# On the lifting of deterministic convergence rates for inverse problems with stochastic noise

Daniel Gerth*, Andreas Hofinger†, Ronny Ramlau‡

July 14, 2017

**Abstract**

Both for the theoretical and practical treatment of Inverse Problems, the modeling of the noise is a crucial part. One either models the measurement via a deterministic worst-case error assumption or assumes a certain stochastic behavior of the noise. Although some connections between both models are known, the communities develop rather independently. In this paper we seek to bridge the gap and show convergence and convergence rates for Inverse Problems with stochastic noise by lifting the theory established in the deterministic setting into the stochastic one. This opens the wide field of deterministic regularization methods for stochastic problems without having to do an individual stochastic analysis for each problem.

In Inverse Problems, the model of the inevitable data noise is of utmost importance. In most cases, an additive noise model

$$y^{\text{noisy}} = y + \epsilon \tag{1}$$

is assumed. In (1), $y \in \mathcal{Y}$ is the true data of the unknown $x \in \mathcal{X}$ under the action of the (in general) nonlinear operator $F : \mathcal{X} \to \mathcal{Y}$,

$$F(x) = y, \tag{2}$$

and $\epsilon$ in (1) corresponds to the noise. The spaces $\mathcal{X}, \mathcal{Y}$ are assumed to be Banach- or Hilbert spaces. When speaking of Inverse Problems, we assume that (2) is ill-posed. In particular this means that solving (2) for $x$ with noisy data (1) is unstable in the sense that "small" errors in the data may lead to arbitrarily large errors in the solution. Hence, (1) is not a sufficient description

*Corresponding author. Faculty of Mathematics, TU Chemnitz, 09107 Chemnitz, Germany. E-mail: daniel.gerth@mathematik.tu-chemnitz.de. Supported in part by the Austrian Science Fund (FWF): W1214-N15 and by the German Research Foundation (DFG) under grants HO1454/8-2 and HO1454/10-1

†Radon Institute for Computational and Applied Mathematics (RICAM), Altenbergerstraße 69, A-4040 Linz, Austria

‡Industrial Mathematics Institute, Johannes Kepler University Linz), Altenbergerstraße 69, A-4040 Linz, Austria

of the noise. More information is needed in order to compute solutions from the data in a stable way. In the *deterministic* setting, one assumes

$$d_{\mathcal{Y}}(y, y^\delta) \leq \delta \tag{3}$$

for some $\delta > 0$ where $d_{\mathcal{Y}}(\cdot, \cdot)$ is an appropriate distance functional. Typically, $d_{\mathcal{Y}}$ is induced by a norm such that (3) reads $||y - y^\delta|| \leq \delta$. Here and further on we use the superscript $\cdot^\delta$ to indicate the deterministic setting. Solutions of (2) under the assumption (1),(3) are often computed via a Tikhonov-type variational approach

$$x_\alpha^\delta = \min_{x \in \mathcal{D}(F)} d_{\mathcal{Y}}(F(x), y^\delta) + \alpha\Phi(x) \tag{4}$$

where again $d_{\mathcal{Y}}$ is a distance function and $\Phi(\cdot)$ is the penalty term used to stabilize the problem and to incorporate a-priori knowledge into the solution. The regularization parameter $\alpha$ is used to balance between data misfit and the penalty and has to be chosen appropriately. The literature in the deterministic setting is rich, at this point we only refer to the monographs [1, 2, 3] for an overview.

The deterministic worst-case error stands in contrast to *stochastic* noise models where a certain distribution of the noise $\epsilon$ in (1) is assumed. We shall indicate the stochastic setting by the superscript $\cdot^\eta$. In this paper, $\eta$ will be the parameter controlling the variance of the noise. Depending on the actual distribution of $\epsilon$, $d_{\mathcal{Y}}(y, y^\eta)$ may be arbitrarily large, but with low probability. In the Inverse Problems community, the Bayesian approach appears to be the most common method to find a solution of (2). For more detailed information, we refer to [4, 5, 6, 7, 8]. In the Bayesian setting, the solution of the Inverse Problem is given as a distribution of the random variable of interest, the *posterior distribution* $\pi_{post}$, determined by Bayes formula

$$\pi_{post}(x|y^\eta) = \frac{\pi_\epsilon(y^\eta|x)\pi_{pr}(x)}{\pi_{y^\eta}(y^\eta)}. \tag{5}$$

That is, roughly spoken, all values $x$ are assigned a probability of being a solution to (2) given the noisy data $y^\eta$. In (5), the *likelihood function* $\pi_\epsilon(y^\eta|x)$ represents the model for the measurement noise whereas the *prior distribution* $\pi_{pr}(x)$ represents a-priori information about the unknown. The data distribution $\pi_{y^\eta}(y^\eta)$ as well as the normalization constants are usually neglected since they only influence the normalization of the posterior distribution. In practice one is often more interested in finding a single representation as solution instead of the distribution itself. Popular point estimates are the *conditional expectation* (conditional mean, CM)

$$\mathbb{E}(\pi_{post}(x|y^\eta)) = \int x\pi_{post}(x|y^\eta)dx \tag{6}$$

and the *maximum a-posteriori (MAP)* solution

$$x^{\mathrm{MAP}} = \operatorname*{argmax}_x \quad \pi_{post}(x|y^\eta), \tag{7}$$

i.e., the most likely value for $x$ under the prior distribution given the data $y^\eta$. Both point estimators are widely used. The computation of the CM-solution is often slow since it requires repeated sampling of stochastic quantities and the evaluation of high-dimensional integrals. The MAP-solution, however, essentially leads to a Tikhonov-type problem. Namely, assuming $\pi_\epsilon(y^\eta|x) \propto \exp(-d_\mathcal{Y}(F(x), y^\eta))$ and $\pi_{pr}(x) \propto \exp(-\alpha\Phi(x))$, one has

$$\begin{aligned} x^{\mathrm{MAP}} &= \operatorname*{argmax}_x \exp(-d_\mathcal{Y}(F(x), y^\eta)) \exp(-\alpha\Phi(x)) \\ &= \operatorname*{argmin}_x d_\mathcal{Y}(F(x), y^\eta) + \alpha\Phi(x) \end{aligned}$$

analogously to (4).

Also non-Bayesian approaches for Inverse Problems often seek to minimize a functional (4), see e.g. [9, 10] or use techniques known from deterministic theory such as filter methods [11, 12]. Finally, Inverse Problems appear in the context of statistics. Hence, the statistics community has developed methods to solve (2), partly again based on the minimization of (4). We refer to [13] for an overview.

In summary, Tikhonov-type functionals (4) and other deterministic methods frequently appear also in the stochastic setting. From a practical point of view, one would expect to be able to use deterministic regularization methods for (2) even when the noise is stochastic. Indeed, the main question for the actual computation of the solution, given a particular sample of noisy data $y^\eta$, is the choice of the regularization parameter. A second question, mostly coming from the deterministic point of view, is the one of convergence of the solutions when the noise approaches zero. In the stochastic setting these questions are answered often by a full stochastic analysis of the problem. In this paper we present a framework that allows to find appropriate regularization parameters, prove convergence of regularization methods and find convergence rates for Inverse Problems with a stochastic noise model by directly using existing results from the deterministic theory.

The paper takes several ideas from the dissertation [14], which to our best knowledge is only publicly available as book [15]. It is organized as follows. In Section 1 we discuss an issue occurring in the transition from deterministic to stochastic noise for infinite dimensional problems. The Ky Fan metric, which will be the main ingredient of our analysis, and its relation to the expectation will be introduced in Section 2. We present our framework to lift convergence results from the deterministic setting into the stochastic setting in Section 3. Examples for the lifting strategy are given in Section 4.

# 1 On the noise model

Before addressing the convergence theory, we would like to discuss stochastic noise modeling and its intrinsic conflict with the deterministic model. Here and throughout the rest of the paper, assume

$$(\Omega, \mathcal{F}, \mathbb{P}) \tag{8}$$

to be a complete probability space with a set $\Omega$ of outcomes of the stochastic event, $\mathcal{F}$ the corresponding $\sigma$-algebra and $\mathbb{P}$ a probability measure, $\mathbb{P} : (\Omega, \mathcal{F}) \to [0, 1]$. We restrict ourselves here to probability measures for the sake of simplicity. Extensions to more general measures are straightforward. In the Hilbert-space setting, the noise is typically modeled as follows, see for example [16, 1, 12, 17]. Let $\xi : \Omega \to \mathcal{Y}$ be a stochastic process. Then for $y \in \mathcal{Y}$

$$\langle y, \xi \rangle \tag{9}$$

defines a real-valued random variable. Assuming that

$$\mathbb{E}(\langle \tilde{y}, \xi \rangle^2) < \infty \tag{10}$$

for all $\tilde{y} \in \mathcal{Y}$ and that this expectation is continuous in $\tilde{y}$,

$$\mathbb{E}(\langle \tilde{y}, \xi \rangle \langle y, \xi \rangle)$$

defines a continuous, symmetric nonlinear bilinearform. In particular, there exists the *covariance operator*

$$\mathcal{C} : \mathcal{Y} \to \mathcal{Y}$$

with

$$\langle \mathcal{C}\tilde{y}, y \rangle = \mathbb{E}(\langle \tilde{y}, \xi \rangle \langle y, \xi \rangle).$$

For the stochastic analysis of infinite dimensional problems via deterministic results, (9) is problematic. Namely, if $\{u_n\}_{n \in \mathbb{N}}$ is an orthonormal basis in $\mathcal{Y}$, the set $\{\langle u_n, \xi \rangle\}_{n \in \mathbb{N}}$ consists of infinitely many identically distributed random variables with $0 < \mathbb{E}|\langle u_n, \xi \rangle|^2 = \text{const} < \infty$ [1]. Thus

$$\mathbb{E}\left( \sum_{n=1}^{\infty} |\langle u_n, \xi \rangle|^2 \right) \tag{11}$$

is almost surely infinite and a realization of the noise is an element of the Hilbert space $\mathcal{Y}$ with probability zero. Let us take the common example of Gaussian white noise which can be modeled via the above construction. Namely, with

$$\mathbb{E}(\langle y, \xi \rangle) = 0 \quad \forall y \in \mathcal{Y}$$

and the covariance operator

$$\mathcal{C} = \eta^2 I,$$

where $I$ is the identity and $\eta > 0$ the variance parameter, the Gaussian white noise is described [1, 17]. As consequence of (11) and explained for example in [17], a realization of such a Gaussian random variable is an element of an infinite dimensional $L_2$-space with probability zero. It is therefore inappropriate to use an $L_2$-norm for the residual in case of an infinite dimensional problem. Since in this case a realization of Gaussian white noise only lies (almost surely) in any Sobolev space $H^s$ with $s < -d/2$ where $d$ is the dimension of the domain, one should adjust the norm for the residual accordingly. Except for the paper [17]

this issue seems not to have been addressed in the literature. A main reason for this might be that for the practical solution of the Inverse Problem this is not a severe issue since in reality the measurements are finite dimensional and, in order to use a computer to solve the problem, a finite dimensional approximation of the unknown has to be used. In this case the sum in (11) is finite and the noise lies within the finite dimensional space almost surely. However, difficulties arise whenever one seeks to investigate convergence of the discretized problem to its underlying infinite dimensional problem. We will not address this issue and assume throughout the whole work that $\mathbb{E}||\epsilon|| < \infty$ or use the slightly weaker bound on the Ky-Fan metric (see Section 2). In order to handle the Ky Fan metric we need to be able to evaluate probabilities $\mathbb{P}(||y - y^\eta|| > \varepsilon)$, $0 \leq \varepsilon \leq 1$, which is only meaningful if $y - y^\eta =: \epsilon \in \mathcal{Y}$. Assuming that $\mathcal{Y}$ is finite dimensional, then this is clear. For infinite dimensional problems, however, we have to assume that the noise is smooth enough for the sum in (11) to converge. Examples for this are Brownian noise ($1/f^2$-noise) or pink noise ($1/f$-noise), see e.g. [18, 19]. At this point we would also like to mention that as a consequence of our rather generic noise model we might not make use of some specific properties of the noise as would be possible when focusing on a particular distribution of the noise. However, we are able to show convergence for a large variety of regularization methods.

