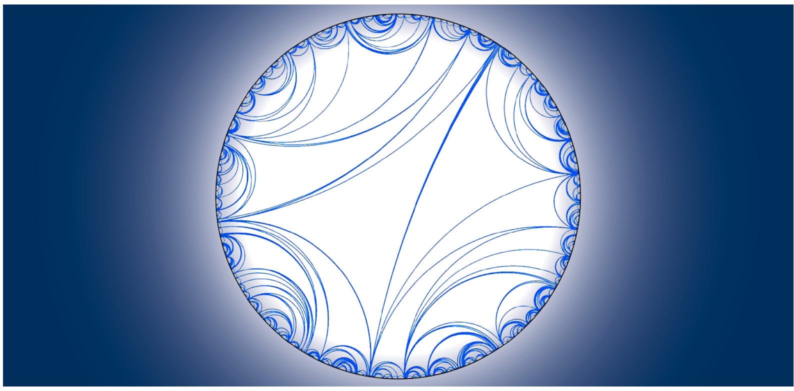
Universität Potsdam



Gilles Blanchard | Peter Mathé

Discrepancy Principle for Statistical Inverse Problems with Application to Conjugate Gradient Iteration

Preprints des Instituts für Mathematik der Universität Potsdam
1 (2012) 7

Preprints des Instituts für Mathematik der Universität Potsdam

Preprints des Instituts für Mathematik der Universität Potsdam
1 (2012) 7

Gilles Blanchard | Peter Mathé

Discrepancy Principle for Statistical Inverse Problems with Application to Conjugate Gradient Iteration

Universitätsverlag Potsdam

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind
im Internet über http://dnb.de abrufbar.

Universitätsverlag Potsdam 2012
http://info.ub.uni-potsdam.de/verlag.htm

Am Neuen Palais 10, 14469 Potsdam
Tel.: +49 (0)331 977 2533 / Fax: 2292
E-Mail: verlag@uni-potsdam.de

Die Schriftenreihe **Preprints des Instituts für Mathematik der Universität Potsdam**
wird herausgegeben vom Institut für Mathematik der Universität Potsdam.

ISSN (online) 2193-6943

Kontakt:
Institut für Mathematik
Am Neuen Palais 10
14469 Potsdam
Tel.: +49 (0)331 977 1028
WWW: http://www.math.uni-potsdam.de

Titelabbildungen:
1. Karla Fritze | Institutsgebäude auf dem Campus Neues Palais
2. Nicolas Curien, Wendelin Werner | Random hyperbolic triangulation
Published at: http://arxiv.org/abs/1105.5089

Online veröffentlicht auf dem Publikationsserver der Universität Potsdam
URL http://pub.ub.uni-potsdam.de/volltexte/2012/5711/
URN urn:nbn:de:kobv:517-opus-57117
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-57117

DISCREPANCY PRINCIPLE FOR STATISTICAL INVERSE PROBLEMS WITH APPLICATION TO CONJUGATE GRADIENT ITERATION

G. BLANCHARD AND P. MATHÉ

Dedicated to Ulrich Tautenhahn, a friend and co-author,
who passed away too early at the age of 60.

Abstract. The authors discuss the use of the discrepancy principle for statistical inverse problems, when the underlying operator is of trace class. Under this assumption the discrepancy principle is well-defined, however a plain use of it may occasionally fail and it will yield sub-optimal rates. Therefore, a modification of the discrepancy is introduced, which takes into account both of the above deficiencies. For a variety of linear regularization schemes as well as for conjugate gradient iteration it is shown to yield order optimal a priori error bounds under general smoothness assumptions. A posteriori error control is also possible, however at a sub-optimal rate, in general. This study uses and complements previous results for bounded deterministic noise.

1. Setting and notation

We consider the following inverse problem:

$$y^\sigma = Tx^\dagger + \sigma\xi \,, \tag{1}$$

for some bounded operator $T\colon X \to Y$ between Hilbert spaces, and noisy data y^σ with a Gaussian white noise ξ with noise level (standard deviation) $\sigma > 0$. The goal is to recover $x^\dagger$ as accurately as possible from the observation of y^σ. Note that equation (1) is actually abusive, since a Gaussian white noise ξ cannot be represented as a random element of Y if the latter space is infinite-dimensional. The formal meaning of (1) is the following: we observe a Gaussian random field y^σ over the Hilbert space Y such that for any $w \in Y$,

$$y^\sigma(w) = \langle Tx^\dagger, w\rangle + \sigma\xi(w) \,,$$

and having covariance structure $\mathrm{Cov}\left[y^\sigma(w), y^\sigma(w')\right] = \sigma^2\mathbb{E}\left[\xi(w)\xi(w')\right] = \sigma^2\langle w, w'\rangle$.

Remark 1. Such statistical inverse problems often arise in practical applications; the seminal study on optimal errors for regression (when T is the identity) in the present context is [12]. The analysis extends to inverse problems under Gaussian white noise, since then (1) is equivalent to a sequence model

$$y_j^\sigma = s_j x_j^\dagger + \sigma\xi_j, \quad j = 1, 2, \dots,$$

with s_j being the singular numbers of T. Optimal recovery rates for such statistical inverse problems are known for many cases, starting from [12], and we refer to [5] for a recent survey. We will recall some of these results below.

In general the solution of statistical inverse problems uses discretization, in many cases as described in Remark 1. However, if we turn to the "normal equations" associated to (1) by formally multiplying (1) by T^* on the left side,

$$z^\sigma := T^* y^\sigma = T^* T x^\dagger + \sigma T^* \xi := A x^\dagger + \sigma\zeta, \tag{2}$$

then the Gaussian noise $\zeta := T^*\xi$ has the covariance structure $\mathbb{E}\left[\zeta(w), \zeta(w')\right] = \langle w, Aw'\rangle$. Here we introduced the non-negative self-adjoint operator $A := T^*T$. This will be our main model from now on.

We will make the following assumption throughout the paper:

Assumption 1. The operator A has a finite trace $\mathrm{Tr}\left[A\right] < \infty$.

Remark 2. This assumption is similar to the assumptions considered in [1]. If the underlying operator does not have a finite trace then discretization is required to find a solution to the inverse problem (1). This approach has been studied in [10] for Tikhonov regularization based on discrete random data.

Under Assumption 1 the operator A maps weakly random elements to random elements, i.e., we can then represent the noise $\zeta = T^*\xi$ as a Gaussian random variable taking values in X, and we have that

$$\mathbb{E}\left[\|\zeta\|^2\right] = \mathrm{Tr}\left[A\right]. \tag{3}$$

In particular, the discrepancy $\|Ax - z^\sigma\|$ is almost surely well defined for any $x \in X$, so one might wish to use the *discrepancy principle* for this class of statistical inverse problems. In this study we analyze several regularization schemes, both linear regularization as well as conjugate gradient (hereafter **cg**) iteration to obtain sequences of approximate (candidate) solutions.

We first briefly indicate a priori error bounds, Theorem 1, which may serve as benchmark error bounds in the later analysis. This already highlights the role of the *effective dimension*, which takes into account the assumption of the statistical noise.

Then we turn to studying a posteriori error bounds by using the discrepancy. We propose a modification of the discrepancy principle, suited for statistical noise. The main result is a general theoretical error bound, given in Theorem 2, for both classes of regularization schemes. As corollaries, we derive a posteriori error bounds, but we also show that the modified discrepancy principle will give order optimal a priori error bounds provided the solution smoothness is known to us. The argument for proving Theorem 2 is essentially a reduction to a deterministic case through a concentration bound. The key observation is to use the discrepancy principle not in the original norm as stated above, but to apply it when the size of the residual is measured in some weighted norm, as presented in Definition 7.

Accordingly, the theoretical analysis is based on previous work on regularization under large (deterministic) noise, as carried out by the present authors in [2], and by P. Mathé, U. Tautenhahn in [11]. The same arguments can also be used for proving that the plain (unweighted) discrepancy principle can be applied, and that convergence holds with high probability. However, the obtained rates are sub-optimal in this case, which justifies the introduction of the proposed modification. Additionally, a complementary condition called "emergency stopping" is introduced to account for the small probability situation where the noise is not controlled through the concentration bound.

As far as we know this is the first rigorous study for using the discrepancy principle under statistical noise assumptions.

2. Regularization schemes

2.1. **General linear regularization.**

We recall the notion of linear regularization, see e.g. [7], in a slightly modified version.

Definition 1 (cf. [7, Def. 2.2])**.** A family of functions $g_\alpha(0, \|A\|) \mapsto \mathbb{R},\ 0 < \alpha < \infty$, of bounded measurable functions is called regularization if they are piece-wise continuous in α and the following properties hold, where $r_\alpha(t) := 1 - tg_\alpha(t),\ 0 < t \le \|A\|$, denotes the residual function.

(1) For each $0 < t \le \|A\|$ there is convergence $|r_\alpha(t)| \to 0$ as $\alpha \to 0$.
(2) There is a constant γ_1 such that $|r_\alpha(t)| \le \gamma_1$ for all $0 < \alpha < \infty$ and all $t \in (0, \|A\|]$.
(3) There is a constant $\gamma_* \ge \frac{1}{2}$ such that $\sup_{0<\alpha<\infty} \sup_{0<t\le\|A\|} \alpha\,|g_\alpha(t)| \le \gamma_*$.

Remark 3. First, we emphasize that item (3) is stronger than the assumption which is usually made. However, all relevant regularization schemes fulfill this stronger assumption. Such stronger condition is typical when applying linear regularization to general noise assumptions. Items (3) and (2) taken together also imply that there exists a numerical constant $C_{1/2}$ such that $|g_\alpha(t)|\sqrt{t} \le C_{1/2}/\sqrt{\alpha}$. Without loss of generality, we shall henceforth assume that γ_* is chosen with $\gamma_* \ge C_{1/2}$.

Secondly, we assumed that g_α is defined for all $0 < \alpha < \infty$. Most of the classical schemes are given in this way. If a scheme is initially defined for $0 < \alpha \le \bar{\alpha}$ for some $\bar{\alpha} > 0$ then it can be extended by letting $g_\alpha(t) = 1/\alpha,\ \alpha > \bar{\alpha},\ 0 < t \le \|A\|$. If this $\bar{\alpha} \ge \|A\|$ then item (2) still holds provided that $\gamma_1 \ge 1$. Moreover, the qualification, see Definition 4, below, is not affected in this case.

Given a regularization g_α we assign the approximate solution

$$x_\alpha^\sigma := g_\alpha(A)T^*y^\sigma, \quad \alpha > 0, \tag{4}$$

and for technical reasons, given $y = Tx$, the (unknown)

$$x_\alpha := g_\alpha(A)T^*y, \quad \alpha > 0. \tag{5}$$

Conditions (1) and (2) in Definition 1 ensure that for each $x \in X$ the family x_α converges to the true solution x as $\alpha \searrow 0$.

2.2. **Conjugate gradient iteration.** As a possible alternative to linear regularization, we also consider applying the standard **cg** iterations, defined as

$$\|z^\sigma - Ax_k^\sigma\| = \min\left\{\|z^\sigma - Ax\|, \quad x \in \mathcal{K}_k(z^\sigma, A)\right\}, \tag{6}$$

where the Krylov subspace $\mathcal{K}_k(z^\sigma, A)$ consists of all elements $x \in X$ of the form $x = \sum_{j=0}^{k-1} c_j A^j z^\sigma$.

