Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Parameter choice methods using minimization schemes

Frank Bauer [1], Peter Mathé[2]

submitted: December 21, 2009

[1] Department of Knowledge-Based Mathematical Systems
University of Linz,
Softwarepark 21,
4232 Hagenberg,
Austria
E-Mail: frank.bauer.de@gmail.com

[2] Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstraße 39,
D-10117 Berlin,
Germany
E-mail: mathe@wias-berlin.de

No. 1474
Berlin 2009

Key words and phrases. Inverse Problems, Heuristic Parameter Choice, Minimization Schemes.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract. In this paper we establish a generalized framework, which allows to prove convergenence and optimality of parameter choice schemes for inverse problems based on minimization in a generic way. We show that the well known quasi-optimality criterion falls in this class. Furthermore we present a new parameter choice method and prove its convergence by using this newly established tool.

1. Introduction

Regularization theory, see e.g., [5] studies the stable reconstruction of ill-posed or badly conditioned problems

$$Ax = y, \tag{1}$$

from noisy data y^δ near y. Such a task is always divided into two parts. Firstly, one must devise a (parametric) family of regularized solutions; popular schemes for this are *spectral cut-off*, *Tikhonov* or *Landweber* regularization, see [5] for details. Secondly, and this is of crucial importance, one must properly choose some element from this parametric family of candidate solutions. Such parameter choice may or may not depend on the given noisy data y^δ and/or the noise level δ. The classical parameter choice rules are the *discrepancy principle* and variants thereof [10, 11, 7], which explicitly make use of the noise level δ, a quantity which is rarely known in practice.

As A. Bakushinsky pointed out in [1] any purely data-driven parameter choice must fail for certain sets of data, and hence cannot be convergent. However, several such rules are successfully used in practice, as for instance the *quasi-optimality criterion*, or the L-curve. The question is why such parameter choice rules work well despite of the *Bakushinsky veto*.

Only recently several authors addressed this problem, probably starting with [4] in a Bayesian framework, and continued with [3, 12] in deterministic or mixed settings. As these studies reveal, heuristic methods may work well when excluding pathological behavior of the solution and the noise. We generalize this to parameter choice which is based on functional minimization.

In this study we approach the problem in a coherent way. The different settings may be described in a unified (probabilistic) way, and this is shown in Section 2. Then we formulate a set of general assumptions on the deviation of some functions from their mean, and on decay rates in Section 3. The convergence of certain heuristic parameter choice rules is also established there. We discuss relations to previous work in detail in Section 4. Finally we will present a new parameter choice, called *residual principle*, in Section 5.

2. Setting and minimization schemes

2.1. **The model.** We suppose that instead of the exact data $y = Ax$ we are given noisy data

$$y^\delta = Ax + \delta\xi, \tag{2}$$

where at the moment the noise ξ may be deterministic or stochastic. In this study we use the singular value decomposition of the operator A as

$$Ax = \sum_{k=1}^{\infty} t_k \langle x, u_k\rangle v_k, \quad x \in \mathcal{X}, \tag{3}$$

where $\{u_k\}_{k\in\mathbb{N}}$ is an orthonormal basis system of $\mathcal{X}$, $\{v_k\}_{k\in\mathbb{N}}$ an orthonormal basis of $\mathcal{Y}$. The sequence $\{t_k\}_{k\in\mathbb{N}}$ of singular values is assumed to be monotonously decreasing to 0, i.e. $\lim_{k\to\infty} t_k = 0$. Regularization (reconstruction) methods $R\colon \mathcal{Y} \to \mathcal{X}$ must use the data y^δ from (2). We recall the following regularization schemes.

Example 1 (Spectral cut-off). Given some integer m we denote the projections onto spaces spanned by the first m basis vectors $\{u_k,\ k = 1, \dots, m\}$ and $\{v_k,\ k = 1, \dots, m\}$ by P_m and Q_m, respectively. We denote by

$$A_m := \mathsf{Q}_m A \mathsf{P}_m \colon \mathcal{X} \to \mathcal{Y} \tag{4}$$

be the corresponding discretization. In these terms spectral cut-off is given as

$$y^\delta \longrightarrow A_m^{-1} \mathsf{Q}_m y^\delta. \tag{5}$$

We restrict the discretization levels to an exponential spacing $m\colon = l(n) = \lfloor n_0 q^n \rfloor$, for some $q > 1$ and for $n = 1, \dots, N$, and we thus obtain as approximate solution

$$x_n^\delta := A_{l(n)}^{-1} \mathsf{Q}_{l(n)} y^\delta, \quad n = 1, \dots, N. \tag{6}$$

Example 2 (Tikhonov regularization). This is the parametric family

$$x_\alpha^\delta := \left(A^*A + \alpha I\right)^{-1} A^* y^\delta, \quad \alpha > 0. \tag{7}$$

Below we shall use Tikhonov regularization with regularization parameters $\alpha_n = \alpha_0 q^{-n}$, $n = 1, 2, \dots$, for some $q > 1$, and we let (with a slight abuse of notation) $x_n^\delta := x_{\alpha_n}^\delta$, $n = 1, 2, \dots$

If now $R\colon \mathcal{Y} \to \mathcal{X}$ is any method of reconstruction based on data y^δ then its error at the instance $x \in \mathcal{X}$ and noise ξ is

$$e(x, \xi, R, \delta) := \|x - R(y^\delta)\|. \tag{8}$$

2.2. **The setting.** Within the model (2) the noise may be (unknown) deterministic or drawn from some probability distribution. This also holds true for the solution element x, which may be either deterministic within some class $\{\subset\}\, \mathcal{X}$ or randomly chosen, such that in principle we can distinguish four different settings (Table 1), as these are

Worst case setting: both, solution and noise are deterministic from some sets,

Bayesian setting: both, solution and the noise are given by single probabilities,

	deterministic noise ξ	stochastic noise ξ
deterministic solution x	worst case	statistical case
randomly drawn solution x	average case	Bayesian analysis

TABLE 1. Various settings for the analysis of inverse problems

Statistical setting: the noise is represented by a singleton measure on $\mathcal{Y}$, and
Average case setting: the solution is drawn from some probability distribution on $\mathcal{X}$.

To cover all these settings simultaneously we will consider the following abstract framework. Recall that in the Bayesian setup there are probability measures on both the spaces $\mathcal{X}$ and $\mathcal{Y}$, whereas in the deterministic framework we take suprema over sets $\mathcal{M}$, $\mathcal{N}$ of solution elements and noise. This can be unified by considering certain sets of probabilities as follows. Notice, that due to the error representation (8) we have to assume that the probability on $\mathcal{X}$ is Radon (concentrated on $\mathcal{X}$). In contrast, randomness in the data enters the error only through the reconstruction mapping R.

Assumption 2.1. The spaces $\mathcal{X}$ and $\mathcal{Y}$ are endowed with the Borel σ-algebras, respectively. The set $\mathcal{P} = \mathcal{P}(\mathcal{X})$ consists of *Radon* probabilities on $\mathcal{X}$, and the set $\mathcal{Q} = \mathcal{Q}(\mathcal{Y})$ is a collection of *cylindrical* probabilities on $\mathcal{Y}$.

The reconstruction R maps the cylindrical probabilities, driving $y^\delta = y^\delta(x, \xi)$ on $\mathcal{Y}$ to Radon ones on $\mathcal{X}$.

The error is measured *uniformly* over $\mathcal{P},\ \mathcal{Q}$, i.e.,

$$\begin{aligned} e(\mathcal{P}, \mathcal{Q}, R, \delta) &:= \sup_{P\in\mathcal{P}} \sup_{Q\in\mathcal{Q}} \left(\mathbb{E}_P \mathbb{E}_Q e(x, \xi, R, \delta)^2\right)^{1/2} \\ &= \sup_{P\in\mathcal{P}} \sup_{Q\in\mathcal{Q}} \left(\mathbb{E}_P \mathbb{E}_Q \|x - R(y^\delta)\|^2\right)^{1/2}. \end{aligned}$$

This covers both the Bayesian setting, with singleton sets, and the worst case setting, in which case we let $\mathcal{P} = \{\delta_x,\ x \in \mathcal{M}\}$, $\mathcal{Q} = \{\delta_y,\ x \in \mathcal{N}\}$, respectively.

