the associated loss $\tilde{l} := c_m \|\Pi(x^* - \hat{x}_\alpha)\|_2^2$ but with different associated rates of convergence c_m, dependent on the singular values.

Theorem 5. *Let Assumption 1 be satisfied and in addition to (14), let $\gamma_m \to 0$. Then, as $m \to \infty$,*

$$\sup_{\alpha \in [0,\infty)} \left| \mathrm{GSURE}(\alpha, y) - c_m^{-1} \tilde{l} \right| = O_{\mathbb{P}} \left(\sqrt{\sum_{i=1}^{m} \frac{x_i^*}{\gamma_i^2}} + \sqrt{\sum_{i=1}^{m} \frac{1}{\gamma_i^4}} \right),$$

where

$$c_m := \left(\sum_{j=1}^{m} \frac{1}{\gamma_i^2} \right)^{-1/2}.$$

Proof: For $m = n$ and invertible matrices A the GSURE-loss $\tilde{l}$ is given by

$$\begin{aligned} c_m^{-1}\tilde{l} &= \|x^* - \hat{x}_\alpha\|_2^2 \\ &= \sum_{i=1}^m \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{(\gamma_i^2+\alpha)}\right)^2 y_i^2 - \sum_{j=1}^m \frac{\tilde{\varepsilon}_i^2}{\gamma_i^2} + 2\alpha \sum_{j=1}^m \frac{x_i^* \tilde{\varepsilon}}{\gamma_i(\gamma_i^2+\alpha)} + 2 - 2\sum_{j=1}^m \frac{\tilde{\varepsilon}_i^2}{\gamma_i^2+\alpha}, \end{aligned}$$

since, in this special case, the projection π satisfies $\pi = \mathrm{id}$. Hence

$$\begin{aligned} \mathrm{GSURE}(\alpha, y) - c_m^{-1}\tilde{l} &= -\sum_{i=1}^m \frac{1}{\gamma_i^2}(\tilde{\varepsilon}_i^2 - \sigma^2) + 2\sum_{i=1}^m \frac{1}{\gamma_i^2}(\tilde{\varepsilon}_i^2 - \sigma^2) - 2\sum_{i=1}^m \frac{1}{\gamma_i^2+\alpha}(\tilde{\varepsilon}_i^2 - \sigma^2) \\ &\quad + 2\alpha \sum_{i=1}^m \frac{x_i^*}{\gamma_i(\gamma_i^2+\alpha)}\tilde{\varepsilon}_i \\ &= 2\alpha \sum_{i=1}^m \frac{x_i^*}{\gamma_i(\gamma_i^2+\alpha)}\tilde{\varepsilon}_i + \sum_{i=1}^m \frac{\alpha^2 - \gamma_i^4}{\gamma_i^2(\gamma_1^2+\alpha)^2}(\tilde{\varepsilon}_i^2 - \sigma^2) \\ &=: GSl_1(\alpha) + GSl_2(\alpha), \end{aligned}$$

where $GSl_1(m, \alpha)$ and $GSl_2(m, \alpha)$ are defined in an obvious manner. Furthermore,

$$\begin{aligned} \sup_{\alpha\in[0,\infty)} |GSl_1(\alpha)| &= \sup_{\alpha\in[0,\infty)} \left| 2\alpha \sum_{i=1}^m \frac{x_i^*}{\gamma_i(\gamma_i^2+\alpha)}\tilde{\varepsilon}_i \right| = \sup_{1\geq c_1\geq\ldots\geq c_m\geq 0} \left| 2\sum_{i=1}^m c_i \frac{x_i^*}{\gamma_i}\tilde{\varepsilon}_i \right| \\ &= \sup_{1\leq j\leq m} \left| 2\sum_{i=1}^j \frac{x_i^*}{\gamma_i}\tilde{\varepsilon}_i \right| = O_{\mathbb{P}}\left(\sqrt{\sum_{i=1}^m \frac{x_i^*}{\gamma_i^2}}\right), \end{aligned}$$

where the last estimate follows from Kolmogorov's maximal inequality. Now we estimate the term $GSl_2(\alpha)$. Consider the functions $\psi_i : \alpha \mapsto (\alpha^2 - \gamma_i^4)/(\gamma_i^2+\alpha)^2,\ i = 1, \ldots, m$. Notice that

$$\psi_i(0) = -1 \quad \text{and} \quad \lim_{\alpha\to\infty} \psi_i(\alpha) = 1 \quad \text{for } i = 1, \ldots, m$$

and that for each i the function ψ_i increases monotonically in α. This implies

$$\begin{aligned} \sup_{\alpha\in[0,\infty)} |GSl_2(\alpha)| &= \sup_{\alpha\in[0,\infty)} \left| \sum_{i=1}^m \frac{\alpha^2 - \gamma_i^4}{\gamma_i^2(\gamma_1^2+\alpha)^2}(\tilde{\varepsilon}_i^2 - \sigma^2) \right| \\ &\leq \sup_{1\geq c_1\geq\ldots\geq c_m\geq 0} \left| \sum_{i=1}^m \frac{c_i}{\gamma_i^2}(\tilde{\varepsilon}_i^2 - \sigma^2) \right| + \sup_{0\leq c_1\leq\ldots\leq c_m\leq 1} \left| \sum_{i=1}^m \frac{c_i}{\gamma_i^2}(\tilde{\varepsilon}_i^2 - \sigma^2) \right| \\ &\leq \sup_{1\leq j\leq m} \left| \sum_{i=1}^j \frac{1}{\gamma_i^2}(\tilde{\varepsilon}_i^2 - \sigma^2) \right| + \sup_{1\leq j\leq m} \left| \sum_{i=j}^m \frac{1}{\gamma_i^2}(\tilde{\varepsilon}_i^2 - \sigma^2) \right|. \end{aligned}$$

A further application of Kolmogorov's maximal inequality yields the desired result.

□

We can now proceed to an estimate between GSURE and $\mathrm{R}_{\mathrm{GSURE}}$ similar to the ones for the SURE risk, however we observe a main difference due to the appearance of the condition number of the forward matrix A:

Theorem 6. *Let $A \in \mathbb{R}^{n\times m}$ be a full rank matrix. In addition to Assumption 1, let $\gamma_m \to 0$. Then, as $m \to \infty$,*

$$\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m\ \mathrm{cond}(A)^2} \big(\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \big) \Big| = O_{\mathbb{P}}\Big(\frac{1}{\sqrt{m}} \Big)$$

and

$$\mathbb{E}\Big(\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m\ \mathrm{cond}(A)^2} \big(\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \big) \Big| \Big)^2 = O_{\mathbb{P}}\Big(\frac{1}{m} \Big). \tag{17}$$

Proof: For full rank matrices $A \in \mathbb{R}^{m\times m}$ we have

$$\mathrm{GSURE}(\alpha, y) = \sum_{i=1}^{m} \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{(\gamma_i^2 + \alpha)} \right)^2 y_i^2 - \sigma^2 \sum_{i=1}^{m} \frac{1}{\gamma_i^2} + 2\sigma^2 \sum_{i=1}^{m} \frac{1}{\gamma_i^2 + \alpha}.$$

This gives

$$\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) = \sum_{i=1}^{m} \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{(\gamma_i^2 + \alpha)} \right)^2 \big(y_i^2 - \mathbb{E}[y_i^2] \big) = \sum_{i=1}^{m} \left(\frac{1}{\gamma_i} - \frac{\gamma_i}{(\gamma_i^2 + \alpha)} \right)^2 \check{\varepsilon}_i.$$

As in the proof of Theorem 2 we set $\check{\varepsilon}_i := y_i^2 - \mathbb{E}[y_i^2]$. Recall that the random variables $\check{\varepsilon}_i$ are centered, independent with $\mathrm{Var}[\check{\varepsilon}_i] = 4\gamma_i^2 x_i^{*2}\sigma^2 + 2\sigma^4$. We find

$$\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) = \frac{1}{\gamma_m^2} \sum_{i=1}^{m} \frac{\gamma_m^2}{\gamma_i^2} \frac{\alpha^2}{(\gamma_i^2 + \alpha)^2} \check{\varepsilon}_i.$$

With the same arguments as in the proofs of Theorems 1 and 2 we obtain

$$\sup_{\alpha\in[0,\infty)} \Big| \mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \Big| \le \sup_{0\le c_1\le c_2\le \ldots\le 1} \Big| \frac{1}{\gamma_m^2} \sum_{i=1}^{m} c_i \check{\varepsilon}_i \Big| \le \max_{1\le j\le m} \Big| \frac{1}{\gamma_m^2} \sum_{i=j}^{m} \check{\varepsilon}_i \Big|.$$