# 2 The Ky Fan metric

The Ky Fan metric (cf. [20]) will be the main tool for our stochastic convergence analysis. It is defined as follows.

**Definition 2.1.** *Let $X_1$ and $X_2$ be random variables in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a metric space $(\chi, d_\chi)$. The distance between $X_1$ and $X_2$ in the Ky Fan metric is defined as*

$$\rho_K(X_1, X_2) := \inf_{\varepsilon > 0} \{\mathbb{P}(\{\omega \in \Omega : d_\chi(X_1(\omega), X_2(\omega)) > \varepsilon\}) < \varepsilon\}. \tag{12}$$

We will often drop the explicit reference to $\omega$. This metric essentially allows to lift results from a metric space to the space of random variables as the connection to the deterministic setting is inherent via the metric $d_\chi$ used in its definition. The deterministic metric is often induced by a norm $||\cdot||$. We will implicitly assume that equation (2) is scaled appropriately since $\rho_K(X_1, X_2) \leq 1$ $\forall X_1, X_2$ by definition. Note that one can use definition (12) also if $d_\mathcal{X}$ is a more general distance function than a metric. Then the construction (12) itself is no longer a metric, however, the techniques used in later parts of the paper can readily be expanded to this setting.

An immediate consequence of (12) is that $\rho_K(X_1, X_2) = 0$ if and only if $X_1 = X_2$ almost surely. Convergence in the Ky Fan metric is equivalent to convergence in probability, i.e., for a sequence $\{X_k\}_{k\in\mathbb{N}} \in \mathcal{X}$ and $X \in \mathcal{X}$ one has

$$\rho_K(X_k, X) \overset{k\to\infty}{\longrightarrow} 0 \quad \Leftrightarrow \quad \forall \varepsilon > 0 : \quad \mathbb{P}(d_\mathcal{X}(X_k, X) > \varepsilon) \overset{k\to\infty}{\longrightarrow} 0.$$

Hence convergence in the Ky Fan metric also leads to pointwise (almost sure) convergence of certain subsequences in the metric $d_\chi$ [21].

A somewhat more intuitive and more frequently used metric is the expectation, or more general, a (stochastic) $L_p$ metric. Assuming its existence, for random variables $Y_1$ and $Y_2$ with values in a metric space $(\chi, d_\mathcal{Y})$,

$$\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^p) = \int_\Omega d_\mathcal{Y}(Y_1, Y_2)^p d\mathbb{P}(\omega)$$

defines the $p$-th moment of $d_\mathcal{Y}(Y_1, Y_2)$ for $p \geq 1$. We will use $p = 1$ and refer to it as *convergence in expectation*. Note that since the variance is defined as

$$\mathrm{Var}(d_\mathcal{Y}(Y_1, Y_2)) = \mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^2) - E(d_\mathcal{Y}(Y_1, Y_2))^2 \geq 0$$

one always has

$$\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)) \leq \sqrt{\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^2)}. \tag{13}$$

We will show later that for parameter choice rules the expectation of the noise has to be slightly overestimated, hence estimating $\mathbb{E}(d_\mathcal{Y}(y, y^\eta))$ via the popular and often easier to compute $L_2$-norm $E(d_\mathcal{Y}(y, y^\eta)^2)$ with (13) is not problematic.

It is well-known that convergence in expectation implies convergence in probability, see for example [21]. Hence, convergence in the Ky Fan metric is implied by convergence in expectation (and also by convergence of higher moments). Namely, with Markovs inequality one has, for an arbitrary nonnegative random variable $X$ with $E(X) < \infty$ and $C > 0$

$$\mathbb{P}(X > C) \leq \frac{\mathbb{E}(X)}{C}. \tag{14}$$

Under an additional assumption, one can show conversely that convergence in probability implies convergence in expectation. We have the following definition.

**Definition 2.2** ([22], Definition A.3.1.)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. A family $\mathcal{G} \subset L_1(\mathbb{P})$ is called uniformly integrable if*

$$\lim_{C \to \infty} \sup_{x \in \mathcal{G}} \int_{|x| > C} |x(t)| \mathbb{P}(dt) = 0$$

**Theorem 2.1** ([22], Theorem A.3.2.)**.** *Let $\{x_k\}_{k \in \mathbb{N}} \subset L_1(\mathbb{P})$ be a sequence convergent almost everywhere (or in probability) to a function $x$. If the sequence $\{x_k\}_{k \in \mathbb{N}}$ is uniformly integrable, then it converges to $x$ in the norm of $L_1(\mathbb{P})$.*

From a practical point of view, uniform integrability of a sequence of regularized solutions to an Inverse Problem is a rather natural condition. Since Inverse Problems typically arise from some real-world application, it is to be expected that the true solution is bounded. For example, in Computer Tomography, the density of the tissue inside the body cannot be arbitrarily high. Although for an Inverse Problem with a stochastic noise model, boundedness of the regularized solutions can not be guaranteed due to the possibly huge measurement error, one can enforce the condition from a priori knowledge of the solution.

**Assumption 2.2.** *Assume that the true solution $x^\dagger$ fulfills $||x^\dagger|| \leq \varrho$ and $|x^\dagger| \leq C$ globally for some fixed $\varrho, C > 0$.*

Under this assumption, let $\{x_k^{\eta(k)}\}_{k\in\mathbb{N}}$ be a sequence of regularized solution for noisy data with variance $\eta(k) \overset{k\to\infty}{\to} 0$. Let $C_1, C_2 > 1$ and define

$$\tilde{x}_k^\eta := \begin{cases} x_k^\eta, & ||x_k^\eta|| \leq C_1\varrho, |x_k^\eta| \leq C_2 C \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

Then the sequence $\{\tilde{x}_k^\eta\}_{k\in\mathbb{N}}$ is uniformly integrable. In other words, by discarding solutions that must be far away from the true solution in regard of a priori knowledge, convergence in the Ky Fan metric implies convergence in expectation.

To close this section, let us remark on the computation of the Ky Fan distance. It can be estimated via the moments of the noise.

**Theorem 2.3.** *Let $Y_1, Y_2$ be random variables in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^s) < \infty$ for some $s \in \mathbb{N}$. Then*

$$\rho_K(Y_1, Y_2) \leq \sqrt[s+1]{\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^s)} \tag{16}$$

*Proof.* One has, due to Markov's inequality (14) and the monotonicity of the mapping $z \mapsto z^s$ for $z \geq 0$,

$$\mathbb{P}(d_\mathcal{Y}(Y_1, Y_2) > C) = \mathbb{P}(d_\mathcal{Y}(Y_1, Y_2)^s > C^s) \leq \frac{\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^s)}{C^s}$$

for $C \geq 0$. Solving $C = \frac{\mathbb{E}(d_\mathcal{Y}(Y_1,Y_2)^s)}{C^s}$ for $C$ yields the claim. □

Note that even if moments exist for all $s \in \mathbb{N}$

$$\lim_{s\to\infty} \sqrt[s+1]{\mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)^s)} \neq \mathbb{E}(d_\mathcal{Y}(Y_1, Y_2)),$$

see [14, 23], due to the tail of the distributions. In the Gaussian case, a direct estimate has been derived in [24, 25]. We present it in Proposition 3.6.

# 3 Convergence in the stochastic setting

## 3.1 Deterministic Inverse Problems with stochastic noise

As mentioned previously, the intention of this paper is to show convergence for Inverse Problems under a stochastic noise model using results from the deterministic setting. Assume we have at hand a deterministic regularization method of our liking for the solution of (2) under the noisy data (1) where now $d_\mathcal{Y}(y, y^\delta) \leq \delta$ for some $\delta > 0$. Under regularization method we understand a (again possibly) nonlinear mapping

$$R_\alpha : \mathcal{D}(R_\alpha) \subset \mathcal{Y} \to \mathcal{X}, \quad y^\delta \mapsto x_\alpha^\delta \tag{17}$$

where $x_\alpha^\eta = R_\alpha(y^\eta)$ is the regularized solution to the regularization parameter $\alpha$ given the data $y^\eta$. Often, $x_\alpha^\delta$ is obtained via the minimization of functionals of the type (4). In order to deserve the name regularization we require $R_\alpha$ to fulfill

$$\lim_{\delta \to 0} ||R_\alpha(y^\delta) - x^\dagger|| = 0 \tag{18}$$

under a certain choice of the regularization parameter $\alpha$ chosen either a priori $\alpha = \alpha(\delta)$ or a posteriori $\alpha = \alpha(\delta, y^\delta)$. In our notation $x^\dagger$ is the true solution, usually the minimum norm solution with respect to the penalty $\Phi$ in (4), i.e.,

$$\Phi(x^\dagger) \leq \Phi(\bar{x}) \quad \text{for all} \quad \bar{x} : F(\bar{x}) = y.$$

Note that, in particular for nonlinear problems, $x^\dagger$ does not need to be unique. In [14, 15] it was pointed out that this is problematic for the lifting arguments. A standard argument in the deterministic theory is to prove convergence of subsequences to the desired solution, and then deduces convergence of the whole series of regularized solutions if possible. In the stochastic setting, this is not possible in general since subsequences for different $\omega$ do not have to be related. A constructed example for this behavior can be found in Section 4.1. of [14, 15]. In order to lift general deterministic regularization methods into the stochastic setting we must therefore require that $x^\dagger$ is unique. We formulate our convergence results allowing the noise to be bounded in the Ky fan metric or in expectation. As we will see, in the latter case we have to "inflate" the expectation for decreasing variance $\eta$ in order to obtain convergence. For the analysis we mainly use a lifting argument using deterministic theory. In [14, 15, Theorem 4.1], it was proved how by means of the Ky Fan metric deterministic results can be lifted to the space of random variables for nonlinear Tikhonov regularization. Since the theorem is based solely on the fact that there is a deterministic regularization theory and that the probability space $\Omega$ can be decomposed into a part where the deterministic theory holds and a small part where it does not, it is easily generalized. Before we state the Theorem, we need the following Lemmata.