Example 3. The Gaussian white noise case is covered by assuming that ξ is given by a centered cylindrical Gaussian probability with identical covariance, i.e.,

$$\begin{aligned} \mathbb{E}\langle \xi, v\rangle &= 0 \qquad \forall v \in \mathcal{Y} \\ \mathbb{E}\langle \xi, v\rangle^2 &= \|v\|^2 \qquad \forall v \in \mathcal{Y}. \end{aligned}$$

Example 4. The case of deterministic noise is covered by fixing a set $\mathcal{N} \subset \mathcal{Y}$ and letting $\mathcal{Q} = \{\delta_\xi, \quad \xi \in \mathcal{N}\}$ the corresponding set of degenerate probabilities. In particular, for any $g : \mathcal{Y} \to \mathbb{R}$ it holds that $\mathbb{E}_Q g(\xi) = g(\xi)$.

Remark 1. If all measures from $\mathcal{P}$ and $\mathcal{Q}$ are Radon, and R is linear continuous, then the distribution of y^δ is Radon, and any such mapping R may be used as reconstruction. This applies to the worst case setting.

If the reconstruction mapping $R\colon \mathcal{Y} \to \mathcal{X}$ is linear, then it transforms arbitrary cylindrical probabilities (with weak second moments) into Radon ones (it is *radonifying*)

if and only if its singular numbers are square summable, i.e., it is a Hilbert–Schmidt operator. As standard reference for cylindcical and Radon probabilities we mention [13].

Finite-rank linear mappings have square summable singular numbers, and hence spectral cut-off schemes as in Example 1 fulfill this requirement.

Tikhonov regularization from Example 2 represents a Hilbert–Schmidt operator if and only if the underlying operator A has square summable singular numbers, and we sketch this: The singular numbers of the mapping in (7) are $t_k/(t_k^2+\alpha),\ k=1,2,\dots$ There are only finitely many k for which $t_k^2 \geq \alpha > 0$, and hence the summability is determined by those with $t_k^2 \leq \alpha$, in which case

$$\sum_{k:t_k^2<\alpha} \frac{t_k^2}{(t_k^2+\alpha)^2} \asymp \frac{1}{\alpha^2} \sum_{k:t_k^2<\alpha} t_k^2,$$

which is finite exactly if the t_k are square summable. The latter explains the difficulties for using Tikhonov regularization for statistical ill-posed problems.

Assumption 2.1 is considered a *standing assumption*, and in the analysis below this will be tacitly assumed without further mentioning.

We highlight the corresponding error criteria in the four settings mentioned above, where we identify sets $\mathcal{M},\ \mathcal{N}$ of degenerate probabilities with the describing sets.

Worst case setting: $e(\mathcal{M},\mathcal{N},R,\delta) = \sup_{x\in\mathcal{M}} \sup_{\xi\in\mathcal{N}} \|x - R(y^\delta)\|$,
Bayesian setting: $e(P,Q,R,\delta) = \left(\mathbb{E}_P\mathbb{E}_Q\|x - R(y^\delta)\|^2\right)^{1/2}$,
Statistical setting: $e(\mathcal{M},Q,R,\delta) = \sup_{x\in\mathcal{M}} \left(\mathbb{E}_Q\|x - R(y^\delta)\|^2\right)^{1/2}$, and
Average case setting: $e(P,\mathcal{N},R,\delta) = \sup_{\xi\in\mathcal{N}} \left(\mathbb{E}_P\|x - R(y^\delta)\|^2\right)^{1/2}$.

Notice that in the worst case, Bayesian and the statistical settings the errors are given as usual.

Remark 2. The average case setting has not been treated so far, but the change from the worst case to the average case setting could also result in

$$e^{avg}(P,\mathcal{N},R,\delta) = \left(\mathbb{E}_P \sup_{\xi\in\mathcal{N}} \|x - R(y^\delta)\|^2\right)^{1/2}.$$

This criterion is stronger than the one from above, but there are issues of measurability. Nonetheless, there are initial studies [9], where such error criterion is used.

2.3. **Parameter choice.** Suppose that, given data y^δ, we have constructed (by some means) a collection $\{x_n^\delta\}_n$ of regularized (i.e. approximate) solutions $x_n^\delta = R(\alpha_n, y^\delta)$, and α_n is the regularization parameter controlling the regularization.

Our goal is to choose the best (or a near to best) representer within the family $\{x_n^\delta\}_n$ which means we would like to minimize or control the error $\|x_n^\delta - x\|$. This will be

done by minimization of some carefully chosen function

$$f : \mathbb{N} \to \mathbb{R}^+, \tag{9}$$

and the chosen regularization parameter n_* is defined as

$$n_* = \operatorname*{argmin}_n f(n). \tag{10}$$

Example 5 (Quasi-Optimality). For the quasi-optimality criterion we let

$$f(n) = \|x_n^\delta - x_{n+1}^\delta\|, \quad n \in \mathbb{N}.$$

Remark 3. We stress the following. If either x or ξ are deterministic, and the sets $\mathcal{M}$, $\mathcal{N}$ are 'small' then the issue of finding *optimal* reconstructions is useless. Within the present model, for instance if $P = \delta_x$, then we could take as reconstruction $y^\delta \mapsto R(y^\delta) = x$, regardless of the data y^δ. Thus the present analysis focuses of the *performance* of given regularization strategies rather than on *finding* optimal methods. Of course, the performance is then measured *uniformly* with respect to $x \in \mathcal{M} \subset \mathcal{X}$, or $\xi \in \mathcal{N} \subset \mathcal{Y}$, respectively. The sets $\mathcal{M}$ and $\mathcal{N}$ then represent *typical instances* for data or noise.

Having fixed a function f as in (9), which is $P \times Q$-measurable, we assign its expected variant

$$F(n) := F_{P,Q}(n) = \left(\mathbb{E}_{P,Q} f(n)^2\right)^{1/2}, \quad n \in \mathbb{N}, \tag{11}$$

which should be close uniformly for $P \in \mathcal{M}$, $Q \in \mathcal{N}$. Finally, we also introduce the expected error

$$E(n) := E_{P,Q}(n) = \left(\mathbb{E}_{P,Q} \|x - x_n^\delta\|^2\right)^{1/2}, \quad n \in \mathbb{N}. \tag{12}$$

Remark 4. The proofs would not change (for properly modified assumptions) if we considered general moments for describing the expected error. However, as the knowledge gained is rather limited we will for now sacrifice generality for simplicity and just consider the second moments.

The key idea can be comprised as follows: If the expected functional F behaves like the error function E, and if the functional f is close to its expectation, then the minimization of f yields a convergent (and optimal) parameter choice. To make this precise we will impose a set of assumptions in § 3.1, and we will also provide more intuitive conditions in § 4.1.

3. Main result

The objective in this section is provide a set of assumptions for which a variety of parameter choices will have regularizing properties. Under an additional assumption such choices will even obey an oracle inequality.

We first exhibit the following useful calculus for the functions $E(n)$ and $F(n)$, respectively, in case that the minimization functional f corresponds to quasi-optimality,

i.e., as in Example 5. We use the splitting

$$x - x_n^\delta = (x - x_n) + (x_n - x_n^\delta),$$

and

$$x_n^\delta - x_{n+1}^\delta = (x_n - x_{n+1}) + \left((x_n^\delta - x_n) - (x_{n+1}^\delta - x_{n+1})\right). \tag{13}$$

Thus, if either the solution element x is drawn from a centered random element in the Hilbert space $\mathcal{X}$ or the noise probability Q is a centered (generalized) random element in the Hilbert space $\mathcal{Y}$ then

$$E_{P,Q}(n)^2 = \mathbb{E}_P\|x - x_n\|^2 + \mathbb{E}_Q\|x_n^\delta - x_n\|^2, \tag{14}$$

$$F_{P,Q}(n)^2 = \mathbb{E}_P\|x_n - x_{n+1}\|^2 + \mathbb{E}_Q\| \left(x_n^\delta - x_n\right) - \left(x_{n+1}^\delta - x_{n+1}\right) \|^2. \tag{15}$$

Therefore, in either of the above settings we must bound these summands uniformly for the probabilities from $\mathcal{P},\ \mathcal{Q}$.