Remark 4. Much more properties of the iterates x_k^σ are known. In particular there is a strong connection to the theory of orthogonal polynomials. Our subsequent analysis will not use those intrinsic techniques, since we will rely on a previous study [2] by the present authors, where the analysis for bounded deterministic noise was carried out.

The regularizing properties of **cg** are achieved by a stopping criterion, and we shall use (a version of the) discrepancy principle, below.

3. A priori error bounds for linear regularization

We shall present some general error bound which will later be used to discuss a posteriori parameter choice methods.

3.1. **Bias-variance decomposition.** Under statistical noise we have a natural bias-variance decomposition, which we are going to discuss next. Since the noise ζ is centered, so is the random element $g_\alpha(T^*T)\zeta$, which yields

$$\mathbb{E}\left[\|x - x_\alpha^\sigma\|^2\right] = \|x - x_\alpha\|^2 + \mathbb{E}\left[\|x_\alpha - x_\alpha^\sigma\|^2\right], \tag{7}$$

since the (squared) bias is deterministic, and it allows for a bound as usual, provided that x obeys some source condition, and that the regularization has the appropriate qualification.

We turn to bounding the variance. Tight bounds are crucial for order optimal reconstruction. Plainly, we have the representation

$$\mathbb{E}\left[\|x_\alpha - x_\alpha^\sigma\|^2\right] = \sigma^2 \mathrm{Tr}\left[g_\alpha^2(T^*T)T^*T\right] = \sigma^2 \mathrm{Tr}\left[g_\alpha^2(A)A\right]. \tag{8}$$

However, in order to treat this we shall use the *effective dimension.*

Definition 2 (effective dimension, see [3, 13])**.** The function $\mathcal{N}(\lambda)$ as

$$\mathcal{N}(\lambda) := \mathrm{Tr}\left[(A + \lambda I)^{-1} A\right], \ \lambda > 0 \tag{9}$$

is called effective dimension of the operator A under white noise.

Under Assumption 1 the operator A has a finite trace, and the operator $(A + \lambda I)^{-1}$ is bounded, thus the function $\mathcal{N}$ is finite. We shall provide further details below. By using the function $\mathcal{N}(\lambda)$ we can bound the trace in (8) as

$$\begin{aligned}\operatorname{Tr}\left[g_\alpha^2(A)A\right] &= \operatorname{Tr}\left[g_\alpha^2(A)(\alpha I + A)\left(\alpha I + A\right)^{-1} A\right] \\ &\leq \|g_\alpha^2(A)(\alpha I + A)\|\mathcal{N}(\alpha),\end{aligned}$$

and we shall further bound the norm on the right.

Lemma 3.1. *Let g_α be any linear regularization. Then we have*

$$\|g_\alpha^2(A)(\alpha I + A)\| \leq 2\frac{\gamma_*^2}{\alpha}.$$

Proof. This is a simple consequence of Definition 1. Indeed, we estimate

$$\|g_\alpha^2(A)(\alpha I + A)\| \leq \alpha\|g_\alpha^2(A)\| + \|g_\alpha^2(A)A\| \leq \alpha\|g_\alpha^2(A)\| + \|g_\alpha(A)A^{1/2}\|^2.$$

An application of item (3) in Definition 1, taking into account Remark 3, allows to complete the proof. □

Summarizing the above discussion we have the following.

Proposition 1. *Under Assumption 1, and for any linear regularization g_α we have the error bound*

$$\mathbb{E}\left[\|x - x_\alpha^\sigma\|^2\right]^{1/2} \leq \|x - x_\alpha\| + \sqrt{2}\gamma_*\sigma\sqrt{\frac{\mathcal{N}(\alpha)}{\alpha}}, \quad \alpha > 0.$$

The above variance bound is explicit, and it is known to yield the optimal order in many cases. This is possible by using the effective dimension from Definition 2. We will see later in this study that it naturally appears in error estimates.

Remark 5. The authors in [1] went another way, and run into additional assumptions, which are required to derive order optimal bounds. We briefly sketch a variation of that approach. The above way to describe the spectral distribution of the operator A through the effective dimension from (9) seems to be related to Tikhonov regularization which is given by the function $g_\alpha(A) := \left(A + \alpha I\right)^{-1}$, $\alpha > 0$, see the formal introduction of regularization schemes in Definition 1. If we replace this by *spectral cut-off* with $g_\alpha(\lambda) := \chi_{[\alpha,\infty)}(\lambda)\frac{1}{\lambda}$, $\alpha > 0$, then we obtain a different function

$$\mathcal{N}_{SC}(\lambda) := \operatorname{Tr}\left[\chi_{[\lambda,\infty)}(A)\right], \quad \lambda > 0. \tag{10}$$

This can be rewritten as $\mathcal{N}_{SC}(\lambda) := \#\left\{j,\ s_j \geq \lambda\right\}$, if s_j denote the singular numbers of A. As was discussed in [1, Rem. 1 & § 4.1], for our setup this just coincides with the function $R(\lambda)$, ibid.

The function $\mathcal{N}_{SC}$ enjoys similar properties as $\mathcal{N}$, it is finite under Assumption 1, non-increasing, and $\mathcal{N}(\lambda) \nearrow \infty$ as $\lambda \to 0$, if the operator has infinite-dimensional range. In particular $\lim_{\lambda\to 0} \frac{\operatorname{Tr}[A]}{\mathcal{N}_{SC}(\lambda)} = 0$.

This function also allows for a bound of the variance, since we have

$$\begin{aligned}\operatorname{Tr}\left[g_\alpha^2(A)A\right] &= \operatorname{Tr}\left[g_\alpha^2(A)A\chi_{[\alpha,\infty)}(A)\right] + \operatorname{Tr}\left[g_\alpha^2(A)A\chi_{(0,\alpha)}(A)\right]\\ &\le \frac{\gamma_*^2}{\alpha}\left(\mathcal{N}_{SC}(\alpha) + \operatorname{Tr}\left[A\right]\right)\\ &= \frac{\gamma_*^2}{\alpha}\mathcal{N}_{SC}(\alpha)\left(1 + o(1)\right), \quad \text{as } \alpha \to 0.\end{aligned}$$

Therefore, the assertion of Proposition 1 extends by replacing the function $\mathcal{N}$ by the function $\mathcal{N}_{SC}$. Since we have that $\mathcal{N}_{SC}(\lambda) \le 2\mathcal{N}(\lambda)$, because $\chi_{[\lambda,\infty)}(t) \le 2\frac{t}{\lambda+t}$, the latter bound is (formally) sharper, although for typical decay rate of the singular numbers of A the growth rates coincide.

3.2. **A priori error bound.** It will be convenient to introduce the function

$$\varrho_{\mathcal{N}}(t) := \frac{1}{\sqrt{2t\mathcal{N}(t)}}, \quad t > 0. \tag{11}$$

We agree to assign to each function $f\colon (0,\infty) \to (0,\infty)$ the function

$$\Theta_f(t) := tf(t),\ t > 0, \tag{12}$$

in particular we consider the function $\Theta_{\varrho_{\mathcal{N}}}$, which is a bijection from $[0,\infty)$ onto itself. With this function at hand we have

$$\sqrt{2}\sqrt{\frac{\mathcal{N}(\alpha)}{\alpha}} = \sqrt{\frac{2\alpha\mathcal{N}(\alpha)}{\alpha^2}} = \frac{1}{\alpha\varrho_{\mathcal{N}}(\alpha)} = \frac{1}{\Theta_{\varrho_{\mathcal{N}}}(\alpha)}, \quad \alpha > 0.$$

Next, we introduce solution smoothness in terms of general source conditions.

Definition 3. We call a function $\psi\colon [0,\infty) \to [0,\infty)$ an *index function* if ψ is an non-decreasing, continuous, positive function on $(0,\infty)$ such that $\psi(0) = 0$.

Assumption 2 (general source condition, see e.g. [7])**.** There is an index function ψ with $\psi(t) = \|A\|$ for all $t \ge \|A\|$, and such that

$$x \in A_\psi := \{\psi(A)v, \quad \|v\| \le 1\}.$$

Remark 6. Definition 3 of an index function modifies previous definitions, see e.g. [7] by formally extending the domain of definition to all of $\mathbb{R}_+$ instead of just $[0,\|A\|]$. In Assumption 2 it is also assumed that $\psi(t)$ is constant for $t \ge \|A\|$, implying in particular that $\psi(t) \le \psi(\|A\|)$ for all $t \ge 0$. The reason for this is merely to avoid technical problems, as it allows to define an inverse defined on the whole positive real line to the related functions $\Theta_{\varrho_{\mathcal{N}}\psi}$. Obviously, this has no impact on the strength of the source condition, which only depends on the value of ψ on $[0,\|A\|]$. Furthermore, the convergence rates (as $\sigma \to 0$) only depend on the behavior of ψ near zero.

Finally, in order for the chosen regularization to take the given smoothness into account, we need the classical assumption that it has sufficient *qualification*, defined as follows:

Definition 4. Let φ be an index function. The linear regularization g_α is said to have qualification φ if there is a constant $\gamma > 0$ for which

$$\sup_{0<t\leq\|A\|} |r_\alpha(t)|\,\varphi(t) \leq \gamma\varphi(\alpha), \quad \alpha > 0. \tag{13}$$

We are now in a position to formulate and prove the a priori error bound.

Theorem 1. *Suppose that assumptions 1 and 2 hold, and that the regularization g_α has qualification ψ. Then*

$$\mathbb{E}\left[\|x - x_\alpha^\sigma\|^2\right]^{1/2} \leq \gamma\psi(\alpha) + \gamma_* \frac{\sigma}{\Theta_{\varrho_{\mathcal{N}}}(\alpha)},\ \alpha > 0.$$

Given $\sigma > 0$, let α_* be the a priori parameter choice from

$$\Theta_{\varrho_{\mathcal{N}}\psi}(\alpha_*) = \sigma. \tag{14}$$

This choice satisfies

$$\mathbb{E}\left[\|x - x_{\alpha_*}^\sigma\|^2\right]^{1/2} \leq (\gamma + \gamma_*)\psi(\Theta_{\varrho_{\mathcal{N}}\psi}^{-1}(\sigma)).$$

Proof. The error decomposition is a reformulation of the estimate in Proposition 1. The second assertion follows from balancing both summands in that bound. □

Remark 7. In view of Remark 5 the bound in Theorem 1 remains true by replacing the effective dimension $\mathcal{N}$, used in $\varrho_{\mathcal{N}}$, by the variation $\mathcal{N}_{SC}$ from (10).

In contrast to the deterministic noise setting, the obtained rate thus also depends on the spectral decay properties of the operator A through the effective dimension function $\mathcal{N}$. In specific cases, the bound from Theorem 1 is known to be of optimal order, and we exhibit this.