**Lemma 3.1.** *([26], see also [21]) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Let $x_k$ and $x$ be measurable functions from $\Omega$ into a metric space $\chi$ with metric $d_\chi$. Suppose $x_k(\omega) \xrightarrow{d_\chi} x(\omega)$ for $\mathbb{P}$-almost all $\omega \in \Omega$. Then for any $\varepsilon > 0$ there is a set $\tilde{\Omega}$ with $\mathbb{P}(\Omega \backslash \tilde{\Omega}) < \varepsilon$ such that $x_k \xrightarrow{d_\chi} x(\omega)$ uniformly on $\tilde{\Omega}$, that is*

$$\lim_{k \to \infty} \sup\{d_\chi(x_k(\omega), x(\omega)) : \omega \in \tilde{\Omega}\} = 0.$$

**Lemma 3.2** ([14, 15], Proposition 1.10)**.** *Let $\{x_k\}_{k \in \mathbb{N}}$ be a sequence of random variables that converges to $x$ in the Ky Fan metric. Then for any $\nu > 0$ and $\varepsilon > 0$ there exist $\tilde{\Omega} \subset \Omega$, $\mathbb{P}(\tilde{\Omega}) \geq 1 - \varepsilon$, and a subsequence $x_{k_j}$ with*

$$d_{\mathcal{X}}(x_{k_j}(\omega), x(\omega)) \leq (1 + \nu)\rho_K(x_{k_j}, x) \qquad \forall \omega \in \tilde{\Omega}.$$

*Furthermore there exists a subsequence that converges to $x$ almost surely.*

*Proof.* We give a sketch of the proof for the first statement taken from [14, 15]. Set $\sigma_k := (1+\nu)\rho_K(x_k, x)$. By definition of the Ky Fan metric (12), for given $\sigma_k$, there exists a set $\Omega_{\sigma_k}$ with $\mathbb{P}(\Omega_{\sigma_k}) \geq 1 - \sigma_k$ and $\omega \in \Omega_{\sigma_k}$ such that $d_{\mathcal{X}}(x(\omega), x_k(\omega)) \leq \sigma_k$. For arbitrary $\varepsilon > 0$ and $\sigma_k \to 0$ we pick a subsequence $(\sigma_{k^j})$ with $\sum_{j=1}^{\infty} \sigma_{k^j} \leq \varepsilon$ and introduce the set $\tilde{\Omega} := \bigcap_{j=1}^{\infty} \Omega_{\sigma_{k^j}}$. One can check that $\mathbb{P}(\tilde{\Omega}) \geq 1 - \varepsilon$. Since $\tilde{\Omega}$ is a subset of every $\Omega_{\sigma_{k^j}}$ we have

$$\forall \omega \in \tilde{\Omega} \subseteq \Omega_{\sigma_{k^j}} : d_{\mathcal{X}}(x(\omega), x_{k^j}(\omega)) \leq \sigma_{k^j},$$

which proves he first statement. The second one follows since convergence in Ky Fan metric is equivalent to convergence in probability, which itself implies almost-sure convergence of a subsequence, cf [21]. □

With this, we are ready for the convergence theorem which we shall split in two parts, one for the Ky Fan metric as error measure and one for the expectation.

**Theorem 3.3.** *Let $R_\alpha$ be a regularization method for the solution of* (2) *in the deterministic setting under a suitable choice of the regularization parameter. Let now $y^\eta = y + \epsilon(\eta)$ where $\epsilon(\eta)$ is a stochastic error such that $\rho_K(y, y^\eta) \to 0$ as $\eta \to 0$. Then, assuming* (2) *has a unique solution $x^\dagger$ and all necessary assumptions for the deterministic theory (except the bound on the noise) hold with probability one, the regularization method $R_\alpha$ fulfills*

$$\lim_{\eta \to 0} \rho_K(x^\dagger, R_\alpha(y^\eta)) = 0$$

*under the same parameter choice rule as in the deterministic setting with $\delta$ replaced by $\rho_K(y, y^\eta)$. If the regularized solutions are defined by* (15) *with regard to Assumption 2.2, then it holds that*

$$\lim_{\eta \to 0} \mathbb{E}(d_{\mathcal{X}}(x^\dagger, R_\alpha(y^\eta))) = 0.$$

*Proof.* Denote $x_\alpha(\eta) := R_\alpha(y^\eta)$. Define $\theta := \limsup_{k\to\infty} \rho_K(x^\dagger, x_{\alpha(\eta_k)})$. (Note that $0 \leq \theta \leq 1$ due to the properties of the Ky Fan metric). We show in the following that for arbitrary $\varepsilon > 0$ we have $\theta/2 \leq \varepsilon$ and hence

$$\limsup_{k\to\infty} \rho_K(x^\dagger, x_{\alpha(\eta_k)}) = \lim_{k\to\infty} \rho_K(x^\dagger, x_{\alpha(\eta_k)}) = 0.$$

As a first step we pick a "worst case" subsequence $\{y^{\eta_{k^j}}\}$ of $\{y^{\eta_k}\}$, a subsequence for which the corresponding solutions satisfy $\rho_K(x^\dagger, x_{\alpha(\eta_{k^j})}) \geq \theta/2$. We now show that even from this "worst case" sequence we can pick a subsequence $\{y^{\eta_{k_l^j}}\}$ for which we have $\limsup \rho_K(x^\dagger, x_{\alpha(\eta_{k_l^j})}) \leq \varepsilon$ for arbitrary $\varepsilon > 0$.

Let $\varepsilon > 0$. According to Lemma 3.2 we can pick a subsequence $\{y^{\eta_{k_l^j}}\}$ and a set $\tilde{\Omega}$ with $\mathbb{P}(\tilde{\Omega}) \geq 1 - \frac{\varepsilon}{2}$ as well as $d_{\mathcal{Y}}(y(\omega), y^{\eta_{k_l^j}}(\omega)) \leq (1+\nu)\rho_K(y, y^{\eta_{k_l^j}})$, $\nu > 0$ arbitrarily small, on $\tilde{\Omega}$. For all $\omega \in \tilde{\Omega}$, the noise tends to zero. We can therefore

use the deterministic result with $\delta = \rho_K(y(\omega), y^{\eta_{k_l^j}})$ and deduce that $x_{\alpha(\eta_{k_l^j})}(\omega)$ converges to the unique solution $x^\dagger(\omega)$ for $\eta_{k_l^j} \to 0$, $\omega \in \tilde{\Omega}$ where in the choice of the regularization parameter $\delta$ is substituted by $\rho_K(y(\omega), y^{\eta_{k_l^j}})$. The convergence is not uniform in $\omega$; nevertheless, pointwise convergence implies uniform convergence except on sets of small measure according to Lemma 3.1. Therefore there exist $\tilde{\Omega}' \subset \tilde{\Omega}$, $\mathbb{P}(\tilde{\Omega}') < \frac{\varepsilon}{2}$ and $j_0 \in \mathbb{N}$ such that $d_{\mathcal{X}}(x_{\alpha(\eta_{k_l^j})}(\omega), x^\dagger(\omega)) < \varepsilon$ $\forall \omega \in \tilde{\Omega} \backslash \tilde{\Omega}'$ and $j \geq j_0$. We thus have

$$\mathbb{P}\left(\left\{\omega \in \tilde{\Omega} : d_{\mathcal{X}}(x_{\alpha(\eta_{k_l^j})}(\omega), x^\dagger(\omega)) > \varepsilon\right\}\right) \leq \mathbb{P}(\tilde{\Omega}') \leq \varepsilon/2.$$

Since we split $\Omega = \Omega \backslash \tilde{\Omega} \cup \tilde{\Omega} \backslash \Omega'_\varepsilon \cup \Omega'_\varepsilon$ with $\mathbb{P}(\Omega \backslash \tilde{\Omega}) < \frac{\varepsilon}{2}$, $\mathbb{P}(\Omega \backslash \tilde{\Omega}) + \mathbb{P}(\tilde{\Omega}') \leq \varepsilon$ we have shown existence of a subsequence $\eta_{k_l^j}$ such that

$$\mathbb{P}\left(\left\{\omega \in \Omega : d_{\mathcal{X}}(x_{\alpha(\eta_{k_l^j})}(\omega), x^\dagger(\omega)) > \varepsilon\right\}\right) \leq \varepsilon$$

for $\eta_{k_l^j}$ sufficiently small. This $\varepsilon$ is by definition of the Ky Fan metric an upper bound for the distance between $x_{\alpha(\eta_{k_l^j})}$ and $x^\dagger$. Therefore we have $\limsup_{l\to\infty} \rho_K(x_{\alpha(\eta_{k_l^j})}, x^\dagger) \leq \varepsilon$. On the other hand, the original sequence satisfied $\liminf_{j\to\infty} \rho_K(x^\dagger, x_{\alpha(\eta_{k^j})}) \geq \theta/2$. Since $\liminf_{j\to\infty} \rho_K(x^\dagger, x_{\alpha(\eta_k)}) \leq \limsup_{l\to\infty} \rho_K(x_{\alpha(\eta_{k_l^j})}, x^\dagger)$ it follows $\theta/2 \leq \varepsilon$. Because $\varepsilon > 0$ was arbitrary, this implies $\theta = 0$, which concludes the proof of convergence in the Ky Fan metric. Convergence in expectation follows from Theorem 2.1 noting that by (15) the sequence of regularized solutions is uniformly integrable. □

**Theorem 3.4.** *Let $R_\alpha$ be a regularization method for the solution of (2) in the deterministic setting under a suitable choice of the regularization parameter. Let now $y^\eta = y + \epsilon(\eta)$ where $\epsilon(\eta)$ is a stochastic error such that $\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta)) \to 0$ as $\eta \to 0$. Then, assuming (2) has a unique solution $x^\dagger$ and all necessary assumptions for the deterministic theory (except the bound on the noise) hold with probability one, the regularization method $R_\alpha$ fulfills*

$$\lim_{\eta\to 0} \rho_K(x^\dagger, R_\alpha(y^\eta)) = 0$$

*under the same parameter choice rule as in the deterministic setting with $\delta$ replaced by $\tau(\eta)\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta))$ where $\tau(\eta)$ fulfills*

$$\tau(\eta) \overset{\eta\to 0}{\to} \infty \quad \text{and} \quad \lim_{\eta\to 0} \tau(\eta)\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta)) = 0. \tag{19}$$

*If the regularized solutions are defined by (15) with regard to Assumption 2.2, then it holds that*

$$\lim_{\eta\to 0} \mathbb{E}(d_{\mathcal{X}}(x^\dagger, R_\alpha(y^\eta))) = 0.$$

*Proof.* As previously we pick a "worst case" subsequence $\{y^{\eta_{k^j}}\}$ of $\{y^{\eta_k}\}$, a subsequence for which the corresponding solutions satisfy $\rho_K(x^\dagger, x_{\alpha(\eta_{k^j})}) \geq \theta/2$. Let $\varepsilon > 0$. We can now pick a subsequence which we again denote by $\{y^{\eta_{k_l^j}}\}$ fulfilling $\frac{2}{\tau(\eta_{k_l^j})} \leq \varepsilon$, where without loss of generality $t(\eta_{k_l^j}) > 1$, such that

$$\mathbb{P}(\omega : d_{\mathcal{Y}}(y(\omega) - y^{\eta_{k_l^j}}(\omega)) > \tau(\eta_{k_l^j})\mathbb{E}(d_{\mathcal{Y}}(y, y^{\eta_{k_l^j}}))) \leq \frac{1}{\tau(\eta_{k_l^j})} \leq \frac{\varepsilon}{2}.$$

This again defines, via the complement in $\Omega$, $\tilde{\Omega}$ with $\mathbb{P}(\tilde{\Omega}) \geq 1 - \frac{\varepsilon}{2}$ on which $d_{\mathcal{Y}}(y(\omega), y^{\eta_{k_l^j}}(\omega)) \leq \tau(\eta_{k_l^j})\mathbb{E}(d_{\mathcal{Y}}(y, y^{\eta_{k_l^j}}))$. As before, we can now apply the deterministic theory by substituting $\delta$ with $\tau(\eta_{k_l^j})\mathbb{E}(d_{\mathcal{Y}}(y, y^{\eta_{k_l^j}}))$. The remainder of the proof is identical to the one of Theorem 3.4. □

The theorems justify the use of deterministic algorithms under a stochastic noise model. Since the proof is solely based on relating the stochastic noise to a deterministic one on subsets of $\Omega$ and does not use any specific properties of the regularization methods or the underlying spaces, it opens most of the deterministic methods for the a stochastic noise model. In particular, the parameter choice rules from the deterministic setting are easily adapted.