Again, an application of Kolmogorov's maximal inequality yields

$$\sup_{\alpha\in[0,\infty)} \Big| \mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \Big| = O_{\mathbb{P}}\left(\left(4 \sum_{i=1}^{m} \gamma_i^2 x_i^{*2} \sigma^2 + 2m\sigma^4 \right)^{\frac{1}{2}} \right)$$

and the first claim of the theorem follows with $\mathrm{cond}(A) = \gamma_1/\gamma_m = 1/\gamma_m$. Moreover, in a similar manner as in the proofs of the previous theorems, we find

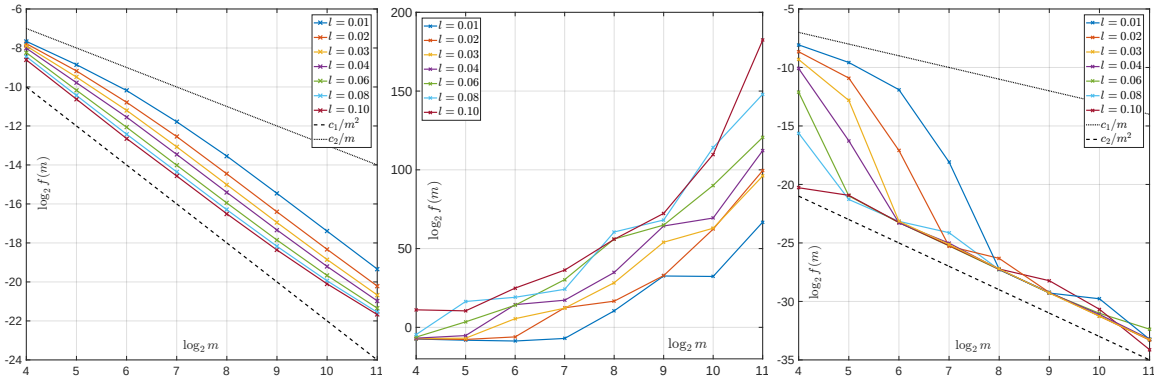
$$\mathbb{E}\Big(\sup_{\alpha\in[0,\infty)} \Big| \frac{1}{m\ \mathrm{cond}(A)^2} \big(\mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha) \big) \Big| \Big)^2 \le \mathbb{E} \sup_{1\le j\le m} \Big| \frac{1}{\gamma_m^2} S_j \Big|^2$$

and by the L^p maximal inequality the second claim now follows as

$$\mathbb{E} \sup_{1\le j\le m} \Big| \frac{1}{\gamma_m^2} S_j \Big|^2 \le \frac{1}{\gamma_m^4} \mathbb{E} S_m^2 = O\big(m/\gamma_m^4 \big).$$

□

We finally note that in the best case the convergence of GSURE is slower than that of SURE. However, since for ill-posed problems the condition number of A will grow with m the typical case is rather divergence of $\frac{\mathrm{cond}(A)^2}{\sqrt{m}}$, hence the empirical estimates of the regularization parameters might have a large variation, which will be confirmed by the numerical results below.



(a) SURE (b) GSURE, without cond(A) (c) GSURE, with cond(A)

Figure 6: Illustration of Theorems 2 and 6 for ℓ_2-regularization: The left hand side of (16)/(17) was estimated by the sample mean and plotted vs. m. For (17), the normalization with cond(A) was omitted in (b) and included in (c). The black dotted lines were added to compare the order of convergence.

4 Numerical Studies for Quadratic Regularization

4.1 Setup

As in the illustrative example in Section 2.1, we computed the empirical distributions of the different parameter choice rules for the same scenario (cf. Section 2.2) for each combination of $m = n = 16, 32, 64, 128, 256, 512, 1024, 2048$, $l = 0.01, 0.02, 0.03, 0.04, 0.06, 0.08, 0, 1$ and $\sigma = 0.1$. For $m = 16, \dots, 512$, $N_\varepsilon = 10^6$ and for $m = 1024, 2048$, $N_\varepsilon = 10^5$ noise realizations were sampled. The computation was, again, based on a logarithmical α-grid, i.e., $\log_{10} \alpha$ is increased linearly in between -40 and 40 with a step size of 0.01. In addition to the distributions of α, the expressions

$$\sup_\alpha \left| \mathrm{SURE}(\alpha, y) - \mathrm{R}_{\mathrm{SURE}}(\alpha, y) \right|, \quad \text{and} \quad \sup_\alpha \left| \mathrm{GSURE}(\alpha, y) - \mathrm{R}_{\mathrm{GSURE}}(\alpha, y) \right| \tag{18}$$

were computed over the α-grid. As in some cases, the supremum is obtained in the limit $\alpha \to \infty$, and hence, on the boundary of our computational grid, we also evaluated (18) for $\alpha = \infty$ in these cases.

4.2 Illustration of Theorems

We first illustrate Theorems 2 and 6 by computing (16) and (17) based on our samples. The results are plotted in Figure 6 and show that the asymptotic rates hold. For GSURE, the comparison between Figures 6(b) and 6(c) also shows that the dependence on cond(A) is crucial.

4.3 Dependence on the Ill-Posedness

We then demonstrate how the empirical distributions of $\hat{\alpha}$ and the corresponding ℓ_2-error, $\|x^* - x_{\hat{\alpha}}\|_2^2$, such as those plotted in Figure 3, depend on the ill-posedness of the inverse problem.

Dependence on m In Figures 7 and 8, m is increased while the width of the convolution kernel is kept fix. The impact of this on the singular value spectrum is illustrated in Figure 2. Most notably, smaller singular values are added and the condition of A increases (cf. Table 1). Figures 7(a) and 8(a) suggest that the distribution of the optimal α^* is Gaussian and converges to a limit for increasing m. The distribution of the corresponding ℓ_2-error looks Gaussian as well and seems to concentrate while shifting to larger mean values. For the discrepancy principle, Figures 7(b) and 8(b) show that the distribution of $\hat{\alpha}_{\mathrm{DP}}$ widens for increasing m, and the distribution of the corresponding ℓ_2-error develops a tail while shifting to larger mean values. Figures 7(c) and 8(c) show that the distribution of $\hat{\alpha}_{\mathrm{SURE}}$ seems to converge to a limit for increasing m. The distribution of the corresponding ℓ_2-error also develops a tail while shifting to larger mean values. For GSURE, Figures 7(d) and 8(d) reveal that increasing m leads to erratic, multimodal distributions: Compared to the other α-distributions, the distribution of $\hat{\alpha}_{\mathrm{GSURE}}$ includes a significant amount of very small values, and the corresponding ℓ_2-error distributions range over very large values.

Dependence on l In Figures 9 and 10, the width of the convolution kernel, l, is increased while $m = 64$ is kept fix (cf. Figure 2 and Table 1). It is worth noticing that as $l = 0.02$ corresponds to a very well-posed problem, the optimal α^* is often extremely small or even 0, as can be seen from Figure 9(a). The general tendencies are similar to those observed when increasing m. For GSURE, Figures 9(d) and 10(d) illustrate how the multiple modes of the distributions slowly evolve and shift to smaller vales of α (and larger corresponding ℓ_2-errors).

4.4 Linear vs Logarithmical Grids

One reason why the properties of GSURE exposed in this work have not been noticed so far is that they only become apparent in very ill-conditioned problems (cf. Section 1). Another reason is the way the risk estimators are typically computed: Firstly, for high dimensional problems, (3) often needs to be solved by an iterative method. For very small α, the condition of $(A^*A+\alpha I)$ is very large and the solver will need a lot of iterations to reach a given tolerance. If, instead, a fixed number of iterations is used, an additional regularization of the solution to (1) is introduced which alters the risk function. Secondly, again due to the computational effort, a coarse, linear α-grid excluding $\alpha = 0$ instead of a fine, logarithmic one is often used for evaluating the risk estimators. For two of the risk estimations plotted in Figure 5(c), Figure 11 demonstrates that this insufficient coverage of small α values by the grid can lead to missing the global minimum and other misinterpretations.