Example 1 (moderately ill-posed problem)**.** Suppose that the operator T has singular numbers $s_j(T) \asymp j^{-r}$ for some $r > 0$. Then the singular numbers $s_j(A)$ obey $s_j(A) \asymp j^{-2r}$. In order for A to be a trace class operator we need that $r > 1/2$. We assign $s := 1/(2r) \in (0, 1)$. Then we get the following asymptotics: $\mathcal{N}(t) \asymp t^{-s}$, see [4]. Notice that the modified function $\mathcal{N}_{SC}$ exhibits the same asymptotics.

Suppose that smoothness is given by $\psi(t) = t^{\nu/(2r)} = t^{s\nu}$. It is easy to see that $\Theta_{\varrho_{\mathcal{N}}}(t) \asymp t^{s\nu+1/2+s/2}$, so that

$$\psi\left(\Theta_{\varrho_{\mathcal{N}}\psi}^{-1}(\delta)\right) \asymp \delta^{\frac{s\nu}{s\nu+1/2+s/2}} = \delta^{\frac{\nu}{\nu+1/(2s)+1/2}} = \delta^{\frac{\nu}{\nu+r+1/2}}, \quad \text{as } \delta \to 0.$$

This is of optimal order (for $\nu \leq 2r$) under Gaussian white noise, as was e.g. shown in [10], but this also is known to hold for general $\nu > 0$, and this follows as in [12, 5].

Example 2 (severely ill-posed problem)**.** If $s_j(A) \asymp \exp(-cj)$, then $\mathcal{N}(t) \asymp \mathcal{N}_{SC}(t) \asymp \frac{1}{c}\log(1/t)$ in a neighborhood of the origin. Hence, for smoothness $\psi(t) \asymp \log^{-s} 1/t$ we obtain that $\psi\left(\Theta_{\varrho_{\mathcal{N}}\psi}^{-1}(\delta)\right) \asymp \log^{-s} 1/\delta$, which is the same rate as for severely ill-posed problems under bounded deterministic noise, known to be the optimal order of reconstruction.

4. Discrepancy principle

4.1. Discrepancy principle under bounded deterministic noise.

Our analysis of the discrepancy principle for random noise will rely heavily on existing results for the discrepancy principle in the classical (deterministic) setup, albeit under general noise assumptions. These results are quite recent, and for the convenience of the reader we state those. For the duration of this section, we assume that the noise is deterministic. We therefore consider the following deterministic counterparts of (1) and (2):

$$y^\delta = Tx^\dagger + \eta\,, \tag{15}$$

$$z^\delta = T^*y^\delta = T^*Tx^\dagger + T^*\eta = Ax^\dagger + \varepsilon\,. \tag{16}$$

In this display, η and ε play the role of $\sigma\xi$ and $\sigma\zeta$ in (1) and(2). The notation change is to emphasize that ε, η are deterministic. Similarly to the random setting, we will mainly focus on the formulation (16) of the problem (normal equations).

The reason to recall here known facts on deterministic noise regularization is that we will apply these in the random setting when ε is a fixed *realization* of the random noise $\sigma\zeta$. Correspondingly, results available in the deterministic setting will hold in the random setting with high probability, namely on the event that the noise realization has controlled amplitude.

In the deterministic setting, we assume the following control of the noise amplitude is known:

Assumption 3. There is a non-negative non-increasing function ϱ such that the function $t \mapsto t\varrho(t)$ is (strictly) increasing from 0 to ∞, and for which

$$\|\varrho(A)\varepsilon\| \le \delta.$$

Remark 8. First, the limiting cases for such function ϱ are $\varrho(t) \equiv 1$, which corresponds to *large noise*, and the function $\varrho(t) = 1/t$, in which case the problem is well-posed.

Assumption 3 as formulated here is taken from [2]. Regularization under general noise was also treated in [11], where a different scaling was taken. Here we control the noise ε, whereas in that study the assumption was imposed on η. Of course, both scalings can be aligned through an appropriate choice of the function ϱ, and we shall hence constrain ourselves to the one made above. Below, we will further specify the weights, parametrized by a parameter $\lambda > 0$ as

$$\varrho_\lambda(t) := \frac{1}{\sqrt{t+\lambda}}, \quad t, \lambda > 0. \tag{17}$$

We recall the discrepancy principle (DP) as this was used in the above studies [2] and [11]. Let $(x_k^\delta := \phi_k(z^\delta))_{k\ge 0}$, a sequence of reconstructions, where ϕ_k is a fixed sequence of reconstruction functions.The discrepancy principle in its generic form is given as follows.

Definition 5 (discrepancy principle (DP))**.** Let ϱ be an index function and $\delta > 0$. Given $\tau > 1$ we let $k_{DP}(\tau, \varrho, \delta)$ be the first $k \geq 0$ for which

$$\|\varrho(A)(z^\delta - Ax_k^\delta)\| \leq \tau\delta,$$

The control of the residual $z^\delta - Ax_k^\delta$ as expressed above can be reformulated in terms of the original residual $y^\delta - Tx_k^\delta$, since

$$\|\varrho(A)(z^\delta - Ax_k^\delta)\| = \|\varrho(A)T^*(y^\delta - Tx_k^\delta)\| = \|(TT^*)^{1/2}\varrho(TT^*)(y^\delta - Tx_k^\delta)\|.$$

Thus, the above definition is equivalent to a bound for the original residual $y^\delta - Tx_k^\delta$ using the corresponding weight function $\tilde{\varrho}(t) := \sqrt{t}\varrho(t)$ and operator $B := TT^*$.

The above discrepancy principle is instantiated as follows in the two regularization schemes we are interested here. For linear regularization, the regularization parameter α is discretized geometrically by defining $\alpha_k := \alpha_0 p^k$ (where $p < 1$ is a user-determined constant) and posing $x_k^\delta := x_{\alpha_k}^\delta$ (with a slight overloading of notation). For **cg** iterations, x_k^δ is unambiguously the output of the k-th iteration (with the convention $x_0^\delta := 0$). For the statements below we also need in this case to assign to each step $k \geq 1$ of **cg** the formal regularization parameter $\alpha_k := |r_k'(0)|^{-1}$, where r_k denotes the kth degree polynomials r_k, corresponding to the minimization as given in (6). The role of the sequence $|r_k'(0)|$ is discussed in [2] in more detail. Here we only mention that it is increasing, and the corresponding decreasing sequence of α_k for **cg** plays (to some extent) a role comparable to that of the regularization parameter for linear regularization. We refer to [6, 2] for details. We emphasize that for **cg** , since the polynomials r_k depend on the data, so does the sequence (α_k); it is therefore itself random. Moreover we cannot control the rate of decay of the α_k as k increases: this makes the analysis more involved.

In order to give a single point of view over these two cases we introduce the following definition.

Definition 6. Let $(\phi_k)_{k\geq 0}$ be a sequence of reconstruction functions. Let $(x_k^\delta := \phi_k(z^\delta))$ the associated reconstruction sequence, (α_k) the associated non-increasing sequence of regularization parameters (defined by the regularization method used and possibly depending on the observed data z^δ), and τ a fixed constant. We say that the discrepancy principle stopping rule $k_{DP}(\tau, ., .)$ is *ψ-optimal and regular* for this regularization scheme if there exist numbers $C, c_1, c_2, \bar{\delta}$ such that, for all $x^\dagger \in A_\psi$, all $0 < \delta \leq \bar{\delta}$, and any function ϱ such that Assumption 3 is satisfied, the following inequalities hold for $k_{DP} := k_{DP}(\tau, \varrho, \delta)$:

$$\|x^\dagger - x_{k_{DP}}^\delta\| \leq C\psi(\Theta_{\varrho\psi}^{-1}(\delta)) \qquad \text{(optimality)}, \tag{18}$$

and, if $k_{DP} \geq 1$ (no immediate stop) then

$$\alpha_{k_{DP}} \geq c_1 \Theta_{\varrho\psi}^{-1}(c_2\delta) \qquad \text{(regularity)}. \tag{19}$$

Remark 9. The property in bound (18) is called optimality, because it is known from [2, 11] that the right hand side reflects the optimal order of reconstruction under $x^\dagger \in A_\psi$. The regularity property from (19) recourses to the known fact that stopping under the discrepancy principle yields *lower bounds* for the chosen regularization parameter; a fact which is important to control the noise propagation.

For general linear regularization the following result was proved in [11, Thm. 5].

Fact 1. *Let g_α be a linear regularization from Definition 1 which has qualification $\Theta_{\varrho\psi}$, and $\tau > \gamma_1$. Let $\alpha_k := \alpha_0 p^k$, for some $\alpha_0 \geq 2\gamma_* \|A\|$ and $p \in (0,1)$, and $x_k^\delta := x_{\alpha_k}^\delta$. Then the discrepancy stopping rule $k_{DP}(\tau, ., .)$ is ψ-optimal and regular. The constant C in* (18) *depends on the parameters $(\gamma, \gamma_1, \gamma_*, \tau, \alpha_0, p)$ only. Also we have the explicit expressions $c_1 = c_1(p) = p$ and $c_2 = c_2(\tau, \gamma, \gamma_1) = (\tau - \gamma_1)/\gamma$.*

We emphasize that higher qualification ($\Theta_{\varrho\psi}$) is required, compared to the a priori bound as established in 1 (qualification ψ). This is in agreement with known results for the classical results for bounded deterministic noise ($\varrho \equiv 1$). For conjugate gradient regularization the following is proved in [2, Main result].

Fact 2. *Suppose that ψ is an index function majorized by a power $\mu > 0$, that is, such that $t \mapsto \psi(t)/t^\mu$ is non-increasing for $t > 0$. Let $(x_k^\delta)_{k\geq 1}$ be the output of* **cg** *iterations, and $\tau > 2$. Then the discrepancy stopping rule $k_{DP}(\tau, ., .)$ is ψ-optimal and regular. The constants C, c_1, c_2 in* (18) *depend on the parameters (μ, τ) only.*

Remark 10. We comment on both facts. A frequent issue for the analysis of the discrepancy principle is *immediate stopping*, i.e, when the criterion (DP) is fulfilled for x_0^δ. For iterative regularization as **cg** , which starts at zero this amounts to the analysis of *small data*, $\|\varrho(A)z^\delta\| \leq \tau\delta$, and it was shown in [2, Lem. 4.6] that order optimal reconstruction is achieved by this. For linear regularization the situation is different, and the authors in [11] treat only the case that there is *no immediate stop* at α_0. Immediate stop may occur for two reasons. First, criterion (DP) is fulfilled *no matter what* α_0 *is.* Then, if α_0 may tend to infinity, and because $\lim_{\alpha\to\infty} x_\alpha^\delta = 0$ (see Definition 1(3)), this is again covered by the small data case, provided we consider $x_0^\delta = 0$ as solution. Secondly, we are faced with the case that α_0 is chosen too small from the very beginning. Then nothing can be said about optimality of the reconstruction. One safeguard instruction is given by the following observation.