As usual in deterministic literature, the general convergence theorem is followed by convergence rates which are obtained under additional assumptions. Often these conditions ensure at least local uniqueness of the true solution. If not, we have to require such a property for the same reason as previously.

**Theorem 3.5.** *Let $R_\alpha$ be a regularization method for the solution of (2) in the deterministic setting such that, under a set of assumptions on the operator $F$ and the solutions $x^\dagger$ and a suitable choice of the regularization parameter,*

$$d_{\mathcal{X}}(x^\dagger, R_\alpha(y^\delta)) \leq \varphi(d_{\mathcal{Y}}(y, y^\delta))$$

*with a monotonically increasing right-continuous function $\varphi$.*

*Let now $y^\eta = y + \epsilon(\eta)$ where $\epsilon(\eta)$ is a stochastic error such that*

*a) $\rho_K(y, y^\eta) \to 0$ or*

*b) $\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta)) \to 0$*

*as $\eta \to 0$. Then, assuming all necessary assumptions for the deterministic theory (except the bound on the noise) hold with probability one and that there is (either by the deterministic conditions or by additional assumption) a (locally) unique solution $x^\dagger$ to (2), the regularization method $R_\alpha$ fulfills*

$$\rho_K(x^\dagger, R_\alpha(y^\eta)) = \mathcal{O}(\max\{\varphi(\rho_K(y, y^\eta)), \rho_K(y, y^\eta)\})$$

*in case a) or, respectively, in case b),*

$$\rho_K(x^\dagger, R_\alpha(y^\eta)) = \mathcal{O}(\max\{\varphi(\tau(\eta)\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta))), \mathbb{P}(d_{\mathcal{Y}}(y, y^\eta) \geq \tau(\eta)\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta)))\})$$

*under the same parameter choice rule as in the deterministic setting with $\delta$ replaced by $\rho_K(y, y^\eta)$ (case a)) or $\tau(\eta)\mathbb{E}(d_\mathcal{Y}(y, y^\eta))$ where $\tau(\eta)$ fulfills* (19) *(case b)).*

*Proof.* We start again with the Ky Fan distance as noise measure. Since we have the deterministic theory at hand, we know that $d_\mathcal{X}(x^\dagger, R_\alpha(y^\eta)) \leq C\varphi(\delta)$ whenever $d_\mathcal{Y}(y, y^\eta) \leq \delta$. With $\delta = \rho_K(y, y^\eta)$ we have, since $\varphi$ is monotonically increasing and right continuous,

$$\begin{aligned}\mathbb{P}(d_\mathcal{X}(x^\dagger, R_\alpha(y^\eta)) > \varphi(\rho_K(y, y^\eta))) &\leq \mathbb{P}(d_\mathcal{Y}(y, y^\eta) \geq \rho_K(y, y^\eta)) \\ &\leq 1 - \mathbb{P}(d_\mathcal{Y}(y, y^\eta) \leq \rho_K(y, y^\eta)) \\ &\leq 1 - (1 - \rho_K(y, y^\eta)) = \rho_K(y, y^\eta)\end{aligned}$$

and hence by definition $\rho_K(x^\dagger, R_\alpha(y^\eta)) = \mathcal{O}(\max\{\varphi(\rho_K(y, y^\eta)), \rho_K(y, y^\eta)\})$.

If the expectation is used as measure for the data error, we have

$$\mathbb{P}(d_\mathcal{Y}(y, y^\eta) \geq \tau(\eta)\mathbb{E}(d_\mathcal{Y}(y, y^\eta))) \leq \frac{1}{\tau(\eta)}$$

by Markovs inequality. Hence, with probability $1 - \frac{1}{\tau(\eta)}$ we are in the deterministic setting with $\delta = \tau(\eta)\mathbb{E}(d_\mathcal{Y}(y, y^\eta))$ and

$$\begin{aligned}&\mathbb{P}(d_\mathcal{X}(x^\dagger, R_\alpha(y^\delta)) > \varphi(\tau(\eta)\mathbb{E}(d_\mathcal{Y}(y, y^\eta)))) \\ &\qquad\qquad \leq \mathbb{P}(d_\mathcal{Y}(y, y^\eta) > \tau(\eta)\mathbb{E}(d_\mathcal{Y}(y, y^\eta))) \leq \frac{1}{\tau(\eta)}.\end{aligned}$$

The convergence rate follows by the definition of the Ky Fan metric. □

For Inverse Problems, the convergence rates are most often given by functions which decay at most linearly fast, i.e.,

$$\max\{\varphi(\rho_K(y, y^\eta)), \rho_K(y, y^\eta)\} = \varphi(\rho_K(y, y^\eta)).$$

Hence in this case the convergence rates are preserved in the Ky Fan metric. For the expectation this is not the case. We have to gradually inflate the expectation by the parameter $\tau$ in order to obtain convergence (and rates). Let us discuss the simple example of Gaussian noise in the finite dimensional setting, i.e. $\epsilon$ from (1) consists of $m \in \mathbb{N}$ i.i.d. random variables $\epsilon_i \sim \mathcal{N}(0, \eta^2 I_m)$ with zero mean and variance $\eta^2$. Then it has been shown in [23] that for any $\tau > 1$

$$\mathbb{P}(\|\epsilon\|_2 \geq \tau\mathbb{E}(\|\epsilon\|_2)) = \frac{\Gamma(\frac{m}{2}, (\tau\Gamma(\frac{m+1}{2})/\Gamma(\frac{m}{2}))^2)}{\Gamma(\frac{m}{2})} \tag{20}$$

with the gamma functions $\Gamma(\cdot)$ and $\Gamma(\cdot, \cdot)$ defined as

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt, \quad \Gamma(a, z) = \int_z^\infty t^{a-1}e^{-t}dt.$$

In particular, (20) is independent of the variance $\eta^2$. In order to to decrease the probability to zero, we therefore have to link $\tau$ with the variance. For Gaussian noise of the above kind the following estimate for the Ky Fan distance between true and noisy data has been given in [24].

**Proposition 3.6.** *Let $\xi$ be a random variable with values in $\mathbb{R}^m$. Assume that the distribution of $\epsilon$ is $\mathcal{N}(0, \eta^2 I_m)$ with $\sigma > 0$. Then it holds in $(\mathbb{R}^m, ||\cdot||_2)$ that*

$$\rho_K(\epsilon, 0) \leq \min\left\{1, \sqrt{2}\eta\sqrt{m - \min\left\{\ln\left(\eta^2 2\pi m^2 \left(\frac{e}{2}\right)^m\right), 0\right\}}\right\}. \tag{21}$$

Recall that

$$\mathbb{E}(||\epsilon||_2) = \eta\Gamma\left(\frac{1+m}{2}\right) / \left(\sqrt{2}\Gamma\left(\frac{m}{2}\right)\right) \leq \eta\sqrt{m}, \tag{22}$$

see e.g. [23]. Comparing (21) and (22), one sees that $\mathbb{E}(||\epsilon||_2) < \rho_K(\epsilon, 0)$ and in particular the decay of $\rho_K(\epsilon, 0)$ slows down with decreasing $\eta$. In other words, the artificial inflation we had to impose on the expectation is automatically included in the Ky Fan distance which we suppose is the reason why the convergence theory carries over in such a direct fashion for the Ky Fan metric.

For many nonlinear Inverse Problems the requirement of a unique solution is too strong. Often one has several solutions of the same quality, in particular there exists more than one minimum norm solution. In this case, Theorem 3.3 is not applicable. In the example [14, 15, Example 4.3 and 4.5] with two minimum norm solutions the noise was constructed such that, while the error in the data converges to zero, for each fixed $\omega \in \Omega$ the regularized solutions jump between both solutions such that no converging subsequence can be found. The main problem there is that the Ky Fan distance cannot incorporate the concept that all minimum norm solutions are equally acceptable. We will now define a pseudo metric that resolves this issue.

**Definition 3.1.** *Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space. Denote with $\mathcal{L}$ the set of minimum-norm solutions to* (2)*. Then*

$$\rho_K^\mathcal{L}(x) := \inf_{\varepsilon > 0}\left\{\mathbb{P}\left(\inf_{x^\dagger \in \mathcal{L}} d_\mathcal{X}(x, x^\dagger) > \varepsilon\right) \leq \varepsilon\right\} \tag{23}$$

*measures the distance between an element $x \in \mathcal{X}$ and the set $\mathcal{L}$, in particular it is*

$$\rho_K^\mathcal{L}(x) = 0 \quad \Leftrightarrow \quad x \in \mathcal{L} \quad \textit{almost surely.}$$

With this, one can define a pseudometric on $(\Omega, \mathcal{F}, \mathbb{P})$ via

$$\rho_K^\mathcal{L}(x_1, x_2) =: \max\{\rho_K^\mathcal{L}(x_1), \rho_K^\mathcal{L}(x_2)\}. \tag{24}$$

Obviously (24) is positive, symmetric and fulfills the triangle inequality. However, $\rho_K^\mathcal{L}(x_1, x_2) = 0$ does not imply $x_1 = x_2$ a.e. but instead $x_1 \wedge x_2 \in \mathcal{L}$ which fixes the aforementioned issue of the Ky Fan metric and allows the following theorems.

**Theorem 3.7.** *Let $R_\alpha$ be a regularization method for the solution of* (2) *in the deterministic setting under a suitable choice of the regularization parameter. Let now $y^\eta = y + \epsilon(\eta)$ where $\epsilon(\eta)$ is a stochastic error such that*

*a)* $\rho_K(y, y^\eta) \to 0$ *or*

*b)* $\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta)) = f(\eta) \to 0$

*as* $\eta \to 0$. *Then, assuming all necessary assumptions for the deterministic theory (except the bound on the noise) hold with probability one, the regularization method* $R_\alpha$ *fulfills*

$$\lim_{\eta \to 0} \rho_K^{\mathcal{L}}(R_\alpha(y^\eta)) = 0$$

*under the same parameter choice rule as in the deterministic setting with* $\delta$ *replaced by* $\rho_K(y, y^\eta)$ *(case a)) or* $\tau(\eta)\mathbb{E}(d_{\mathcal{Y}}(y, y^\eta))$ *where* $\tau(\eta)$ *fulfills* (19) *(case b)). In particular, the series of regularized solutions fulfills*

$$\lim_{\eta_1, \eta_2 \to 0} \rho_K^{\mathcal{L}}(R_\alpha(y^{\eta_1}), R_\alpha(y^{\eta_2})) = 0$$

*Proof.* The proof follows the lines of the one of Theorem 3.3 with $\rho_K(\cdot, x^\dagger)$ replaced by $\rho_K^{\mathcal{L}}(\cdot)$. Also Lemma 3.1 is easily adjusted to incorporate multiple solutions. □

So far we assumed that only the noise is stochastic whereas the operator $F$ and the unknown $x$ were assumed to be deterministic. In [14, 15] general stochastic Inverse Problems

$$F(x(\omega), \omega) = y(\omega)$$

were considered. It was shown how deterministic conditions such as source conditions can be incorporated into the stochastic setting by assuming that the deterministic conditions hold with a certain probability. However, additional conditions may occur when lifting these in order to ensure the deterministic requirements up to a certain probability. Since this is easier seen given an example, we move the discussion of the complete stochastic formulation in the next section. Although we will address only one particular example, the technique can be applied to general approaches.