Finally, by using the Hölder inequality, we have that

$$E(n) \leq \left(\mathbb{E}_{P,Q}\|x - x_n^\delta\|^{2\alpha}\right)^{1/(2\alpha)},$$

for $\alpha \geq 1$, and below we shall require that some converse also holds true.

3.1. **Assumptions.** Next, we gather the assumptions for the general analysis. We start with

Assumption 3.1 (existence of minimizer). For each $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ the function $F_{P,Q}$ is finite and there is a minimizing point $n_\#$, i.e.,

$$n_\# = \operatorname*{argmin}_n F_{P,Q}(n).$$

Furthermore, the following set of properties holds *uniformly* for $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$. We assume that there are constants

- $\alpha > 1$, controlling an additional moment, and we assign the corresponding dual index β satisfying $\alpha^{-1} + \beta^{-1} = 1$,
- r, controlling the decay of probabilities, and
- $\overline{n}$, controlling the region of uncertainty, and
- constants $c_1, \ldots, c_6 > 0$,

such that the following holds.

Assumption 3.2 (concentration of f). For $|n - n_\#| > \bar{n}$ it holds uniformly

$$\mathbb{P}_{P,Q}\left\{\frac{f(n_\#)}{F(n_\#)} > \sqrt{\frac{F(n)}{F(n_\#)}}\right\} \leq \frac{c_1}{2}\left(\frac{F(n_\#)}{F(n)}\right)^r,$$

and

$$\mathbb{P}_{P,Q}\left\{\frac{f(n)}{F(n)} < \sqrt{\frac{F(n_\#)}{F(n)}}\right\} \leq \frac{c_1}{2}\left(\frac{F(n_\#)}{F(n)}\right)^r.$$

Assumption 3.3 (moments of E). For each $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ the function $E_{P,Q}$ is finite, and there is a constant c_2 for which

$$\left(\mathbb{E}_{P,Q}\|x - x_n^\delta\|^{2\alpha}\right)^{1/(2\alpha)} \leq c_2 E(n)$$

Assumption 3.4 (concentration of E).

$$E(n) \leq c_3^{|n-m|} E(m).$$

Assumption 3.5 (decay rate for F). For $n > n_\# + \overline{n}$ it holds

$$\sum_{n=n_\#+\overline{n}}^{\infty} \left(\frac{F(n_\#)}{F(n)}\right)^r \leq c_4 < \infty$$

Assumption 3.6 (combined decay rate).

$$\sum_{|n-n_\#|\geq\overline{n}} \left(\frac{E(n)}{E(n_\#)}\right)\left(\frac{F(n_\#)}{F(n)}\right)^{r/(2\beta)} \leq c_5$$

For an optimality assertion we need the additional

Assumption 3.7 (oracle bound for $n_\#$).

$$E(n_\#) \leq c_6 \min_n E(n).$$

We shall provide sufficient conditions which ensure the validity of the assumptions when discussing relation to previous work, below.

3.2. **Statement of the main result.** We turn to state and prove the main result and we start with the following assertion.

Lemma 3.1. *Let $n_\#$ be a minimizer of F. Under assumptions 3.2 and 3.5 it holds that*

$$\mathbb{P}_{P,Q}\{n_* = n\} \leq c_1 \left(\frac{F(n_\#)}{F(n)}\right)^r, \quad n = 1, 2, \ldots,$$

and hence the minimizer n_* of f exists (P,Q)-almost surely. [1]

Proof. By the definition of $n_\#$ and n_* we can bound

$$\mathbb{P}_{P,Q}\{n_* = n\} = \mathbb{P}_{P,Q}\{f(m) \geq f(n) \text{ for all } m \in \mathbb{N}\} \leq \mathbb{P}_{P,Q}\{f(n_\#) \geq f(n)\}.$$

[1] If either the noise or the solution (or both) are deterministic then, according to Remark 3 the notion (P,Q)-almost surely means that this is to hold uniformly for the corresponding set $\mathcal{M}$, $\mathcal{N}$, respectively.

By Assumption 3.2 and for $n \geq n_\# + \bar{n}$ we deduce from $F(n_\#) \leq F(n)$ that

$$\begin{aligned}
\mathbb{P}_{P,Q}\{f(n_\#) > f(n)\} &\leq \mathbb{P}_{P,Q}\left\{f(n_\#) > \sqrt{F(n)F(n_\#)}\right\} \\
&\qquad + \mathbb{P}_{P,Q}\left\{f(n) < \sqrt{F(n)F(n_\#)}\right\} \\
&= \mathbb{P}_{P,Q}\left\{\frac{f(n_\#)}{F(n_\#)} > \sqrt{\frac{F(n)}{F(n_\#)}}\right\} + \mathbb{P}_{P,Q}\left\{\frac{f(n)}{F(n)} < \sqrt{\frac{F(n_\#)}{F(n)}}\right\} \\
&\leq c_1 \left(\frac{F(n_\#)}{F(n)}\right)^r,
\end{aligned}$$

which proves the first assertion. Furthermore, by Assumption 3.5 it holds

$$\sum_{n=n_\#+\bar{n}}^{\infty} \mathbb{P}_{P,Q}\{f(n) < f(n_\#)\} \leq c_1 \sum_{n=n_\#+\bar{n}}^{\infty} \left(\frac{F(n_\#)}{F(n)}\right)^r \leq c_4 < \infty$$

Due to the Borell-Cantelli Lemma, see e.g. [6, Chapt. VIII.3], we have

$$\mathbb{P}_{P,Q}\{\text{number of } n \text{ with } f(n) < f(n_\#) \text{ is finite}\} = 1$$

and hence the parameter n_* is finite with probability 1. □

This yields the main result.

Theorem 1. *Let $n_\#$ be as in Assumption 3.1. Under assumptions 3.2–3.6 there is a constant $C < \infty$ such that uniformly for $P \in \mathcal{P},\ Q \in \mathcal{Q}$ we have that*

$$E_{P,Q}(n_*) \leq C E_{P,Q}(n_\#). \tag{16}$$

If in addition the oracle bound from Assumption 3.7 holds then

$$E_{P,Q}(n_*) \leq \widetilde{C} \min_n E_{P,Q}(n). \tag{17}$$

Proof. By Lemma 3.1 the minimizer n_* exists with probability 1. For any pair P, Q, and using the Hölder inequality we get