Proposition 2. *Let γ_1, γ_* be as in Definition 1, and assume $\alpha_0 \geq 2\gamma_* \|A\|$. Suppose that the solution satisfies Assumption 2 with function ψ, and let $\bar\delta := \Theta_{\varrho\psi}(\alpha_0)$. If $k_{DP} = 0$ (immediate stop) then*

$$\|x - x_{\alpha_0}^\delta\| \leq (2\gamma_1(1 + 2\tau) + \gamma_*)\, \psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right), \quad 0 < \delta \leq \bar\delta.$$

Proof. We first establish that the operator $r_{\alpha_0}(A)\colon \ker^{\perp}(A) \to X$ has a bounded inverse, specifically that we have

$$\|r_{\alpha_0}(A)^{-1}\colon \ker^{\perp}(A) \to X\| \leq 2. \tag{20}$$

Indeed, we use the properties of the regularization as captured in Definition 1 to argue that for $0 < t \leq \|A\|$ we have

$$|r_{\alpha_0}(t)| \geq 1 - t\,|g_{\alpha_0}(t)| \geq 1 - \gamma_* \frac{t}{\alpha_0} \geq 1 - \frac{1}{2}\frac{t}{\|A\|} \geq \frac{1}{2}.$$

This yields (20). With the above value for α_0, and at immediate stop we find that

$$\|\varrho(A)z^{\delta}\| = \|r_{\alpha_0}(A)^{-1}\varrho(A)r_{\alpha_0}(A)z^{\delta}\| \leq 2\tau\delta,$$

and we are in the position of having small data. Applying [2, Lem. 4.6] we infer that the solution x then obeys $\|x\| \leq 2(1+2\tau)\psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right)$, and we get the following bound for the bias

$$\|x - x_{\alpha_0}\| = \|r_{\alpha_0}(A)x\| \leq \gamma_1\|x\| \leq 2\gamma_1(1+2\tau)\psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right).$$

We turn to bounding the noise term under the model (16) as

$$\|x_{\alpha_0} - x_{\alpha_0}^{\delta}\| = \|g_{\alpha_0}(A)\varepsilon\| = \|g_{\alpha_0}(A)\varrho(A)^{-1}\varrho(A)\varepsilon\| \leq \|g_{\alpha_0}(A)\varrho(A)^{-1}\|\delta.$$

Under Assumption 3 the above norm bound can be estimated by

$$\|g_{\alpha_0}(A)\varrho(A)^{-1}\| \leq \frac{\gamma_*}{\alpha_0\varrho(\|A\|)} \leq \frac{\gamma_*}{\Theta_{\varrho}(\alpha_0)},$$

using that $\gamma_* \geq \frac{1}{2}$ (see Definition 1), and therefore $\alpha_0 \geq \|A\|$. Overall this results in

$$\|x - x_{\alpha_0}^{\delta}\| \leq \|x - x_{\alpha_0}\| + \|x_{\alpha_0} - x_{\alpha_0}^{\delta}\| \leq 2\gamma_1(1+2\tau)\psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right) + \frac{\gamma_*}{\Theta_{\varrho}(\alpha_0)}\delta.$$

The right hand side can further be estimated as

$$\begin{aligned}
2\gamma_1(1+2\tau)&\psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right) + \frac{\gamma_*}{\Theta_{\varrho}(\alpha_0)}\delta \\
&= \psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right)\left(2\gamma_1(1+2\tau) + \frac{\gamma_*}{\Theta_{\varrho}(\alpha_0)}\frac{\delta}{\psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right)}\right) \\
&= \psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right)\left(2\gamma_1(1+2\tau) + \frac{\gamma_*}{\Theta_{\varrho}(\alpha_0)}\Theta_{\varrho}\left(\Theta_{\varrho\psi}^{-1}(\delta)\right)\right) \\
&\leq \psi\left(\Theta_{\varrho\psi}^{-1}(\delta)\right)\left(2\gamma_1(1+2\tau) + \gamma_*\right), \quad 0 < \delta \leq \bar{\delta}
\end{aligned}$$

because the function $t \mapsto \Theta_{\varrho}\left(\Theta_{\varrho\psi}^{-1}(t)\right)$ is non-decreasing, and by the choice of $\bar{\delta}$. The proof is complete. □

4.2. **Modified discrepancy principle for statistical inverse problems.** We now turn back to *statistical* inverse problems and discuss how to use two modifications of the discrepancy principle in order to deal with the random nature of the noise. Suppose that we agreed upon a regularization scheme resulting in a sequence of approximations x_k^σ. A plain use of the discrepancy principle would suggest monitoring $\|z^\sigma - Ax_k^\sigma\| \leq \tau\sigma$, since this norm is almost surely finite by Assumption 1. As already mentioned in the introduction, this plain use of the DP is possible, however this usually will result in sub-optimal rates of approximation (this case will be discussed more precisely at the end of Section 5). Instead, it is favorable to use the discrepancy principle in some weighted norm. For any weight ϱ_λ of the form (17) we have, cf. Definition 2, that

$$\mathbb{E}\left[\|\varrho_\lambda(A)\zeta\|\right] \leq \mathbb{E}\left[\|\varrho_\lambda(A)\zeta\|^2\right]^{1/2} = \left(\operatorname{Tr}\left[(\lambda I + A)^{-1} A\right]\right)^{1/2} = \sqrt{\mathcal{N}(\lambda)}\,. \tag{21}$$

Such a bound holds only on average, but entails that, with high probability, a control of the form $\|\varrho_\lambda(A)\,(z^\sigma - Ax_k^\sigma)\,\| \leq (1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}$ holds. Nevertheless, the parameter choice must still take into account those realizations for which such a bound fails to hold. We therefore impose an additional *emergency stop.* From the definition, whenever the discrepancy principle is optimal and regular, there is a (scheme dependent) lower bound for the parameter $\alpha_{k_{DP}}$, provided the noise norm bound of Assumption 3 holds. For realizations of the noise where this norm bound fails to hold (which is an event of small probability), we impose some default lower bound on $\alpha_{k_{MDP}}$, which is chosen such that it does not interfere with the original parameter $\alpha_{k_{DP}}$ on the "good" realizations of the noise. This lower bound on $\alpha_{k_{MDP}}$ will ensure that the contribution of the "bad cases" to the *overall averaged error* does not become too large.

Definition 7 (Modified discrepancy principle (MDP))**.** Given positive constants $\tau, \eta, \lambda, \kappa$, the modified discrepancy parameter choice $k_{MDP}(\tau, \eta, \lambda, \sigma, \kappa)$ is the smallest $k \geq 0$ for which either of the following conditions is satisfied:

$$\|\varrho_\lambda(A)(z^\sigma - Ax_k^\sigma)\| \leq \tau(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}, \qquad \text{(regular stop);} \tag{22}$$

$$\text{or} \quad \Theta_{\varrho_\lambda}(\alpha_{k+1}) < \eta(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}, \qquad \text{(emergency stop).} \tag{23}$$

(Note that the parametrization of k_{MDP} is deliberately redundant in order to ease the reading further on.)

The following lemma confirms the intuition that if the usual DP is optimal and regular for the considered reconstruction sequence, then (23) is indeed an emergency stop, that is, the modified DP coincides with the usual DP (given by condition (22) alone) whenever the noise realization has a controlled norm and the noise amplitude σ is small enough.

Lemma 4.1. *Let (x_k^δ) be a reconstruction sequence. Fix $0 < \lambda \leq \|A\|$ and $\tau > 1, \eta, \sigma, \kappa$ positive constants. Let $k_{MDP} := k_{MDP}(\tau, \eta, \lambda, \sigma, \kappa)$ be obtained from the* *modified discrepancy principle. Then*

$$\frac{\eta(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}}{\Theta_{\varrho_\lambda}(\alpha_{k_{MDP}})} \leq 1. \tag{24}$$

Assume that the (usual) discrepancy principle stopping rule $k_{DP}(\tau,.,.)$ is ψ-optimal and regular for the considered regularization method, and that $\eta \leq c_1 c_2$ where c_1, c_2 are given by (19)*. Let ζ be a realization of the noise for which Assumption 3 holds with parameters $\varepsilon := \sigma\zeta$, $\varrho := \varrho_\lambda$, and $\delta := (1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}$. Assume furthermore that the following inequality is satisfied:*

$$\psi\left(\Theta^{-1}_{\varrho_\lambda\psi}(c_2(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)})\right) \leq 1. \tag{25}$$

Then $k_{MDP}(\tau,\eta,\lambda,\sigma,\kappa) = k_{DP}(\tau,\varrho_\lambda,\delta)$.

Remark 11. Because the function $\psi \circ \Theta^{-1}_{\varrho_\lambda\psi}$ is is continuously decreasing to zero as $\sigma \searrow 0$, for each index function ψ there exists $\bar\sigma(\psi,\lambda,\kappa,c_2) > 0$ such that for all $\sigma \in [0,\bar\sigma]$, condition (25) is satisfied. Hence, when the parameters (λ,κ) of the procedure are fixed independently of σ, condition (25) is satisfied for small enough noise level. Below (in Sections 5.2 and 5.3), we will also consider cases where λ or κ depend on the noise level σ; some extra care will then be required to check that condition (25) is also satisfied for small enough σ in those situations.

5. Error bounds for the modified discrepancy principle

5.1. **Main error bound.**

We now study the modified discrepancy principle (MDP) for some fixed $\lambda > 0$. The rate which we can actually establish will be be a maximum of two rates with equality, if λ is chosen 'optimally'. More precisely, we introduce the following rate function:

$$F_\psi(t) := \max\left\{\psi\left(\Theta^{-1}_{\psi}(\sqrt{\lambda}t)\right), \psi\left(\Theta^{-1}_{\varrho_0\psi}(t)\right)\right\}, \tag{26}$$

Observe that F_ψ depends on λ, although this is not explicitly indicated in the notation. For the understanding of this rate function, we present the following result, with proof postponed to § 7.1.

Lemma 5.1. *Let $\mathcal{N}$ be as in (9) with corresponding weight function $\varrho_\mathcal{N}$ from (11), and let $\lambda > 0$ be arbitrary.*

(1) *We have*

$$\psi\left(\Theta^{-1}_{\varrho_\lambda\psi}\left(\sigma\sqrt{\mathcal{N}(\lambda)}\right)\right) \leq F_\psi\left(\sigma\sqrt{2\mathcal{N}(\lambda)}\right).$$

(2) *We have that*

$$\psi\left(\Theta^{-1}_{\varrho_0\psi}(\sigma\sqrt{2\mathcal{N}(\lambda)})\right) \leq \psi\left(\Theta^{-1}_{\psi}(\sigma\sqrt{2\lambda\mathcal{N}(\lambda)})\right),$$

exactly if $\Theta_{\varrho_\mathcal{N}\psi}(\lambda) \geq \sigma$.

(3) *If we define*

$$\lambda_* := \Theta^{-1}_{\varrho_{\mathcal{N}}\psi}(\sigma)\,, \tag{27}$$

we have

$$F_\psi(\sigma\sqrt{2\mathcal{N}(\lambda_*)}) = \psi\left(\Theta^{-1}_\psi(\sigma\sqrt{2\lambda_*\mathcal{N}(\lambda_*)})\right) = \psi\left(\Theta^{-1}_{\varrho_{\mathcal{N}}\psi}(\sigma)\right).$$

(4) *The following identity holds:*

$$\Theta_{\varrho_\lambda\psi}(\lambda) = \Theta_{\varrho_{\mathcal{N}}\psi}(\lambda)\sqrt{\mathcal{N}(\lambda)}, \quad \lambda > 0\,. \tag{28}$$

To give some interpretation of the above rates, let us compare them to known convergence rates in the deterministic case. Remember from Section 4.1 that, under the deterministic noise model (15)–(16), and when the deterministic noise ε is controlled as in Assumption 3, that is, $\|\varrho(A)\varepsilon\| \leq \delta$, then the worst-case convergence rate of an optimal procedure is of order $\psi(\Theta^{-1}_{\varrho\psi}(\delta))$.