## 3.2 Fully stochastic Inverse Problems

Due to the possible multiplicity of stochastic conditions which might appear in this context it seems not possible to develop a lifting strategy in such a general fashion as in the previous section. We will therefore consider two classical examples, namely nonlinear Tikhonov regularization and Landweber's method for nonlinear Inverse Problems. The theory is taken completely from [14, 15].

### 3.2.1 Nonlinear Tikhonov Regularization

We seek the solution of a nonlinear ill-posed problem (2) via the variational problem

$$x_\alpha^\delta = \text{argmin}||F(x) - y^\delta||^2 + \alpha||x - x^*||^2$$

with a reference point $x^* \in X$ and given noisy data $y^\eta$ according to (1) where the stochastic distribution of the noise is assumed to be known. We shall skip the general convergence theorem (which follows as in the previous section) and move to convergence rates directly. In the deterministic theory, i.e. when $y^\delta$ is the noisy data with $||y - y^\delta|| \leq \delta$, we have the following theorem from [2].

**Theorem 3.8.** *Let $\mathcal{D}(\mathcal{F})$ be convex, $y^\delta \in \mathcal{Y}$ such that $||y - y^\delta|| \leq \delta$ and $x^\dagger$ denote the $x^*$-minimum norm solution of (2). Furthermore let the following conditions hold.*

*a) $F$ is Fréchet-differentiable*

*b) There exists $\gamma \geq 0$ such that $||F'(x^\dagger) - F'(x)|| \leq \gamma ||x^\dagger - x||$ in a sufficiently large ball $\mathcal{B}_\theta(x^\dagger) \cap \mathcal{D}(F)$*

*c) $x^\dagger - x^*$ satisfies the source condition $x^\dagger - x^* = F'(x^\dagger)^* v$ for some $v \in \mathcal{Y}$.*

*d) The source element satisfies $\gamma ||v|| < 1$.*

*Then for the choice $\alpha = c\delta$ with some fixed $c > 0$ we obtain*

$$||x^\dagger - x_\alpha^\delta|| \leq \frac{\delta + \alpha||v||}{\sqrt{\alpha}\sqrt{1 - \gamma||v||}} = \mathcal{O}(\sqrt{\delta}) \text{ and } ||F(x_\alpha^\delta) - y^\delta)|| = \mathcal{O}(\delta). \tag{25}$$

As given in Theorem 4.6 of [14], the following stochastic formulation of Theorem 3.8 holds.

**Theorem 3.9.** *Let $\mathcal{D}(\mathcal{F})$ be convex, let $y^\eta$ be such that $0 \leq \rho_K(y, y^\eta) < \infty$ and $x^\dagger$ denote the $x^*$-minimum norm solution of (2) for almost all $\omega$. Furthermore let the following conditions hold.*

*a) $F(., \omega)$ is Frechet-differentiable for almost all $\omega$*

*b) $F'(\cdot, \omega)$ satisfies*

$$||F'(x^\dagger(\omega), \omega) - F'(x, \omega)|| \leq \gamma(\omega)||x^\dagger(\omega) - x||$$

*in a sufficiently large ball $\mathcal{B}_\theta(x^\dagger(\omega)) \cap \mathcal{D}(F)$*

*c) (smoothness) $\mathbb{P}(\Omega_{sc}) = 1$ where*

$$\Omega_{sc} := \{\omega : \exists v(\omega), x^\dagger(\omega) - x^*(\omega) = F'(x^\dagger(\omega), \omega)^* v(\omega)\}.$$

*d) (closedness) $\mathbb{P}(\omega \in \Omega sc : \gamma(\omega)||v(\omega)|| > \xi) < \phi_{cl}(\xi), \quad \lim_{\xi \to 1-} \phi_{cl}(\xi) = 0$*

*e) (decay) $\mathbb{P}(\omega \in \Omega sc : ||v(\omega)|| > \tau) < \varphi_{de}(\tau), \quad \lim_{\tau \to \infty} \varphi_{de}(\tau) = 0.$*

*Then for the choice $\alpha \sim \rho_K(y, y^\eta)$ we obtain*

$$\rho_K(x^\dagger, x_\alpha^\eta) \leq \inf_{\substack{\tau < \infty \\ \xi \in (0,1)}} \max \left\{ \rho_K(y, y^\eta) + \varphi_{cl}(\xi) + \varphi_{de}(\tau), \sqrt{\rho_K(y, y^\eta)} \frac{\mathcal{O}(1 + \tau)}{\sqrt{1 - \xi}} \right\}. \tag{26}$$

*Proof.* We have $||y - y^\eta|| \leq \rho_K(y, y^\eta)$ with probability $1 - \rho_k(y, y^\eta)$. Now fix $\xi < 1$ and $0 < \tau < \infty$. Then with probability $1 - (\varphi_{cl}(\xi) + \varphi_{de}(\tau))$ conditions d) and e) are fulfilled. Thus for the corresponding values of $\omega$ we can apply Theorem 3.8 and obtain

$$||x^\dagger(\omega) - x_\alpha^\eta(\omega)|| \leq \frac{\rho_K(y, y^\eta) + \alpha\tau}{\sqrt{\alpha}\sqrt{1-\xi}}$$

or, fixing the parameter $\alpha \sim \rho_K(y, y^\eta)$,

$$||x^\dagger(\omega) - x_\alpha^\eta(\omega)|| \leq \sqrt{\rho_K(y, y^\eta)}\frac{\mathcal{O}(1+\tau)}{\sqrt{1-\xi}}.$$

This estimate holds on a set with probability greater or equal $1 - (\rho_K(y, y^\eta) + \varphi_{cl}(\xi) + \varphi_{de}(\tau))$. The Ky Fan distance can therefore be bounded as

$$\rho_K(x^\dagger, x_\alpha^\eta) \leq \max\left\{\rho_K(y, y^\eta) + \varphi_{cl}(\xi) + \varphi_{de}(\tau), \sqrt{\rho_K(y, y^\eta)}\frac{\mathcal{O}(1+\tau)}{\sqrt{1-\xi}}\right\}.$$

This estimate is valid for arbitrary choices of $\xi$ and $\tau$ above, therefore we may bound the Ky fan distance of $x^\dagger$ and $x_\alpha^\eta$ by taking the infimum with respect to $\xi$ and $\tau$. □

The core principle of the lifting strategy is to ensure that there exists a subset $\tilde{\Omega} \subset (\Omega)$ such that all deterministic assumptions hold with probability one on $\tilde{\Omega}$. This may lead to the introduction of new conditions such as the decay condition in Theorem 3.9. Namely, since $\gamma(\omega)$ and $||v(\omega)||$ may vary with $\omega$, it may be possible that for a sequence $\{\omega_k\}_{k\in\mathbb{N}}$ $\gamma(\omega_k) \to 0$ and $||v(\omega_k)|| \to \infty$ such that still for all $k \in \mathbb{N}$ $\gamma(\omega(k))||v(\omega_k)|| < 1$. In this case the parameter $\tau$ cannot be treated as a constant in the convergence rate, but it influences it to a significant degree. The decay condition had to be imposed in order to control the growth of $\tau$. It is, however, possible to avoid condition e) by imposing other conditions. For example, one could require that $\gamma(\omega)$ is bounded below by some $0 < c < 1$. In this case condition d) implies e). A more detailed discussion is given in [14].

Accordingly, in order to lift other deterministic convergence rate results into the fully stochastic setting, a careful examination of the conditions necessary for convergence in the stochastic setting, understanding their cross-connections and dependencies is important. However, once the conditions have been translated to the stochastic setting, convergence rates follow immediately using the Ky Fan metric. We will close this example by showing how particular choices of the stochastic parameters in Theorem 3.9 influence the convergence rate. To this end, we cite Remark 4.8 of [14].

Let in the first examples the operator be deterministic, i.e., $F(\cdot, \omega) = F(\cdot)$ where $\gamma(\omega) = \gamma = 1$.

First consider the case that $||v|| \in U[0, 1]$, i.e., it is uniformly distributed on the interval $[0, 1]$. We therefore have $\varphi_{cl}(\xi) = 1 - \xi$, as well as $\varphi_{de} = 0$ for $\tau > 1$.

Thus Theorem 3.9 implies

$$\rho_K(x^\dagger, x_\alpha^\eta) \leq \inf_{0<\alpha<\infty} \inf_{\xi\in(0,1)} \max\left\{\rho_K(y,y^\eta)+1-\xi, \sqrt{\rho_K(y,y^\eta)}\frac{\rho_K(y,y^\eta)+\alpha}{\alpha\sqrt{1-\xi}}\right\}$$

which gives for $\alpha \sim \rho_K(y, y^\eta)$ the optimal rate

$$\rho_K(x^\dagger, x_\alpha^\eta) = \mathcal{O}(\rho_K(y,y^\eta)^{1/3}).$$

For the second case suppose that $\varphi_{de}(\tau) = c\tau^{-e}$ for some exponent $e > 0$. Since now we do not have $\varphi_{cl}(\xi) \to 0$, but $\varphi_{cl} \geq c > 0$ we obtain

$$\rho_K(x^\dagger, x_\alpha^\eta) \leq \inf_{0<\alpha<\infty} \inf_{\substack{t<\infty \\ \xi\in(0,1)}} \max\left\{c+c\tau^{-e}, \sqrt{\rho_K(y,y^\eta)}\frac{\rho_K(y,y^\eta)+\alpha}{\alpha\sqrt{1-\xi}}\right\}.$$

Since the right hand side does not converge to zero we do not obtain a convergence rate anymore. However, convergence itself still follows from Theorem 3.3.

Finally, consider the case when both d) and e) from Theorem 3.9 influence the convergence behavior, because $F$ is stochastic with varying $\gamma(\omega)$. For instance in the case the for some $\omega \in U[0,1]$ we have $x^\dagger(\omega) = \omega x^\dagger$ and $\gamma(\omega) = 1-\omega$, we find that $\varphi_{cl}(\xi) = 1-\xi$ and $\varphi_{de}(\tau) = c/(1+\tau)$ are compatible realizations of $\varphi_{cl}(\cdot)$ and $\varphi_{de}(\cdot)$. With this one can show

$$\rho_K(x^\dagger, x_\alpha^\eta) = \mathcal{O}(\rho_K(y,y^\eta)^{1/4})$$

under the parameter choice $\alpha \sim \rho_K(y,y^\eta)^{5/4}$. From the given examples it is evident that the convergence speed is heavily influenced by the conditions d) and e) in Theorem 3.9. Therefore, although the general formula for the convergence rate (26) may suggest that the convergence rate is close to the deterministic one, it may be significantly slower due to the additional stochastic properties.

