Regularising linear inverse problems under unknown non-Gaussian noise

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

Vorgelegt beim Fachbereich Informatik und Mathematik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Tim Nikolas Jahn
aus Bensheim

Frankfurt am Main (2020)
(D 30)

vom Fachbereich Informatik und Mathematik

der Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr.-Ing. Lars Hedrich

Gutachter: Prof. Dr. Bastian von Harrach
Prof. Dr. Bangti Jin
Prof. Dr. Bernd Hofmann

Datum der Disputation: 30.04.2021

An dieser Stelle möchte ich ganz herzlich allen danken, die zum Gelingen dieser Arbeit beigetragen haben. Ich danke insbesondere meinem Doktorvater Prof. Dr. Bastian von Harrach für die hervorragende Betreuung, sowie Prof. Dr. Bangti Jin, Prof. Dr. Roland Potthast und Dr. Peter Mathé für ihre Gastfreundschaft und Hilfe. Professor Jin und Prof. Dr. Bernd Hofmann danke ich weiter für ihre Bereitschaft, als Zweit- und Drittgutachter zur Verfügung zu stehen.

Schließlich gilt mein ganz besonderer Dank meiner Frau Sabrina und unseren Kindern, denen ich diese Arbeit widmen möchte, und meinen Eltern und Geschwistern.

Contents

Zusammenfassung **1**

Introduction **7**

1 The case with finite variance **15**

1.1 A priori regularisation . . . 16

1.2 The discrepancy principle . . . 19

1.2.1 A counter example for convergence . . . 20

1.2.2 Convergence in probability of the discrepancy principle . . . 21

1.3 Optimality . . . 23

1.4 Connection to heuristic regularisation . . . 26

1.4.1 Counter example for validity of the Mouckenhoupt condition . 27

1.5 Almost sure convergence . . . 28

1.6 Proofs . . . 29

1.6.1 Proofs of Theorem 1.2.2 and 1.2.4 . . . 29

1.6.1.1 Proof of Theorem 1.2.4 without emergency stop . . . 34

1.6.1.2 Proof of Theorem 1.2.2 without emergency stop . . . 38

1.6.1.3 Proofs for the emergency stop case . . . 39

1.6.2 Proof of Corollary 1.2.5 . . . 40

1.6.3 Proofs of Theorem 1.3.1, 1.3.2 and 1.3.3 . . . 42

1.7 Numerical demonstration . . . 49

1.7.1 Differentiation of binary option prices . . . 49

1.7.1.1 Rescaling of the measurements and the operator . . . 51

1.7.2 Inverse heat equation . . . 51

1.8 Concluding remarks . . . 54

2 The white noise case **55**

2.1 Approach with finite-dimensional residuum . . . 58

2.1.1 A priori regularisation with finite-dimensional residuum . . . 60

2.1.2 A posteriori regularisation with finite dimensional residuum . 61

2.2 Approach with infinite-dimensional residuum . . . 63

2.2.1 A priori regularisation for infinite-dimensional residuum . . . 66

2.2.2 A posteriori regularisation for infinite-dimensional residuum . 66

2.3 Proofs . . . 68

2.3.1 Proofs for finite-dimensional residuum . . . 68

2.3.1.1 Proof of Proposition 2.1.4 . . . 70

2.3.1.2 Proof of Proposition 2.1.5 70
2.3.1.3 Proof of Proposition 2.1.6 70
2.3.1.4 Proof of Theorem 2.1.8 72
2.3.1.5 Proof of Theorem 2.1.11 75
2.3.2 Proofs for infinite-dimensional residuum 83
2.3.2.1 Proof of Proposition 2.2.2 83
2.3.2.2 Proof of Proposition 2.2.3 84
2.3.2.3 Proof of Proposition 2.2.4 84
2.3.2.4 Proof of Proposition 2.2.5 84
2.3.2.5 Proof of Proposition 2.2.6 85
2.3.2.6 Proof of Theorem 2.2.7 86
2.3.2.7 Proof of Theorem 2.2.9 87
2.4 Numerical Demonstration 96
2.4.1 Convergence of finite-dimensional residuum approach 97
2.4.2 (Semi-)Convergence of infinite-dimensional residuum approach 97
2.4.3 Comparison of the both approaches 99
2.5 Concluding remarks 101

3 The discrepancy principle for stochastic gradient descent 103
3.1 Convergence and a finite termination property 104
3.2 Proofs 106
3.2.1 The proof of Theorem 3.1.2 106
3.2.2 The proof of Theorem 3.1.4 110
3.2.3 The proofs of Propositions 3.2.1 and 3.2.2 113
3.2.3.1 The proof of Proposition 3.2.1 113
3.2.3.2 Proof of Proposition 3.2.2 114
3.3 Stochastic error 117
3.3.1 The case with finite variance 118
3.3.2 The white noise case 120
3.4 Numerical experiments and discussions 121
3.4.1 Optimality 122
3.4.2 How influential is the variance? 124
3.4.3 Independent run 126
3.5 Concluding remarks 130

References 131

Lebenslauf 139

Zusammenfassung

Im Zentrum der naturwissenschaftlichen Forschung steht das Verständnis und die Modellierung realer Phänomene. Hierbei basieren die Modelle auf empirischen Beobachtungen sowie grundlegenden physikalischen Prinzipien und werden verwendet, um konkrete praktische Probleme zu lösen. Häufig verknüpfen sie hierbei eine unzugängliche Ursache mit einer messbaren Wirkung. Daraus ergeben sich zwei fundamentale, zueinander duale Probleme: einmal das direkte Problem, aus der Kenntnis der Ursache die Wirkung vorherzusagen, sowie das inverse Problem, die eine gemessene Wirkung bedingende Ursache zu bestimmen. Beispielsweise werden im Rahmen tomographischer Verfahren die physikalischen Eigenschaften von Röntgenstrahlen verwendet, um das Zusammenspiel zwischen der Dichteverteilung eines Körpers sowie dem Intensitätsabfall durch diesen Körper gesandter Strahlen zu beschreiben. Hierbei besteht das inverse Problem darin, Röntgenstrahlen aus verschiedenen Richtungen durch den Körper zu schicken und die entsprechenden Intensitätsabfälle zu messen, und schließlich daraus die unbekannte Dichteverteilung im Körper zu ermitteln. Inverse Probleme zeichnen sich dadurch aus, dass sie typischerweise schlecht gestellt sind, in erster Linie die Stabilität betreffend. Selbst kleinste Messfehler haben enorme Konsequenzen für die Rekonstruktion der Ursache.

Neben der Röntgentomographie lassen sich noch viele weitere inverse Probleme als abstrakte lineare Gleichung

$$Kx = y$$

in unendlichdimensionalen Hilberträumen formulieren, wobei K ein kompakter, linearer Operator ist, x die unbekannte Ursache und y die zu messende Wirkung. Aufgrund unvermeidbarer Mess- und Modellierungsfehler ist auch die Wirkung y unbekannt, stattdessen ist aus der (gestörten) Messung $y^\delta \approx y$ eine Approximation an x zu bestimmen. Der naive Lösungsansatz, der darin besteht eine Approximation x^δ so zu bestimmen, dass möglichst $Kx^\delta = y^\delta$ gilt, scheitert im Allgemeinen an der intrinsischen Schlechtgestelltheit inverser Probleme. Unkontrollierte Fehlerverstärkung macht die so gewonnenen Rekonstruktionen selbst bei kleinsten Messungenauigkeiten unbrauchbar.

Um dennoch eine robuste Rekonstruktion zu ermöglichen, muss das Problem regularisiert werden. Dabei ersetzt man es zunächst durch eine ganze Familie abgeänderter Probleme. Diese Familie soll zum einen beliebig genaue Approximationen des schlechtgestellten Problems enthalten, und zum anderen soll jede Approxima-

tion für sich stabil sein. Die konkrete Wahl aus der Familie, die sogenannte Parameterwahlstrategie, stützt sich dann auf zusätzliche a priori Kenntnisse über den Messfehler $y^\delta - y$.

In der klassischen deterministischen Theorie setzt man dabei zunächst die Kenntnis einer oberen Schranke $\delta \geq \|y^\delta - \hat{y}\|$ für die Norm des Datenfehlers voraus. Dies ist entscheidend für die Wahl des Ersatzproblems: je besser ein solches das wahre instabile Problem approximiert, desto unstabiler wird es zwingenderweise (bzw. desto größer wird die Stetigkeitskonstante). Aus der Stetigkeitskonstante lässt sich nun wiederum ablesen, wie stark Fehler verstärkt werden und um Konvergenz gegen die wahre Lösung zu erhalten ist nun nur noch darauf zu achten, dass die maximale Fehlerverstärkung mal der oberen Schranke der Norm des Datenfehlers gegen Null geht (im Limes immer genauerer Messungen $\delta \to 0$).

In praktischen Anwendungen liegt eine präzise obere Fehlerschranke allerdings ad hoc nicht vor. Tatsächlich besagt das fundamentale Bakushinskii-Veto, dass die Kenntnis einer oberen Fehlerschranke nicht nur hinreichend, sonder auch notwendig ist. Aus der Existenz einer von der Schranke δ unabhängigen Rekonstruktionsmethode würde direkt die stetige Invertierbarkeit obiger Gleichung folgern, was der Schlechtgestelltheit widerspräche.

Eine häufige Annahme ist, dass die Messungen y^δ Realisierungen einer Zufallsvariablen sind. Auf Grund der Zufälligkeit hat man keine sichere Kenntnis einer oberen Schranke für die Norm des Fehlers. Stattdessen trifft man quantitative Vorabannahmen über die Fehlerverteilung und häufig schränkt man sich auf Gaußverteilungen ein, um untypisches Verhalten der (zufälligen) Messungen durch starke Konzentrationsungleichungen zu kontrollieren. Auch hierbei ist in Anwendungen die Annahme einer Gaußverteilung sowie die Vorabkenntnis der Verteilung überhaupt oftmals schwierig zu rechtfertigen, beziehungsweise nicht erfüllt.

Dieser Problematik wird in dieser Arbeit wie an folgendem Beispiel demonstriert begegnet. Mit einem Teleskop sollen Informationen über eine entfernte Galaxie gewonnen werden. Dabei werden die Bilder unweigerlich durch z.B. atmosphärische Turbulenzen gestört. Allerdings ändern sich besagte Turbulenzen sehr schnell, wohingegen die relevanten Eigenschaften der Galaxie auf einer entsprechend kleinen Zeitskala als konstant angesehen werden können. Es liegt also nahe, mehrere Bilder aufzunehmen und dann die Messfehler herauszumitteln.

Die Verwendung mehrfacher Messungen ist tatsächlich gängige Praxis in Anwendungen, bekannt als "Signalmittelung". In dieser Arbeit wird dieser Prozess in die Analyse integriert. Die Daten werden dabei aus mehreren Messungen gemittelt, welche einer beliebigen, unbekannten Verteilung folgen, wobei die zur Lösung des Problems unweigerlich notwendige Fehlerschranke geschätzt wird. Auf Mittelwert und Schätzer wird dann ein klassisches Regularisierungsverfahren angewandt.