5 Numerical Studies for Non-Quadratic Regularization

In this section, we consider the popular sparsity-inducing $R(x) = \|x\|_1$ as a regularization functional (LASSO penalty) to examine whether our results also apply to non-quadratic regularization functionals. For this, let I be the support of $\hat{x}_\alpha(y)$ and J its complement. Let further $|I| = k$ and $P_I \in \mathbb{R}^{k\times n}$ be a projector onto I and A_I the restriction of A to I. We have that

$$\mathrm{df}_\alpha = \|\hat{x}_\alpha(y)\|_0 = k \qquad \text{and} \qquad \mathrm{gdf}_\alpha = \mathrm{tr}(\Pi B^{[J]}), \quad B^{[J]} := P_I(A_I^*A_I)^{-1}P_I^*,$$

as shown, e.g., in [29, 12, 10], which allows us to compute SURE (6) and GSURE (7). Notice that while $\hat{x}_\alpha(y)$ is a continuous function of α [6], SURE and GSURE are discontinuous at

all α where the support I changes.
To carry out similar numerical studies as those presented the last section, we have to overcome several non-trivial difficulties: While there exist various iterative optimization techniques to solve (2) nowadays (see, e.g., [7]), each method typically only works well for certain ranges of α, cond(A) and tolerance levels to which the problem should be solved. In addition, each method comes with internal parameters that have to be tuned for each problem separately to obtain fast convergence. As a result, it is difficult to compute a consistent series of $\hat{x}_\alpha(y)$ for a given logarithmical α-grid, i.e., that accurately reproduces all the change-points in the support and has a uniform accuracy over the grid. Our solution to this problem is to use an all-at-once implementation of ADMM [5] that solves (2) for the whole α-grid simultaneously, i.e., using exactly the same initialization, number of iterations and step sizes. See Appendix A for details. In addition, an extremely small tolerance level ($tol = 10^{-14}$) and 10^4 maximal iterations were used to ensure a high accuracy of the solutions.
Another problem for computing quantities like (18) is that we cannot compute the expectations defining the real risks $\mathrm{R}_{\mathrm{SURE}}$ (6) and $\mathrm{R}_{\mathrm{GSURE}}$ (7) anymore: We have to estimate them as the sample mean over SURE and GSURE in a first run of the studies, before we can compute (18) in a second run (wherein $\mathrm{R}_{\mathrm{SURE}}$ and $\mathrm{R}_{\mathrm{GSURE}}$ are replaced by the estimates from the first run).
We considered scenarios with each combination of $m = n = 16, 32, 64, 128, 256, 512$, $l = 0.02, 0.04, 0.06$ and $\sigma = 0.1$. Depending on m, $N_\varepsilon = 10^5, 10^4, 10^4, 10^4, 10^3, 10^3$ noise realizations were examined. The computation was based on a logarithmical α-grid where $\log_{10}\alpha$ is increased linearly in between -10 and 10 with a step size of 0.01.

Risk plots Figure 12 shows the different risk functions and estimates thereof. The jagged form of the SURE and GSURE plots evaluated on this fine α-grid indicates that the underlying functions are discontinuous. Also note that while SURE and GSURE for each individual noise realization are discontinuous, $\mathrm{R}_{\mathrm{SURE}}$ and $\mathrm{R}_{\mathrm{GSURE}}$ are smooth and continuous, as can be seen already from the empirical means over $N_\varepsilon = 10^4$.

Empirical Distributions Figure 13 shows the empirical distibutions of the different parameter choice rules for α. Here, the optimal α^* is chosen as the one minimizing the ℓ_1-error $\|x^* - x_{\hat{\alpha}}\|_1$ to the true solution x^*. We can observe similar phenomena as for ℓ_2-regularization. In particular, the distributions for GSURE, also have multiple modes at small values of α and at large values of ℓ_1-error.

Sup-Theorems Due to the lack of explicit formulas for the ℓ_1-regularized solution $x_\alpha(y)$, carrying out similar analysis as in Section 3 to derive theorems such as Theorems 2 and 6 is very challenging. In this work, we only illustrate that similar results may hold for the case of ℓ_1-regularization by computing the left hand side of (16) and (17) based on our samples. The results are shown in Figure 14 and are remarkably similar to those shown in Figure 6.

Linear Grids and Accurate Optimization All the issues raised in Section 4.4 about why the properties of GSURE revealed in this work are likely to be overlooked when working on high dimensional problems are even more crucial for the case of ℓ_1-regularization: For computational reasons, the risk estimators are often evaluated on a coarse, linear α-grid using a small, fixed number of iterations of an iterative method such as ADMM. Figure 15

illustrates that this may obscure important features of the real GSURE function, such as the strong discontinuities for small α, or even change it significantly.

6 Conclusion

From the results presented in this work, we see that unbiased risk estimators encounter enormous difficulties for the parameter choice in variational regularization methods for ill-posed problems. While the discrepancy principle yields a quite unimodal distribution of regularization parameters resembling the optimal one with slightly increased mean value, the SURE estimates start to develop multimodality, and the additional modes consist of underestimated regularization parameters, which may lead to significant errors in the reconstruction.
For the case of GSURE, which is based on a presumably more reliable risk, the estimates produce quite wide distributions (at least in logarithmic scaling) for increasing ill-posedness, in particular there are many highly underestimated parameters, which clearly yield bad reconstructions. We expect that this behavior is rather due to the bad quality of the risk estimators than the quality of the risk. These findings may be explained by Theorem 6, which indicates that the estimated GSURE risk might deviate strongly from the true risk function when the condition number of A is large, i.e. the problem is asymptotically ill-posed as $m \to 0$. Consequently one might expect a strong variation in the minimizers of GSURE with varying y compared to the ones of $\mathrm{R}_{\mathrm{GSURE}}$. A potential way to cure those issues is to develop novel risk estimates for $\mathrm{R}_{\mathrm{GSURE}}$ that are not based on Stein's method, possibly it might even be useful not to insist on the unbiasedness of the estimators.
We finally mention that for problems like sparsity-promoting regularization, the GSURE risk leads to additional issues, since it is based on a Euclidean norm. While the discrepancy principle and the SURE risk only use the norms appearing naturally in the output space of the inverse problem (or in a more general setting the log-likelihood of the noise), the Euclidean norm in the space of the unknown is rather arbitrary. In particular, it may deviate strongly from the Banach space geometry in ℓ^1 or similar spaces in high dimensions. Thus, different constructions of GSURE risks are to be considered in such a setting, e.g. based on Bregman distances.

A A Consistent LASSO Solver

We want to solve (2) with $R(x) = \|x\|_1$ for a large number of different values of α but need to ensure that the results are comparable and consistent. For this, we rely on an implementation of the scaled version of ADMM [5] that carries out the iterations for all α simultaneously, with the same penalty parameter ρ for all α and a stop criterion based on the maximal primal and dual residuum over all α. Online adaptation of ρ is also performed based on primal and dual residua for all α. While ensuring the consistency of the results, this leads to sub-optimal performance for individual α's which has to be countered by using a large number of iterations to obtain high accuracies.

Algorithm 1 (All-At-Once ADMM)**.** *Given $\alpha_1, \dots, \alpha_{N_\alpha}$, $\rho > 0$ (penalty parameter), $\tau > 1$, $\mu > 1$ (adaptation parameters), $K \in \mathbb{N}$ (max. iterations) and $\varepsilon \geqslant 0$ (stopping tolerance), initialize $X^0, Z^0, U^0 \in \mathbb{R}^{n \times N_\alpha}$ by 0, and $Y = y \otimes \mathbb{1}_{N_\alpha}^T$, $\Lambda = [\alpha_1, \dots, \alpha_{N_\alpha}] \otimes \mathbb{1}_n$, where $\mathbb{1}_l$ denotes an all-one column vector in $\mathbb{R}^l$. Further, let $\odot$ denote the component-wise multipli-*

cation between matrices (Hadamard product).