$$\begin{aligned}
E(n_*) &= \left(\mathbb{E}_{P,Q}\|x - x^\delta_{n_*}\|^2\right)^{1/2} \\
&= \left(\sum_{n=1}^{\infty} \mathbb{E}_{P,Q}\left(\|x - x^\delta_n\|^2 \mathbf{1}_{n=n_*}\right)\right)^{1/2} \\
&\leq \left(\sum_{n=1}^{n_\#-\overline{n}} \left(\mathbb{E}_{P,Q}\|x - x^\delta_n\|^{2\alpha}\right)^{1/(2\alpha)} \left(\mathbb{E}_{P,Q}\mathbf{1}^{2\beta}_{n=n_*}\right)^{1/(2\beta)}\right. \\
&\qquad + c_3^{\overline{n}} \max_{n_\#-\overline{n}\leq n\leq n_\#+\overline{n}} E(n) \\
&\qquad \left. + \sum_{n=n_\#+\overline{n}}^{\infty} \left(\mathbb{E}_{P,Q}\|x - x^\delta_n\|^{2\alpha}\right)^{1/(2\alpha)} \left(\mathbb{E}_{P,Q}\mathbf{1}^{2\beta}_{n=n_*}\right)^{1/(2\beta)}\right) \\
&\leq c_2 \left(c_3^{\overline{n}} c_1^{1/(2\beta)} \sum_{n=1}^{n_\#-\overline{n}} E(n_\#)\frac{E(n)}{E(n_\#)}\left(\frac{F(n_\#)}{F(n)}\right)^{r/(2\beta)}\right. \\
&\qquad + c_3^{\overline{n}} E(n_\#) \\
&\qquad \left. + c_3^{\overline{n}} c_1^{1/(2\beta)} \sum_{n=n_\#+\overline{n}}^{\infty} E(n_\#)\frac{E(n)}{E(n_\#)}\left(\frac{F(n_\#)}{F(n)}\right)^{r/(2\beta)}\right) \\
&= c_2 \left(c_3^{\overline{n}} + c_3^{\overline{n}} c_1^{1/(2\beta)} \sum_{|n-n_\#|\geq\overline{n}} \left(\frac{E(n)}{E(n_\#)}\right)\left(\frac{F(n_\#)}{F(n)}\right)^{r/(2\beta)}\right) E(n_\#) \\
&\leq c_2 \left(c_3^{\overline{n}} + c_3^{\overline{n}} c_1^{1/(2\beta)} c_5\right) E(n_\#)
\end{aligned}$$

which proofs the claim with $C = c_2\left(c_3^{\overline{n}} + c_3^{\overline{n}} c_1^{1/(2\beta)} c_5\right)$. The oracle bound is an immediate application of Assumption 3.7, and the proof is complete. □

Remark 5. The assumptions 3.1–3.7 as well as the proofs of Lemma 3.1 and Theorem 1 do *not* depend on the linearity of the operator A, i.e., the analysis extends (in principle) also to non-linear inverse problems.

4. Discussion and relation to previous work

Next we provide a set of conditions which imply the assumptions made in § 3.1. These conditions have partly been set up in previous studies, in particular [4, 3], and we discuss the relations to such in some detail.

4.1. **Sufficient conditions.**

We start with using a general moment inequality (Kahane's Inequality) for Gaussian centered variables.

Lemma 4.1. *If, in the settings from § 2.2, the non-degenerate probabilities are centered Gaussian, then Assumption 3.3 holds true whenever regularization is linear in the data.*

Proof. We first treat the Bayesian setting with two Gaussian probabilities P and Q. In this case, and for $p > 1$, we use Kahane's Inequality, see [8, Cor. 3.2], to conclude that

$$\begin{aligned}\left(\mathbb{E}_{P,Q}\|x - x_n^\delta\|^p\right)^{1/p} &\leq \left(\mathbb{E}_{P,Q}\|x - x_n\|^p\right)^{1/p} + \left(\mathbb{E}_{P,Q}\|x_n - x_n^\delta\|^p\right)^{1/p} \\ &\leq C_p\left(\left(\mathbb{E}_{P,Q}\|x - x_n\|^2\right)^{1/2} + \left(\mathbb{E}_{P,Q}\|x_n - x_n^\delta\|^2\right)^{1/2}\right) \\ &\leq \sqrt{2}C_p\left(\left(\mathbb{E}_{P,Q}\|x - x_n\|^2\right) + \left(\mathbb{E}_{P,Q}\|x_n - x_n^\delta\|^2\right)\right)^{1/2} \\ &= \sqrt{2}C_p E_{P,Q}(n).\end{aligned}$$

In the worst case setting the statement is trivial. Finally, in either the statistical or the average case setting with centered Gaussian probability, the proof is similar, and we provide the one for the statistical setting, i.e., when Q is centered Gaussian. Then, similar as above, we bound

$$\begin{aligned}\left(\mathbb{E}_{Q}\|x - x_n^\delta\|^p\right)^{1/p} &\leq \|x - x_n\| + \left(\mathbb{E}_{Q}\|x_n - x_n^\delta\|^p\right)^{1/p} \\ &\leq \|x - x_n\| + C_p\left(\mathbb{E}_{Q}\|x_n - x_n^\delta\|^2\right)^{1/2} \\ &\leq C_p\sqrt{2}\left(\|x - x_n\|^2 + \mathbb{E}_{Q}\|x_n - x_n^\delta\|^2\right)^{1/2} \\ &= C_p\sqrt{2}\left(\mathbb{E}_{Q}\|x - x_n^\delta\|^2\right)^{1/2}.\end{aligned}$$

For any degenerate $P = \delta_x,\ x \in \mathcal{M}$ we may thus conclude that

$$\begin{aligned}\left(\mathbb{E}_{P,Q}\|x - x_n^\delta\|^p\right)^{1/p} &= \left(\mathbb{E}_{Q}\|x - x_n^\delta\|^p\right)^{1/p} \\ &\leq C_p\sqrt{2}\left(\mathbb{E}_{Q}\|x - x_n^\delta\|^2\right)^{1/2} = C_p\left(\mathbb{E}_{P,Q}\|x - x_n^\delta\|^2\right)^{1/2},\end{aligned}$$

and the proof is complete. □

Further conditions are more specific, and tied to the used parameter choice, and the following set of conditions was highlighted in [4, Ass..1], and [3, Eq. (10), (14) & (21)], and can be used for the quasi-optimality criterion.

Assumption 4.1 (bias decay). There are constants $w_1, w_2 > 1$ such that uniformly for $P \in \mathcal{P}$ it holds that

$$w_1^2\mathbb{E}_P\|x - x_{n+1}\|^2 \leq \mathbb{E}_P\|x - x_n\|^2 \leq w_2^2\mathbb{E}_P\|x - x_{n+1}\|^2 < \infty.$$

Assumption 4.2 (noise propagation). There are constants $w_3, w_4 > 1$ such that uniformly for $Q \in \mathcal{Q}$ it holds that

$$w_3^2\mathbb{E}_Q\|x_n^\delta - x_n\|^2 \leq \mathbb{E}_Q\|x_{n+1}^\delta - x_{n+1}\|^2 \leq w_4^2\mathbb{E}_Q\|x_n^\delta - x_n\|^2 < \infty.$$

Remark 6. We stress that Assumption 4.1 only depends on the set $\mathcal{P}$. It is constant with respect to $Q \in \mathcal{Q}$, and hence we may replace the above expectations $\mathbb{E}_P$ by $\mathbb{E}_{P,Q}$. A similar remark applies to Assumption 4.2, which only depends on $Q \in \mathcal{Q}$ and is constant with respect to $P \in \mathcal{P}$.

To proceed it is convenient to introduce the following set of abbreviations (for fixed probabilities P, Q):

$$b(n) := \|x - x_n\|, \qquad B(n) := \left(\mathbb{E}_{P,Q} b^2(n)\right)^{1/2}, \tag{18}$$

$$s(n) := \|x_n^\delta - x_n\|, \qquad S(n) := \left(\mathbb{E}_{P,Q} s^2(n)\right)^{1/2}, \tag{19}$$

$$d(n) := \left\|\left(x_{n+1}^\delta - x_{n+1}\right) - \left(x_n^\delta - x_n\right)\right\|, \quad D(n) := \left(\mathbb{E}_{P,Q} d^2(n)\right)^{1/2}, \tag{20}$$

$$g(n) := \|x_{n+1} - x_n\|, \qquad G(n) := \left(\mathbb{E}_{P,Q} g^2(n)\right)^{1/2}. \tag{21}$$

With this notation the assumptions 4.1 and 4.2 translate to

$$1 < w_1 \leq \frac{B(n)}{B(n+1)} \leq w_2, \quad \text{and} \quad 1 < w_3 \leq \frac{S(n+1)}{S(n)} \leq w_4. \tag{22}$$

Lemma 4.2. *The following assertions hold true.*
Under Assumption 4.1

- *the bias B is an exponentially decreasing sequence, and*
- *Uniformly for $P \in \mathcal{P}$ we have that $B(n) \asymp G(n)$.*

Under Assumption 4.2

- *the noise propagation S is exponentially increasing, and*
- *uniformly for $Q \in \mathcal{Q}$ we have that $D(n) \asymp S(n)$.*

Proof. Let $n_2 > n_1$. By iterating the left hand side in (22) $(n_2 - n_1)$-times we have that $B(n_2) \leq w_1^{-(n_2-n_1)} B(n_1)$, which proves the first assertion. Moreover, we have under Assumption 4.1 that

$$(w_2 - 1)B(n) \leq B(n) - B(n+1) \leq G(n) \leq B(n) + B(n+1) \leq (1 + w_1)B(n),$$

proving the asymptotic equivalence. The conclusions under Assumption 4.2 are proved similarly, and hence omitted. □

The crucial observation is comprised in the following lemma.