Als Regularisierungen werden größtenteils Filter-basierte Verfahren behandelt, die sich auf die Spektralzerlegung des Operators K stützen. Als Parameterwahlstrategien werden dabei zunächst einfache Strategien betrachtet, die sich nur auf die (geschätzte) Fehlerschranke stützen (a priori-Wahlen). Mit besagten a priori-Strategien erzielt man im Allgemeinen jedoch nur suboptimale Konvergenzraten. Die Konvergenzgeschwindigkeit hängt von abstrakten Glattheitseigenschaften der wahren Lösung ab, welche in der Regel unbekannt sind. Adaptive Parameterwahlstrategien, welche üblicherweise neben der (geschätzten) Fehlerschranke von den konkret gemessenen Daten abhängen, passen sich automatisch an die unbekannten Glattheitseigenschaften an und liefern somit optimale Konvergenzraten. Adaptive Verfahren werden auch als a posteriori Verfahren bezeichnet. Als Prototyp eines adaptiven Verfahrens betrachten wir in dieser Arbeit das Diskrepanzprinzip. Hierbei wird das Beispielproblem (beziehungsweise der Regularisierungsparameter) so ausgewählt, dass die Norm des Residuums, dass heißt der Abstand zwischen den gemessenen Daten und der (abgebildeten) Rekonstruktion, in etwa dem geschätzten Datenfehler entspricht. Dieses Verfahren ist eines der am häufigsten genutzten, da es einfach zu implementieren ist und in vielen Fällen optimale Konvergenzraten liefert.

Konkret ist im ersten Kapitel eine Folge $Y_1, Y_2, ...$ unabhängiger und identisch verteilter Messungen der wahren Daten $\hat{y}$ gegeben, d.h. jede Messung entspricht beispielsweise einer ganzen zufälligen Funktion. Die Verteilung der Messungen ist beliebig, es ist einzig vorausgesetzt, dass sie unverzerrt sind, $\mathbb{E}[Y_i] = \hat{y}$, und endliche Varianz haben, $\mathbb{E}\|Y_i - \hat{y}\|^2 < \infty$. Für n Messungen bezeichnet

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

den Mittelwert als Schätzer für $\hat{y}$. Der zentrale Grenzwertsatz (für Hilbertraumwertige Zufallsvariablen) besagt nun, dass

$$\sqrt{n}(\bar{Y}_n - \hat{y}) \to Z$$

für $n \to \infty$ schwach konvergiert, wobei Z eine normalverteilte Zufallsvariable ist. Demnach sind

$$\delta_n^{est} := \frac{1}{\sqrt{n}} \quad \text{oder} \quad \delta_n^{est} := \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\|Y_i - \bar{Y}_n\|^2}}{\sqrt{n}}$$

naheliegende Schätzer für den unbekannten wahren Datenfehler $\delta_n^{true} := \|\bar{Y}_n - \hat{y}\|$. Auf die gemittelten Daten $\bar{Y}_n$ und den geschätzten Datenfehler δ_n^{est} wird nun ein deterministisches Regularisierungsverfahren angewandt. Man zeigt ohne Probleme die Konvergenz in L^2, auch genannt Konvergenz im quadratischen Mittel, für a priori Verfahren. Schwierigkeiten bereitet dann die Analyse des Diskrepanzprinzips, welches bekanntermaßen sensitiv auf ein Unterschätzen des Datenfehlers reagiert, was hier unvermeidlich immer wieder auftritt (mit uniform in n von der Null weg beschränkter Wahrscheinlichkeit). Tatsächlich wird gezeigt, dass das Diskrepanzprinzip nicht in L^2 konvergiert. Dementsprechend ist überraschend, dass Konvergenz in Wahrscheinlichkeit gilt. Hierbei nutzt man aus, dass die Richtung, aus der $\bar{Y}_n$ gegen $\hat{y}$ konvergiert, durch die Kovarianzstruktur einer Messung festgelegt ist. Weiterhin adaptiert das Diskrepanzprinzip in dem Sinne, dass mit gegen Eins strebender Wahrscheinlichkeit die optimale deterministische Rate gilt.

Schließlich wird noch der Zusammenhang zu minmax-optimalen Schätzern sowie zu heuristischen Regularsierungsverfahren diskutiert. Auch hierbei ist entscheidend, dass durch die spezielle Struktur des Schätzers als Mittelwert aus vielen Einzelmessungen, viele worst-case Fehlerszenarien ausgeschlossen sind. Es wird unter anderem gezeigt, dass durch eine Reskalierung der Messungen und des Operators potentiell eine bessere (als die deterministisch erwartete) Konvergenzrate erreichbar ist, sowie dass die in der Theorie heuristischer Verfahren populäre Mouckenhoupt-Bedingung hier in der Regel nicht erfüllt ist.

Die Bedingung der beschränkten Varianz $\mathbb{E}\|Y_i - \hat{y}\|^2 < \infty$ schließt Weißes Rauschen als Fehlerverteilung aus. Um die Resultate auf diesen Fall zu erweitern, wird im zweiten Kapitel ein semi-diskretes Modell betrachtet. Die Messungen sind nicht mehr Elemente des Hilbertraums, sondern lineare Funktionale $l_1, l_2, ...$, bezeichnet als Messkanäle. Man denke im Falle von Funktionenräumen beispielsweise an Punktauswertungen oder Integrale über kleine Bereiche. Es sind dann wiederholte Messungen auf jedem Kanal gegeben, wobei Y_{ij} die i-te Messung des j-ten Kanals bezeichnet. Damit ist

$$\begin{pmatrix} Y_{11} - l_1(\hat{y}) \\ ... \\ Y_{1m} - l_m(\hat{y}) \end{pmatrix}, \begin{pmatrix} Y_{21} - l_1(\hat{y}) \\ ... \\ Y_{2m} - l_m(\hat{y}) \end{pmatrix}, ... \subset \mathbb{R}^m$$

unabhängiges und identisch verteiltes Weißes Rauschen (unbekannter Verteilung), mit $\mathbb{E}[Y_{ij} - l_j(\hat{y})] = 0$ und $\mathbb{E}\left(Y_{ij} - l_j(\hat{y})\right)^2 < \infty$. Entsprechend ist

$$\bar{Y}_n^{(m)} := \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} Y_{i1} \\ ... \\ Y_{im} \end{pmatrix}$$

der komponentenweise Mittelwert und der unbekannte Datenfehler

$$\left\| Y_n^{(m)} - \begin{pmatrix} l_1(\hat{y}) \\ ... \\ l_m(\hat{y}) \end{pmatrix} \right\|$$

wird durch

$$\delta_{m,n}^{est} := \sqrt{\frac{m}{n}} \quad \text{oder} \quad \delta_{m,n}^{est} := \sqrt{\frac{m}{n}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{ij} - \frac{1}{n}\sum_{l=1}^{n} Y_{lj}\right)^2}$$

geschätzt. Wieder gilt L^2-Konvergenz für a priori Verfahren und Konvergenz in Wahrscheinlichkeit für das Diskrepanzprinzip, im Limes unendlich vieler Messungen ($n \to \infty$) und unendlich feiner Diskretisierung ($m \to \infty$). Hierbei ist bemerkenswert, dass die Diskretisierung $l_1, l_2, ...$ nur zwei sehr einfache, qualitative Bedingungen zu erfüllen hat, namentlich dass sie vollständig und quadratsummierbar ist, d.h. dass für beliebige $y \neq 0$ aus dem Hilbertraum gilt

$$\exists l_j \text{ with } l_j(y) \neq 0 \quad \text{und} \quad \sum_{j=1}^{\infty} l_j(y)^2 < \infty.$$

Im dritten Kapitel wird schließlich ein einfaches stochastisches Gradientenverfahren für die Lösung inverser Probleme untersucht. Dies ist ein Prototyp einer ganzen Klasse neuartiger Verfahren, die für im Kontext maschinellen Lernens auftretende extrem hoch-dimensionale Probleme entstanden sind, für welche sich Filter-basierte Verfahren auf Grund des enormen numerischen Aufwands als unpraktikabel erweisen.

Diese Verfahren sind im Rahmen der klassischen Regularisierungstheorie kaum untersucht. Hier wird mit dem Diskrepanzprinzip erstmals eine adaptive Stoppregel für das stochastische Gradientenverfahren untersucht und rigoros Konvergenz gezeigt.

Introduction

At the heart of science lies the modeling of natural phenomena. The models, derived from empirical observations and principal physical laws, are used to tackle problems in practical applications. The models are connecting a cause with an observation, and therefore yield two fundamental problems. The direct problem is the question, given a cause what is the observation, whereas the inverse problem is to determine the cause, when given the observation. E.g. in x-ray tomography the physical properties of x-rays are used to describe the interplay between the absorptive properties of a body and the resulting damping of the intensity of a x-ray passing through the body. Here the inverse problem is to determine the mass density of the body from measurements of the intensity decays of x-rays passing through the body from various directions. Inverse problems are often ill-posed, i.e. they fail to fulfill Hadamard's criteria of well-posedness. Hadamard called a problem well-posed, if there exists a solution for arbitrary data, the solution is unique and moreover depends continuously on the data. In our case, typically the solution is unstable with respect to small perturbations of the observation, which are intrinsically inevitable in practical applications.

Many inverse problems, as the one mentioned above, can be mathematically stated as the equation

$$K\hat{x} = \hat{y}, \tag{0.1}$$

where K is a compact linear operator between infinite-dimensional Hilbert spaces, $\hat{x}$ is the unknown quantity of interest which has to be determined from a noisy measured observation y^δ of the true data $\hat{y}$. We assume for a moment, that K is injective with dense range. Clearly, on $\mathcal{R}(K)$ the problem (0.1) has a solution given by $K^{-1}\hat{y}$. The ill-posedness stems now from the compactness of K and manifests in the fact, that $\mathcal{R}(K) \subsetneq \mathcal{Y}$ and hence, not for all $\hat{y} \in \mathcal{Y}$ there is a solution of (0.1). Even worse it holds that the restriction of K^{-1} to $\mathcal{R}(K)$ is not continuous. Thus for noisy measurements y^δ of $\hat{y}$ with $\|y^\delta - \hat{y}\| \leq \delta$, $\delta \to 0$ does not imply that $K^{-1}y^\delta \to K^{-1}\hat{y}$ (if the former exists at all). That is even for arbitrarily precise measurements no stable reconstruction is possible. From that arises the need of regularisation. Going back to the works of Tikhonov [Tik63] and Phillips [Phi62], (0.1) is replaced with the minimisation problem

$$\min_{x \in \mathcal{X}} \|Kx - y^\delta\|^2 + \alpha\|x\|^2 \tag{0.2}$$

where the so called regularisation parameter $\alpha > 0$ balances how good a candidate x fits the observed data y^δ with is regularity $\|x\|$. In contrast to (0.1), a solution of (0.2) exists for all $y^\delta \in \mathcal{Y}$ and $\alpha > 0$, and it depends for fixed α continuously (and linearly) on y. In fact, via the Gaussian normal equations, one obtains the following explicit presentation

$$R_\alpha y := (K^*K + \alpha I_\mathcal{Y})^{-1} K^* y$$

for the solution of (0.2). For the exact data $\hat{y} \in \mathcal{R}(K)$ it thus holds that $R_\alpha \hat{y} \to K^{-1}\hat{y} = \hat{x}$ as $\alpha \to 0$. In order to approximate the unknown solution $\hat{x}$ it remains to determine the regularisation parameter $\alpha = \alpha(\delta, y^\delta)$ for noisy measurements y^δ with $\|y^\delta - \hat{y}\| \leq \delta$. From the decomposition

$$\begin{aligned} \|R_{\alpha(\delta,y^\delta)}y^\delta - \hat{x}\| &\leq \|R_{\alpha(\delta,y^\delta)}y^\delta - R_{\alpha(\delta,y^\delta)}\hat{y}\| + \|R_{\alpha(\delta,y^\delta)}\hat{y} - K^{-1}\hat{y}\| &\quad (0.3)\\ &\leq \|R_{\alpha(\delta,y^\delta)}\|\delta + \|R_{\alpha(\delta,y^\delta)}\hat{y} - K^{-1}\hat{y}\| &\quad (0.4)\end{aligned}$$

one directly deducts the following convergence result. Let $\alpha = \alpha(\delta)$ be such that

$$\alpha(\delta) \quad \text{and} \quad \|R_{\alpha(\delta)}\|\delta \to 0, \tag{0.5}$$

then

$$\|R_{\alpha(\delta)}y^\delta - K^{-1}\hat{y}\| \to 0,$$

as $\delta \to 0$, where $\hat{y} \in \mathcal{R}(K)$ and $(y^\delta)_{\delta>0} \subset \mathcal{Y}$ with $\|y^\delta - \hat{y}\| \leq \delta$. Such choices α, which do not depend on the observation y^δ, are called a priori.