For $k = 1, \dots, K$ *do:*

$$X^{k+1} = (A^*A + \rho I)^{-1}(A^*Y + \rho(Z^k - U^k)) \qquad (x - update)$$

$$Z^{k+1} = \text{sign}\left(X^{k+1} + U^k\right) \odot \max\left(X^{k+1} + U^k - \Lambda/\rho, 0\right) \qquad (z - update)$$

$$U^{k+1} = U^k + X^{k+1} - Z^{k+1} \qquad (u - update)$$

$$r_i^{k+1} = X_{(\cdot,i)}^{k+1} - Z_{(\cdot,i)}^{k+1} \quad \forall \quad i = 1, \dots, N_\alpha \qquad (primal\ residuum)$$

$$s_i^{k+1} = -\rho(Z_{(\cdot,i)}^{k+1} - Z_{(\cdot,i)}^{k}) \quad \forall \quad i = 1, \dots, N_\alpha \qquad (dual\ residuum)$$

$$(U^{k+1}, \rho) = \begin{cases} (U^{k+1}/\tau, \tau\rho) & if \quad \#\left\{i \;\middle|\; \|r_i^{k+1}\|_2 > \mu\|s_i^{k+1}\|_2\right\} > N_\alpha/2 \\ (\tau U^{k+1}, \rho/\tau) & if \quad \#\left\{i \;\middle|\; \|s_i^{k+1}\|_2 > \mu\|r_i^{k+1}\|_2\right\} > N_\alpha/2 \\ (U^{k+1}, \rho) & else. \end{cases} \qquad (\rho - adaptation)$$

$$\epsilon_i^{pri} = \varepsilon\left(\sqrt{n} + \max(\|X_{(\cdot,i)}^{k+1}\|_2, \|Z_{(\cdot,i)}^{k+1}\|_2)\right) \quad \forall \quad i = 1, \dots, N_\alpha \qquad (primal\ stop\ tol)$$

$$\epsilon_i^{dual} = \varepsilon\left(\sqrt{n} + \rho\|U_{(\cdot,i)}^{k+1}\|_2\right) \quad \forall \quad i = 1, \dots, N_\alpha \qquad (dual\ stop\ tol)$$

$$stop\ if \quad \|r_i^{k+1}\|_2 < \epsilon_i^{pri} \ \wedge \ \|s_i^{k+1}\|_2 < \epsilon_i^{dual} \quad \forall \quad i = 1, \dots, N_\alpha$$

The algorithm returns both $X_{(\cdot,i)}^{k+1}$ and $Z_{(\cdot,i)}^{k+1}$ as approximations of the solution to (2) with $R(x) = \|x\|_1$ and $\alpha = \alpha_i$ of which we use $Z_{(\cdot,i)}^{k+1}$ for our purposes as it is exactly sparse due to the soft-thresholding step (*z-update*). In the computations, we furthermore initialized $\rho = 1$ and used $\tau = 2$, $\mu = 1.1$, $\varepsilon = 10^{-14}$ and $K = 10^4$.

Acknowledgements. This work of N. Bissantz, H. Dette and K. Proksch has been supported by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Project A1, C1, C4) of the German Research Foundation (DFG).

References

[1] M. S. C. Almeida and M. a T. Figueiredo, *Parameter estimation for blind and non-blind deblurring using residual whiteness measures.*, IEEE Transactions on Image Processing **22** (2013), no. 7, 2751–63. 2

[2] F. Bauer and T. Hohage, *A Lepskij-type stopping rule for regularized Newton methods*, Inverse Problems **21** (2005), no. 6, 1975–1991. 2

[3] Gilles Blanchard and Peter Mathé, *Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration*, Inverse problems **28** (2012), no. 11, 115011. 2

[4] Peter Blomgren and Tony F Chan, *Modular solvers for image restoration problems using the discrepancy principle*, Numerical linear algebra with applications **9** (2002), no. 5, 347–358. 3

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning **3** (2011), no. 1, 1–122. 20, 21

[6] Björn Bringmann, Daniel Cremers, Felix Krahmer, and Michael Möller, *The homotopy method revisited: Computing solution paths of ℓ_1-regularized problems*, arXiv preprint arXiv:1605.00071 (2016). 19

[7] M. Burger, A. Sawatzky, and G. Steidl, *First Order Algorithms in Variational Image Processing*, arXiv (2014), no. 1412.4237, 60. 20

[8] E. J. Candes, C. a. Sing-Long, and J. D. Trzasko, *Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators*, IEEE Transactions on Signal Processing **61** (2013), no. 19, 4643–4657. 2

[9] C. Deledalle, S. Vaiter, J. Fadili, and G. Peyré, *Stein Unbiased GrAdient estimator of the Risk (SUGAR) for Multiple Parameter Selection*, SIAM Journal on Imaging Sciences **7** (2014), no. 4, 2448–2487. 2

[10] C. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal, *Proximal Splitting Derivatives for Risk Estimation*, Journal of Physics: Conference Series **386** (2012), 012003. 2, 19

[11] C. Deledalle, S. Vaiter, G. Peyré, J. Fadili, and C. Dossal, *Unbiased risk estimation for sparse analysis regularization*, 2012 19th IEEE International Conference on Image Processing, IEEE, September 2012, pp. 3053–3056. 2

[12] C. Dossal, M. Kachour, J. Fadili, G. Peyré, and C. Chesneau, *The degrees of freedom of the lasso for general design matrix*, Statistica Sinica **23** (2013), no. 2, 809–828. 2, 19

[13] Y. C. Eldar, *Generalized SURE for Exponential Families: Applications to Regularization*, IEEE Transactions on Signal Processing **57** (2009), no. 2, 471–481. 2

[14] S. K. Ghoreishi and M. R. Meshkani, *On SURE estimates in hierarchical models assuming heteroscedasticity for both levels of a two-level normal hierarchical model*, Journal of Multivariate Analysis **132** (2014), 129–137. 11

[15] R. Giryes, M. Elad, and Y.C. Eldar, *The projected GSURE for automatic parameter tuning in iterative shrinkage methods*, Applied and Computational Harmonic Analysis **30** (2011), no. 3, 407–422. 2

[16] H. Haghshenas Lari and A. Gholami, *Curvelet-TV regularized Bregman iteration for seismic random noise attenuation*, Journal of Applied Geophysics **109** (2014), 233–241. 2

[17] P. C. Hansen, *Analysis of Discrete Ill-Posed Problems by Means of the L-Curve*, SIAM Review **34** (1992), no. 4, pp. 561–580. 2

[18] P. C. Hansen and D. P. OLeary, *The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems*, SIAM Journal on Scientific Computing **14** (1993), no. 6, 1487–1503. 2

[19] Bangti Jin, Jun Zou, et al., *Iterative parameter choice by discrepancy principle*, IMA Journal of Numerical Analysis (2012), drr051. 3

[20] O. V. Lepskii, *On a Problem of Adaptive Estimation in Gaussian White Noise*, Theory of Probability & Its Applications **35** (1991), no. 3, 454–466 (en). 2

[21] K.-C. Li, *From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation*, The Annals of Statistics **13** (1985), no. 4, The Annals of Statistics. 12

[22] F. Luisier, T. Blu, and M. Unser, *Image denoising in mixed Poisson-Gaussian noise.*, IEEE Transactions on Image Processing **20** (2011), no. 3, 696–708. 2

[23] J.-C. Pesquet, A. Benazza-Benyahia, and C. Chaux, *A SURE Approach for Digital Signal/Image Deconvolution Problems*, IEEE Transactions on Signal Processing **57** (2009), no. 12, 4616–4632. 2

[24] Peng Qu, Chunsheng Wang, and Gary X Shen, *Discrepancy-based adaptive regularization for grappa reconstruction*, Journal of Magnetic Resonance Imaging **24** (2006), no. 1, 248–255. 3

[25] S. Ramani, T. Blu, and M. Unser, *Monte-Carlo sure: a black-box optimization of regularization parameters for general denoising algorithms.*, IEEE Transactions on Image Processing **17** (2008), no. 9, 1540–54. 2

[26] S. Ramani, Z. Liu, J. Rosen, J.-F. Nielsen, and J. A. Fessler, *Regularization parameter selection for nonlinear iterative image restoration and MRI reconstruction using GCV and SURE-based methods.*, IEEE Transactions on Image Processing **21** (2012), no. 8, 3659–72. 2

[27] C. M. Stein, *Estimation of the Mean of a Multivariate Normal Distribution*, The Annals of Statistics **9** (1981), no. 6, pp. 1135–1151. 2

[28] Gennadii M Vainikko, *The discrepancy principle for a class of regularization methods*, USSR computational mathematics and mathematical physics **22** (1982), no. 3, 1–19. 2

[29] S. Vaiter, C. Deledalle, and G Peyré, *The Degrees of Freedom of Partly Smooth Regularizers*, arXiv (2014), no. 1404.5557v1. 2, 19