Lemma 4.3. *Let f be as in Example 5 and $n_\#$ be from Assumption 3.1. Suppose that assumptions 4.1 and 4.2 hold. Then there is a number $\overline{n}$ such that for $n > n_\# + \overline{n}$:*

$$F(n) \asymp E(n) \asymp D(n) \asymp S(n), \tag{23}$$

for $n < n_\# - \overline{n}$:

$$F(n) \asymp E(n) \asymp B(n) \asymp G(n). \tag{24}$$

Consequently, it holds that $E(n) \asymp F(n) \asymp \max\{B(n), S(n)\}$, and $E(n)$, $F(n)$ decrease for $n < n_\# - \overline{n}$ and increase for $n > n_\# + \overline{n}$ exponentially, with constants uniformly for $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Proof. Fix probabilities P and Q. First, assume that $S(1) < B(1)$. In this case we let n_+ be the last point where $S(n) \leq B(n)$, i.e., where

$$\gamma B(n_+ + 1) \leq S(n_+) \leq B(n_+),$$

for some constant $\gamma > 0$. This parameter n_+ allows for the following implications. If $m > n_+$ then, by Lemma 4.2, we have that

$$\frac{S(m)}{S(n_+)} \geq w_3^{m-n_+}, \quad \text{and} \quad \frac{B(n_+ + 1)}{B(m)} \geq w_1^{m-n_+-1}.$$

Therefore, we conclude

$$\begin{aligned} S(m) &\geq w_3^{m-n_+} S(n_+) \geq \gamma w_3^{m-n_+} B(n_+ + 1) \geq \gamma w_3^{m-n_+} w_1^{m-n_+-1} B(m) \\ &= \frac{\gamma}{w_1} (w_3 w_1)^{m-n_+} B(m). \end{aligned}$$

Similarly, we conclude for $k < n_+$ that

$$B(k) \geq w_1^{n_+-k} B(n_+) \geq w_1^{n_+-k} S(n_+) \geq w_1^{n_+-k} w_3^{n_+-k} S(k).$$

Thus, there are $\widetilde{n}$ and $c > 1$ for which

$$S(m) \geq cB(m), \quad \text{and} \quad B(k) \geq cS(k), \text{ whenever } k + \widetilde{n} < n_+ < m - \widetilde{n}.$$

Plainly, by enlarging $\widetilde{n}$ we may increase $c > 1$, and hence we may assume that $\frac{c+1}{c-1} < \min\{w_3, w_2\}$.

In the second case, when $S(1) > B(1)$, then $S(n) \geq S(1) \geq B(1) \geq B(n)$, and $S(m) \geq cB(m)$ for $m \geq \widetilde{n}$, and we may let $n_+ := 1$.

We turn to proving the assertions. To establish (23) with constants n_+ and $\widetilde{n}$ we first consider the case that $n > n_+ + \widetilde{n}$. Using the triangle inequality we bound

$$E(n) \leq B(n) + S(n) \leq \frac{c+1}{c} S(n).$$

Similarly, we have that

$$E(n) \geq S(n) - B(n) \geq \frac{c-1}{c} S(n).$$

Also, it holds

$$F(n) \leq E(n) + E(n+1) \leq \frac{c+1}{c}(1 + w_4) S(n),$$

and finally that

$$\begin{aligned} F(n) \geq E(n+1) - E(n) &\geq \frac{c-1}{c} S(n+1) - \frac{c+1}{c} S(n) \\ &\geq \frac{1}{c} \left(w_3 (c-1) - (c+1) \right) S(n). \end{aligned}$$

Notice that $w_3 (c-1) - (c+1) > 0$.

The case $n < n_+ - \widetilde{n}$ is treated similarly, allowing to establish assertion (24) with constants n_+ and $\widetilde{n}$.

Obviously, n_+ and $n_\#$ do not need to coincide, however $n_+ - \widetilde{n} \leq n_\# \leq n_+ + \widetilde{n}$ and hence (23) and (24) hold for $\overline{n} = 2\widetilde{n}$. $\square$

The above lemma allows for the following implication.

Corollary 1. *The assumptions 4.1 and 4.2 imply the validity of the assumptions 3.1, 3.4, 3.5, and 3.7.*

Proof. Clearly, Lemma 4.3 implies the existence of the minimizer $n_{\#}$. This assertion also yields assumptions 3.4 and 3.5 (for every $r > 0$). It remains to show that Assumption 3.7 is valid. By using Lemma 4.3 we bound (for some constants $0 < C, \tilde{C} < \infty$) as follows:

$$E(n_{\#}) \leq CF(n_{\#}) = C \min_n F(n) \leq \tilde{C} \min_n E(n),$$

and the proof is complete. □

So far we have provided sufficient conditions to yield all but assumptions 3.2 and 3.6. By the exponential decay proved in Lemma 4.3 the latter assumption would follow from the first if $r > 0$ is large enough. Thus, we aim at providing conditions which ensure that Assumption 3.2 holds for every $r > 0$, uniformly for $P \in \mathcal{P},\ Q \in \mathcal{Q}$. To this end we draw another conclusion from the assumptions 4.1, 4.2. For quasi-optimality, the function f is related to the auxiliary functions d, g, see (13), and therefore the following requirement helps us.

Assumption 4.3 (tail behavior). There are constants C_1, C_2, and $r > 0$ for which

$$\mathbb{P}_P\left(\frac{g(n)}{G(n)} > \eta\right) \leq C_1\eta^{-2r}, \quad \mathbb{P}_Q\left(\frac{d(n)}{D(n)} > \eta\right) \leq C_2\eta^{-2r},\ \eta > 1,$$

and there is n_0 such that for $n \geq n_0$ it holds

$$\mathbb{P}_P\left(\frac{g(n)}{G(n)} < \frac{1}{\eta}\right) \leq C_1\eta^{-2r}, \quad \mathbb{P}_Q\left(\frac{d(n)}{D(n)} < \frac{1}{\eta}\right) \leq C_2\eta^{-2r},\ \eta > 1,$$

uniformly for $P \in \mathcal{P},\ Q \in \mathcal{Q}$.

Remark 7. Assumption 4.3 is trivially fulfilled whenever the probabilities are degenerate, and can be proved to hold for single centered Gaussian ones. The latter can be seen from Lemma 4.5, below.

Lemma 4.4. *Suppose that the assumptions 4.1, 4.2, and 4.3 hold. If f is the quasi-optimality function as in Example 5 then Assumption 3.2 holds for r.*

Proof. First, if both probabilities P and Q are degenerate, then $f(n) = F(n)$ and Assumption 3.2 holds trivially for all r and with constant 1.