The assumption for α in (0.5) is rather unspecific and fulfilled by a wide range of choices, so the question arises, to find the optimal one. A quick look at (0.3) reveals, that for δ fixed, the first and second term, called the data propagation and the approximation error, become larger respectively smaller for decreasing α. Thus one has to find the value of α which balances the both terms. Since the true data y is unknown, this cannot be done directly. However, the following intuitive and simple choice in essence automatically balances the both terms. The discrepancy principle, due to Morozov [Mor68], which postulates to determine the regularisation parameter such that the discrepancy between the measured data and the candidate approximately matches the noise level δ

$$\|KR_{\alpha(\delta,y^\delta)}y^\delta - y^\delta\| \approx \delta. \tag{0.6}$$

Both the a priori choice (0.5) and the discrepancy principle (0.6) share one drawback, as they require the explicit knowledge of the noise level δ, which is usually not justified in applications. Therefore, it would be desirable to have rules for determining the regularisation parameter, which are purely data driven (i.e. independent of δ) and guarantee convergence for $\delta \to 0$. The following famous result, the Bakushinskii veto [Bak84], states, that such rules do not exist.

Theorem 0.0.1 (Bakushinskii veto). *There are no parameter choice rules $\alpha : \mathcal{Y} \to (0, \infty)$, such that for all $\hat{y} \in \mathcal{R}(Y)$ and all $(y^\delta)_{\delta>0} \subset \mathcal{Y}$ with $\|y^\delta - \hat{y}\| \leq \delta$, there holds*

$$\lim_{\delta \to \infty} R_{\alpha(y^\delta)} y^\delta = K^{-1}\hat{y} = \hat{x}.$$

Proof. We give the short standard proof and argue by contradiction. Assume there exists such a parameter choice rule α. For $y \in \mathcal{R}(K)$ we set $y^\delta = y$ for all $\delta > 0$. Then it follows that $R_{\alpha(y)} y = \lim_{\delta \to 0} R_{\alpha(y^\delta)} y^\delta = K^{-1}y$. Now let $(y^\delta)_{\delta>0} \subset \mathcal{R}(K)$ be arbitrary with $\|y^\delta - y\| \leq \delta$. Then it holds that

$$K^{-1}y^\delta = R_{\alpha(y^\delta)} y^\delta \to K^{-1}y$$

as $\delta \to 0$. Hence K^{-1} is continuous on $\mathcal{R}(K)$, contradicting the ill-posedness of K.

□

The framework presented so far (where one has an upper bound for the noise level), is usually titled deterministic inverse problems, which is complemented by so called stochastic or statistical inverse problems. In the latter scenario, the measurement y^δ is seen as a random variable. The assumption of knowing the noise level δ is then replaced with knowledge about the error distribution. Very often, one restricts to certain classes of distributions, e.g. Gaussian distributions.

In this thesis, we rigorously explore a natural setting which requires no a priori knowledge, neither of an upper bound of the noise level nor about the (arbitrary and usually non-Gaussian) error distribution. In applications, the a priori knowledge of the error stems, in many cases, from the estimation with multiple repeated measurements. Thus the key assumption we impose, is that a measurement can be repeated and a crucial requirement is that the solution does not change at least on small time scales. Let us stress, that using multiple measurements to decrease the data error is actually a standard engineering practice called 'signal averaging', see, e.g., [Lyo04] for an introducing monograph or [HA10] for a survey article. Examples with low or moderate numbers of measurements (up to a hundred) can be found in [BLMT09] or [MBLW04] on image averaging or [GSS14] on satellite radar measurements. For the recent first image of a black hole, even up to 10^9 samples were averaged, cf. [AAA$^+$19].

So instead of one single measurement y^δ we have multiple unbiased, identically distributed and independent measurements $Y_1, Y_2, Y_3, ...$ of the exact data y (e.g., every measurement is a whole random function). The multiple measurements naturally yield

$$\bar{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

as an estimator of y. Indeed, the law of large numbers (see [LT91] for Hilbert space valued random variables), states that $\bar{Y}_n \to y$ in probability (and a.s. and in L^2) as $n \to \infty$. So, in the light of the Bakushinskii veto (Theorem 2.1.8) we need a reasonable guess for the data error $\delta_n^{true} := \|\bar{Y}_n - y\|$. By the central limit theorem,

$$\sqrt{n}\left(\bar{Y}_n - y\right) \to Z$$

in distribution, as $n \to \infty$, where Z is a Gaussian random variable. Therefore, a natural estimator of δ_n^{true} would be

$$\delta_n^{est} = \frac{1}{\sqrt{n}} \quad \text{or} \quad \delta_n^{est} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left\|Y_i - \bar{Y}_n\right\|^2}}{\sqrt{n}}.$$

Consequently, a natural attempt to solve the inverse problem, is to apply some deterministic regularisation method with data $\bar{Y}_n$ and (estimated) noise level δ_n^{est} and to investigate whether and in which sense the resulting approximation converges against the true solution of the problem.

In Chapter 1 we thoroughly analyse the above approach under the assumption, that the measurements have a strongly bounded second moment, i.e. $\mathbb{E}\|Y_1\|^2 < \infty$. We show convergence in L^2 (a.k.a. convergence in mean square) for a priori regularisation (0.5) and convergence in probability for the discrepancy principle (0.6). In case of the discrepancy principle it is moreover shown, that the optimal deterministic rate holds with a probability tending to 1 as the number of repetitions n goes to infinity. Thus, one can solve the inverse problem by estimating the noise level from multiple measurements of unknown distribution. Further, we also discuss optimality in a statistical context and show, how one may obtain a better rate (than the one from a deterministic worst case scenario). Finally we relate to popular heuristic methods.

In general, there are two approaches to tackle an ill-posed problem with stochastic noise. The Bayesian setting considers the solution of the problem itself as a random quantity, on which one has some a priori knowledge (see [KS06], [NP15]). This opposes the frequentist setting, where the inverse problem is assumed to have a deterministic, exact solution ([Cav11],[BHMR07],[NP15]). We are working in the frequentist setting, but we stay close to the classic deterministic theory of linear inverse problems ([EHN96],[Rie13],[TA77], [IJ15]). E.g., in statistical inverse problem optimality results are usually of the form, that one shows that a given estimator is minmax (eventually using an oracle inequality). We on the other hand show, that our approach yields asymptotically the optimal deterministic rate with probability 1. However, we will relate to minmax estimators in Section 1.3. In [BHMR07] a priori error bounds under general noise distributions are given. Popular adaptive methods to determine the regularisation parameter are cross validation [Wah77], Lepski's balancing principle [MP03b] or penalised empirical risk minimisation [CG$^+$06]. These works are restricted to Gaussian noise, which is usually due to the need for strong concentration inequalities and control of large deviations. Moreover, the implementation of these methods is typically computationally much more demanding compared to the discrepancy principle, which is the main parameter choice rule considered in this thesis. Recently, modifications of the discrepancy principle, which require the explicit knowledge of the singular value composition of K, were studied in the statistical setting under Gaussian noise ([BM12],[LM14],[BHR18],[LPB$^+$18]). In [G$^+$11], [BR08],[Bec11],[Wer18] various ways are described how to solve a given statistical inverse problem under Gaussian noise without knowing the exact noise level. We extend this results to arbitrary error distributions without any knowledge of the singular value decomposition of K. Finally we want to mention ([Hof06],[GHR17]), where results from the classical deterministic theory are transfered using the Ky-Fan metric, which induces convergence in probability. Here the crucial requirement is, that one knows the Ky-Fan distance between the measurements and the true data.

In the references mentioned above, the error is often modelled as a Hilbert space process (such as Gaussian white noise, [Don95],[CT02]) and thus violates the condition $\mathbb{E}\|Y_1\|^2 < \infty$. Under the popular assumption that the operator K is Hilbert-Schmidt, one could in principle extend the results of Chapter 1 to a general Hilbert space process error model (considering the symmetric equation $K^*K\hat{x} = K^*\hat{y}$ instead of $K\hat{x} = \hat{y}$, as it is done for example in [BM12]). However, this usually impairs the relative smoothness of the true solution and yields worse error bounds. In order to extend the above results to white noise scenarios we thus proceed differently and investigate a semi-discrete model under arbitrary unknown white noise. As an arbitrary element of an infinite-dimensional space, y cannot be measured directly. Instead we assume that one may measure $l_1(y), l_2(y), ...$ for various (normed) linear functionals, which we refer to as measurement channels.

So we assume that we have multiple unbiased, identically distributed and independent measurements on each measurement channel. We denote by Y_{ij} the i-th sample on the j-th measurement channel. Then

$$\begin{pmatrix} Y_{i1} - l_1(y) \\ ... \\ Y_{im} - l_m(y) \end{pmatrix}_{i\in\mathbb{N}} \subset \mathbb{R}^m$$

are i.i.d. white noise vectors with unknown distribution. With the component wise average

$$\bar{Y}_n^{(m)} := \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Y_{i1} \\ \dots \\ Y_{im}\end{pmatrix},$$

the application of Tikhonov's method 0.2 yields the following optimisation problem

$$\min_{x\in\mathcal{X}} \left\| \begin{pmatrix} l_1(Kx) \\ \dots \\ l_m(Kx)\end{pmatrix} - \bar{Y}_n^{(m)} \right\|^2 + \alpha\|x\|^2.$$

The regularisation parameter α has to be chosen accordingly to the data error $\|\bar{Y}_n^{(m)} - \begin{pmatrix} l_1(y) & \dots & l_m(y)\end{pmatrix}^T\|$. Based on the samples we estimate the latter by

$$\delta_{m,n}^{est} = \sqrt{\frac{m}{n}} \quad \text{or} \quad \delta_{m,n}^{est} = \sqrt{\frac{m}{n}\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{ij} - \frac{1}{n}\sum_{l=1}^{n} Y_{lj}\right)^2}.$$

Again the approach is to use a deterministic regularisation method together with $\bar{Y}_n^{(m)}$ and $\delta_{m,n}^{est}$.

In Chapter 2 we analyse the above approach in detail for arbitrary error distributions fulfilling $\mathbb{E}Y_{11}^2 < \infty$. Regarding the measurement channels $(l_j)_{j\in\mathbb{N}}$ we only impose two natural restrictions, namely, that it is complete and l^2-summable, i.e. that for all $y \in \mathcal{Y}\setminus\{0\}$ there is a l_j with $l_j(y) \neq 0$, and that $\sum_{j=1}^{\infty} l_j(y)^2 < \infty$. Again we obtain convergence in L^2 for a priori regularisation and convergence in probability for the discrepancy principle, as the number of measurements channels m and the number of repetitions n tend to infinity (such that $m/n \to 0$). We also investigate a related approach and show how to obtain the optimal deterministic rate under additional knowledge of a distretisation error.