[30] S. Vaiter, C. Deledalle, G. Peyré, C. Dossal, and J. Fadili, *Local behavior of sparse analysis regularization: Applications to risk estimation*, Applied and Computational Harmonic Analysis **35** (2013), no. 3, 433–451. 2

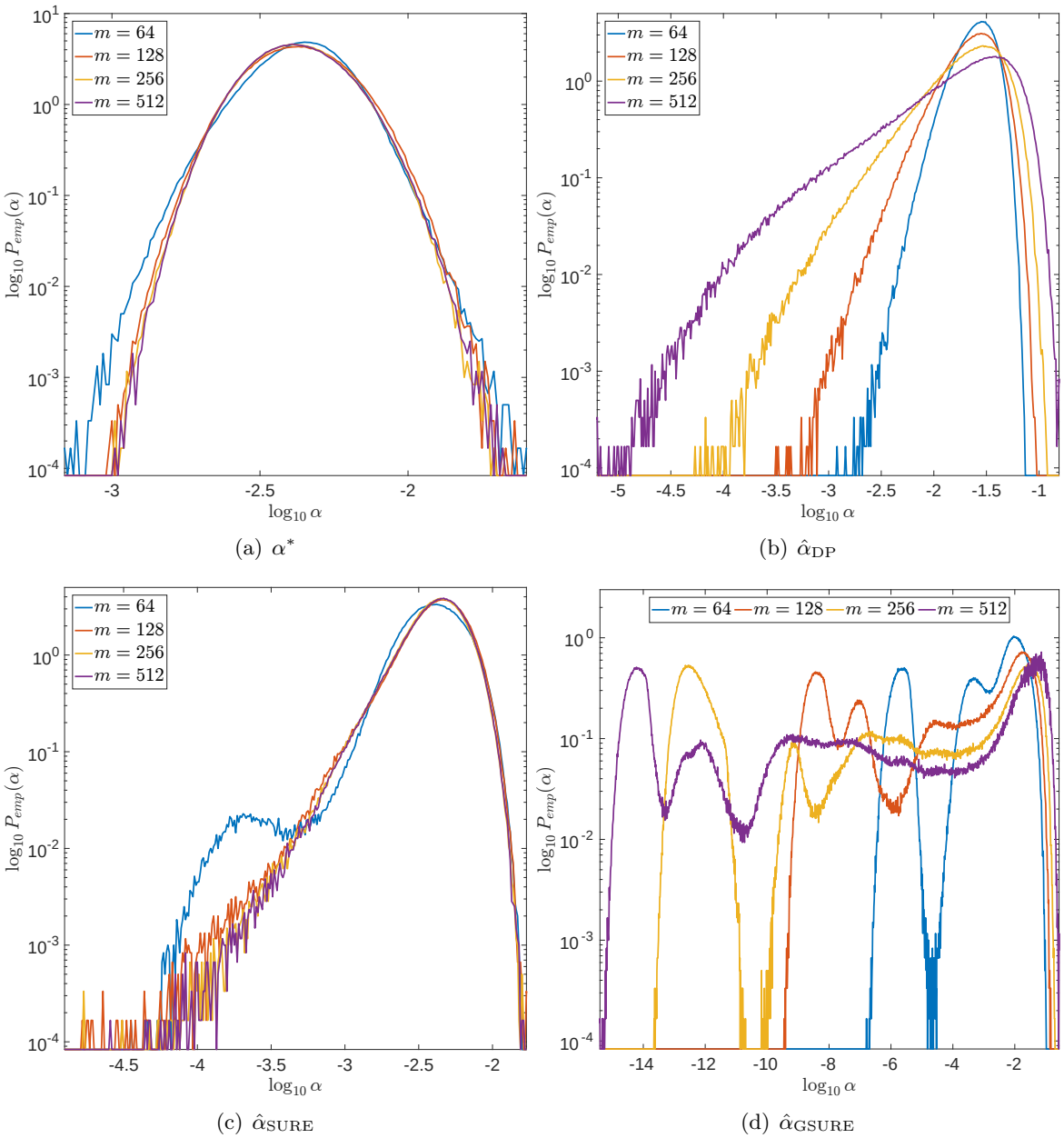
[31] D. Van De Ville and M. Kocher, *SURE-Based Non-Local Means*, IEEE Signal Processing Letters **16** (2009), no. 11, 973–976. 2

[32] D. Van De Ville and M. Kocher, *Nonlocal means with dimensionality reduction and SURE-based parameter selection.*, IEEE Transactions on Image Processing **20** (2011), no. 9, 2683–90. 2

[33] Y.-Q. Wang and J.-M. Morel, *SURE Guided Gaussian Mixture Image Denoising*, SIAM Journal on Imaging Sciences **6** (2013), no. 2, 999–1034 (en). 2

[34] D. S. Weller, S. Ramani, J.-F. Nielsen, and J. A. Fessler, *Monte Carlo SURE-based parameter selection for parallel magnetic resonance imaging reconstruction*, Magnetic Resonance in Medicine **71** (2014), no. 5, 1760–1770. 2

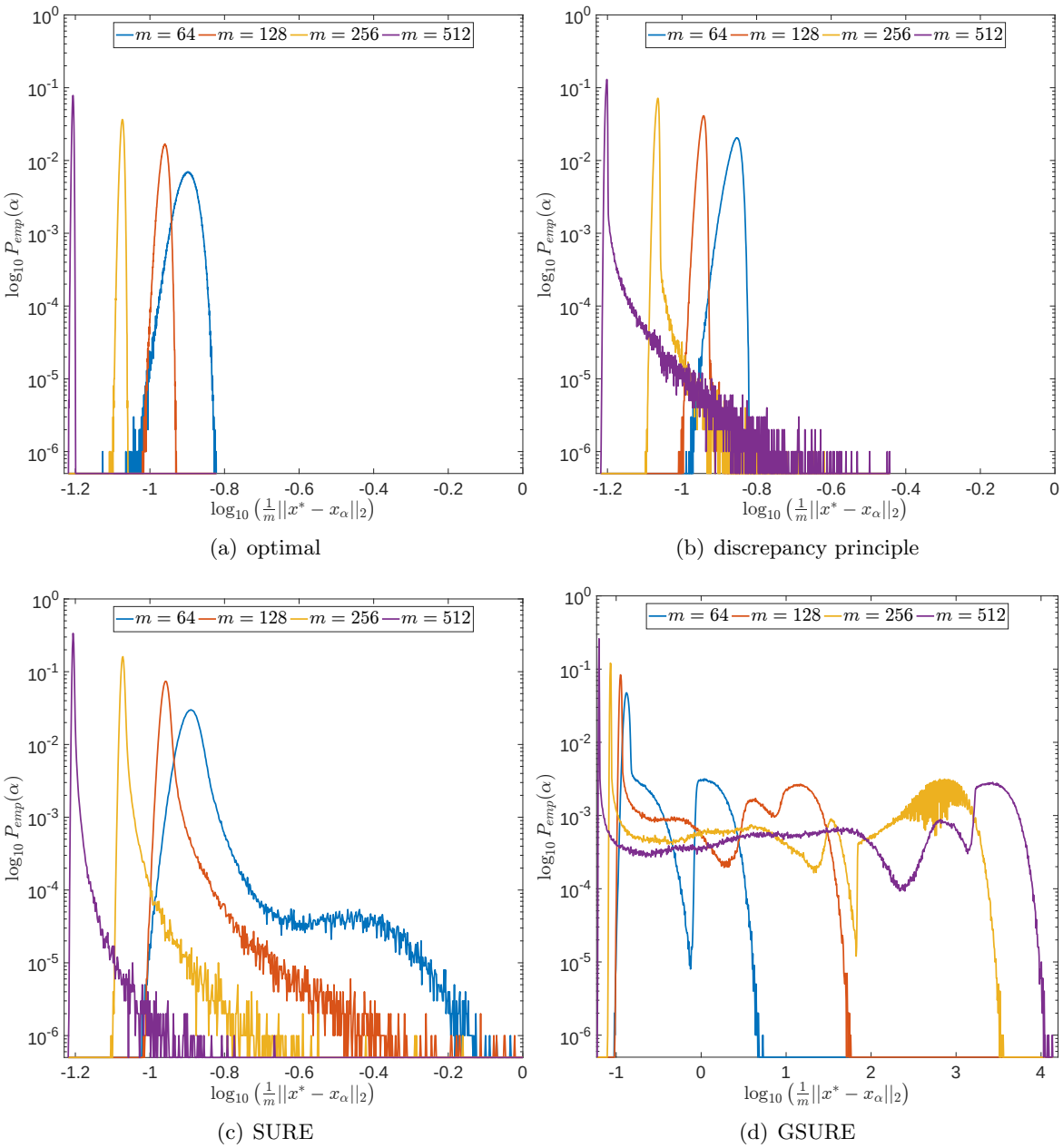
[35] X. Xie, S. C. Kou, and L. D. Brown, *SURE Estimates for a Heteroscedastic Hierarchical Model*, Journal of the American Statistical Association **107** (2012), no. 500, 1465–1479. 11



(a) α^* (b) $\hat{\alpha}_{\mathrm{DP}}$

(c) $\hat{\alpha}_{\mathrm{SURE}}$ (d) $\hat{\alpha}_{\mathrm{GSURE}}$

Figure 7: Empirical probabilities of α for ℓ_2-regularization and different parameter choice rules for $l = 0.06$ and varying m.