Next, using the triangle inequality we have that $f(k) \leq d(k) + g(k)$, and therefore we can deduce that

$$\begin{aligned}
\mathbb{P}\left(\frac{f(k)}{F(k)} > \eta\right) &\leq \mathbb{P}\left(\frac{d(k)}{D(k)} > \frac{F(k)}{D(k)}\frac{\eta}{2}\right) + \mathbb{P}\left(\frac{g(k)}{G(k)} > \frac{F(k)}{G(k)}\frac{\eta}{2}\right) \\
&\leq \mathbb{P}_Q\left(\frac{d(k)}{D(k)} > \frac{\eta}{2}\right) + \mathbb{P}_P\left(\frac{g(k)}{G(k)} > \frac{\eta}{2}\right) \\
&\leq 2^{2r}\left(C_1 + C_2\right)\eta^{-2r},
\end{aligned}$$

which proves the first bound in Assumption 3.2 by letting $\eta := \sqrt{F(n)/F(n_\#)}$ and $k = n_\#$.

Furthermore, if $n > n_\# + \overline{n}$ we bound, by using that $f(n) \geq d(n) - g(n)$, as

$$\begin{aligned}
&\mathbb{P}\left(\frac{f(n)}{F(n)} < \sqrt{\frac{F(n_\#)}{F(n)}}\right) \leq \mathbb{P}\left(\frac{d(n)}{F(n)} < \sqrt{\frac{F(n_\#)}{F(n)}} + \frac{g(n)}{F(n)}\right) \\
&\leq \mathbb{P}\left(\frac{d(n)}{F(n)} < 2\sqrt{\frac{F(n_\#)}{F(n)}}\right) \mathbb{P}\left(\frac{g(n)}{F(n)} < \sqrt{\frac{F(n_\#)}{F(n)}}\right) \\
&\qquad\qquad + \mathbb{P}\left(\frac{d(n)}{F(n)} < 2\frac{g(n)}{F(n)}\right) \mathbb{P}\left(\frac{g(n)}{F(n)} > \sqrt{\frac{F(n_\#)}{F(n)}}\right) \\
&\leq \mathbb{P}\left(\frac{d(n)}{F(n)} < 2\sqrt{\frac{F(n_\#)}{F(n)}}\right) + \mathbb{P}\left(\frac{g(n)}{F(n)} > \sqrt{\frac{F(n_\#)}{F(n)}}\right).
\end{aligned}$$

We continue as follows. First, by using the asymptotics (23) we have that

$$\frac{D^2(n)}{F(n)F(n_\#)} \asymp \frac{F(n)}{F(n_\#)} > 1.$$

Therefore we bound

$$\mathbb{P}\left(\frac{d(n)}{F(n)} < 2\sqrt{\frac{F(n_\#)}{F(n)}}\right) = \mathbb{P}\left(\frac{d(n)}{D(n)} < 2\sqrt{\frac{F(n)F(n_\#)}{D(n)^2}}\right) \leq CC_2\left(\frac{F(n_\#)}{F(n)}\right)^r.$$

Similarly we use Lemma 4.2 to infer that $G(n) \asymp B(n) \leq B(n_\#) \leq F(n_\#)$, hence

$$\frac{F(n)F(n_\#)}{G^2(n)} \geq \frac{F(n)}{F(n_\#)} > 1,$$

and we bound

$$\mathbb{P}\left(\frac{g(n)}{F(n)} > \sqrt{\frac{F(n_\#)}{F(n)}}\right) = \mathbb{P}\left(\frac{g(n)}{G(n)} > \sqrt{\frac{F(n)}{F(n_\#)}}\right) \leq C_1\left(\frac{F(n_\#)}{F(n)}\right)^r.$$

The case that $k < n_\# - \overline{n}$ is treated similarly, and hence we omit the proof. □

Since we have established that the assumptions from Section 3.1 are fulfilled for quasi-optimality, we conclude the following

Corollary 2. *Suppose that the set of probabilities $\mathcal{P}$ and $\mathcal{Q}$ are such that the assumptions 4.1, 4.2, and 4.3 hold. Then quasi-optimality yields an oracle bound*

$$E(n_*) \leq \widetilde{C} \min_n E(n).$$

4.2. **Relation to previous work.** Here we shall exhibit how previous analysis is covered by the present setup. As already mentioned, in previous analysis both some worst case, Bayesian and 'intermediate' settings were considered. The parameter choice was quasi-optimality.

For the *worst case setting* the authors in [3] introduce assumptions, called (P) and (Q), with corresponding sets $\mathcal{M}$, $\mathcal{N}$. Then they show in Lemma 2.4, ibid., that (P) implies the growth bounds in Assumption 4.1, and in Lemma 2.6, ibid., that (Q) yields Assumption 4.2. Therefore, all the assumptions in § 4.1 are fulfilled, and Theorem 1 reproves their Theorem 3.9 for the worst case setting.

The authors in [4] study the *Bayesian setting*, and they impose as Ass. 2.1, ibid. exactly the growth constraints in assumptions 4.1 and 4.2. In Example 2.2, ibid., they verify the following: if either the solution element x is drawn from a centered Gaussian distribution, or the noise is distributed according to some Gaussian law, and if the variances decrease polynomially then the assumptions 4.1 and 4.2 are fulfilled.

The key to show, that Assumption 4.3 and hence Assumption 3.2 holds for every $r > 0$ within the Bayesian framework is the following result, see [4, Lem. 6.1].

Lemma 4.5. *Let* $Z = \sum_{k=1}^{\infty} \alpha_k^2 \zeta_k^2$ *with* $\sum_{k=1}^{\infty} \alpha_k^2 = 1$ *and* $\zeta_k \sim N(0,1)$ *i.i.d. Assume that* $\max_k \alpha_k > 0$. *Then*

$$\forall\, z \in (0,1):\ \mathbb{P}(Z \le z) \le \exp\Big(\frac{1 - z + \log(z)}{2 \max_k \alpha_k^2}\Big),\ \ \forall\, z > 0:\ \mathbb{P}(Z \ge z) \le \sqrt{2} e^{-z/4}.$$

Therefore, all the assumptions in § 4.1 are fulfilled, and Theorem 1 reproves their results for the Bayesian setting.

The authors in [3] also study the statistical setting with Gaussian white noise, and Theorem 3.9, ibid., corresponds to Theorem 1.

As we have shown here, the quasi-optimality principle can be applied in all four cases, the "Worst case setting", "Bayesian setting", "Statistical setting" and the "Average case setting". Even more, instead of imposing a single Gaussian probability for the statistical or Bayesian settings, our results extend to classes of such, provided that the bounds in Assumption 4.3 hold uniformly.

Example 6. We briefly sketch this for families $\{P_\gamma, \gamma \in \Gamma\}$ of centered Gaussian priors for the solution, i.e., we assume that each P_γ has diagonal covariance operator with respect to the $u_1, u_2, \dots$, and the diagonal elements, denoted by $\gamma_1, \gamma_2, \dots$ decay like $\gamma_k \asymp k^{-\nu_\gamma},\ \gamma \in \Gamma$. Suppose furthermore, that regularization is spectral cut-off. Then, for fixed $\gamma \in \Gamma$, as an easy calculation shows, we have for $g(n)$ from (21) that

$$g^2(n) = \sum_{k=l(n)+1}^{l(n+1)} \gamma_k^2 \eta_k^2,$$

with i.i.d. standard normal $\eta_1, \eta_2, \ldots$. In particular this yields that $G^2(n) = \sum_{j=l(n)+1}^{l(n+1)} \gamma_j^2$, and hence that

$$\frac{g^2(n)}{G^2(n)} = \sum_{k=l(n)+1}^{l(n+1)} \frac{\gamma_k^2}{\sum_{j=l(n)+1}^{l(n+1)} \gamma_j^2},$$

and Lemma 4.5 applies, with $\alpha_k^{(\gamma)} := \gamma_k^2 / \sum_{j=l(n)+1}^{l(n+1)} \gamma_j^2$. In order to fulfill (the first couple of bounds in) Assumption 4.3 for arbitrarily large $r > 0$ we need that

$$\max_{l(n)+1 \le k \le l(n+1)} \alpha_k^{(\gamma)} \to 0, \text{ as } n \to \infty,$$

uniformly for $\gamma \in \Gamma$. This results in a requirement for the exponents ν_γ, $\gamma \in \Gamma$, and leads to require that there are $1/2 < \nu_{min} \le \nu_{max} < \infty$ such that $\nu_\gamma \in [\nu_{min}, \nu_{max}]$, $\gamma \in \Gamma$. The same requirement ensures that Assumption 4.1 is fulfilled.