In the random noise case, we will obtain a control of this form, wherein we take $\varrho := \varrho_\lambda$ and $\delta := \sigma\sqrt{\mathcal{N}(\lambda)}$ (where $\lambda > 0$ is a parameter of the procedure which can be freely chosen). Point (1) of the above Lemma relates the resulting rate to more familiar ones through the function F_ψ, which involves two different rates.

The first rate, $g^{(1)}(\delta) := \psi\left(\Theta^{-1}_\psi(\sqrt{2\lambda}\delta)\right)$, is the optimal rate under bounded deterministic noise ε with noise level given by $\|\varepsilon\| \asymp \sqrt{\lambda}\delta$ (large noise). The second rate, $g^{(2)}(\delta) := \psi\left(\Theta^{-1}_{\varrho_0\psi}(\delta)\right)$, is the optimal rate under bounded deterministic noise satisfying $\|\eta\| = \|A^{-\frac{1}{2}}\varepsilon\| \asymp \delta$ (small noise). For large values of λ, the term $g^{(1)}(\sqrt{2\mathcal{N}(\lambda)}\sigma)$ will be the dominant one, while for small values of λ, the second rate $g^{(2)}(\sqrt{2\mathcal{N}(\lambda)}\sigma)$ will be the largest.

A balance between these two rates is obtained exactly if λ_* is chosen according to (27), and yields the rate expressed in point (3) of the Lemma, thus recovering the same rate as stated in Theorem 1 for a priori rules for linear regularization, see discussion and examples there about optimality properties.

We now turn to the statement of the fundamental error estimate, for arbitrary fixed $\lambda > 0$, as follows. We explicitly highlight the specific assumptions for either linear regularization scheme or **cg** .

linear regularization: The family g_α is a linear regularization from Definition 1, and $x^\sigma_k := x^\sigma_{\alpha_k}$ where $\alpha_k := \alpha_0 p^k$ for some $\alpha_0 > 2\gamma_*\|A\|$ and $p \in (0,1)$. It is assumed to have qualification $\Theta_{\varrho\psi}$. The constants c_1, c_2 are given from from Fact 1. It is assumed $\tau > \gamma_1$, $\eta \leq c_1c_2$, and $\alpha_0 \geq 2\gamma_*\|A\|$.

cg iteration: It is assumed $\tau > 2$ and $\eta \leq c_1c_2$, where the constants are from from Fact 2. The function ψ is majorized by a power $\mu > 0$, that is, the function $\lambda \mapsto \psi(\lambda)/\lambda^\mu$ is non-increasing on $(0,\infty)$.

Theorem 2. *Consider the statistical inverse problem model from* (2) *with noise level* $\sigma > 0$. *Suppose Assumptions 1 and 2 hold. Let* x^σ_k *denote the reconstruction* *sequence obtained either through linear regularization of* **cg** *iteration, as described just above. Fix parameters $\lambda > 0$ and $\kappa > 0$. For τ, η satisfying the conditions described above, let $k_{MDP} := k_{MDP}(\tau, \eta, \lambda, \kappa, \sigma)$ be obtained from the modified discrepancy principle. Assume the following is satisfied:*

$$\psi\left(\Theta^{-1}_{\varrho_\lambda \psi}(c_2(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)})\right) \leq 1. \tag{29}$$

Let $q > 1$ be a positive number; then the following bound holds for any $x^\dagger \in A_\psi$:

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k_{MDP}}\|^q\right]^{\frac{1}{q}} \leq C \max\left\{(1+\kappa), \psi(\|A\|)\right\} \times \left(F_\psi(\sigma\sqrt{2\mathcal{N}(\lambda)}) + \exp\left(-\frac{\kappa^2 \mathcal{N}(\lambda)}{4q}\right)\right). \tag{30}$$

The factor C in (30) *depends on the parameters $(\gamma, \gamma_1, \gamma_*, p, \tau, \eta, q)$, for linear regularization. For* **cg** *the factor C in* (30) *depends on the parameters (μ, τ, η, q), only.*

By virtue of Remark 11 condition (29) will be satisfied for $0 < \sigma \leq \bar{\sigma}(\psi, \lambda, \kappa, c_2)$. Thus in the asymptotic setting, when (λ, κ) are fixed independently of σ, and $\sigma \to 0$, the bound given in (30) will hold. Next, we address the cases where λ or κ may depend on σ.

5.2. Optimizing the parameter λ.

We can optimize the rate given in Theorem 2 by a an appropriate choice of λ (depending on σ and on the underlying smoothness ψ), see Lemma 5.1.

Corollary 1. *Consider the same conditions as in Theorem 2. There exists $\bar{\sigma}(\psi, \kappa, c_2) > 0$ such that for all $\sigma \in [0, \bar{\sigma}]$, $\lambda := \lambda_*$ satisfies* (29)*. Assume σ satisfies this condition, and consider x^σ_k obtained from the linear regularization or* **cg** *iterations; let $k_{MDP} := k_{MDP}(\tau, \eta, \lambda_*, \sigma, \kappa)$ be picked by (MDP). For any $q \geq 1$ the following inequality holds:*

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k^*_{MDP}}\|^q\right]^{\frac{1}{q}} \leq C \max\left\{(1+\kappa), \psi(\|A\|)\right\} \times \left(\psi\left(\Theta^{-1}_{\varrho_\mathcal{N}\psi}(\sigma)\right) + \exp\left(-\frac{\kappa^2 \mathcal{N}(\Theta^{-1}_{\varrho_\mathcal{N}\psi}(\sigma))}{4q}\right)\right), \tag{31}$$

where the factor C depends on the same parameters as those stated in Theorem 2. If the functions ψ and $\mathcal{N}$ furthermore satisfy

$$\psi(t) \geq e^{-\widetilde{C}\mathcal{N}(t)}, \quad \textit{as } t \to 0, \tag{32}$$

for some $\widetilde{C} > \kappa^2/(4q)$, then

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k_{DP}}\|^q\right]^{\frac{1}{q}} \leq C'\left(\psi\left(\Theta^{-1}_{\varrho_\mathcal{N}\psi}(\sigma)\right)\right),$$

where the factor C' depends on the same parameters as earlier, and additionally on $(\kappa, \widetilde{C}, \|A\|)$.

Proof. First we use equation (28), leading to

$$\sigma\sqrt{\mathcal{N}(\lambda_*)} = \Theta_{\varrho_\lambda\psi}(\lambda_*)\,;$$

since $\psi \circ \Theta^{-1}_{\varrho_\lambda\psi}$ is majorized by the power 1 (see Lemma 7.2), we can bound the LHS of (29) as follows:

$$\psi\left(\Theta^{-1}_{\varrho_\lambda\psi}(c_2(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda_*)})\right) \leq c_2(1+\kappa)\psi(\lambda_*)\,;$$

finally, since $\lambda_* \searrow 0$ as $\sigma \searrow 0$, we deduce that there exists $\bar{\sigma}(\psi,\kappa,c_2) > 0$ such that for all $\sigma \in [0,\bar{\sigma}]$, the RHS of the last display is bounded by 1. In this case, we can apply Theorem 2; plugging in (30) the definition of λ_* as well as identity (28), we obtain (31). The last statement of the corollary is straightforward: condition (32) ensures that the first term in the RHS of (31) is dominant as $\sigma \searrow 0$. □

Examples. Corollary 1 applies to the examples considered in Section 3.2. If the decay rate of the singular numbers and the smoothness are both power functions, condition (32) holds and we recover the rates exhibited in Example 1.

In the severely ill-posed case of Example 2, condition 32 holds for any smoothness $\psi(t) \asymp \log^{-s} 1/t$. Even for monomial smoothness $\psi(t) \asymp t^s$, corresponding in this setting to a "supersmooth" signal, condition 32 holds as long as $s < \kappa^2/(4q)$.

5.3. **The modified discrepancy principle as a posteriori parameter choice.**

Another consequence of Theorem 2 is the convergence of linear regularization or **cg** for any fixed parameter $\lambda > 0$, however at a sub-optimal rate.

Corollary 2. *Consider the same conditions as in Theorem 2, but fix*

$$\kappa_* := 2\sqrt{\frac{q}{\mathcal{N}(\lambda)}}\sqrt{\log 1/(1 \wedge \sigma\sqrt{2\mathcal{N}(\lambda)})}.$$

Then there exists $\bar{\sigma}(\psi, \|A\|, \lambda, c_2, q) \in (0,1)$ such that (29) (with $\kappa = \kappa^*$) is satisfied for all $\sigma \in [0,\bar{\sigma}]$. Assume σ satisfies this condition, consider x^σ_k obtained from the linear regularization or **cg** *iterations; let $k_{MDP} := k_{MDP}(\tau,\eta,\lambda,\sigma,\kappa_*)$ be picked by (MDP). Then for any $q \geq 1$ the following inequality holds:*

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k_{MDP}}\|^q\right]^{\frac{1}{q}} \leq CF_\psi(\sigma)\sqrt{\log 1/\sigma},$$

where the factor C depends on the same parameters as those stated in Theorem 2 and additionally on λ, ψ and $\|A\|$.

Proof. By plugging in $\kappa := \kappa_*$, we see that the LHS of (29) is a continuous function of σ that vanishes when $\sigma \searrow 0$ (all other parameters (ψ,λ,c_2,q) being held fixed). This justifies the existence of $\bar{\sigma}$ as stated. Moreover, by taking $\bar{\sigma}$ small enough we can also ensure that $\bar{\sigma} \leq \sqrt{2\mathcal{N}(\lambda)}$; and that for $0 < \sigma \leq \bar{\sigma}$ it holds $\kappa^* \geq 1$ and $\psi(\|A\|) \leq 1 + \kappa^*$.