It is widely known that discretisation has a regularising effect, see for example [MP01],[Han10] for the discretisation in the deterministic setting, [MP01], [MP03a] for the statistical frequentist setting and [KS07] for the Bayesian approach. In general, one can either first regularise the infinite-dimensional problem and then discretise, or, as it is done here one first discretises and then regularises the finite-dimensional problem. So far, inverse problems under white noise are treated the first way, and the white noise is modeled as a Hilbert space process operating on $\mathcal{Y}$, see [BHMR07], [Cav11]. The major challenge of this modeling is, that then the measurements are not elements of $\mathcal{Y}$. This implies some drawbacks, e.g. one has to restrict to sufficiently smoothing operators and to include correction terms in the convergence rates (compared to the classical deterministic rates). Most importantly, the discrepancy principle cannot be applied directly due to the unboundedness of the noise. These technical difficulties are not present in the semi-discretised setting considered here. It is notable, that the main convergence result in this chapter guarantees convergence for arbitrary unknown distribution, as long as one is able to measure repeatedly, under quite general assumptions on the discretisation, which are only of qualitative nature and most importantly are independent of the unknown exact right hand side.

To summarise the connection of the first two chapters to the Bakushinskii veto let us state the following. The Bakushinskii veto states that the inverse problem can only be solved with a deterministic regularisation, if the noise level of the data is known. Here we show, that if one has access to multiple i.i.d. measurements of an unknown distribution, one may use as data the average together with the estimated noise level and one (eventually) obtains the optimal deterministic rate with high probability, as the number of measurements tends to infinity. That is one can estimate the measurement error from the data.

Finally, when it comes to practically solve an inverse problem one has to fully discretise (0.1). This yields a finite-dimensional equation

$$Ax = y, \tag{0.7}$$

with $x \in \mathbb{R}^m, y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times m}$. The ill-posedness now resembles in the fact that these problems are extremely bad conditioned, thus standard methods for solving linear equations may it be direct methods (e.g. LR/LU decomposition) or iterative methods (Gauß-Seidel or Jacobi) cannot be used due to stability issues and have to be replaced with stable regularisation methods. For extremely high dimensional discretisations, the direct application of Tikhonov's method as in (0.3) (or other classical methods) may become infeasible due to computational complexity and one rather relies on iterative methods. Hereby, a computationally particularly cheap method is stochastic gradient descent [RM51, BCN18], which directly scales to the dimension of the problem and uses only vector-vector multiplications. In fact, stochastic gradient descent and its variants (e.g., minibatch and accelerated) have been established as the workhorse behind many challenging training tasks in deep learning [Bot10, BCN18], and they are also popular for image reconstruction in computed tomography [GBH70, Nat86]. Due to its popularity in machine learning and big data applications, there exists a considerable amount of literature about the convergence properties of stochastic gradient descent as an optimisation algorithm. However, the mathematical theory in the lens of classical regularisation theory is rather incomplete, as it does not fit in the framework of filter-based regularisation. In the work [JL19] the regularising property of stochastic gradient descent was explored for the first time. Further, a convergence rate in the mean squared norm was derived, under suitable source type condition on the true solution $\hat{x}$. These results were recently extended to mildly nonlinear inverse problems, further assisted with suitable nonlinearity conditions of the forward map [JZZ20a]. However, in these works, the convergence rate can only be achieved under a knowledge of the smoothness parameter of $\hat{x}$, which is usually not directly accessible in practice. Therefore, it is of enormous practical importance and theoretical interest to develop *a posteriori* stopping rules that do not require such a knowledge.

In Chapter 3 we give a first rigorous analysis of the discrepancy principle as an adaptive stopping rule for stochastic gradient descent. We prove the convergence of the approach and a finite termination property. Also, a partial result on optimality is given.

Chapter 1

The case with finite variance

Section 1.1, 1.2 and 1.5 are up to minor changes published in [HJP20a]. Sections 1.3 and 1.4 contain yet unpublished results. All the main proofs for the results from the aforementioned sections are collected in section 1.6. Accompanying numerical results and a short outlook are presented in sections 1.7 and 1.8.

We start by recapping and slightly generalising the setting as presented in the introduction. The goal is to solve the ill-posed equation $K\hat{x} = \hat{y}$, where $\hat{x} \in \mathcal{X}$ and $\hat{y} \in \mathcal{Y}$ are elements of infinite-dimensional Hilbert spaces and K is either linear and bounded with non-closed range, or more specifically compact. We do not know the right hand side $\hat{y}$ exactly, but we are given several measurements $Y_1, Y_2, ...$ of it, which are independent, identically distributed and unbiased ($\mathbb{E}Y_i = \hat{y}$) random variables. Thus we assume, that we are able to measure the right hand side multiple times. The given multiple measurements naturally lead to an estimator of $\hat{y}$, namely the sample mean

$$\bar{Y}_n := \frac{\sum_{i\leq n} Y_i}{n}.$$

But, in general $K^+\bar{Y}_n \not\to K^+\hat{y}$ for $n \to \infty$, because the generalised inverse (Definition 2.2 of [EHN96]) of K is not continuous. So the inverse is replaced with a family of continuous approximations $(R_\alpha)_{\alpha>0}$, called regularisation, e.g. the Tikhonov regularisation $R_\alpha := (K^*K + \alpha Id)^{-1} K^*$, where $Id : \mathcal{X} \to \mathcal{X}$ is the identity, as motivated (0.2) in the introduction. The regularisation parameter α has to be chosen accordingly to the data $\bar{Y}_n$ and the true data error

$$\delta_n^{true} := \|\bar{Y}_n - \hat{y}\|,$$

which is a random variable. Since $\hat{y}$ is unknown, δ_n^{true} is also unknown and has to be guessed. Natural guesses are

$$\delta_n^{est} := \frac{1}{\sqrt{n}} \quad \text{or} \quad \delta_n^{est} := \frac{\sqrt{\sum_{i\leq n} \|Y_i - \bar{Y}_n\|^2/(n-1)}}{\sqrt{n}}.$$

A natural approach is now to use a (deterministic) regularisation method together with $\bar{Y}_n$ and δ_n^{est}. We are in particular interested in the discrepancy principle (0.6), which is known to provide optimal convergence rates (for some $\hat{y}$) in the classical deterministic setting. The main result of this chapter states, that the approach converges in probability for the discrepancy principle, and the optimal deterministic (worst case) error bound holds with a probability converging to 1 (as the number of measurements n tends to infinity). Moreover it is shown, that the approach in general does not yield L^2-convergence[1] for a naive use of the discrepancy principle, but it does for a priori regularisation.
In the following section we apply our approach to a priori regularisations and in the main part we consider the widely used discrepancy principle. Then we compare the obtained rates with the rates attained by the optimal oracle and show how to obtain a better rate than the deterministic one with a modified rescaled discrepancy principle for spectral cut-off. Then we quickly discuss relations to heuristic parameter choice rules and show how to choose δ_n^{est} to obtain almost sure convergence. We conclude with some numerical experiments.

1.1 A priori regularisation

We use the usual definition that $R_\alpha : \mathcal{Y} \to \mathcal{X}$ is called a linear regularisation, if R_α is a bounded linear operator for all $\alpha > 0$ and if $R_\alpha y \to K^+ y$ for $\alpha \to 0$ for all $y \in \mathcal{D}(K^+)$. A regularisation method is a combination of a regularisation and a parameter choice strategy $\alpha : \mathbb{R}^+ \times \mathcal{Y} \to \mathbb{R}^+$, such that $R_{\alpha(\delta, y^\delta)} y^\delta \to K^+ y$ for $\delta \to 0$, for all $y \in \mathcal{D}(K^+)$ and for all $(y^\delta)_{\delta>0} \subset \mathcal{Y}$ with $\|y^\delta - y\| \leq \delta$. The method is called a priori, if the parameter choice does not depend on the data, that is if $\alpha(\delta, y) = \alpha(\delta)$.
The measurements can be modelled as realisations of an independent and identically distributed sequence $(Y_i)_{i\in\mathbb{N}}$ of (integrable) random variables with values in $\mathcal{Y}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. This requires $\mathbb{E}\|Y_1\|^2 < \infty$ (so that the measurements lie (almost surely) in the Hilbert space) and the unbiasedness assumption can simply be stated as $\mathbb{E}[Y_1] = \hat{y} \in \mathcal{D}(K^+)$. Moreover we assume that $\mathbb{E}\|Y_1\|^2 > 0$, to avoid the trivial case of (almost surely) constant measurements. Finally, measureability of all involved quantities (e.g. of $\delta_n^{est}, \delta_n^{true}, \alpha_n, ...$) can be derived by standard arguments from the measureability of $(Y_i)_{i\in\mathbb{N}}$ and we will not comment further on this issue throughout the thesis, but rather refer to [LT91].
In the following we apply the above approach to a priori parameter choice strategies $\alpha(y^\delta, \delta) = \alpha(\delta)$. We restrict to $\delta_n^{est} = 1/\sqrt{n}$ here, that is we do not estimate the variance (otherwise the parameter choice would depend on the data). Since then δ_n^{est} and hence $\alpha(\delta_n^{est})$ are deterministic, the situation is very easy here and the results are not surprising (see Remark 1.1.5).

[1] also called convergence of the integrated mean squared error or root mean squared error

Theorem 1.1.1 (Convergence of a priori regularisation)**.** *Assume that $K : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator with non-closed range between Hilbert spaces and that $Y_1, Y_2, ...$ are i.i.d. $\mathcal{Y}-$valued random variables which fulfill $\mathbb{E}[Y_1] = \hat{y} \in \mathcal{D}(K^+)$ and $0 < \mathbb{E}\|Y_1\|^2 < \infty$. Take an a priori regularisation scheme, with $\alpha(\delta) \xrightarrow{\delta\to 0} 0$ and $\|R_{\alpha(\delta)}\|\delta \xrightarrow{\delta\to 0} 0$. Set $\bar{Y}_n := \sum_{i\leq n} Y_i/n$ and $\delta_n^{est} := n^{-1/2}$. Then $\lim_{n\to\infty} \mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - K^+\hat{y}\|^2 = 0$.*

Proof. Because of linearity, $\mathbb{E}\left[R_\alpha Y_1\right] = R_\alpha \mathbb{E}\left[Y_1\right] = R_\alpha \hat{y}$ and thus by (1.5) below

$$\mathbb{E}\|R_\alpha \bar{Y}_n - R_\alpha \hat{y}\|^2 = \frac{1}{n^2}\mathbb{E}\left\|\sum_{i=1}^n R_\alpha\left(Y_i - \hat{y}\right)\right\|^2 = \frac{\mathbb{E}\|R_\alpha Y_1 - R_\alpha \hat{y}\|^2}{n},$$

since $R_\alpha Y_i \in \mathcal{R}(K^*)$ where the latter is separable. Therefore, by the bias-variance-decomposition,

$$\begin{aligned}
\mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - K^+\hat{y}\|^2 &= \mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - R_{\alpha(\delta_n^{est})}\hat{y} + R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&= \mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - R_{\alpha(\delta_n^{est})}\hat{y}\|^2 + \|R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&= \frac{\mathbb{E}\|R_{\alpha(\delta_n^{est})}Y_1 - R_{\alpha(\delta_n^{est})}\hat{y}\|^2}{n} + \|R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&\leq \frac{\|R_{\alpha(\delta_n^{est})}\|^2}{n}\mathbb{E}\|Y_1 - \hat{y}\|^2 + \|R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&= \|R_{\alpha(\delta_n^{est})}\|^2 \delta_n^{est2}\mathbb{E}\|Y_1 - \hat{y}\|^2 + \|R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&\to 0 \qquad \text{for} \quad n \to \infty.
\end{aligned}$$

□

As in the deterministic case, under additional source conditions for $\hat{x}$ we can prove convergence rates. In this thesis we will consider classical Hölder-type source conditions, which date back at least to [Lav62]. These allow explicit presentation of the derived rates. However, not all $\hat{x}$ fulfill such a condition. In this context, the study of general source conditions (see e.g. [Heg92],[Tau98],[HM07] and [MH08]) would be interesting, allowing to derive convergence rates for arbitrary $\hat{x}$. We leave this as a possible future work. In the following we restrict to regularisations $R_\alpha := F_\alpha\left(K^*K\right)K^*$ defined via the spectral decomposition (see [EHN96]) with the following assumptions for the generating filter.