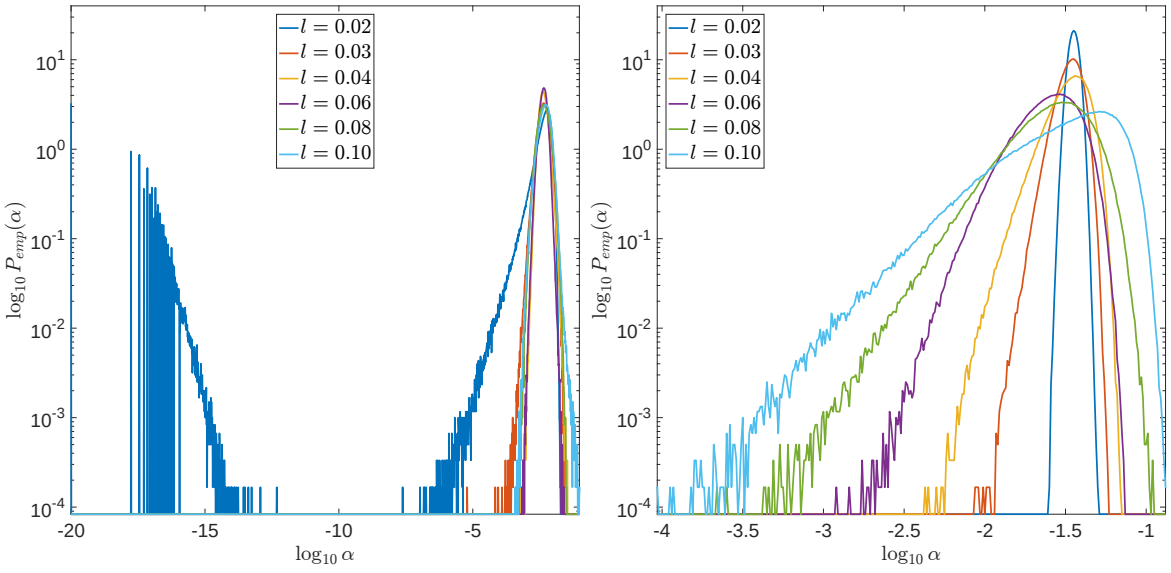
(a) optimal

(b) discrepancy principle

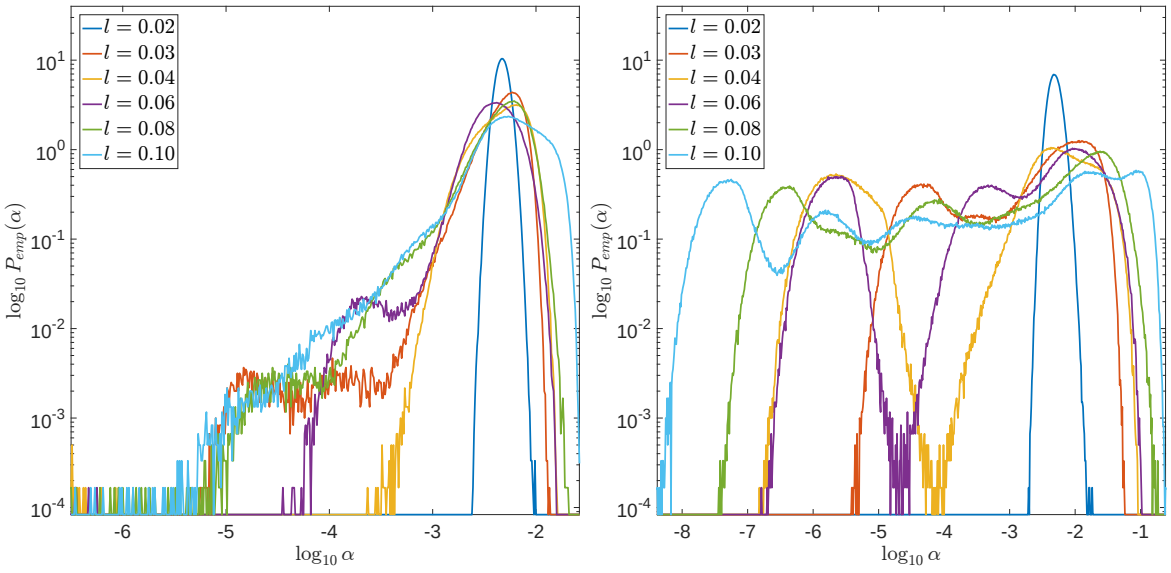
(c) SURE

(d) GSURE

Figure 8: Empirical probabilities of $\log_{10}\left(\frac{1}{m}\|x^* - x_\alpha\|_2^2\right)$ for ℓ_2-regularization and different parameter choice rules for $l = 0.06$ and varying m.

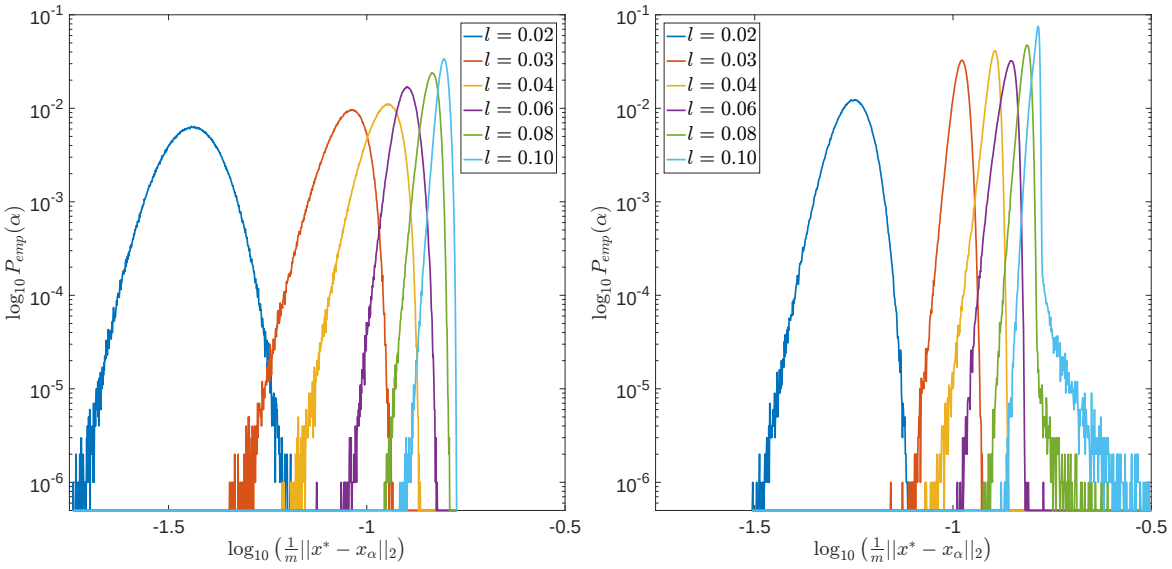


(a) α^* (b) $\hat{\alpha}_{\mathrm{DP}}$



(c) $\hat{\alpha}_{\mathrm{SURE}}$ (d) $\hat{\alpha}_{\mathrm{GSURE}}$

Figure 9: Empirical probabilities of α for ℓ_2-regularization and different parameter choice rules for $m = 64$ and varying l.



(a) optimal

(b) discrepancy principle

(c) SURE

(d) GSURE

Figure 10: Empirical probabilities of $\log_{10}\left(\frac{1}{m}\|x^* - x_\alpha\|_2^2\right)$ for ℓ_2-regularization and different parameter choice rules for $m = 64$ and varying l.