For $\sigma \in [0, \bar{\sigma}]$, we have

$$\exp\left(-\frac{\kappa_*^2 \mathcal{N}(\lambda)}{4q}\right) \leq \sigma\sqrt{2\mathcal{N}(\lambda)} \leq \frac{\bar{\sigma}}{F_\psi(\bar{\sigma})}\sqrt{2\mathcal{N}(\lambda)}F_\psi(\sigma)\,,$$

since the function F_ψ is majorized by the power 1, as a maximum of two functions which are majorized by the power 1 (see Section 7.2.2 for more details). We therefore deduce from Theorem 2 and the requirement $(1+\kappa^*) \geq \psi(\|A\|)$ that

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k_{MDP}}\|^q\right]^{\frac{1}{q}} \leq 2C(1+\kappa^*)\sqrt{2\mathcal{N}(\lambda)}F_\psi(\sigma)\,,$$

where the factor C depends on the parameters stated in Theorem 2 and additionally on λ, ψ and $\|A\|$. Taking into account the requirements put above for $\bar{\sigma}$, we have that

$$(1+\kappa^*)\sqrt{2\mathcal{N}(\lambda)} \leq 2\kappa^*\sqrt{2\mathcal{N}(\lambda)} \leq 8\sqrt{q}\sqrt{\log 1/\sigma}\,,$$

from which the proof can easily be completed. □

Remark 12 (Comparison with unweighted DP). We briefly compare the rates derived in Corollaries 1 and 2 to rates that can be obtained from similar argumentation when using the unweighted (modified) DP, that is, using a weighting function $\varrho \equiv 1$ instead of ϱ_λ. In this case, the reasoning in expectation (21) indicates we should replace $\mathcal{N}(\lambda)$ by $\mathrm{Tr}\,[A]$. Modifying accordingly the stopping criterion by performing the above replacements ($\varrho_\lambda \to 1$ and $\mathcal{N}(\lambda) \to \mathrm{Tr}\,[A]$) in (22) and (23), the MDP then dictates to stop when either $\|z^\sigma - Ax^\sigma_k\| \leq \tau\sigma\sqrt{\mathrm{Tr}\,[A]}$ or $\alpha_{k+1} < \eta(1+\kappa)\sigma\sqrt{\mathrm{Tr}\,[A]}$. (This corresponds, informally, to taking $\lambda \to \infty$ in the original Definition 7.) In this situation, the arguments of the proof of Theorem 2 go through with the above replacements. Namely, in the deterministic setting Facts 1 and 2 hold for the choice $\varrho \equiv 1$; also, the concentration bound in Lemma 7.5 holds with those changes. Finally, analogously to Corollary 2, we get under the condition that $\kappa \asymp \sqrt{\log 1/\sigma}$ the bound

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k_{MDP}}\|^q\right]^{\frac{1}{q}} \leq C\psi(\Theta_\psi^{-1}(\sigma\sqrt{\mathrm{Tr}\,[A]}))\sqrt{\log 1/\sigma},$$

where C depends on many parameters but not on σ. Once again, this corresponds informally to taking $\lambda \to \infty$ in Corollary 2. The above rate, which is the rate within the deterministic noise model $\|\varepsilon\| \leq \sqrt{\mathrm{Tr}\,[A]}$ with ε as in (16), is suboptimal in the random noise case; this justifies the interest of introducing the additional weighting through ϱ_λ.

Example 3. If the operator A is of finite rank, then $\mathcal{N}(t)$ is bounded near the origin and Corollary 1 will not give an explicit rate. However, we can apply Corollary 2 with $\lambda = 0$. Then if $0 < \sigma \leq \bar{\sigma}$, by virtue of Lemma 5.1 we obtain that

$$\mathbb{E}\left[\|x^\dagger - x^\sigma_{k_{MDP}}\|^q\right]^{\frac{1}{q}} \leq C\psi(\Theta^{-1}_{\varrho_0\psi}(\sigma))\sqrt{\log 1/\sigma}, \quad \text{as } \sigma \to 0.$$

This covers the deterministic case, as established in [2, Cor. 1], and it is known to be (up to the logarithmic factor) best possible, since here $\Theta_{\varrho_0\psi}(t) = \sqrt{t}\psi(t)$.

To conclude this section, we discuss without entering into formal details a further application of Theorem 2, namely if we keep κ constant and let λ go slowly to zero as a function of σ. Assume that the singular numbers $s_j(T) \geq \underline{c} j^{-r}$, $j = 1, 2, \dots$ for some $r > 1/2$ (the lower bound because of Assumption 1), and that the smoothness function ψ is an index function majorized by a power $\mu > 0$, that is $t \mapsto \psi(t)/t^{\mu}$ is non-increasing for $t > 0$ (where μ does not have to be known). Fix $\kappa > 0$ and put $\lambda(\sigma) := (\log 1/\sigma)^{-t}$, whith $t > 1/(2r)$, e.g. $t = 1$. It is easily checked that condition (29) is satisfied for σ small enough, and we can apply Theorem 2 in this setting. It can be seen readily that the remainder exponential term in the main bound (30) decreases faster to zero than any fixed power of σ. Since the other terms in the bound converge to zero at most at some monomial rate, the remainder term is then negligible. Observe that $\lambda_*(\sigma)$ as given by (27) decreases to zero at least as some power of σ. For σ small enough, it will therefore hold that $\lambda(\sigma) \geq \lambda_*(\sigma)$ and, by point (2) of Lemma 5.1, we have

$$F_{\psi}(\sigma\sqrt{2\mathcal{N}(\lambda)}) = \psi(\Theta_{\psi}^{-1}(\sigma\sqrt{2\lambda\mathcal{N}(\lambda)})) = \mathcal{O}(\psi(\Theta_{\psi}^{-1}(\sigma))).$$

Hence without a priori knowledge of the smoothness function, this ensures that the convergence rate, though not optimal in the white noise setting, is at least asymptotically as good as the (optimal) deterministic rate. If the smoothness function ψ is a power function, the rate function F_{ψ} is upper bounded by some monomial and the above $\mathcal{O}(.)$ is actually an $o(.)$; that is, we are ensured to get a convergence rate strictly better than the deterministic rate.

6. Extension to colored noise

The analysis extends to colored noise, i.e., when the covariance structure is given by some non-negative self-adjoint operator $K\colon Y \to Y$, in which case $\mathbb{E}\left[\xi(w)\xi(w')\right] = \langle Kw, w'\rangle$, $w, w' \in Y$. We may and do restrict this to operators K with norm bound $\|K\| \leq 1$, since the level is measured by the additional parameter σ. Next, we highlight the modification of the effective dimension for correlated noise.

In principle, one can reduce the case of correlated noise, say with covariance operator $K\colon Y \to Y$, to the uncorrelated one by *pre-whitening*. Here we mean that we 'formally' apply the operator $K^{-1/2}$ to the equation (1), which leads to an equation with operator $S := K^{-1/2}T$ under white noise. Then the corresponding effective dimension should be

$$\operatorname{Tr}\left[(S^*S + \lambda I)^{-1}S^*S\right], \quad \lambda > 0.$$

We provide a more intuitive representation by using the *principle of related operators*: Suppose that $S_0\colon Y \to Y$ admits a factorization $S_0 = S_1 S_2$ with $S_1\colon Z \to Y$, $S_2\colon Y \to Z$. If $S_0 = S_1 S_2$ has finite trace then so has $S_2 S_1$ and $\operatorname{Tr}[S_1 S_2] = \operatorname{Tr}[S_2 S_1]$.

Lemma 6.1. *Suppose that the operator $S = K^{-1/2}T$ obeys Assumption 1. Then we have*

$$\operatorname{Tr}\left[(S^*S+\lambda I)^{-1}S^*S\right] = \operatorname{Tr}\left[\left(TT^*+\lambda K\right)^{-1}TT^*\right], \quad \lambda > 0.$$

Proof. First, we let $B := TT^* \colon Y \to Y$. Notice that then $SS^* = K^{-1/2}BK^{-1/2}$, and this operator has a finite trace by Assumption 1; consequently, the left hand side above is finite. We use the principle of related operators to infer that

$$\begin{aligned}\operatorname{Tr}\left[\left(B+\lambda K\right)^{-1}B\right] &= \operatorname{Tr}\left[\left(K^{1/2}\left(K^{-1/2}BK^{-1/2}+\lambda I\right)K^{1/2}\right)^{-1}B\right]\\ &= \operatorname{Tr}\left[K^{-1/2}\left(K^{-1/2}BK^{-1/2}+\lambda I\right)^{-1}K^{-1/2}B\right]\\ &= \operatorname{Tr}\left[(SS^*+\lambda I)^{-1}SS^*\right] = \operatorname{Tr}\left[(S^*S+\lambda I)^{-1}S^*S\right],\end{aligned}$$

the latter from functional calculus. This completes the proof. □

Therefore the effective dimension under colored noise is introduced as follows.

Definition 8. Let $K \colon Y \to Y$ be the covariance operator for the noise ξ in equation (1). If the mapping $K^{-1/2}T$ has a finite trace then the effective dimension is given as

$$\mathcal{N}_K(\lambda) := \operatorname{Tr}\left[\left(TT^*+\lambda K\right)^{-1}TT^*\right], \quad \lambda > 0.$$

Remark 13. It is worth-wile to notice that the assumption on $K^{-1/2}T$ requires a "minimum distance" between the decay rate of the singular numbers of A and the ones of K. This will be more intuitive in the example, given below.

Furthermore, for white noise $K = I$ we have that $\mathcal{N}_I(\lambda) = \mathcal{N}(\lambda),\ \lambda > 0$. Indeed, by using that $\left(TT^*+\lambda I\right)^{-1}T = T\left(T^*T+\lambda I\right)^{-1}, \quad \lambda > 0$, we deduce that

$$\begin{aligned}\mathcal{N}_I(\lambda) &= \operatorname{Tr}\left[\left(TT^*+\lambda I\right)^{-1}TT^*\right] = \operatorname{Tr}\left[T^*\left(TT^*+\lambda I\right)^{-1}T\right]\\ &= \operatorname{Tr}\left[T^*T\left(T^*T+\lambda I\right)^{-1}\right] = \mathcal{N}(\lambda), \quad \lambda > 0.\end{aligned}$$

Therefore, the above definition for the effective dimension under colored noise is consistent with the one given in Definition 2 in the white noise case.

Example 4. We shall indicate that in 'typical' cases one can obtain optimal rates under colored noise by using the corresponding effective dimension $\mathcal{N}_K$. Indeed, assume that $s_j(A) \asymp j^{-2r}$, and $s_j(K) \asymp j^{-2\mu}$ for some $r, \mu > 0$, and that both operators A and K commute. The assumption in Lemma 6.1 requires that $r > \mu + 1/2$, which means that the distance is at least $1/2$. An easy calculation shows that then the weight function $\varrho_{\mathcal{N}_K}$ obeys $\varrho_{\mathcal{N}_K}(t) \asymp t^{(1/2+(\mu-r))/(2r-2\mu)}$. Hence, Proposition 1 holds with function $\mathcal{N}$ replaced by $\mathcal{N}_K$. However, we have to recalculate the smoothness, previously given in terms of the operator A, say $\psi(t) = t^{\nu/(2r)}$ for some $\nu > 0$, to the new smoothness, say $\tilde{\psi}$ with respect to the operator S^*S. This gives $\tilde{\psi}(t) = t^{\nu/(2r-2\mu)}$. Putting things together we infer an a priori rate, according to Theorem 1, as

$$\mathbb{E}\left[\|x - x_\alpha^\sigma\|^2\right]^{1/2} \leq C\tilde{\psi}(\Theta^{-1}_{\varrho_{\mathcal{N}_K}\tilde{\psi}}(\delta)) \asymp \delta^{\frac{\nu}{\nu+r+1/2-\mu}}, \text{ as } \delta \to 0. \tag{33}$$

This is known to be optimal for $\mu < 1/2$, as shows a comparison with results given in [10, Thm. 4]. The decay rate $s_j(K) \asymp j^{-2\mu}$ gives that Assumption 1 in [10] holds with $p = 1/(1-2\mu)$ for $\mu < 1/2$. Then the optimal rate behaves like $\delta^{\nu/(\nu+r+1/(2p))}$, which coincides with the one from (33).