### 3.2.2 Nonlinear Landweber iteration

As before we seek the solution of a nonlinear ill-posed problem (2) given noisy data $y^\eta$ according to (1) where the stochastic distribution of the noise is assumed to be known. Landweber's method can be seen as a descent algorithm for $||F(x)-y^\delta||^2$ and is defined via the iteration

$$x_{k+1}^\delta = x_k^\delta - \gamma F'(x_k^\delta)(F(x_k^\delta)-y^\delta), \quad k=1,2,\ldots, \tag{27}$$

where $\gamma > 0$ is an appropriately chosen stepsize and $x_0^\delta$ an initial guess. Landweber's method constitutes a regularization method if it is stopped early enough [2]. In the deterministic theory, i.e. when $y^\delta$ is the noisy data with $||y-y^\delta|| \leq \delta$, we have the following theorem from [2] for convergence rates of the Landweber method.

**Theorem 3.10.** *Let $\mathcal{D}(\mathcal{F})$ be convex, $y^\delta \in \mathcal{Y}$ such that $||y-y^\delta|| \leq \delta$ and $x^\dagger$ denote the $x^*$-minimum norm solution of (2). Assume (2) has a solution in $\mathcal{B}_\vartheta(x^*)$. Furthermore let the following conditions hold on $\mathcal{B}_{2\vartheta}(x^*)$.*

*a) $F$ is Frechet-differentiable with $||F'(x)|| \leq 1$ and*

$$||F(x) - F(x^\dagger) - F'(x^\dagger)(x - x^\dagger)|| \leq \zeta ||F(x) - F(x^\dagger)||, \quad 0 < \zeta < \frac{1}{2}$$

*b) $F'(x) = R_x F'(x^\dagger)$ where the bounded linear operators $R$ satisfy $||R_x - I|| \leq C||x - x^\dagger||$*

*c) $x^\dagger - x^*$ satisfies the source condition $x^\dagger - x^* = (F'(x^\dagger)^* F'(x^\dagger))^\nu v$ for some $v \in \mathcal{Y}$ and $0 < \nu \leq \frac{1}{2}$.*

*Let $||v||$ be sufficiently small. Then, if the regularization parameter is stopped according to the discrepancy principle, i.e., at the unique index $k_*$ for which for the first time*

$$||F(x_k) - y^\delta|| \leq \hat{\tau}\delta$$

*with $\hat{\tau} > 2\frac{1+\zeta}{1-2\zeta} > 2$, we obtain*

$$||x^\dagger - x_{k_*}^\delta|| \leq c||v||^{1/(2\nu+1)} \delta^{2\nu/(2\nu+1)}. \tag{28}$$

We can obtain a stochastic version of Theorem 3.10 with the same arguments as Theorem 3.9 followed from Theorem 3.9.

**Theorem 3.11.** *Let $\mathcal{D}(\mathcal{F})$ be convex, $y^\delta \in \mathcal{Y}$ with known value that $\rho_K(y, y^\eta)$ and $x^\dagger(\omega)$ denote the $x^*$-minimum norm solution of (2). Assume (2) has a solution in $\mathcal{B}_\vartheta(x^*(\omega))$ for almost all $\omega$. Furthermore let the following conditions hold on $\mathcal{B}_{2\vartheta}(x^*)$.*

*a) $F'(x, \omega) = R_{x,\omega} F'(x^\dagger(\omega), \omega)$ where for almost all $\omega$ the set $\{R_{x,\omega} : x \in \mathcal{B}_\vartheta(x^*)\}$ describes a family of bounded linear operators with*

$$||R_{x,\omega} - I|| \leq C(\omega)||x - x^\dagger(\omega)||$$

*b) $x^\dagger - x^*$ satisfies the source condition*

$$x^\dagger(\omega) - x^*(\omega) = (F'(x^\dagger(\omega), \omega) * F'(x^\dagger(\omega), \omega))^\nu v(\omega)$$

*for some $v(\omega) \in \mathcal{Y}$ and $0 < \nu \leq \frac{1}{2}$.*

*c) $\mathbb{P}(\omega \in \Omega : C(\omega)||v(\omega)|| > c\} < \varphi_{cl}(c)$*

*d) $\mathbb{P}(\omega \in \Omega : ||v(\omega)|| > \tau) < \varphi(\tau)$*

*Then, if the regularization parameter is stopped according to the discrepancy principle, i.e., at the unique index $k_*$ for which for the first time*

$$||F(x_k) - y^\eta|| \leq \hat{\tau}\rho_K(y, y^\eta)$$

*with $\hat{\tau} > 2$, we obtain for $c_0 > 0$ sufficiently small the rate*

$$\rho_K(x^\dagger - x_{k_*}^\eta) \leq$$
$$\inf_{0<\tau\leq\infty} \max\left\{\rho_K(y, y^\eta) + \varphi_{cl}(c_0) + \varphi_{de}(\tau), \tilde{c}\tau^{1/(2\nu+1)} \rho_K(y, y^\eta)^{2\nu/(2\nu+1)}\right\}$$

*where the constant $\tilde{c}$ depends on $\nu$ only.*

In the fully stochastic setting, the source condition b) from Theorem 3.11 need not hold with constant exponent $\nu$ for all $\omega \in \Omega$. There are at least two situations which lead to the power $\nu$ being a stochastic quantity as well, i.e., it holds

$$x^\dagger(\omega) = (F'(x^\dagger(\omega), \omega)^* F'(x^\dagger(\omega), \omega))^{\nu(\omega)} v(\omega) \tag{29}$$

with $0 < \nu(\omega) \leq \frac{1}{2}$.

In the first case all solutions $x^\dagger(\omega)$ come from some initial element $v(\omega) = v \in \mathcal{Y}$, with small $\mathcal{Y}$-norm. Some randomly smoothing operator is acting on this element and generates $x^\dagger(\omega)$. (One could for instance think of some kind of evolution process, e.g., a diffusion process that is applied to some initial value $v$). The smoothness of $x^\dagger(\omega)$ is therefore random.

Secondly, $x^\dagger$ may be a deterministic element satisfying a certain smoothness condition. The data $y(\omega)$ is generated by applying a forward operator $F(\cdot, \omega)$ with random smoothness properties. If the realization of $F(\cdot, \omega)$ is smoothing strongly, this corresponds to a source condition with small $\nu(\omega)$, if $F(\cdot, \omega)$ is smoothing weakly we have the source condition with larger $\nu(\omega)$.

The following proposition shows the convergence rate that results from the source condition (29) for the case that $\nu(\omega)$ is uniformly distributed on the interval $[0, \frac{1}{2}]$.

**Theorem 3.12.** *Let all conditions of Theorem 3.11 hold except for b) and d). Let $x^\dagger(\omega)$ satisfy (29) where $||v(\omega)||$ is uniformly bounded and sufficiently small. Let*

$$\mathbb{P}\left(\omega \in \Omega : 0 \leq \nu(\omega) < \nu \leq \frac{1}{2}\right) = 2\nu.$$

*Then the approximations $x_{k_*}^\eta$ obtained by Landweber's method satisfy the convergence rate*

$$\rho_K(x^\dagger, x_{k_*}^\eta) = \mathcal{O}\left(\frac{W(-\log(\rho_K(y, y^\delta))}{-\log(\rho_K(y, y^\delta)}\right) \tag{30}$$

*where $W$ denotes the Lambert W-function, defined by $W(z)e^{W(z)} = z$, see [27].*

*Proof.* As can be seen from the proof of Theorem 3.1 in [28], the requirement "$||v||$ sufficiently small", becomes stronger, the larger $\nu$ is. Supposing that $||v||$ in (29) is sufficiently small for the case $\nu = \frac{1}{2}$, implies therefore that also the convergence conditions for $\nu \leq \frac{1}{2}$ are satisfied.

Secondly we observe that the convergence rate in Theorem 3.11 contains a constant $\tilde{c}$ that depends on $\nu$. Although it is difficult to state an explicit formula for $\tilde{c}$, investigation of [28] shows, that $\tilde{c}(\nu)$ attains its maximum value when $\nu = \frac{1}{2}$.

After these observations we start with the actual derivation of the convergence rate. For the sake of simplicity we assume that all appearing constants are just equal to 1. Furthermore we may assume that $\varphi_{cl}(\cdot)$ and $\varphi_{de}(\cdot)$ both vanish. Asymptotically, for given $\omega$ we therefore have the estimate

$$||x^\dagger(\omega) - x_{k_*}^\eta(\omega)|| \leq \rho_K(y, y^\eta).$$

Measuring the distance in the Ky Fan metric we must, since we assumed that $\nu(\omega)$ is as in (29), solve the equation

$$\rho_K(y, y^\delta)^{\frac{2\nu}{2\nu+1}} = 2\nu \tag{31}$$

for $\nu$. We first consider the simplified equation

$$\rho_K(y, y^\delta)^{2\tilde{\nu}} = 2\tilde{\nu}$$

which is solved by

$$\tilde{\nu}(\rho_K(y, y^\delta)) = \frac{W(-\log \rho_K(y, y^\delta))}{-2 \log \rho_K(y, y^\delta)}.$$

In the following we show that the above approximate solution is sufficiently accurate. Therefore we construct a better estimate via the ansatz $\nu(\rho_K(y, y^\delta)) = \tilde{\nu}(\rho_K(y, y^\delta))(1 + \varepsilon(\tilde{\nu}(\rho_K(y, y^\delta))))$. The original equation then contains the term $2\tilde{\nu} + 3\tilde{\nu}\varepsilon + 1$. Neglecting the quadratic part, we can replace this term with $2\tilde{\nu} + 1$, and obtain an equation that MATHEMATICA can solve for $\varepsilon(\tilde{\nu})$. The solution for the correction term is given as

$$\varepsilon(\tilde{\nu}) = \frac{\log(\tilde{\nu}) + (2\tilde{\nu} + 1)W\left(-\frac{\log(\tilde{\nu})}{2\tilde{\nu}^2 + \tilde{\nu}}\right)}{-\log(\tilde{\nu})}$$

and tends to zero approximately linearly in $\tilde{\nu}$. Thus this correction becomes small rather quickly, and we can consider the asymptotic bound in (30) as sufficiently accurate due to the asymptotics of the Lambert W-function. □

# 4 Examples

## 4.1 Filter-based regularization methods

Let $A$ be a linear compact operator between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ with singular system $\{\sigma_n, u_n, v_n\}_{n\in\mathbb{N}}$, see e.g. [2]. Then, for $y \in \mathcal{D}(A)$, the generalized inverse $A^\dagger$ to $A$ is given by

$$A^\dagger y = \sum_{\sigma_n > 0} \sigma_n^{-1} \langle y, u_n \rangle v_n. \tag{32}$$

Since for compact operators the singular values approach zero, their inverse blows up and the generalized inverse yields a meaningless solution to (2) for noisy data. A popular class of regularization methods is based on the filtering of the generalized inverse. Introducing an appropriate filter function $F_\alpha(\sigma)$ depending on the regularization parameter $\alpha$ that controls the growth of $\sigma^{-1}$, the regularized solutions are defined by

$$R_\alpha(y) = \sum_{\sigma_n > 0} F_\alpha(\sigma)\sigma_n^{-1} \langle y, u_n \rangle v_n. \tag{33}$$

Examples for filter based methods are for example the classical Tikhonov regularization, truncated singular value decomposition or Landwebers method [1, 2]. The regularization properties are fully determined by the filter functions. In the deterministic setting, the conditions can be found in, e.g.,[1, Theorem 3.3.3.]. Convergence rates can be obtained for a priori and a posteriori parameter choice rules under stricter conditions on the filter functions. We will only comment on an a priori choice here in order to keep it short. An example of the discrepancy principle as a posteriori parameter choice is given in the next section in a different context. Using the smoothness condition

$$x^\dagger \in \mathcal{R}((A^*A)^{\nu/2}) \quad \text{with} \quad \|x^\dagger\|_\nu := \{||z||_{\mathcal{X}} : x^\dagger = (A^*A)^{\nu/2} z, z \in \mathcal{N}(A)^\perp\} \leq \varrho \tag{34}$$

the following theorem can be obtained.