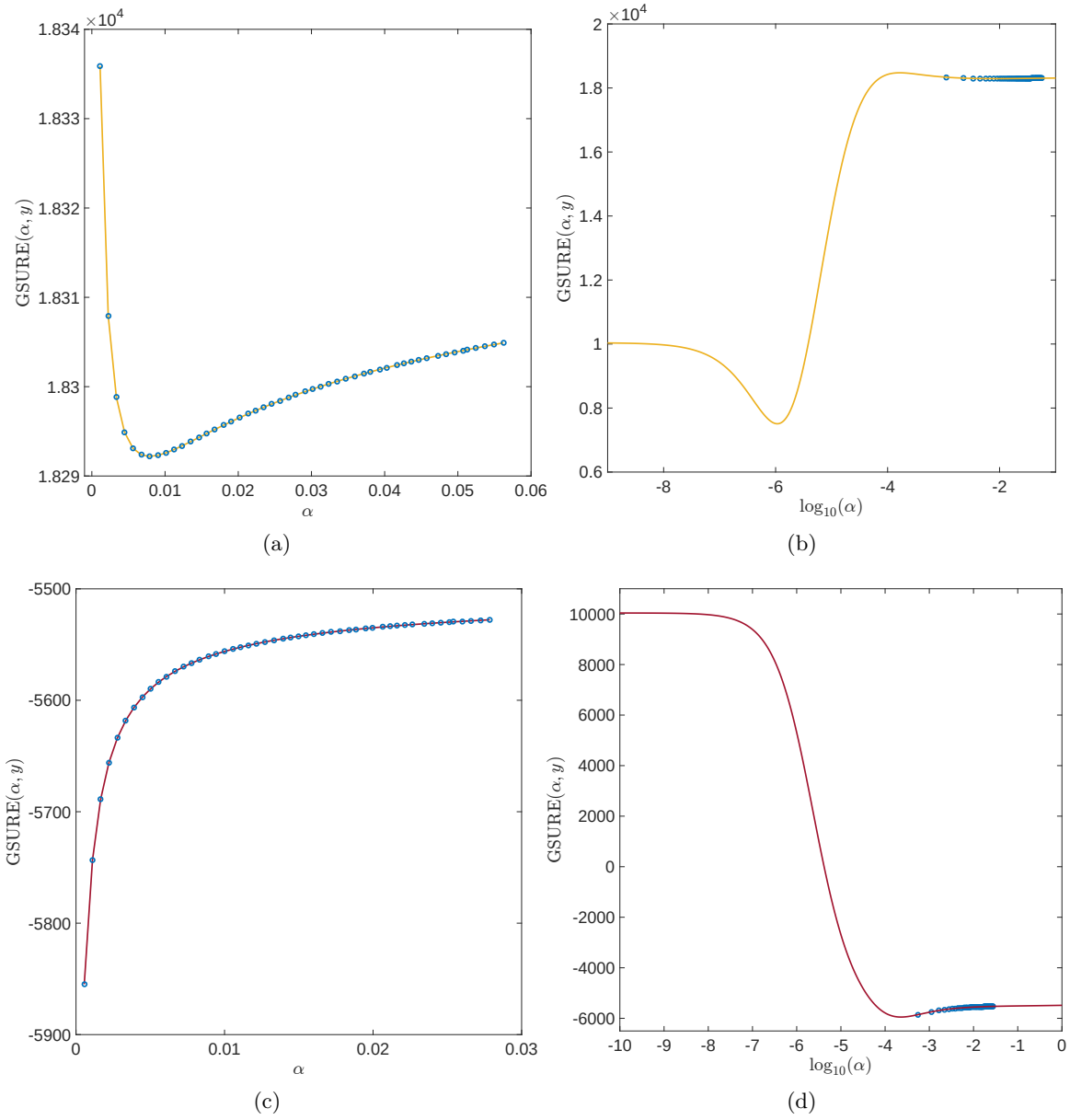
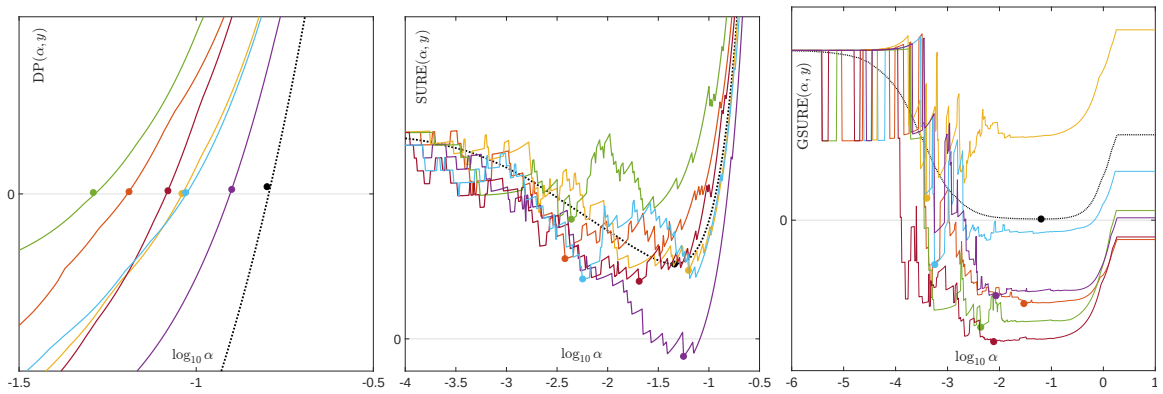
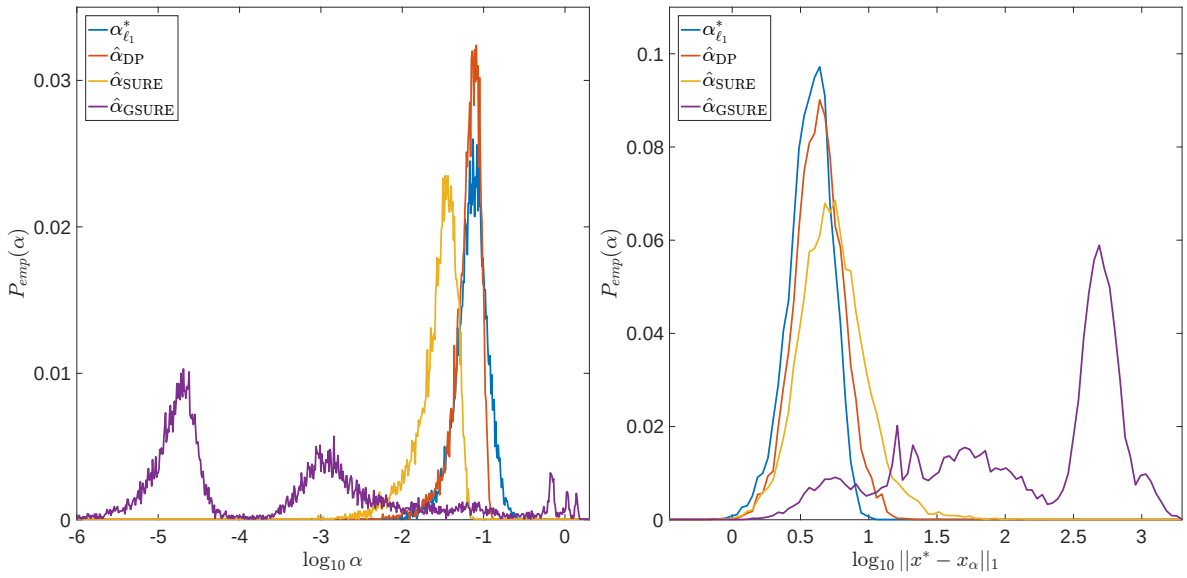


Figure 11: Illustration of the difference between evaluating the GSURE risk on a coarse, linear grid for α as opposed to a fine, logarithmic one: In (a), a linear grid is constructed around $\hat{\alpha}_{\mathrm{DP}}$ as $\alpha = \Delta_\alpha, 2\Delta_\alpha, \ldots, 50\Delta_\alpha$ with $\Delta_\alpha = 2\hat{\alpha}_{\mathrm{DP}}/50$. While the plot suggests a clear minimum, (b) reveals that it is only a sub-optimal local minimum and that the linear grid did not cover the essential parts of $\mathrm{GSURE}(\alpha, y)$. (c) and (d) show the same plots for a different noise realization. Here, a linear grid will not even find a clear minimum. Both risk estimators are the same as those plotted in Figure 5(c) with the same colors.