7. Proofs

7.1. Proof of lemmata 4.1 and 5.1.

Lemma 4.1 is a consequence of the following simple result:

Lemma 7.1. *Let $c_1 \in (0,1]$, $\lambda \in (0, \|A\|]$ and t satisfying*

$$0 < t \leq \Theta_{\varrho_\lambda\psi}(\|A\|) \text{ and } \psi\left(\Theta^{-1}_{\varrho_\lambda\psi}(t)\right) \leq 1. \tag{34}$$

If the parameter $\alpha > 0$ obeys

$$\alpha \geq c_1\Theta^{-1}_{\varrho_\lambda\psi}(t), \tag{35}$$

then

$$\Theta_{\varrho_\lambda}(\alpha) \geq c_1 t.$$

Proof. Under the condition (34) on t, we have that $\Theta^{-1}_{\varrho_\lambda\psi}(t)$ is well-defined and lies in $(0, \|A\|]$. Furthermore, we have

$$t = \Theta_{\varrho_\lambda\psi}\left(\Theta^{-1}_{\varrho_\lambda\psi}(t)\right) = \Theta_{\varrho_\lambda}\left(\Theta^{-1}_{\varrho_\lambda\psi}(t)\right)\psi\left(\Theta^{-1}_{\varrho_\lambda\psi}(t)\right) \leq \Theta_{\varrho_\lambda}\left(\Theta^{-1}_{\varrho_\lambda\psi}(t)\right),$$

using (34) in the last inequality. Therefore $\Theta^{-1}_{\varrho_\lambda}(t) \leq \Theta^{-1}_{\varrho_\lambda\psi}(t)$. Finally, we recall that for $0 < c \leq 1$ we have $\Theta^{-1}_{\varrho_\lambda}(ct) \leq c\Theta^{-1}_{\varrho_\lambda}(t)$, which allows us deduce from (35) that

$$\alpha \geq c_1\Theta^{-1}_{\varrho_\lambda\psi}(t) \geq c_1\Theta^{-1}_{\varrho_\lambda}(t) \geq \Theta^{-1}_{\varrho_\lambda}(c_1 t).$$

□

Proof of Lemma 4.1. Condition (23) in the definition of the MDP ensures that inequality (24) is satisfied in all cases at step $k_{MDP} := k_{MDP}(\tau, \eta, \lambda, \sigma, \kappa)$.

For the second part of the claim, we only have to check that under the postulated bound on the noise norm, the usual discrepancy stopping step $k_{DP} := k_{DP}(\tau, \varrho_\lambda, \delta)$, wherein $\delta := (1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}$, violates (23). Since the sequence of regularization parameters (α_k) is non-increasing, this will imply that condition (22) is first satisfied before condition (23) is, namely at iteration k_{DP}, so that $k_{MDP} = k_{DP}$.

Let ζ be a realization of the noise such that Assumption 3 holds with the stated parameters, that is:

$$\|\varrho_\lambda(A)\sigma\zeta\| \leq (1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}.$$

Since $k_{DP}(\tau,.,.)$ is assumed ψ-optimal and regular, we conclude from (19) that

$$\alpha_{k_{DP}} \geq c_1 \Theta^{-1}_{\varrho_\lambda \psi}\left(c_2(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}\right).$$

This together with inequality (25) implies that Lemma 7.1 applies to $t := c_2(1+\kappa)\sigma\sqrt{\mathcal{N}(\lambda)}$ and yields that $\alpha_{k_{DP}}$ violates (23), since $\eta \leq c_1 c_2$. Thus $k_{DP} = k_{MDP}$. □

Proof of Lemma 5.1. To prove the first assertion we start from $(\lambda+t)\min\left\{\lambda^{-1},t^{-1}\right\} \leq 2,\ t > 0$, which is easily checked. Therefore, $\varrho_\lambda(t) \geq 2^{-1/2}\min\left\{\lambda^{-1/2},t^{-1/2}\right\}$, and a fortiori

$$\Theta_{\varrho_\lambda \psi}(t) \geq \frac{1}{\sqrt{2}}\min\left\{\Theta_{\varrho_0 \psi}(t), \lambda^{-1/2}\Theta_\psi(t)\right\}.$$

This yields

$$\psi\left(\Theta^{-1}_{\varrho_\lambda \psi}(u)\right) \leq \max\left\{\psi\left(\Theta^{-1}_{\varrho_0\psi}(\sqrt{2}u)\right), \psi\left(\Theta^{-1}_{\psi}(\sqrt{2\lambda}u)\right)\right\}$$

Letting $u := \sigma\sqrt{\mathcal{N}(\lambda)}$ we obtain the bound as stated. The proof of the second assertion is along the following chain of equivalent reformulations, with u as before.

$$\Theta_{\varrho_{\mathcal{N}}\psi}(\lambda) \geq \sigma \tag{36}$$

$$\Theta_\psi(\lambda) = \lambda\psi(\lambda) \geq \sigma\sqrt{2\lambda\mathcal{N}(\lambda)} = \sqrt{\lambda}u \tag{37}$$

$$\left(\Theta^{-1}_\psi\left(\sqrt{\lambda}u\right)\right)^{1/2} \leq \sqrt{\lambda} \tag{38}$$

$$u \leq \frac{\sqrt{\lambda}u}{\left(\Theta^{-1}_\psi\left(\sqrt{\lambda}u\right)\right)^{1/2}} = \Theta_{\varrho_0\psi}\left(\Theta^{-1}_\psi\left(\sqrt{\lambda}u\right)\right) \tag{39}$$

$$\Theta^{-1}_{\varrho_0\psi}(u) \leq \left(\Theta^{-1}_\psi\left(\sqrt{\lambda}u\right)\right), \tag{40}$$

from which the assertion is an immediate consequence. The last two assertions can easily be checked by straightforward calculations. □

7.2. **Proof of Theorem 2.** As already mentioned earlier, one specific feature of statistical inverse problems is the following: There is no uniform noise bound, and hence our approach will distinguish between the 'good cases', which shall take place in the majority of cases and the 'bad cases' which happen only rarely. For the a priori bound from Theorem 1 the parameter choice did not depend on the realizations z^σ, and such distinction was not necessary there. Therefore we shall need a point-wise error bound, regardless of the size of the noise ζ.

7.2.1. *Point-wise error decomposition.* From Assumption 1 we know that almost surely the norm $\|\varrho_\lambda(A)\zeta\|$ is finite, and the corresponding noise level is $\sigma\|\varrho_\lambda(A)\zeta\|$. For both linear regularization and **cg** error decompositions under bounded deterministic noise are known.

Fact 3. *Assume that x is any element which obeys Assumption 2, and that $x^\sigma_{\alpha_k}$ is any reconstruction, based on data z^σ. We denote $\tilde{\delta} := \sigma\|\varrho_\lambda(A)\zeta\|$.*

1. *If $x^\sigma_{\alpha_k}$ is obtained from linear regularization g_α which has qualification ψ, cf. (13) then*
$$\|x - x^\sigma_{\alpha_k}\| \leq \gamma\psi(\alpha_k) + \frac{\max\{\gamma_*, 1+\gamma_1\}}{\Theta_{\varrho_\lambda}(\alpha_k)}\tilde{\delta}.$$
2. *If $x^\sigma_{\alpha_k}$ is obtained from* **cg** *then*
$$\|x - x^\sigma_{\alpha_k}\| \leq C(\mu)\psi\left(\Theta^{-1}_{\varrho_\lambda,\psi}\left(\tilde{\delta} + (2\mu+1)^{\mu+1}\,\Theta_{\varrho_\lambda\psi}(\alpha_k)\right)\right) + \frac{3}{\Theta_{\varrho_\lambda}(\alpha_k)}\tilde{\delta}.$$

Remark 14. We comment on the above assertions. Item (2) is obtained from [2, Prop. 4.3], after using [2, lem. 4.1].. The decomposition from Item (1) is based on
$$\|x - x^\sigma_{\alpha_k}\| \leq \|x - x_{\alpha_k}\| + \|x_{\alpha_k} - x^\sigma_{\alpha_k}\| \leq \gamma\psi(\alpha_k) + \|x_{\alpha_k} - x^\sigma_{\alpha_k}\|,$$
where x_{α_k} is as in (5). The latter norm difference was bounded in [11, Prop. 2] by $\max\{\gamma_*, \gamma_0\}\,\frac{\tilde{\delta}}{\Theta_{\varrho_\lambda}(\alpha_k)}$, where we mention that $\gamma_0 := \sup_{0<t\leq\|A\|} t\,|g_\alpha(t)| \leq 1+\gamma_1$, and the function ϱ_λ in which the noise is controlled corresponds to $\psi(t)/\sqrt{t}$, ibid.

7.2.2. *Auxiliary results on functions majorized by the power 1.* We recall the following notion as used in [2, Def. 1.4]. If f, g are two positive functions defined on $(0,\infty)$, we say that the function f is majorized by the function g, denoted $f \prec g$, if the function g/f is non-decreasing. If $g(x) = x^\mu$, we say that f is majorized by the power μ. (This definition was used in Fact 2.)

We gather here a couple of somewhat general lemmata concerning functions bounded by the power 1.

Lemma 7.2. *Let $\varrho : (0, \|A\|] \to \mathbb{R}_+$ be a function such that $v^{-1} \prec \varrho(v) \prec 1$, and ψ an increasing function. Then $\psi \circ \Theta^{-1}_{\varrho\psi}$ is majorized by the power 1.*

Proof. Minor variation on [2, Section 3.1]. Taking $v = \Theta_{\varrho\psi}(s)$, we have
$$\frac{v}{\psi\left(\Theta^{-1}_{\varrho\psi}(v)\right)} = \frac{\Theta_{\varrho\psi}(s)}{\psi(s)} = s\varrho(s),$$
which is a non-decreasing function of v. □

Functions which are majorized by the power 1 enjoy many properties of *moduli of continuity*, in particular they are sub-additive and can be upper bounded by a concave increasing function up to a factor of 2, see [8, Chapt. 6].