Assumption 1.1.2. *$(F_\alpha)_{\alpha>0}$ is a regularising filter, i.e. a family of bounded real valued functions on $(0, \|K\|^2]$ with $\lim_{\alpha\to 0} F_\alpha(\lambda) = \frac{1}{\lambda}$ and $\lambda F_\alpha(\lambda) \leq C_R$ for all $\alpha > 0$ and all $\lambda \in (0, \|K\|^2]$, where $C_R > 0$ is some constant. Moreover, it has qualification $\nu_0 > 0$, i.e. ν_0 is maximal such that for all $\nu \in [0, \nu_0]$ there exists a constant $C_\nu > 0$ with*

$$\sup_{\lambda\in(0,\|K\|^2]} \lambda^{\nu/2}|1 - \lambda F_\alpha(\lambda)| \leq C_\nu \alpha^{\nu/2}.$$

Finally, there is a constant $C_F > 0$ *such that* $|F_\alpha(\lambda)| \leq C_F/\alpha$ *for all* $0 < \lambda \leq \|K\|^2$.

Remark 1.1.3. The generating filter of the following regularisation methods fulfill the Assumption 1.1.2:

1. Tikhonov regularisation (qualification 2)
2. n-times iterated Tikhonov regularisation (qualification $2n$),
3. truncated singular value regularisation (infinite qualification),
4. Landweber iteration (infinite qualification).

Theorem 1.1.4 (Rate of convergence of aprioi regularisation)**.** *Assume that* $K : \mathcal{X} \to \mathcal{Y}$ *is a bounded linear operator with non-closed range between Hilbert spaces and that* $Y_1, Y_2, ...$ *are i.i.d.* $\mathcal{Y}-$*valued random variables which fulfill* $\mathbb{E}[Y_1] = \hat{y} \in \mathcal{D}(K^+)$ *and* $0 < \mathbb{E}\|Y_1\|^2 < \infty$. *Let* R_α *be induced by a filter fulfilling Assumption 1.1.2. Set* $\bar{Y}_n := \sum_{i\leq n} Y_i/n$ *and* $\delta_n^{est} = n^{-1/2}$. *Assume that for* $0 < \nu \leq \nu_0$ *and* $\rho > 0$ *we have that* $K^+\hat{y} = (K^*K)^{\nu/2}\xi$ *for some* $\xi \in \mathcal{X}$ *with* $\|\xi\| \leq \rho$. *Then if for constants* $0 < c < C$,

$$c\left(\frac{\delta_n^{est}}{\rho}\right)^{\frac{2}{\nu+1}} \leq \alpha(\delta_n^{est}) \leq C\left(\frac{\delta_n^{est}}{\rho}\right)^{\frac{2}{\nu+1}},$$

we have that $\sqrt{\mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - K^+\hat{y}\|^2} \leq C'\delta_n^{est\frac{\nu}{\nu+1}}\rho^{\frac{1}{\nu+1}} = \mathcal{O}(n^{-\frac{\nu}{2(\nu+1)}})$ *for some constant* $C' > 0$.

Proof.

We proceed similar to the proof of Theorem 1.1.1, using additionally Proposition 1.6.2 of section 1.6.1.

$$\begin{aligned}
\mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - K^+\hat{y}\|^2 &= \mathbb{E}\|R_{\alpha(\delta_n^{est})}\bar{Y}_n - R_{\alpha(\delta_n^{est})}\hat{y}\|^2 + \|R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&\leq \|R_{\alpha(\delta_n^{est})}\|^2\delta_n^{est2}\mathbb{E}\|Y_1 - \hat{y}\|^2 + \|R_{\alpha(\delta_n^{est})}\hat{y} - K^+\hat{y}\|^2 \\
&\leq C_R C_F \mathbb{E}\|Y_1 - \hat{y}\|^2 \frac{\delta_n^{est2}}{\alpha(\delta_n^{est})} + C_\nu^2\rho^2\alpha(\delta_n^{est})^\nu \\
&\leq \frac{C_R C_F \mathbb{E}\|Y_1 - \hat{y}\|^2}{c}\delta_n^{est\frac{-2}{\nu+1}}\rho^{\frac{2}{\nu+1}}\delta_n^{est2} \\
&\qquad + C_\nu^2 C^\nu \delta_n^{est\frac{2\nu}{\nu+1}}\rho^{\frac{-2\nu}{\nu+1}}\rho^2 \\
&\leq C'^2\delta_n^{est\frac{2\nu}{\nu+1}}\rho^{\frac{2}{\nu+1}}.
\end{aligned}$$

□

Remark 1.1.5. For separable Hilbert spaces one could alternatively argue as follows: The spaces $\mathcal{X}' := L^2(\Omega, \mathcal{X}) = \{X : \Omega \to \mathcal{X} : \mathbb{E}\|X\|^2 < \infty\}$ and $\mathcal{Y}' := L^2(\Omega, \mathcal{Y})$

are also Hilbert spaces, with scalar products $(X, \tilde{X})_{\mathcal{X}'} := \sqrt{\mathbb{E}(X, \tilde{X})_{\mathcal{X}}}$ and $(\cdot, \cdot)_{\mathcal{Y}'}$ defined similary. Then $K : \mathcal{X} \to \mathcal{Y}$ induces naturally a bounded linear operator $K' : \mathcal{X}' \to \mathcal{Y}', X \mapsto KX$. Clearly we have that $\hat{y} \in \mathcal{Y}'$, and $(\bar{Y}_n)_n$ is a sequence in $\mathcal{Y}'$ which fulfills

$$\|\bar{Y}_n - \hat{y}\|_{\mathcal{Y}'} := \sqrt{(\bar{Y}_n - \hat{y}, \bar{Y}_n - \hat{y})_{\mathcal{Y}'}} = \sqrt{\frac{\mathbb{E}\|Y_1 - \hat{y}\|^2}{n}} = \sqrt{\mathbb{E}\|Y_1 - \hat{y}\|^2}\delta_n^{est}$$

and we can use the classic deterministic results for $K' : \mathcal{X}' \to \mathcal{Y}'$ and $\bar{Y}_n$ and δ_n^{est}.

1.2 The discrepancy principle

In this section we restrict to compact operators with dense range. Note that then $\mathcal{Y} = \overline{\mathcal{R}(K)}$ is separable. In practice the above parameter choice strategies are of limited interest, since they require the knowledge of the abstract smoothness parameters ν and ρ. Here, the classical discrepancy principle (0.6) would be to choose α_n such that

$$\|(KR_{\alpha_n} - Id)\bar{Y}_n\| \approx \delta_n^{true} = \|\bar{Y}_n - \hat{y}\|, \tag{1.1}$$

which is not possible, because of the unknown δ_n^{true}. So we replace it with our estimator δ_n^{est} and implement the discrepancy principle via Algorithm 1 with or without the optional emergency stop.

Algorithm 1 Discrepancy principle with estimated data error (optional: with emergency stop)

```
Given measurements $Y_1, ..., Y_n$;
Set $\bar{Y}_n := \sum_{i \le n} Y_i/n$ and $\delta_n^{est} = 1/\sqrt{n}$ or $\delta_n^{est} = \sqrt{\sum_{i \le n} \|Y_i - \bar{Y}_n\|^2/(n-1)}/\sqrt{n}$.
Choose a $q \in (0,1)$.
$k = 0$;
while $\|(KR_{q^k} - Id)\bar{Y}_n\| > \delta_n^{est}$ (optional: and $q^k > 1/n$) do
    $k = k + 1$;
end while
$\alpha_n = q^k$;
```

Remark 1.2.1. To our knowledge, the idea of an emergency stop first appeared in [BM12]. It provides a deterministic lower bound for the regularisation parameter, which may avoid over fitting. We use an elementary form of an emergency stop here, which does not require the knowledge of the singular value decomposition of K. It would be interesting to see, how more sophisticated versions of the emergency stop worked here, which is not clear to us since in our general setting we cannot rely on the concentration properties of Gaussian noise.

Algorithm 1 will terminate, if we use the emergency stop. Otherwise, we can guarantee that Algorithm 1 terminates, if K has dense image (or equivalently, if K^* is injective) and if $\delta_n^{est} > 0$. This is because then $\lim_{\alpha\to 0} KR_\alpha = P_{\overline{\mathcal{R}(K)}} = Id$ pointwise, so $\|(KR_{q^k} - Id)\bar{Y}_n\| < \delta_n^{est}$ for k large enough . If we decided to use the sample variance, it may happen that $\delta_n^{est} = 0$. But assuming $\mathbb{E}\|Y_1 - \hat{y}\|^2 > 0$, it follows that $\mathbb{P}\left(\delta_n^{est} = 0\right) = \mathbb{P}\left(Y_1 = ... = Y_n\right) \to 0$ for $n \to \infty$ (with exponential rate). If the distribution of Y_1 possess a density (with respect to the Gaussian measure for example), then actually $\mathbb{P}(Y_1 = ... = Y_n) = 0$ for all $n \in \mathbb{N}$.
Unlike in the previous section, here the L^2 error will not converge in general, even if Y_1 has a density. The regularisation parameter α_n is now random, since it depends on the potentially bad random data. With a diminishing probability p we are underestimating the data error significantly, and thus the discrepancy principle gives a too small α and we still have $p\|R_\alpha\| \gg 1$ in such a case.
In the following we will need the singular value decomposition of the compact operator K with dense range (see [Cav11]): there exists a monotone sequence $\|K\| = \sigma_1 \geq \sigma_2 \geq ... > 0$ with $\sigma_l {\to} 0$ for $l \to \infty$. Moreover there are families of orthonormal vectors $(u_l)_{l\in\mathbb{N}}$ and $(v_l)_{l\in\mathbb{N}}$ with $span(u_l : l \in \mathbb{N}) = \mathcal{Y}$, $span(v_l : l \in \mathbb{N}) = \mathcal{N}(K)^\perp$ such that $Kv_l = \sigma_l v_l$ and $K^* u_l = \sigma_l v_l$.