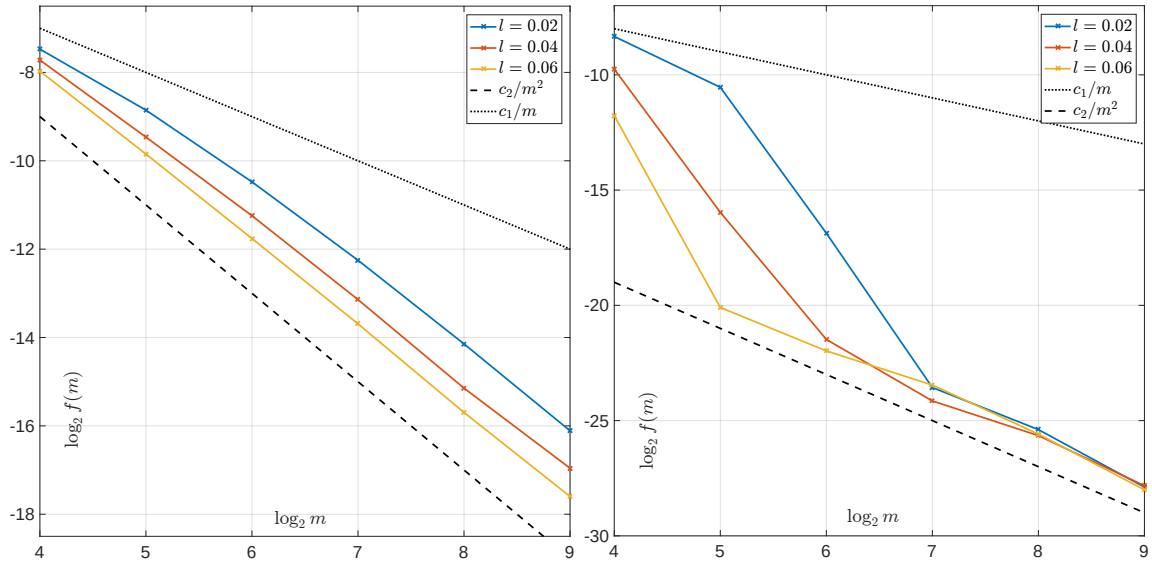
(a) DP (b) SURE (c) GSURE

Figure 12: Risk functions (black dotted line), $k = 1, \ldots, 6$ estimates thereof (solid lines) and their corresponding minima/roots (dots on the lines) in the setting described in Figure 1 using ℓ_1-regularization: (a) DP(α, Ax^*) and DP(α, y^k). (b) $\mathrm{R}_{\mathrm{SURE}}(\alpha)$ (empirical mean over $N_\varepsilon = 10^4$) and SURE(α, y^k). (c) $\mathrm{R}_{\mathrm{GSURE}}(\alpha)$ (empirical mean over $N_\varepsilon = 10^4$) and GSURE(α, y^k).



(a) (b)

Figure 13: Empirical probabilities of (a) α and (b) the corresponding ℓ_1-error for different parameter choice rules using ℓ_1-regularization, $m = n = 64$, $l = 0.06$, $\sigma = 0.1$ and $N = 10^4$ samples of ε.



(a) SURE (b) GSURE

Figure 14: Illustration that Theorems 2 and 6 might also hold for ℓ_1-regularization: The left hand side of (16)/(17) is estimated by the sample mean and plotted vs. m. The black dotted lines were added to compare the order of convergence.

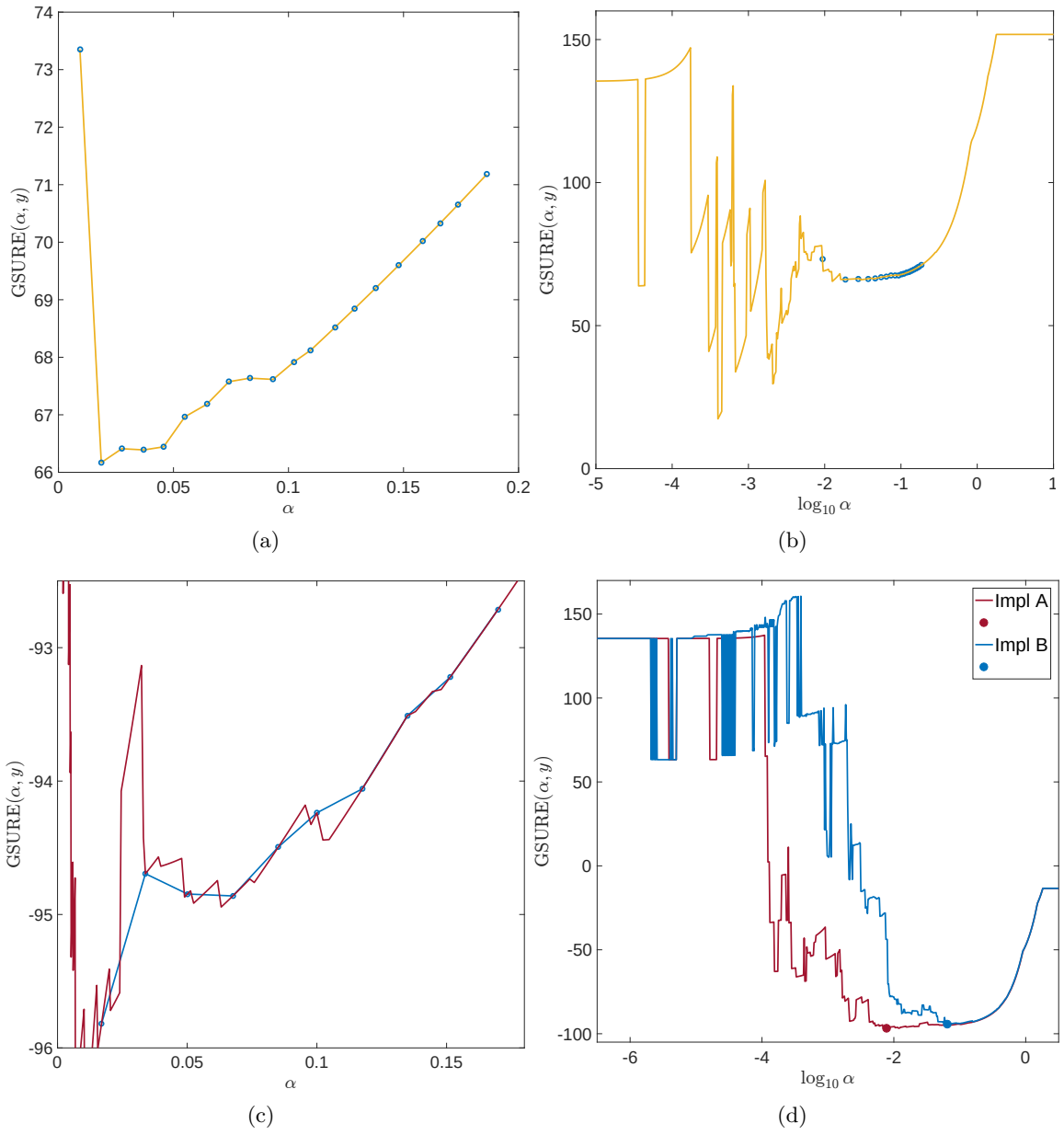


Figure 15: Illustration of the difficulties of evaluating the GSURE risk in the case of ℓ_1-regularization: In (a), a coarse linear grid is constructed around $\hat{\alpha}_{\mathrm{DP}}$ as $\alpha = \Delta_\alpha, 2\Delta_\alpha, \ldots, 20\Delta_\alpha$ with $\Delta_\alpha = \hat{\alpha}_{\mathrm{DP}}/10$. Similar to Figure 11(a) the plot suggests a clear minimum. However, using a fine, logarithmic grid, (b) reveals that it is only a sub-optimal local minimum before a very erratic part of GSURE(α, y) starts. (c) shows how a coarse α-grid can lead to an arbitrary projection of GSURE(α, y) that is likely to miss important features. Both risk estimators are the same as those plotted in Figure 12(c) with the same colors. In (d), the difference between computing GSURE(α, y) with the consistent and highly accurate version of ADMM (Impl A) and with a standard ADMM version using only 20 iterations (Impl B) is illustrated.