In a similar way one can treat families of Gaussian distributions for the noise, and we leave the details to the reader.

5. Residual Principle

Here we present a new minimization scheme; this time not operating in the solution space $\mathcal{X}$ but in the data space $\mathcal{Y}$. Numerical simulations and preliminary discussion on this principle can be found in [2]. To formulate this principle, we use the singular value decomposition of the operator A from (3). Regularization is done by spectral cut-off as in Example 1, yielding $x_n^\delta := A_{l(n)}^{-1} \mathsf{Q}_{l(n)} y^\delta, \quad n = 1, \ldots, N$.

According to this spacing we let $y_n^\delta := \mathsf{Q}_{l(n)} y^\delta$, $n = 1, \ldots, N$. The parameter choice now consists of choosing some discretization level, and we base our parameter choice (called Residual method) on the projected data, and we let

$$f(n) := \frac{\|y_N^\delta - y_n^\delta\|^2}{\sqrt{\operatorname{trace}(B_n^* B_n)}} \tag{25}$$

where

$$B_n := A(\mathsf{P}_{l(N)} - \mathsf{P}_{l(n)}), \ n = 1, \ldots, N. \tag{26}$$

and we let

$$n_* := \operatorname*{argmin}_{1 \le n \le N} f(n), \tag{27}$$

denote any minimizer of f.

5.1. **Bayesian framework.** We will analyze this parameter choice in a Bayesian setting, similar to the one presented in [4]. We assume the following prior probability P for x: All Fourier coefficients $\langle x, u_k\rangle,\ k = 1, \dots,$ are independently normally distributed according to $\mathcal{N}(0, \gamma_k^2)$, with a decreasing and square summable sequence $\gamma : \mathbb{N}_0^+ \to \mathbb{R}_0^+$.

The normalized error elements are also randomly chosen according to Q, i.e. the (formal) Fourier coefficients $\langle \xi, v_k\rangle,\ k = 1, \dots,$ are all independent and distributed according to the normal distribution $\mathcal{N}(0, \sigma_k^2)$. Below, the expectation is always with respect to the product probability, i.e., $\mathbb{E} := \mathbb{E}_{P,Q}$.

To keep computations simple we will restrict ourselves to the following model.

Assumption 5.1. Let $\mu > 0$, $\nu > 1/2$ and $-1/2 < \tau$.
We let $t_k := k^{-\mu},\ \gamma_k := k^{-\nu}$ and $\sigma_k := k^{\tau},\ k = 1, \dots$

Remark 8. Obviously it holds that $\mathbb{E}\|y^\delta - y_n^\delta\|^2 = \infty$ as long as σ is not decreasing too fast. Therefore we assume to know an upper bound $N > n_{opt}$ for the 'optimal' parameter n_{opt}. Hence instead of the residual $\|y^\delta - y_n^\delta\|$ we just consider the finite dimensional version $\|y_N^\delta - y_n^\delta\|$ in order to perform the analysis.

Let us briefly motivate the above weighting in (25) in the Bayesian framework. By a reasoning as for the representations (14) and (15) we have that

$$\begin{aligned}\mathbb{E}\|y_N^\delta - y_n^\delta\|^2 &= \mathbb{E}_P\|(\mathsf{P}_{l(N)} - \mathsf{P}_{l(n)})Ax\|^2 + \delta^2\mathbb{E}_Q\|(\mathsf{P}_{l(N)} - \mathsf{P}_{l(n)})\xi\|^2 \\ &= \sum_{k=l(n)+1}^{l(N)} t_k^2\gamma_k^2 + \delta^2 \sum_{k=l(n)+1}^{l(N)} \sigma_k^2. \qquad (28)\end{aligned}$$

The first term in (28), now called $\varphi(n) := \sum_{k=l(n)+1}^{l(N)} t_k^2\gamma_k^2$, is a fast decreasing function whereas the second term stays comparably stable, depending on the noise variances. In order to detect the change point more reliably it is advisable to multiply both summands with a moderately increasing function. The function $\varphi(\cdot)^{-1/2}$ would be preferable, however this is unknown to us. Since γ is supposed to be decreasing we can instead use that

$$\varphi(n) = \sum_{k=l(n)+1}^{l(N)} t_k^2\gamma_k^2 \geq c \sum_{k=l(n)+1}^{l(N)} t_k^2 = \operatorname{trace}\left(B_n^* B_n\right),$$

with operator B_n as in (26). In this case the function $\varphi(n)\left(\operatorname{trace}\left(B_n^* B_n\right)\right)^{-1/2}$ is still decreasing, and as long as the color of the noise is not too bad and the ill-posedness of the operator is moderate, the function

$$\left(\operatorname{trace}\left(B_n^* B_n\right)\right)^{-1/2} \sum_{k=l(n)+1}^{l(N)} \sigma_k^2$$

is increasing. The required trace can be either calculated directly or by using fast trace estimators. For the geometric discretization scheme $l(n),\ n = 1, \dots, N$, and

under Assumption 5.1 the representation (28) leads to

$$\mathbb{E}\|y_N^\delta - y_n^\delta\|^2 = \delta^2 \sum_{k=l(n)+1}^{l(N)} k^{2\tau} + \sum_{k=l(n)+1}^{l(N)} k^{-2\nu-2\mu} \text{ and } \operatorname{trace}(B_n^* B_n) = \sum_{l(n)}^{l(N)} k^{-2\mu}.$$

Elementary calculus allows to upper and lower bound both of these expressions, using the spacing $q > 1$, and yielding that

$$\begin{aligned}\mathbb{E}\|y_N^\delta - y_n^\delta\|^2 &\asymp q^{n(-2\nu-2\mu+1)} + \delta^2 q^{N(2\tau+1)},\\ \operatorname{trace}(B_n^* B_n) &\asymp q^{n(-2\mu+1)}.\end{aligned}$$

with constants independent of n, N and δ. Hence we have that

$$F^2(n) = \frac{\mathbb{E}\|y_N^\delta - y_n^\delta\|^2}{\sqrt{\operatorname{trace}(B_n^* B_n)}} \asymp q^{n(-2\nu-\mu+1/2)} + \delta^2 q^{N(2\tau+1)} q^{n(\mu-1/2)}, \tag{29}$$

which is a sum of a decreasing and an increasing function of n, provided that $\mu > 1/2$.

Obviously, the chosen regularization parameter depends on the choice of N: the bigger N, the worse the result. However we need to make sure that $N > n_{opt}$, and this will be discussed, below.

Similarly, the expected squared error can be computed as

$$E^2(n) = \mathbb{E}\|x - x_n^\delta\|^2 \asymp q^{n(-2\nu+1)} + \delta^2 q^{n(2\mu+2\tau+1)}. \tag{30}$$

We notice that the growth and decay rates of E and F are different, in general.

5.2. **Analysis.** The asymptotic error expansion (30) yields the following result, stated without proof.

Lemma 5.1. *For the minimal average error it holds that*

$$\min_n \sqrt{\mathbb{E}\|x - x_n^\delta\|^2} \le C_{13} \delta^{\frac{\nu-1/2}{\nu+\mu+\tau}}$$

This minimum is obtained at n_{opt} with

$$q^{n_{opt}} \asymp \delta^{-1/(\mu+\nu+\tau)}.$$

Remark 9. The value n_{opt} uses complete information about the model parameters and is not available. However, necessarily this yields that $q^N \ge c\delta^{-1/(\mu+\nu+\tau)}$.