1.2.1 A counter example for convergence

We now show that a naive use of the discrepancy principle, as implemented in Algorithm 1 without emergency stop, may fail to converge in L^2. To simplify calculations we pick Gaussian noise and the truncated singular value regularisation and we set $\delta_n^{est} = 1/\sqrt{n}$. We choose $\mathcal{X} := l^2(\mathbb{N})$ with the standard basis $\{u_k := (0, ..., 0, 1, 0, ...)\}$ and consider the diagonal operator

$$K : l^2(\mathbb{N}) \to l^2(\mathbb{N}), \quad u_l \mapsto \left(\frac{1}{100}\right)^{\frac{l}{2}} u_l$$

with $\hat{x} = 0 = \hat{y} = K\hat{x}$. Hence the $\sigma_l = (1/100)^{\frac{l}{2}}$ are the eigenvalues of K and

$$R_\alpha : l^2(\mathbb{N}) \to l^2(\mathbb{N}), \quad y \mapsto \sum_{l:\sigma_l^2 \geq \alpha} \sigma_l^{-1}(y, u_l)u_l.$$

We assume that the noise is distributed along $y := \sum_{l\geq 2} 1/\sqrt{l(l-1)}u_l$, so we have that $\sum_{l>n}(y, u_l)^2 = 1/n$ and thus $y \in l^2(\mathbb{N})$. That is we set $\bar{Y}_n := \sum_{i\leq n} Y_i = \sum_{i\leq n} Z_i y$, where Z_i are i.i.d. standard Gaussians. We define $\Omega_n := \{Z_i \geq 1, i = 1...n\}$, a (very unlikely) event on which we significantly underestimate the true data error. We get that $\mathbb{P}(\Omega_n) := \mathbb{P}(Z_1 \geq 1)^n \geq 1/10^n$. Moreover, by the definition of the discrepancy principle

$$\begin{aligned}
\frac{1}{n}\chi_{\Omega_n} = \delta_n^{est\,2}\chi_{\Omega_n} &\geq \|(KR_{\alpha_n} - Id)\bar{Y}_n\|^2\chi_{\Omega_n} = |\bar{Z}_n|^2\|(KR_{\alpha_n} - Id)y\|^2\chi_{\Omega_n} \\
&\geq \|(KR_{\alpha_n} - Id)y\|^2\chi_{\Omega_n} \\
&= \sum_{l:\sigma_l^2<\alpha_n} (y, u_l)^2\chi_{\Omega_n} = \sum_{l:(1/100)^i<\alpha_n} (y, u_l)^2\chi_{\Omega_n} \\
&= \sum_{l>\frac{\log(\alpha_n)}{\log(1/100)}} (y, u_l)^2\chi_{\Omega_n} \geq \frac{\log(1/100)}{\log(\alpha_n)}\chi_{\Omega_n} \\
\Longrightarrow \alpha_n\chi_{\Omega_n} &< \frac{1}{100^n}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 &= \mathbb{E}\|R_{\alpha_n}\bar{Y}_n\|^2 \geq \mathbb{E}\|R_{\alpha_n}\bar{Y}_n\chi_{\Omega_n}\|^2 \\
&= \mathbb{E}\left[\bar{Z}_n^2\|R_{\alpha_n}y\chi_{\Omega_n}\|\right]^2 \geq \mathbb{E}\|R_{1/100^n}y\chi_{\Omega_n}\|^2 \\
&\geq \sum_{l:\sigma_i^2\geq 1/100^n} \sigma_l^{-2}(y, u_l)^2\mathbb{P}(\Omega_n) \geq \frac{1}{10^n}\sum_{l\leq n}\sigma_l^{-2}(y, u_l)^2 \\
&\geq \frac{1}{10^n}100^n(y, u_n)^2 = \frac{10^n}{n(n-1)} \to \infty.
\end{aligned}$$

That is the probability of the events Ω_n is not small enough to compensate the huge error we have on these events, so in the end $\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 \to \infty$ for $n \to \infty$.

1.2.2 Convergence in probability of the discrepancy principle

In this section we show, that the discrepancy principle yields convergence in probability, matching the optimal deterministic rate with growing probability. The proofs of the Theorems 1.2.2 and 1.2.4 and of Corollary 1.2.5 are moved to section 1.6.

Theorem 1.2.2 (Convergence of the discrepancy principle)**.** *Assume that K is a compact operator with dense range between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and that $Y_1, Y_2, ...$ are i.i.d. $\mathcal{Y}-$valued random variables with $\mathbb{E}Y_1 = \hat{y} \in \mathcal{R}(K)$ and $0 < \mathbb{E}\|Y_1 - \hat{y}\|^2 < \infty$. Let R_α be induced by a filter fulfilling Assumption 1.1.2 with $\nu_0 > 1$. Applying Algorithm 1 with or without the emergency stop yields a sequence $(\alpha_n)_n$. Then we have that for all $\varepsilon > 0$*

$$\mathbb{P}\left(\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\| \leq \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 1,$$

i.e. $R_{\alpha_n}\bar{Y}_n \xrightarrow{\mathbb{P}} K^+\hat{y}$.

Remark 1.2.3. If one tried to argue as in Remark 1 to show L^2 convergence one would have to determine the regularisation parameter not as given by equation (1.1), but such that $\mathbb{E}\|(KR_\alpha - Id)\bar{Y}_n\|^2 \approx \delta_n^{est}$, which is not practicable since we cannot calculate the expectation on the left hand side.

The popularity of the discrepancy principles is a result of the fact that it guarantees optimal convergence rates under an additional source condition: Assuming that there is a $0 < \nu \leq \nu_0 - 1$ (where ν_0 is the qualification of the chosen regularisation method) such that $K^+\hat{y} = (K^*K)^{\frac{\nu}{2}}w$ for a $w \in \mathcal{X}$ with $\|w\| \leq \rho$, then

$$\sup_{y^\delta:\|y^\delta-\hat{y}\|\leq\delta} \|R_{\alpha(y^\delta,\delta)}y^\delta - K^+\hat{y}\| \leq C\rho^{\frac{1}{\nu+1}}\delta^{\frac{\nu}{\nu+1}} \tag{1.2}$$

for some constant $C > 0$. The next theorem shows a concentration result for the discrepancy principle as implemented in Algorithm 1, where the deterministic bound δ in (1.2) is replaced with $1/\sqrt{n}$.

Theorem 1.2.4 (Rate of convergence of the discrepancy principle)**.** *Assume that K is a compact operator with dense range between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. Moreover, $Y_1, Y_2, ...$ are i.i.d. $\mathcal{Y}-$valued random variables with $\mathbb{E}Y_1 = \hat{y} \in \mathcal{R}(K)$ and $0 < \mathbb{E}\|Y_1 - \hat{y}\|^2 < \infty$. Let R_α be induced by a filter fulfilling Assumption 1.1.2 with $\nu_0 > 1$. Moreover, assume that there is a $0 < \nu \leq \nu_0 - 1$ and a $\rho > 0$ such that $K^+\hat{y} = (K^*K)^{\nu/2}\xi$ for some $\xi \in \mathcal{X}$ with $\|\xi\| \leq \rho$. Applying Algorithm 1 with or without the emergency stop yields a sequence $(\alpha_n)_{n\in\mathbb{N}}$. Then there is a constant L, such that*

$$\mathbb{P}\left(\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\| \leq L\rho^{\frac{1}{\nu+1}}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu}{\nu+1}}\right) \xrightarrow{n\to\infty} 1.$$

The ad hoc emergency stop $\alpha_n > 1/n$, additionally assures, that the L^2 error will not explode (unlike in the counter example of the previous subsection). Under the assumption that $\mathbb{E}\|Y_1 - \hat{y}\|^4 < \infty$, one can guarantee, that the L^2 error will converge.

Corollary 1.2.5. *Under the assumptions of Theorem 1.2.2, consider the sequence α_n determined by Algorithm 1 with emergency stop. Then there is a constant C such that $\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 \leq C$ for all $n \in \mathbb{N}$. If additionally $\mathbb{E}\|Y_1 - \hat{y}\|^4 < \infty$, then it holds that $\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 \to 0$ for $n \to \infty$.*

Remark 1.2.6. It would be interesting to quantify in Theorem 1.2.4 how fast the probability converges to 1, which would eventually allow to determine confidence balls. Such so called uncertainty quantification has recently been successfully applied to inverse problems, see e.g. [Ten17], [Bar18] or [BZAJ20]. The main challenge will probably lie in quantifying a lower bound for α_n (and the probability that such a bound holds). One could avoid this in weaken the bound on $\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|$ slightly from $(1/\sqrt{n})^{\frac{\nu}{\nu+1}}$ to $\max\left(\delta_n^{est\,\frac{\nu}{\nu+1}}, \left(\frac{\delta_n^{true}}{\delta_n^{est}}\right)^{\frac{1}{\nu+1}} \delta_n^{true\,\frac{\nu}{\nu+1}}\right)$ as in Theorem 4 of [HJP20a].

1.3 Optimality

The convergence rates given in Theorem 1.1.4 and 1.2.4 depend on the smoothness of the true solution $\hat{x}$ relative to the operator K and were compared to the optimal deterministic convergence rate one would obtain for an arbitrary sequence y_n converging to $\hat{y} = K\hat{x}$ with noise level $\delta_n = \delta_n^{true} \sim 1/\sqrt{n}$ (i.e. $\|y_n - \hat{y}\| \leq \delta_n$). Since with the average $\bar{Y}_n$ one has a rather specific sequence converging to $\hat{y}$, the question naturally arises, if one could actually obtain a better convergence rate than the deterministic one. After fixing a specific regularisation, we therefore define the minimal risk

$$\inf_{\alpha>0} \sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_\alpha\bar{Y}_n - K^+\hat{y}\|^2.$$

and the oracle (if it exists)

$$\alpha_n^o := \arg\min_{\alpha>0} \sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_\alpha\bar{Y}_n - K^+\hat{y}\|^2,$$

as that parameter choice, which minimises the L^2 error under all possible parameters. So called oracle inqualities, which link the error obtained by an estimator to the one of the oracle are an universal tool to prove minmax properties [DJ94], [Can06],[CGP$^+$02]. We illustrate in the following, that the minimal rate attained by the oracle is usually better than the deterministic rate given in Theorem 1.1.4, a possibility which was already noted in [BR08]. We restrict to the singular value decomposition as a regularisation and to mildly ill-posed problems. The notation $\sigma_j^2 \asymp j^{-q}$ means that there exist constants c_q, C_q with $c_q j^{-q} \leq \sigma_j^2 \leq C_q j^{-q}$ for all $j \in \mathbb{N}$. The proofs are deferred to Section 1.6.

Theorem 1.3.1. *Assume that K is a compact operator (with singular value decomposition (σ_j, u_j, v_j)) between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and that $Y_1, Y_2, ...$ are i.i.d. $\mathcal{Y}-$valued random variables with $\mathbb{E}[Y_1] = \hat{y} \in \mathcal{R}(K)$. Moreover, assume that there are $q > 0$ and $p > 1$ such that $\sigma_j^2 \asymp j^{-q}$ and $\mathbb{E}(Y_1 - \hat{y}, u_j)^2 \asymp j^{-p}$ and there are $\nu, \rho > 0$ such that $K^+\hat{y} = (K^*K)^{\nu/2}\xi$ for some $\xi \in \mathcal{X}$ with $\|\xi\| \leq \rho$. Then for R_α the truncated singular value decomposition it holds that*

$$\inf_{\alpha>0} \sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_\alpha \bar{Y}_n - K^+\hat{y}\|^2 \asymp \begin{cases} \frac{1}{n} & q-p<-1 \\ \frac{\log(\rho n)}{n} & q-p=-1 \\ \rho^{\frac{q+1-p}{(\nu+1)q+1-p}} \left(\frac{1}{n}\right)^{\frac{\nu}{\nu+1-\frac{p-1}{q}}} & q-p>-1 \end{cases}.$$

In particular, for the a priori choice

$$\alpha_n \asymp \begin{cases} (\rho n)^{-\frac{1}{\nu}} & q-p\leq -1 \\ (\rho n)^{-\frac{1}{(1+\nu)q+1-p}} & q-p>-1 \end{cases}$$

it holds that

$$\sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_{\alpha_n} \bar{Y}_n - K^+\hat{y}\|^2 \asymp \inf_{\alpha>0} \sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_\alpha \bar{Y}_n - K^+\hat{y}\|^2.$$

So we see that the optimal rate attained by the oracle is in all cases better than the optimal deterministic rate (since for $q-p>-1$ it holds that $\frac{p-1}{q}\in(0,1)$ because of $p>1$). In particular for $q-p<-1$ the problem is in fact wellposed. However, the above optimal choice of α depends on the in general unknown relative smoothness ν of $\hat{x}$ and on the decay of the variances p. At least the latter may be estimated with multiple measurements, so we assume for the moment that we know a $p>1$ with $\mathbb{E}(Y_1-\hat{y},u_j)^2 \asymp j^{-p}$, but the smoothness ν is unknown. We consider the discrepancy principle, and the general idea is to rescale the measurements in order to improve the relative smoothness. So we define

$$\begin{aligned} S :&\mathcal{D}(S) \to \mathcal{Y} \\ &u_i \mapsto j^{\frac{r}{2}} u_i \end{aligned} \tag{1.3}$$

for $r<\min(p-1,q)$ and apply the discrepancy principle to the rescaled operator $SK:\mathcal{X}\to\mathcal{Y}$ and the rescaled measurements $SY_1, SY_2, ...$

$$\|SKR'_\alpha S\bar{Y}_n - S\bar{Y}_n\| \approx \delta_n^{est},$$

with $\delta_n^{est} := \frac{1}{\sqrt{n}}$ and R'_α the truncated singular value decomposition for SK. The restriction on r guarantees, that the rescaled measurements still have finite variances and the rescaled operator is compact.