**Theorem 4.1.** *[1, Theorem 3.4.3] Let $y \in \mathcal{R}(A)$ and $\left\|y - y^\delta\right\|_{\mathcal{Y}} \leq \delta$. Assume that it holds $||x^\dagger||_\nu \leq \varrho$ and for $0 \leq \nu \leq \nu^*$,*

$$\sup_{0<\sigma\leq\sigma_1} \sigma^{-1}|F_\alpha(\sigma)| \leq c\alpha^{-\beta} \tag{35}$$

$$\sup_{0<\sigma\leq\sigma_1} |1 - F_\alpha(\sigma)|\sigma^{\nu^*} \leq c_{\nu^*}\alpha^{\beta\nu^*}, \tag{36}$$

*where $\beta > 0$ and $c, c_{\nu^*}$ are constants independent of $\delta$. Then with the a priori parameter choice*

$$\alpha = C\left(\frac{\delta}{\varrho}\right)^{1/\beta(\nu+1)}, \qquad C > 0 \quad \textit{fixed}, \tag{37}$$

*the method induced by the filter $F_\alpha$ is order optimal for all $0 \leq \nu \leq \nu^*$, i.e.,*

$$||x^\dagger - R_\alpha y^\delta|| \leq c\delta^{\frac{\nu}{\nu+1}}\varrho^{\frac{1}{\nu+1}}$$

*for some constant $c$ independent of $\delta$ and $\varrho$.*

Now we use Theorem 3.5 and obtain convergence rates in the Ky Fan metric.

**Theorem 4.2.** *Let $y \in \mathcal{R}(A)$ and $\rho_K(y, y^\eta)$ be known. Assume that it holds $||x^\dagger||_\nu \leq \varrho$ and for $0 \leq \nu \leq \nu^*$, (35) and (36) hold. Then with the a priori parameter choice*

$$\alpha = C\left(\frac{\rho_K(y, y^\eta)}{\varrho}\right)^{1/\beta(\nu+1)}, \qquad C > 0 \quad \textit{fixed}, \tag{38}$$

*the method induced by the filter $F_\alpha$ fulfills*

$$||x^\dagger - R_\alpha y^\eta|| \leq c\rho_K(y, y^\eta)^{\frac{\nu}{\nu+1}}\varrho^{\frac{1}{\nu+1}}$$

*for some constant $c$ independent of $\delta$ and $\varrho$.*

More about filter methods in the stochastic setting including numerical examples can be found in [23].

## 4.2 Sparsity-regularization for an autoconvolution problem

We consider an autoconvolution equation

$$[F(x)](s) = \int_0^s x(s-t)x(t)\,dt, \qquad 0 \le s \le 1 \tag{39}$$

between Hilbert spaces $\mathcal{X} = L_2[0,1]$ and $\mathcal{Y} = L_2[0,1]$ where $x \in \mathcal{D}(F)$. Such an equation is of great interest in, for example, stochastics or spectroscopy and has been analyzed in detail in [29]. Recently, a more complicated autoconvolution problem has emerged from a novel method to characterize ultra-short laser pulses [30, 31]. Here, we want to show the transition from the deterministic setting to the stochastic setting in a numerical example. We base our results on the deterministic paper [32].

Using the Haar-wavelet basis, the authors of [32] reformulate (39) as an equation from $\ell_2$ to $\ell_2$ by switching to the space of coefficients in the Haar basis. In order to stabilize the inversion, an $\ell_1$ penalty term is used such that the task is to minimize the functional

$$J_\alpha(x) = ||F(x) - y^\delta||_2^2 + \alpha||x||_1. \tag{40}$$

The regularization parameter $\alpha$ in (40) is chosen according to the discrepancy principle. In [32], the following formulation is used: For $1 < \tau_1 \le \tau_2$ choose $\alpha = \alpha(\delta, y^\delta)$ such that

$$\tau_1\delta \le ||F(x_\alpha^\delta) - y^\delta||_2 \le \tau_2\delta \tag{41}$$

holds. The authors show that this leads to a convergence of the regularized solutions against a solution of (39) with minimal $\ell_1$-norm of its coefficients. It was also shown that the regularization parameter fulfills

$$\alpha(\delta, y^\delta) \to 0, \qquad \frac{\delta^2}{\alpha(\delta, y^\delta)} \to 0 \quad \text{as} \quad \delta \to 0. \tag{42}$$

By courtesy of Stephan Anzengruber we were allowed to use the original code for the numerical simulation in [32]. We only changed the parts directly connected to the data noise. Namely, we replaced the deterministic error $||y - y^\delta||_2 \le \delta$ with i.i.d Gaussian noise,

$$y^\eta = y + \epsilon,$$

$\epsilon \sim \mathcal{N}(0, \eta^2 I)$. The discretization is due to the truncation of the expansion of the functions in the Haar-basis after $m$ elements. The parameter choice (41) was realized with $\delta$ replaced by $\tau(\eta)\mathbb{E}(||\epsilon||_2)$ in accord with Theorem 3.3. Instead of the correct expectation

$$\mathbb{E}(||\epsilon||_2) = \frac{\eta}{\sqrt{2}} \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})},$$

see [23], we used the upper bound

$$\mathbb{E}(||\epsilon||_2) \leq \eta\sqrt{m}$$

since, as shown in this chapter, the expectation has to be "blown up" anyway. In a first experiment we let $\tau(\epsilon) = 1.3 = \text{const}$. In this case, the numerical results suggest that the regularization parameter decreases too fast, i.e., $\frac{(\tau(\eta)\mathbb{E}(||\epsilon||_2))^2}{\alpha}$ does not converge to zero as in (42), see Figure 1. For comparison, in a second run we chose $\tau(\eta) = \sqrt{1 - \log(\eta^2 2\pi m^2 (\frac{e}{2})^m)}$ where $m$ is the amount of data points. This way, $\tau(\eta)\mathbb{E}(||\epsilon||_2) \propto \rho_K(y, y^\eta)$. Now $\frac{(\tau(\eta)\mathbb{E}(||\epsilon||_2))^2}{\alpha}$ converges to zero as it should be the case from 42, see Figure 1.

At this point we would like to mention that the discrepancy principle in the stochastic and deterministic setting are not completely equivalent since a different way of measuring the noise is used. Typically the stochastic noise level will be smaller (it need to bound 100% of the possible realizations) and the iteration will be stopped later than in the deterministic setup.

## 4.3 Linear Inverse Problems with Besov-space prior

In [33] the lifting strategy was used in a slightly different way. In particular, the Ky Fan metric was used to obtain a novel parameter choice rule. The convergence rates obtained there, however, can also be viewed in the framework of this work. The scope of that paper was to transfer the deterministic convergence results from [34] into the stochastic setting. The seminal paper [34] initiated the investigation of sparsity-promoting regularization for Inverse Problems. Looking for the solution of the linear ill-posed problem

$$Ax = y \tag{43}$$

between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ with given noisy data $y^\delta = y + \epsilon$, the regularization strategy was to obtain an approximation $x_\alpha^\delta$ to $x^\dagger$ via

$$x_\alpha^\delta = \min_x ||Ax - y^\delta||_2^2 + \sum_{\lambda \in \Lambda} w_\lambda |\langle x, \psi_\lambda \rangle|^p \psi_\lambda, \tag{44}$$

where $\Lambda$ is an appropriate index set, $w_\lambda > 0 \ \forall \lambda \in \Lambda$, $\{\psi_\lambda\}_{\lambda \in \Lambda}$ a dictionary (typically an orthonormal basis or frame) in $\mathcal{X}$ and $1 \leq p \leq 2$. Choosing a sufficiently smooth wavelet basis for $\{\psi_\lambda\}_{\lambda \in \Lambda}$ and setting $w_\lambda = 2^{\zeta|\lambda|p}$ with $\zeta = s - d(\frac{1}{2} - \frac{1}{p}) > 0$, the penalty term in (44) corresponds to a norm in the Besov space $B^s_{p,p}(\mathbb{R}^d)$. Formulating the problem of determining $x$ from noisy data $y^\eta = y + \epsilon$, $\epsilon \sim \mathcal{N}(0, \eta^2 I_m)$, in the Bayesian setting with the distributions $\pi_\epsilon(y^\delta|x) \propto \exp(-\frac{1}{2\eta^2}||Ax - y^\delta||_2^2)$ and $\pi_{pr}(x) \propto \exp(-\frac{\tilde{\alpha}}{2}||x||^p_{B^s_{p,p}(\mathbb{R}^d)})$ and using the maximum a-posteriori solution lead to the formulation

$$x^{\text{MAP}} = \min_x ||Ax - y^\delta||_2^2 + \tilde{\alpha}\eta^2 ||x||^p_{B^s_{p,p}(\mathbb{R}^d)} \tag{45}$$

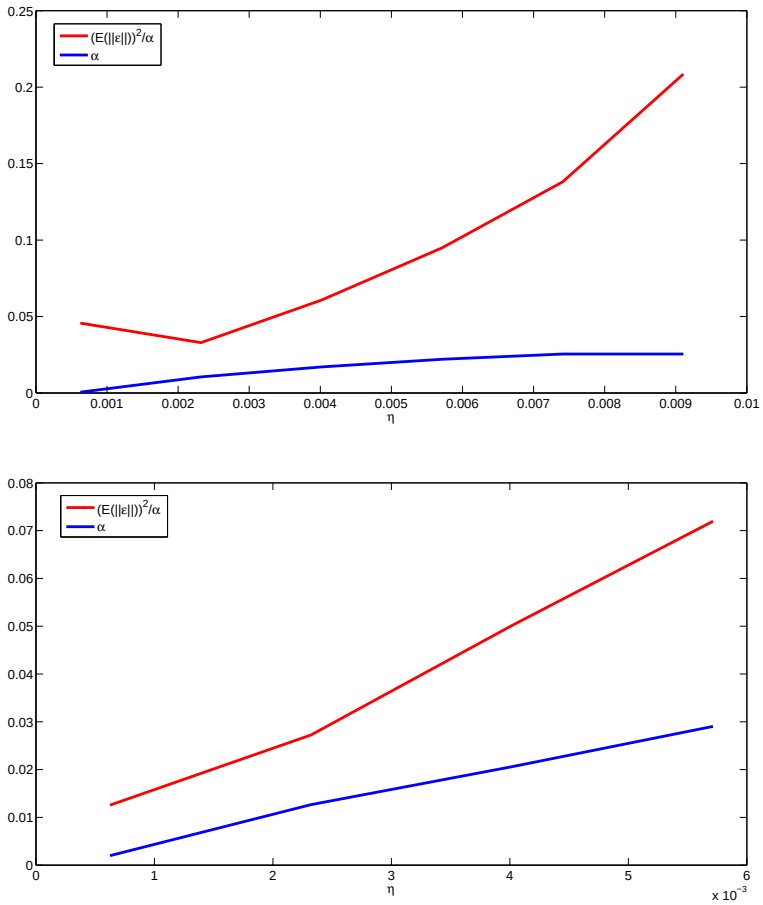


Figure 1: Top: constant $\tau$ in the discrepancy principle with the expectation of the noise leads to the regularization parameter decreasing too fast. Bottom: increasing $\tau$ with decreasing variance appropriately resolves this issue.