We turn to proving some result for the minimizer $n_\#$ of the functional F from (29). As shown below, we can guarantee that $n_{opt} < N$ as long as

$$0 < \delta \le c q^{-N(\tau+1/2)}. \tag{31}$$

Proposition 1. *Suppose that the maximal discretization N is fixed, $\mu > 1/2$, and that assumption 5.1 holds. If (31) holds then there is a non-trivial (interior) minimizer $1 < n_\# < N$ of the functional F. The corresponding error $E(n_\#)$ obeys*

$$E(n_\#) = \left(\mathbb{E}\|x - x_{n_\#}^\delta\|^2\right)^{1/2} \asymp C(q^N)\delta^\kappa, \tag{32}$$

for $\kappa = \frac{\nu-1/2}{\mu+\nu-1/2}$. *The constant* $C(q^N)$ *is of the order*

$$C(q^N) = \left[q^N\right]^{\frac{(\nu-1/2)(\tau+1/2)}{\mu+\nu-1/2}}. \tag{33}$$

Proof. We use the expansion (29) of the functional F. Its minimizer $n_\#$ must obey that

$$q^{n_\#} \asymp \left[\frac{1}{\delta^2 q^{N(2\tau+1)}}\right]^{1/(2\mu+2\nu-1)}. \tag{34}$$

Inserting this into (30) we observe that the first summand is dominating, and hence we obtain that

$$E(n_\#) \asymp C(q^N)\delta^\kappa,$$

With exponent κ and constant $C(q^N)$ as stated. The proof is complete. □

We turn to proving that Theorem 1 applies for the function f of the residual method, and this will be based on Lemma 4.5.

Corollary 3. *Under the assumptions of Proposition 1 the assumptions 3.2–3.6 hold for the residual principle.*

Proof. First, for the Bayesian setting we can apply Lemma 4.1 to see that Assumption 3.3 holds.

Moreover, the assumptions 3.4 and 3.5 hold by the representations (29) and (30). It remains to establish the concentration for f. The quotient $f(n)^2/F(n)^2$ is a sum of squares of Gaussian random variables (non-central χ^2), and hence we shall apply Lemma 4.5. By construction, the coefficients α_k^2 are equal to

$$\alpha_k^2 = \frac{\delta^2\sigma_k^2 + s_k^2\gamma_k^2}{\mathbb{E}\|y_N^\delta - y_n^\delta\|^2},$$

such that the assumption of Lemma 4.5 holds trivially, see (28). In order to establish a rate for Assumption 3.2 we need to upper bound the maximal value $\max_{l(n)\le k\le l(N)} \alpha_k^2$. Since $2\tau > -1$, $n > \widetilde{n}$, and $c < 1/2$, we have that

$$\begin{aligned}
\max_{l(n)\le k\le l(N)} \left\{\frac{\delta^2 k^{2\tau} + k^{-2\nu-2\mu}}{\mathbb{E}\|y_N^\delta - y_n^\delta\|^2}\right\} &\le \max_{l(n)\le k\le l(N)} \left\{\frac{\delta^2 k^{2\tau}}{\mathbb{E}\|y_N^\delta - y_n^\delta\|^2}\right\} \\
&\qquad + \max_{l(n)\le k\le l(N)} \left\{\frac{k^{-2\nu-2\mu}}{\mathbb{E}\|y_N^\delta - y_n^\delta\|^2}\right\} \\
&\le \frac{\delta^2 l(n)^{2\tau} + \delta^2 l(N)^{2\tau} + l(n)^{-2\mu-2\nu}}{\mathbb{E}\|y_N^\delta - y_n^\delta\|^2} \\
&\le C_{11}\frac{\delta^2 q^{n(2\tau)} + \delta^2 q^{N(2\tau)} + q^{n(-2\mu-2\nu)}}{q^{n(-2\nu-2\mu+1)} + \delta^2 q^{N(2\tau+1)}} \\
&\le C_{11}\left(q^{-n} + q^{-N} + q^{-n}\right) \\
&\le 3C_{11}q^{-n}
\end{aligned}$$

(A lower bound of the same order is obtained by letting $k := l(n) = q^n$.) Hence using Lemma 4.5 the assumption 3.2 holds for arbitrarily large values r, provided that $\bar{n}$ is large enough. We conclude that Assumption 3.6 also is valid. The proof is complete. □

As a direct consequence of Theorem 1 we have the following.

Corollary 4. *Under the assumptions of Proposition 1 there is a constant $C < \infty$ such that*

$$\sqrt{\mathbb{E}\|x - x_{n_*}^{\delta}\|^2} \leq CC(q^N)\delta^{\frac{\nu-1/2}{\nu-1/2+\mu}}.$$

Remark 10. The above bound can be interpreted as follows. The probabilistic smoothness index ν corresponds to a deterministic index $\nu - 1/2$, and then the above bound is the optimal order bound for smoothness $\nu - 1/2$ and decay μ of the singular numbers under bounded deterministic noise. However, this is multiplied by a penalty $C(q^N)$. As mentioned in Remark 9 there is a lower bound for q^N, and when inserting this into (33) we see that we do not get optimal rates. However, we still get convergence despite of the presence of unknown colored noise, and the bound is better for τ close to $-1/2$. The latter means that we are close to the case of bounded noise.

Extensive stochastic experiments we performed and are reported in [2, § 4.9]. In contrast to the considerations above the parameter N was chosen independent of δ at machine precision. Both, for the white noise and the (unknown) colored noise cases, the method works considerably well. The choice of an exponential cut-off scheme is, regarding our experiments, not necessary in order to get stable results. This indicates the the given estimates are much too rough; however better results can only be expected when taking the correlation inherent in the structure of $f(\cdot)$ into account.

The authors in [2] also observe that the same parameter choice method works comparably well and stable for Tikhonov regularization, a regularization which has not been covered in the above proof.

Conclusion

The authors introduce a unified framework to understand the typical behavior of parameter choice in inverse problems. If it works in expectation and if it obeys certain stability then it is provably convergent, often even order optimal in a oracle sense.

This subsumes some of the previous studies for classical inverse problems and Bayesian analysis, where the parameter choice was either spectral cut-off or Tikhonov regularization. Here this extends to other parameter choice and to extended settings.

The study concludes with a new parameter choice, which was numerically tested before, and which receives theoretical justification in a Bayesian framework, here.

Acknowledgments

The first author would like to acknowledge gratefully the financial support by the Upper Austrian Technology and Research Promotion. Part of this work has been conducted during the *Mini Special Semester on Inverse Problems*, May 18–July 15, 2009, organized by RICAM (Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Science) Linz, Austria, and the second author thanks RICAM for being invited.

References

[1] A. Bakushinsky. Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion. *Comput. Maths. Math. Phys.*, 24(4):181–182, 1984.

[2] F. Bauer and M. Lukas. Comparing different parameter choice methods for inverse problems. work in progress, 2009.

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

[4] Frank Bauer and Markus Reiß. Regularization independent of the noise level: an analysis of quasi-optimality. *Inverse Problems*, 24(5):055009, 16 pp., 2008.

[5] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publisher, Dordrecht, Boston, London, 1996.

[6] William Feller. *An introduction to probability theory and its applications. Vol. I*. Third edition. John Wiley & Sons Inc., New York, 1968.

[7] U. Hämarik and U. Tautenhahn. On the monotone error rule for parameter choice in iterative and continuous regularization methods. *BIT*, 41(5, suppl.):1029–1038, 2001. BIT 40th Anniversary Meeting.

[8] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[9] Peter Mathé. Average case analysis of ill-posed equations. unpublished, talk presented at "Chemnitz Symposium on Inverse Problems", September 2008.

[10] V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

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

[12] Andreas Neubauer. The convergence of a new heuristic parameter selection criterion for general regularization methods. *Inverse Problems*, 24(5):055005, (10pp), 2008.

[13] Laurent Schwartz. *Radon measures on arbitrary topological spaces and cylindrical measures*. Published for the Tata Institute of Fundamental Research, Bombay by Oxford University Press, London, 1973. Tata Institute of Fundamental Research Studies in Mathematics, No. 6.