Theorem 1.3.2. *Under the assumptions of Theorem 1.3.1 and the additional assumption that K has dense range, let S be given by* (1.3) *for* $r<\min(p-1,q)$*. Then* *for R'_α the truncated singular value decomposition for SK and α_n the output of the discrepancy principle as implemented in Algorithm 1 (for $K = SK$ and $Y_i = SY_i$) it holds that*

$$\mathbb{P}\left(\|R'_{\alpha_n} S\bar{Y}_n - K^+\hat{y}\| \leq L\rho^{\frac{q-r}{q(\nu+1)-r}} \left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu}{\nu+1-\frac{r}{q}}}\right) \to 1$$

as $n \to \infty$ for some $L > 0$.

So we see that, with the choice $r \to \min(p-1, q)$ we in essence recover the optimal rate from Theorem 1.3.1. However, the larger r, the slower will be the convergence of the probability to 1.

Of course, in application we do not ad hoc know the decay rate of the variances. We propose the following algorithm for the implementation of a modified rescaled discrepancy principle where we also estimate the decay of the variances. In contrast to Algorithm 1 we only consider finitely many components for fixed n.

Algorithm 2 Modified discrepancy principle with estimated data error

```
Given measurements $Y_1, ..., Y_n$ with $m_n := \lfloor n^{1-\varepsilon_1} \rfloor$;
Set $s_{j,n}^2 := \frac{1}{n-1}\sum_{i=1}^n \left(Y_1 - \bar{Y}_n, u_j\right)^2$ for $j = 1, ..., m_n$;
Set $d_{j,n} := \sqrt{\min\left(\frac{j^{-(1+\varepsilon_2)}}{s_{j,n}^2}, \sigma_j^{-2}\right)\sum_{j'=1}^{m_n} s_{j',n}^2}$;
Set $\delta_n^{est} := \sqrt{\frac{\sum_{j=1}^{m_n} d_{j,n}^2 s_{j,n}^2}{n}}$;
$k = 0$;
while $\sum_{j=k+1}^{m_n} d_{j,n}^2 \left(\bar{Y}_n, u_j\right)^2 > \delta_n^{est}$ do
    $k = k + 1$;
end while
$k_n = k$;
$\bar{X}_n := \sum_{j=1}^{k_n} \frac{\left(\bar{Y}_n, u_j\right)}{\sigma_j} u_j$;
```

In order to guarantee simultaneous estimation of the component variances, we slightly strengthen our assumption on the error distribution.

Theorem 1.3.3. *Assume that K is a compact operator with dense range between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ and that $Y_1, Y_2, ...$ are i.i.d. $\mathcal{Y}-$valued random variables with $\mathbb{E}Y_1 = \hat{y} \in \mathcal{R}(K)$. Moreover, assume that there are $p > 1$ and $q > p - 1$ and $C_d \geq 1$ such that $\sigma_j^2 \asymp j^{-q}$ and $\mathbb{E}(Y_1 - \hat{y}, u_j)^2 \asymp j^{-p}$ and $\sup_{j \in \mathbb{N}} \frac{\mathbb{E}[(Y_1 - \hat{y}, u_j)^4]}{(\mathbb{E}[(Y_1 - \hat{y}, u_j)^2])^2} \leq C_d$. Finally, assume that there are $\nu, \rho > 0$ such that $K^+\hat{y} = (K^*K)^{\nu/2}\xi$ for some $\xi \in \mathcal{X}$ with $\|\xi\| \leq \rho$ and $(\xi, v_j) \neq 0$ for infinitely many $j \in \mathbb{N}$. For $\varepsilon_1, \varepsilon_2 > 0$ let $\bar{X}_n$ be the output of the rescaled discrepancy principle implemented with Algorithm 2. Then there is a $L > 0$ with*

$$\mathbb{P}\left(\|\bar{X}_n - K^+\hat{y}\| \leq L \max\left(\rho^{\frac{q+1+\varepsilon_2-p}{\nu q+q+1+\varepsilon_2-p}}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu}{\nu+1-\frac{p-1-\varepsilon_2}{q}}}, \rho\left(\frac{1}{\sqrt{n}}\right)^{2(1-\varepsilon_1)q\nu}\right)\right) \to 1$$

as $n \to \infty$.

Remark 1.3.4. Clearly, Algorithm 2 could be applied in a general setting. A similar result to the one of Theorem 1.3.3 could be obtained under relaxed assumptions, e.g. if only $C_{q'}j^{-q'} \leq \sigma_j^2 \leq C_q j^{-q}$ and $\mathbb{E}(Y_1 - \hat{y}, u_j)^2 \leq C_p j^{-p}$ for some $C_q, C_{q'}, C_p > 0$ and $q' \leq q$. Moreover, one could omit to assume $(\xi, v_j) \neq 0$ for infinitely many $j \in \mathbb{N}$ and $q > p - 1$.

The second argument in the maximum is a discretisation error. If the latter is negligible, i.e. if $\frac{\nu}{\nu+1} < 2(1-\varepsilon)q\nu$, the rate from Theorem 1.3.3 is better than the one from Theorem 1.2.4 for $\varepsilon_2 < p - 1$ (if we only consider $\lceil n^{1-\varepsilon} \rceil$ components there as well). The additional assumption $\sup_{j\in\mathbb{N}} \frac{\mathbb{E}[(Y_1-\hat{y},u_j)^4]}{(\mathbb{E}[(Y_1-\hat{y},u_j)^2])^2} < \infty$ assures that the component distribution are not too degenerated. This is clearly fulfilled, if $\mathbb{E}(Y_1 - \hat{y}, u_j) \overset{d}{=} c_j Z$ for some Z with $\mathbb{E}[Z] = 0$, $E[Z^4] < \infty$ and $(c_j)_{j\in\mathbb{N}} \subset \mathbb{R} \setminus \{0\}$ (e.g. this holds under Gaussian noise). In particular no independence between the components is required.

In 1.7.1.1 we consider a small example to confirm numerically, that the above approach can significantly reduce the error of our approximation.

1.4 Connection to heuristic regularisation

We briefly discuss the relation to so called heuristic parameter choice rules. These rules $\alpha = \alpha(y^\delta)$ depend only on the data. The term heuristic indicates that even though they might perform remarkably well in practical applications, such rules will not converge under general (deterministic) noise, as stated by the Bakushinskii veto 0.0.1. Thus rigorous convergence results for heuristic parameter choice rules are possible only under a noise-restricted analysis, see e.g. ([Neu08],[KN08],[KPJP18]). In the aforementioned articles the noise is assumed to fulfill a Mouckenhoupt condition [AM90], i.e. there is a constant $C > 0$ such that for all $k \in \mathbb{N}$ it holds that

$$\sigma_k^4 \sum_{j=1}^{k} \sigma_j^{-2}\left(y^\delta - \hat{y}, u_j\right)^2 \leq C \sum_{j=k+1}^{\infty} \sigma_j^2 \left(y^\delta - \hat{y}, u_j\right)^2.$$

In our case, an interesting question is whether $\bar{Y}_n - \hat{y}$ fulfills the Mouckenhoupt condition, which would allow to directly transfer results for heuristic strategies. So far, there are some result on the validness of the Mouckenhoupt condition under stochastic noise. In particular, Theorem 2 of [KPJP18] states, that the Mouckenhoupt condition holds true with probability 1, if K is mildly ill-posed, i.e. $\sigma_j \asymp j^{-q}$ for some $q > 0$ and if $\mathbb{E}(Y_1 - \hat{y}, u_j)^2 \asymp j^{-p}$ for $p > 1$, and the $(Y_1 - \hat{y}, u_j)_{j\in\mathbb{N}}$ are independent and have infinitely many moments. It is also shown in Theorem 4 of the aforementioned paper, that the Mouckenhoupt condition is not fulfilled with probability 1, if K is severely ill-posed (i.e. $\sigma_j \asymp a^j$ with $a \in (0,1)$) even under Gaussian noise. We now give a counter example (with non independent components) showing that the Mouckenhoupt condition does not hold true in general under our noise model, may it be in the mildly or severely ill-posed case.

1.4.1 Counter example for validity of the Mouckenhoupt condition

Assume that

$$Y_i - \hat{y} \overset{d}{=} \sum_{j=1}^{\infty} B_i a_j u_j \chi_{\{Z_i=j\}}$$

where $(B_i)_{i\in\mathbb{N}}$ are i.i.d. Bernoulli random variables (i.e. $\mathbb{P}\left(B_i = \pm 1\right) = 1/2$) independent from the i.i.d. $\mathbb{N}$-valued random variables $(Z_i)_{i\in\mathbb{N}}$. Let $p_j := \mathbb{P}\left(Z = j\right)$ and $a_j = \sqrt{j^{-p}/p_j}$ with $p > 1$. Then it holds that

$$\mathbb{E}\left[Y_1 - \hat{y}\right] = \sum_{j=1}^{\infty} \mathbb{E}[B_i] a_j u_j p_j = 0$$

and

$$\mathbb{E}\left(Y_1 - \hat{y}, u_k\right)^2 = \sum_{j=1}^{\infty} p_j \mathbb{E}\left[B_i^2\right] a_j^2 \left(u_k, u_j\right)^2 = p_k \mathbb{E}[B_i^2] a_k^2 = \mathbb{E}[B_i^2] k^{-p}$$

$$\Longrightarrow \mathbb{E}\|Y_1 - \hat{y}\|^2 = \mathbb{E}[B_i^2] \sum_{k=1}^{\infty} k^{-p} < \infty.$$

Thus $Y_1, Y_2, ...$ are i.i.d unbiased measurements of $\hat{y}$ with finite variance. However, for all $n \in \mathbb{N}$ it holds that

$$\mathbb{P}\left(\max(Z_1, ..., Z_n) < \infty\right) = 1$$

which implies

$$\mathbb{P}\left((Y_1 - \hat{y}, u_j) = 0, ..., (Y_n - \hat{y}, u_j) = 0, \ \forall j \text{ large enough}\right) = 1.$$

From that we deduce that

$$\mathbb{P}\left((\bar{Y}_n - \hat{y}, u_j) = 0, \ \forall j \text{ large enough}\right) = 1$$

and hence

$$\mathbb{P}\left(\sup_{k \in N} \frac{\sigma_k^4 \sum_{j=1}^k \sigma_j^{-2} \left(\bar{Y}_n - \hat{y}, u_j\right)^2}{\sum_{j=k+1}^\infty \sigma_j^2 \left(\bar{Y}_n - \hat{y}, u_j\right)^2} = \infty\right) = 1,$$

thus the Mouckenhoupt condition is violated with probability 1.