Lemma 7.3. *A function F majorized by the power 1 is sub-additive.*

Proof. Assume $x, y > 0$. Then
$$F(x+y) = x\frac{F(x+y)}{x+y} + y\frac{F(x+y)}{x+y} \leq x\frac{F(x)}{x} + y\frac{F(y)}{y} = F(x) + F(y).$$

If x or y or both are equal to zero then the assertion is trivial. □

Lemma 7.4 (Jensen replacement)**.** *Let F be a non-negative, non-decreasing function majorized by the power 1, and X be a non-negative variable. Then for any $p > 0$, denoting $\|Z\|_p = \mathbb{E}\left[Z^p\right]^{\frac{1}{p}}$, one has*

$$\|F(X)\|_p \leq 2^{\frac{1}{p}} F(\|X\|_p) \,.$$

Proof. If $\|X\|_p = 0$ we have $X = 0$ a.s. and the result is trivial. Otherwise $\|X\|_p > 0$ and

$$\begin{aligned} F^p(X) &\leq F^p(\|X\|_p)\mathbf{1}\{X \leq \|X\|_p\} + F^p(X)\mathbf{1}\{X > \|X\|_p\} \\ &\leq F^p(\|X\|_p) + \frac{X^p}{\|X\|_p^p} F^p(\|X\|_p) \,, \end{aligned}$$

taking expectations we obtain the desired inequality. □

7.2.3. *Proof of the main error bound.* In order to prove our main theorem we shall establish some auxiliary technical results. The main decomposition between "good" and "bad" realizations of the noise is given next:

Proposition 3. *Let $Z \subset X$ be any Borel set and let x_k^σ be a of reconstruction, based on data z^σ, i.e., $x_k^\sigma = x_k^\sigma(z^\sigma)$ with noise $\zeta \in Z$. For every $x \in X$ we have that*

$$\begin{aligned} &\left(\mathbb{E}\left[\|x - x_k^\sigma(z^\sigma)\|^q\right]\right)^{1/q} \\ &\qquad \leq \sup_{\zeta \in Z} \|x - x_k^\sigma(z^\sigma)\| + \left(\mathbb{E}\left[\|x - x_k^\sigma(z^\sigma)\|^{2q}\right]\right)^{1/2q} \left(\mathbb{P}_{z^\sigma}\left[Z^c\right]\right)^{1/2q} . \end{aligned} \tag{41}$$

Proof. The proof is straightforward. We insert the characteristic function $\mathbf{1}_Z$ of the set Z and bound

$$\begin{aligned} \mathbb{E}\left[\|x^\dagger - x_k^\sigma\|^q\right]^{\frac{1}{q}} &\leq \mathbb{E}\left[\|x^\dagger - x_k^\sigma\|^q \mathbf{1}_Z\right]^{\frac{1}{q}} + \mathbb{E}\left[\|x^\dagger - x_k^\sigma\|^q \mathbf{1}_{Z^c}\right]^{\frac{1}{q}} \\ &\leq \mathbb{E}\left[\|x^\dagger - x_k^\sigma\|^q \mathbf{1}_Z\right]^{\frac{1}{q}} + \mathbb{E}\left[\|x^\dagger - x_k^\sigma\|^{2q}\right]^{\frac{1}{2q}} \left(\mathbb{P}_{z^\sigma}\left[Z^c\right]\right)^{\frac{1}{2q}} . \end{aligned}$$

Finally, the first summand above can be bounded by the worst performance of x_k^σ on the set $\zeta \in Z$, which gives the desired bound. □

We shall next identify a family of Borel sets $Z \subset X$ for which Proposition 3 can be applied fruitfully: they have large measure and allow for a (order optimal) bound uniformly for $z^\sigma \in Z$. Let $\kappa > 0$ be a tuning parameter. We denote

$$Z_\kappa := \left\{\zeta, \quad \|\varrho_\lambda(A)\zeta\| \leq (1+\kappa)\sqrt{\mathcal{N}(\lambda)}\right\} . \tag{42}$$

Lemma 7.5. *For any fixed $\lambda > 0$ and $\alpha > 0$, the following holds:*

$$\mathbb{P}\left[\|\varrho_\lambda(A)\zeta\| > \sqrt{\mathcal{N}(\lambda)} + \sqrt{2\log\alpha^{-1}}\right] \leq \alpha. \tag{43}$$

Consequently,

$$\mathbb{P}\left[Z_\kappa^C\right] \leq \exp\left(-\frac{\kappa^2\mathcal{N}(\lambda)}{2}\right). \tag{44}$$

Proof. For any Gaussian random variable X taking values in Y, the following inequality holds:

$$\mathbb{P}\left[\|X\| > \mathbb{E}\left[\|X\|\right] + x\right] \leq \exp\left(-\frac{x^2}{2v^2}\right),$$

where $v^2 = \sup_{\|w\|=1} \mathbb{E}\left[\langle w, X\rangle^2\right]$, see [9, Lemma 3.1], and the discussion around (3.2), ibid. We apply this to $X = \varrho_\lambda(A)\zeta$, so that $\mathbb{E}\left[\langle w, X\rangle^2\right] = \langle w, A\varrho_\lambda(A)^2 w\rangle \leq 1$. Furthermore, we use (21) to complete the proof of the first assertion. For the second one we set $\alpha = \alpha(\kappa,\lambda) := \exp\left(-\frac{\kappa^2\mathcal{N}(\lambda)}{2}\right)$. □

A look at (41) shows that it remains to provide a bound for

$$\left(\mathbb{E}\left[\|x - x^\sigma_{\alpha_{k_{MDP}}}(z^\sigma)\|^{2q}\right]\right)^{1/2q},$$

and we shall use the point wise error bound from Fact 3 together with bound (24)) of Lemma 4.1. Before turning to such bound we recall the following well-known fact. Recall that $\tilde{\delta} = \sigma\|\varrho_\lambda(A)\zeta\|$ the (random) noise level measured in ϱ_λ-norm. The following is a consequence of the equivalence of all moments of a Gaussian variable in a Banach space, see e.g., [9, Cor. 3.2].

Lemma 7.6. *There is a constant $C(q)$ such that*

$$\mathbb{E}\left[\tilde{\delta}^{2q}\right]^{\frac{1}{2q}} \leq C(q)\mathbb{E}\left[\tilde{\delta}^2\right]^{\frac{1}{2}} = C(q)\sigma\sqrt{\mathcal{N}(\lambda)}.$$

Lemma 7.7. *Let x_k^σ be obtained from either* **cg** *or linear regularization. If $\alpha_{k_{MDP}}$ is obtained from the modified discrepancy principle, and if x obeys Assumption 2 then there is a constant C, not depending on σ, κ or λ for which*

$$\left(\mathbb{E}\left[\|x - x^\sigma_{\alpha_{k_{MDP}}}(z^\sigma)\|^{2q}\right]\right)^{1/2q} \leq C + C(q)F_\psi(\sigma\sqrt{\mathcal{N}(\lambda)}).$$

(In case of **cg** *iteration we additionally require that ψ is majorized by the power μ ($\psi \prec t^\mu$), and in this case $C(q) = C(\mu, q)$.)*

Proof. We start with linear regularization, and use Fact 3. This gives

$$\begin{aligned}\mathbb{E}\left[\|x - x^\sigma_{\alpha_{k_{MDP}}}\|^{2q}\right]^{1/2q} &\leq \gamma\psi(\alpha_{k_{MDP}}) + \frac{\max\left\{\gamma_*, 1+\gamma_1\right\}}{\Theta_{\varrho_\lambda}(\alpha_{k_{MDP}})}\left(\mathbb{E}\left[\tilde{\delta}^{2q}\right]\right)^{1/2q} \\ &\leq \gamma\psi(\|A\|) + C(q)\max\left\{\gamma_*, 1+\gamma_1\right\}\frac{1}{\eta\left(1+\kappa\right)} \\ &\leq \max(\psi(\|A\|), 1)C(\gamma, \gamma_1, \gamma_*, \eta, q).\end{aligned}$$

by virtue of Lemma 7.6 and the bound (24) of Lemma 4.1. The second summand is missing in this case.

We turn to proving a bound for **cg** iteration. The proof is similar, although more involved. Again, we start from Fact 3. The function $F := \psi \circ \Theta^{-1}_{\varrho_\lambda \psi}$ is majorized by the power 1, see Lemma 7.2, and hence sub-additive, see Lemma 7.3. Therefore, we deduce from Fact 3, item 2 that

$$\begin{aligned}\|x - x^\sigma_{\alpha_{k_{MDP}}}\| &\leq C(\mu) F\left(\tilde{\delta}\right) + C(\mu) F\left((2\mu+1)^{\mu+1}\,\Theta_{\varrho_\lambda \psi}(\alpha_{k_{MDP}})\right) + \frac{3}{\Theta_{\varrho_\lambda}(\alpha_{k_{MDP}})}\tilde{\delta} \\ &\leq C(\mu) F\left(\tilde{\delta}\right) + C(\mu)\left(2\mu+1\right)^{\mu+1} F\left(\Theta_{\varrho_\lambda \psi}(\alpha_{k_{MDP}})\right) + \frac{3}{\Theta_{\varrho_\lambda}(\alpha_k)}\tilde{\delta} \\ &\leq C(\mu) F\left(\tilde{\delta}\right) + C(\mu)\left(2\mu+1\right)^{\mu+1} \psi(\alpha_{k_{MDP}}) + \frac{3}{\Theta_{\varrho_\lambda}(\alpha_k)}\tilde{\delta}.\end{aligned}$$

The same reasoning as before allows to complete the proof in the **cg** case, where we notice that by Lemma 7.4 we can bound

$$\left(\mathbb{E}\left[F^{2q}\left(\tilde{\delta}\right)\right]\right)^{1/2q} \leq C(q) F(\sigma\sqrt{\mathcal{N}(\lambda)}) \leq C(q) F_\psi(\sigma\sqrt{\mathcal{N}(\lambda)}),$$

by Item (1) of Lemma 5.1. □

Remark 15. The explicit form of the constant shows that it involves, aside from $q, \psi(\|A\|)$ only scheme dependent constants. In case of **cg** iteration the required majorization power μ appears, and the bounds become worse with increasing μ.

We are now in the position to prove Theorem 2.

Proof of Theorem 2. We start with Proposition 3 using the set $Z := Z_\kappa$ from (42). The probability of its complement was bounded in Lemma 7.5. On the set Z_κ we have that $k_{MDP} = k_{DP}$ by the second part of of Lemma 4.1, wherein condition (25) holds by assumption of the Theorem. Therefore, at $k = k_{MDP}$ the uniform error bound over $\zeta \in Z_\kappa$ is given in Facts 1 and 2, respectively for linear regularization or **cg** iteration with $\delta := \sigma\sqrt{\mathcal{N}(\lambda)}$. We notice that for any λ the bounds given there are further bounded by $F_\psi(\sigma\sqrt{\mathcal{N}(\lambda)})$, by virtue of Lemma 5.1. The term with the $2q$th absolute moment in (41) is bounded in Lemma 7.7. Overall this gives the bound as stated in Theorem 2, and the proof is complete. □

References

[1] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636 (electronic), 2007.

[2] G. Blanchard and P. Mathé. Conjugate gradient regularization under general smoothness and noise assumptions. *J. Inverse Ill-Posed Probl.*, 18(6):701–726, 2010.

[3] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical Report CSAIL-TR 2006-062, Massachusetts Institute of Technology, 2006.

[4] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

[5] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3):034004, 19, 2008.

[6] M. Hanke. *Conjugate gradient type methods for ill-posed problems*, volume 327 of *Pitman Research Notes in Mathematics Series*. Longman Scientific & Technical, Harlow, 1995.

[7] B. Hofmann and P. Mathé. Analysis of profile functions for general linear regularization methods. *SIAM J. Numer. Anal.*, 45(3):1122–1141 (electronic), 2007.

[8] N. Korneĭchuk. *Exact constants in approximation theory*, volume 38 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1991. Translated from the Russian by K. Ivanov.

[9] M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

[10] P. Mathé and S. V. Pereverzev. Regularization of some linear ill-posed problems with discretized random noisy data. *Math. Comp.*, 75(256):1913–1929, 2006.

[11] P. Mathé and U. Tautenhahn. Regularization under general noise assumptions. submitted, 2010.

[12] M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission*, 16(2):52–68, 1980.

[13] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098, 2005.

Institut für Mathematik, Universität Potsdam, Am neuen Palais 10, 14469 Potsdam, Germany

Weierstrass-Institut für angewandte Analysis und Stochastik (WIAS), Mohrenstrasse 39, 10117 Berlin, Germany