where $\eta$ is the variance of the noise and $\tilde{\alpha}$ can roughly be described as the inverse variance of the prior. The product $\alpha := \tilde{\alpha}\eta^2$ gives the actual regularization parameter. In direct application of Theorem 3.3, the deterministic condition

$$\alpha \to 0, \quad \frac{\delta^2}{\alpha} \to 0 \text{ as } \delta \to 0,$$

with $\delta$ replaced by $\rho_K(y, y^\eta)$ from (21) translates to the conditions

$$\tilde{\alpha}\eta^2 \to 0, \quad \frac{\log(\eta)}{\tilde{\alpha}} \to 0 \text{ as } \eta \to 0,$$

leading to convergence of $x^{\text{MAP}}$ to the unique (in case $p = 1$ the operator is assumed to be injective) solution $x^\dagger$ of minimal norm $||\cdot||_{B^s_{p,p}(\mathbb{R}^d)}$ in the Ky Fan metric. The proof of convergence rates is based on two assumption:

$$C_l \sum_{\lambda\in\Lambda} 2^{-2|\lambda|\beta}|\langle x, \psi_\lambda\rangle|^p \leq ||Ax|| \leq C_u \sum_{\lambda\in\Lambda} 2^{-2|\lambda|\beta}|\langle x, \psi_\lambda\rangle|^p$$

where $\beta, C_l, C_u > 0$ and

$$||x^\dagger||_{B^s_{p,p}(\mathbb{R}^d)} \leq \rho$$

for some $\rho > 0$. Combining Proposition 4.5, Proposition 4.6, Proposition 4.7 from [34] it is

$$||x^\delta_\alpha - x^\dagger|| \leq C\left(\delta + \sqrt{\delta^2 + \alpha\rho^p}\right)^{\frac{\zeta}{\zeta+\beta}} \left(\rho + \left(\rho^p + \frac{\delta^2}{\alpha}\right)^{1/p}\right)^{\frac{\beta}{\zeta+\beta}}. \tag{46}$$

Translated into the stochastic setting, the right hand side of (46) reads

$$C\mathcal{E}(\eta, m, \tilde{\alpha})^{\frac{\zeta}{\zeta+\beta}} \tilde{\rho}^{\frac{\beta}{\zeta+\beta}} \tag{47}$$

where with $L_m(\eta) = \min\{0, \eta^2 2\pi m^2 (\frac{e}{2})^m\}$,

$$\mathcal{E}(\eta, m, \tilde{\alpha}) := \eta\left(\sqrt{m - L_m(\eta)} + \sqrt{m - L_m(\eta) + \frac{\tilde{\alpha}\rho^p}{2}}\right)$$

and

$$\tilde{\rho} = \rho + \left(\rho^p + \frac{2m - L_m(\eta)}{\tilde{\alpha}}\right)^{1/p}.$$

We know that the deterministic rate is an upper bound to the reconstruction error whenever $||y - y^\eta|| = ||\epsilon|| \leq \rho_K(y, y^\eta)$ and $||x^\dagger||_{B^s_{p,p}(\mathbb{R}^d)} \leq \rho$. Hence, it is

$$\mathbb{P}\left(||x^{\text{MAP}} - x^\dagger|| \geq C\mathcal{E}(\eta, m, \tilde{\alpha})^{\frac{\zeta}{\zeta+\beta}} \tilde{\rho}^{\frac{\beta}{\zeta+\beta}}\right) \leq \frac{\Gamma(\frac{m}{2}, m - L_m(\eta))}{\Gamma(\frac{m}{2})} + \frac{\Gamma(\frac{n}{p}, \frac{\tilde{\alpha}\rho^p}{2})}{\Gamma(\frac{n}{p})} \tag{48}$$

where

$$\mathbb{P}(||y - y^\eta|| > \rho_K(y, y^\eta)) = \frac{\Gamma(\frac{m}{2}, m - L_m(\eta))}{\Gamma(\frac{m}{2})}.$$

and

$$\mathbb{P}(||x^\dagger||_{B^s_{p,p}(\mathbb{R}^d)} \geq \rho) = \frac{\Gamma(\frac{n}{p}, \frac{\tilde{\alpha}\varrho^p}{2})}{\Gamma(\frac{n}{p})}$$

where the Besov-space functions were truncated after the first $n$ basis functions. By Definition of the Ky Fan metric, it follows immediately from (48) that

$$\rho_K(x^{\mathrm{MAP}}) = \max\left\{C\mathcal{E}(\eta, m, \tilde{\alpha})^{\frac{\varsigma}{\varsigma+\beta}}\tilde{\rho}^{\frac{\beta}{\varsigma+\beta}}, \frac{\Gamma(\frac{m}{2}, m - L_m(\eta))}{\Gamma(\frac{m}{2})} + \frac{\Gamma(\frac{n}{p}, \frac{\tilde{\alpha}\varrho^p}{2})}{\Gamma(\frac{n}{p})}\right\}. \tag{49}$$

Since $\tilde{\alpha}$ is a free parameter, we can balance the terms in (49), i.e. solve the nonlinear equation

$$C\mathcal{E}(\eta, m, \tilde{\alpha})^{\frac{\varsigma}{\varsigma+\beta}}\tilde{\rho}^{\frac{\beta}{\varsigma+\beta}} = \frac{\Gamma(\frac{m}{2}, m - L_m(\eta))}{\Gamma(\frac{m}{2})} + \frac{\Gamma(\frac{n}{p}, \frac{\tilde{\alpha}\varrho^p}{2})}{\Gamma(\frac{n}{p})}$$

for $\tilde{\alpha}$. With this parameter choice rule one obtains by construction

$$\rho_K(x^{\mathrm{MAP}}) = \mathcal{O}(\mathcal{E}(\eta, m, \tilde{\alpha})^{\frac{\varsigma}{\varsigma+\beta}}\tilde{\rho}^{\frac{\beta}{\varsigma+\beta}}). \tag{50}$$

In the deterministic setting [34] it was proposed to chose the regularization parameter $\alpha = \delta^2/\varrho^p$. Combining [34, Proposition 4.5] and [34, Proposition 4.7] then yields the rate

$$||x^\delta_\alpha - x^\dagger|| \leq C\left(\frac{\delta}{C_l}\right)^{\frac{\varsigma}{\varsigma+\beta}}\varrho^{\frac{\beta}{\varsigma+\beta}}$$

with $C_l$ from (47) and some $C > 0$. Theorem 3.5 then yields in the stochastic setting the parameter choice $\alpha = \rho_K(y^\eta, y)^2/\varrho^p$ and

$$\rho_K(x^\eta_\alpha, x^\dagger) = \mathcal{O}\left(\left(\frac{\rho_K(y^\eta, y)}{C_l}\right)^{\frac{\varsigma}{\varsigma+\beta}}\varrho^{\frac{\beta}{\varsigma+\beta}}\right).$$

# Acknowledgements

D. Gerth was supported in part by the Austrian Science Fund (FWF): W1214-N15 and by the German Research Foundation (DFG) under grant HO 1454/8-2.

# References

[1] A. K. Louis. *Inverse und schlecht gestellte Probleme.* B. G. Teubner, Stuttgart, 1989.

[2] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[3] B. Hofmann. *Regularization for applied inverse and ill-posed problems*. Teubner, 1986.

[4] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[5] A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[6] K. Mosegaard and M. Sambridge. Monte carlo analysis of inverse problems. *Inverse Problems*, 18, 2002.

[7] A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[8] D. Calvetti and E. Somersalo. *An introduction to Bayesian scientific computing — Ten lectures on subjective computing*. Springer, 2007.

[9] T. Hohage and F. Werner. Convergence rates in expectation for tikhonov-type regularization of inverse problems with poisson data. *Inverse Problems*, 28:104004, 2012.

[10] Nicolai Bissantz, Thorsten Hohage, and Axel Munk. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20(6):1773–1789, 2004.

[11] G. Blanchard and P. Mathé. Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse Problems*, 28(11):115011, 2012.

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

[13] S. N. Evans and P. B. Stark. Inverse problems as statistics. *Inverse Problems*, 18, 2002.

[14] A. Hofinger. *Ill-posed problems: Extending the Deterministic Theory to a Stochastic Setup*. PhD thesis, Johannes Kepler University Linz, 2006.

[15] A. Hofinger. *Ill-posed problems: Extending the Deterministic Theory to a Stochastic Setup*. Trauner Verlag, 2006.

[16] M. Lassas, E. Saksman, and S. Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009.

[17] M. Lassas H. Kekkonen and S. Siltanen. Analysis of regularized inversion of data corrupted by white Gaussian noise. *Inverse Problems*, 30(4):045009, 18, 2014.

[18] M. Gardner. Mathematical games—white and brown music, fractal curves and one-over-f fluctuations. *Scientific American*, 238:16–32, 1978.

[19] S. Kogan. *Electronic Noise and Fluctuations in Solids*. Cambridge University Press, 1996.

[20] Ky Fan. Entfernung zweier zufälligen Grössen und die Konvergenz nach Wahrscheinlichkeit. *Math. Z.*, 49:681–683, 1944.

[21] R. M. Dudley. *Real analysis and probability*. The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1989.

[22] V. I. Bogachev. *Gaussian measures*, volume 62 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998.

[23] D. Gerth. *Problem-adapted Regularization for Inverse Problems in the Deterministic and Stochastic Setting*. dissertation, Johannes Kepler University Linz, Austria, 2015.

[24] A. Neubauer and H. K. Pikkarainen. Convergence results for the Bayesian inversion theory. *J. Inverse Ill-Posed Probl.*, 16(6):601–613, 2008.

[25] A. Hofinger and H. K. Pikkarainen. Convergence rate for the Bayesian approach to linear inverse problems. *Inverse Problems*, 23(6):2469–2484, 2007.

[26] D. Th. Egoroff. Sur les suites de fonctions mesurables. *C. R.*, 152:244–246, 1911.

[27] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert $W$ function. *Adv. Comput. Math.*, 5(4):329–359, 1996.

[28] A. Neubauer M. Hanke and O. Scherzer. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math*, 72:21–37, 1995.

[29] R. Gorenflo and B. Hofmann. On autoconvolution and regularization. *Inverse Problems*, 10(2):353–373, 1994.

[30] Daniel Gerth, Bernd Hofmann, Simon Birkholz, Sebastian Koke, and Günter Steinmeyer. Regularization of an autoconvolution problem in ultrashort laser pulse characterization. *Inverse Probl. Sci. Eng.*, 22(2):245–266, 2014.

[31] Simon Birkholz, Günter Steinmeyer, Sebastian Koke, Daniel Gerth, Steven Bürger, and Bernd Hofmann. Phase retrieval via regularization in self-diffraction-based spectral interferometry. *JOSA B*, 32(5):983–992, 2015.

[32] S. W. Anzengruber and R. Ramlau. Morozov's discrepancy principle for tikhonov-type functionals with nonlinear operators. *Inverse Problems*, 26:025001, 2010.

[33] Daniel Gerth and Ronny Ramlau. A stochastic convergence analysis for Tikhonov regularization with sparsity constraints. *Inverse Problems*, 30(5):055009, 24, 2014.

[34] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.