1.5 Almost sure convergence

The results so far delivered either convergence in probability or convergence in L^2. We give a short remark how one can obtain almost sure convergence. Roughly speaking, one has to multiply a $\sqrt{\log\log n}$ term to δ_n^{est}. This is a simple consequence of the following theorem

Theorem 1.5.1 (Law of the iterated logarithm)**.** *Assume that $Y_1, Y_2, ...$ is an i.i.d sequence with values in some separable Hilbert space $\mathcal{Y}$. Moreover, assume that $\mathbb{E}Y_1 = 0$ and $\mathbb{E}\|Y_1\|^2 < \infty$. Then we have that*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{\|\sum_{i \le n} Y_i\|}{\sqrt{2\mathbb{E}\|Y_1\|^2 n \log\log n}} \le 1\right) = 1.$$

Proof. This is a simple consequence of Corollary 8.8 in [LT91]. □

So if $\mathbb{E}Y_1 = \hat{y} \in \mathcal{Y}$ we have for $\delta_n^{true} = \|\bar{Y}_n - \hat{y}\|$

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{\sqrt{n}\delta_n^{true}}{\sqrt{2\mathbb{E}\|Y_1 - \hat{y}\|^2 \log\log n}} \le 1\right) = 1,$$

that is, with probability 1 it holds that $\delta_n^{true} \le \sqrt{\frac{2\mathbb{E}\|Y_1 - \hat{y}\|^2 \log\log n}{n}}$ for n large enough. Consequently, for some $\tau > 1$ the estimator should be

$$\delta_n^{est} := \tau s_n \sqrt{\frac{2\log\log n}{n}},$$

where s_n is the square root of the sample variance. Since $\mathbb{P}(\lim_{n\to\infty} s_n^2 = \mathbb{E}\|Y_1 - \hat{y}\|^2) = 1$ and $\tau > 1$ it holds that $\sqrt{\mathbb{E}\|Y_1 - \hat{y}\|} \leq \tau s_n$ for n large enough with probability 1 and thus $\delta_n^{true} \leq \delta_n^{est}$ for n large enough with probability 1. In other words, there is an event $\Omega_0 \subset \Omega$ with $\mathbb{P}(\Omega_0) = 1$ such that for any $\omega \in \Omega_0$ there is a $N(\omega) \in \mathbb{N}$ with $\delta_n^{true}(\omega) \leq \delta_n^{est}(\omega)$ for all $n \geq N(\omega)$. So we can use $\bar{Y}_n$ and δ_n^{est} together with any deterministic regularisation method to get almost sure convergence.

1.6 Proofs

1.6.1 Proofs of Theorem 1.2.2 and 1.2.4

Throughout this thesis, we will steadily use without pointing out the following basic facts, i.e.

$$\begin{aligned} A \subset B \quad &\Rightarrow \quad \mathbb{P}(A) \leq \mathbb{P}(B) \quad \text{(monotinicity)} \\ A^C := \Omega \setminus A \quad &\Rightarrow \mathbb{P}(A^C) = 1 - \mathbb{P}(A) \quad \text{(normed to 1)} \\ \mathbb{P}\left(\cup_{i\in\mathbb{N}} A_i\right) &\leq \sum_{i\in\mathbb{N}} \mathbb{P}(A_i) \quad (\sigma\text{-subadditivity}) \\ \left(\cap_{i\in\mathbb{N}} A_i\right)^C &= \cup_{i\in\mathbb{N}} A_i^C \quad \text{(law of de Morgan)} \end{aligned}$$

for arbitrary events $A, B, A_i \in \mathcal{A}$ (where $\mathcal{A}$ is the σ-Algebra on Ω). We will multiple times use the Pythagorean theorem for independent centralised random variables. For real-valued random variables X_i with $\mathbb{E}[X_i^2] < \infty$ and $\mathbb{E}[X_i] = 0$ there holds

$$\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] = \sum_{i,i'=1}^n \mathbb{E}[X_i X_{i'}] = \sum_{i=1}^n \mathbb{E}[X_i^2] + \sum_{\substack{i,i'=1 \\ i\neq i'}}^n \mathbb{E}[X_i]\mathbb{E}[X_{i'}] = \sum_{i=1}^n \mathbb{E}[X_i^2]. \tag{1.4}$$

We deduce, that for separable Hilbert space valued random variables Z_i with $\mathbb{E}\|Z_i\|^2 < \infty$ and $\mathbb{E}Z_i = 0$ it holds that

$$\mathbb{E}\left\|\sum_{i=1}^n Z_i\right\|^2 = \sum_{i=1}^n \sum_{l,l'=1}^\infty \mathbb{E}\left[(Z_i, e_l)(Z_i, e_{l'})\right] = \sum_{i=1}^n \mathbb{E}\left[\sum_{j=1}^\infty (Z_i, e_j)^2\right] = \sum_{i=1}^n \mathbb{E}\left\|Z_i\right\|^2, \tag{1.5}$$

where $(e_l)_{l\in\mathbb{N}}$ is an orthonormal basis. Based on this, one central ingredient will be the following lemma, which strengthens the point wise worst case error bound $\|(KR_\alpha - Id)(\bar{Y}_n - \hat{y})\| \leq C_0 \delta_n^{true}$ in some sense.

Lemma 1.6.1. *For all $\varepsilon > 0$ and (deterministic) sequences $(q_n)_{n\in\mathbb{N}}$ with $q_n > 0$ and $\lim_{n\to\infty} q_n = 0$, it holds that*

$$\mathbb{P}\left(\sup_{0<\alpha\leq q_n} \|(KR_\alpha - Id)(\bar{Y}_n - \hat{y})\| \geq \varepsilon/\sqrt{n}\right) \to 0$$

and

$$\mathbb{P}\left(|\sqrt{n}\delta_n^{est} - \gamma| \geq \varepsilon\right) \to 0$$

for $n \to \infty$, where $\gamma = 1$ or $\gamma = \sqrt{\mathbb{E}\|Y_1 - \hat{y}\|^2}$, depending on if we used the sample variance or not.

Proof.

Let $\varepsilon' > 0$ be arbitrary and $J_{\varepsilon'}$ such that $\sum_{j=J_{\varepsilon'}}^{\infty} C_0^2\mathbb{E}(Y_1 - \hat{y}, u_j)^2 \leq \varepsilon'\varepsilon^2/2$. Then by Markov's inequality

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0<\alpha\leq q_n} \|(KR_\alpha - Id)(\bar{Y}_n - \hat{y})\| \geq \varepsilon/\sqrt{n}\right)\\
\leq&\frac{n}{\varepsilon^2}\mathbb{E}\left[\left(\sup_{0<\alpha\leq q_n} \|(KR_\alpha - Id)(\bar{Y}_n - \hat{y})\|\right)^2\right]\\
=&\frac{n}{\varepsilon^2}\mathbb{E}[\sup_{0<\alpha\leq q_n}\sum_{j=1}^{\infty}\left(F_\alpha(\sigma_j^2)\sigma_j^2 - 1\right)^2(\bar{Y}_n - \hat{y}, u_j)^2]\\
\leq&\frac{n}{\varepsilon^2}\sum_{j=1}^{\infty}\mathbb{E}\left[\sup_{0<\alpha\leq q_n}\left(F_\alpha(\sigma_j^2)\sigma_j^2 - 1\right)^2(\bar{Y}_n - \hat{y}, u_j)^2\right]\\
=&\frac{1}{\varepsilon^2}\sum_{j=1}^{\infty}\sup_{0<\alpha\leq q_n}\left(F_\alpha(\sigma_j^2)\sigma_j^2 - 1\right)^2\mathbb{E}(Y_1 - \hat{y}, u_j)^2\\
\leq&\frac{\mathbb{E}\|Y_1 - \hat{y}\|^2}{\varepsilon^2}\sum_{j=1}^{J_{\varepsilon'}}\sup_{0<\alpha\leq q_n}\left(F_\alpha(\sigma_j^2)\sigma_j^2 - 1\right)^2 + \frac{C_0^2}{\varepsilon^2}\sum_{j=J_{\varepsilon'}}^{\infty}\mathbb{E}(Y_1 - \hat{y}, u_j)^2 \leq \varepsilon'
\end{aligned}$$

for n large enough, where we used Tschebyscheff's inequality in the first, subadditivity of the supremum in the third, linearity of the expectation in the fourth and the point wise convergence of $F_\alpha(\lambda)$ to $1/\lambda$ in the last step. This proves the first assertion. The second assertion only needs a proof for $\gamma = \sqrt{\mathbb{E}\|Y_1 - \hat{y}\|^2}$ and then

$$n\delta_n^{est2} = \frac{1}{n-1}\sum_{i=1}^{n}\|Y_i - \bar{Y}_n\|^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}\|Y_i\|^2 - \|\bar{Y}_n\|^2\right)$$
$$\to \mathbb{E}\|Y_1\|^2 - \|\hat{y}\|^2 = \mathbb{E}\|Y_1 - \hat{y}\|^2 = \gamma^2$$

almost surely (thus in particular in probability) for $n \to \infty$ by the strong law of large numbers (Corollary 7.10 in [LT91]) and the bias-variance-decomposition. Therefore $\sqrt{n}\delta_n^{est} \to \gamma$ in probability for $n \to \infty$.

□

For convergence in probability it does not matter how large the error is on sets with diminishing probability and with Lemma 1.6.1 we will show, that the probability of certain 'good events' is 1 in the limit of infinitely many measurements.

We will also need some well known properties of regularisations defined by filters which fulfill Assumption 1.1.2. These are mostly easy modifications from [EHN96].

Proposition 1.6.2. *The constants in the following are defined as in Assumption 1.1.2. We assume, that K is bounded and linear with non-closed range. Assume that $(R_\alpha)_{\alpha>0}$ is induced by a regularising filter fulfilling $|F_\alpha(\lambda)| \leq C_F/\alpha$ for all $0 < \lambda \leq \|K\|^2$. Then*

$$\|R_\alpha\| \leq \sqrt{C_R C_F}/\sqrt{\alpha} \tag{1.6}$$
$$\|Id - KR_\alpha\| \leq C_0 \tag{1.7}$$

for all $\alpha > 0$, with $C_0 \geq 1$. If moreover, the filter has qualification $\nu_0 > 0$ and there is a $w \in \mathcal{X}$ with $\|w\| \leq \rho$ such that $K^+\hat{y} = (K^*K)^{\frac{\nu}{2}} w$ for some $0 < \nu \leq \nu_0$, then

$$\|R_\alpha\hat{y} - K^+\hat{y}\| \leq C_\nu \rho \alpha^{\nu/2} \tag{1.8}$$
$$\|R_\alpha\hat{y} - K^+\hat{y}\| \leq \|KR_\alpha\hat{y} - KK^+\hat{y}\|^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} \rho^{\frac{1}{\nu+1}} \tag{1.9}$$

for all $\alpha > 0$. If additionally, $\nu_0 \geq \nu + 1 > 1$, then

$$\|KR_\alpha\hat{y} - KK^+\hat{y}\| \leq C_{\nu+1}\rho\alpha^{\frac{\nu+1}{2}}. \tag{1.10}$$

Moreover, if K is compact, than for all $x \in \mathcal{X}$ it holds that

$$\lim_{\alpha\to 0}\|(KR_\alpha - Id)Kx\|/\sqrt{\alpha} = 0. \tag{1.11}$$

Proposition 1.6.2 (1.6) and (1.9) are shown in the proofs of Theorem 4.2 and Theorem 4.17 in [EHN96]. (1.8) and (1.10) are Theorem 4.3 in [EHN96]. (1.7)

follows directly from Assumption 1.1.2.
For (1.11) mimic the proof of Theorem 3.1.17 of [NP15] and set $\varepsilon > 0$. We fix L, such that $C_1^2 \sum_{l=L+1}^{\infty} (\hat{x}, v_j)^2 < \varepsilon$. Then

$$\begin{aligned}
\|(KR_\alpha - Id)K\hat{x}\|^2/\alpha = &\sum_{l=1}^{\infty} \left(F_\alpha(\sigma_l^2)\sigma_l^2 - 1\right)^2 \frac{\sigma_l^2}{\alpha} (\hat{x}, v_l)^2 \\
\leq \left(\sup_{\lambda>0} \lambda^{\frac{\nu_0}{2}} |F_\alpha(\lambda)\lambda - 1|\right)^2 &\|\hat{x}\|^2 \sum_{l=1}^{L} \frac{\sigma_l^{2(1-\nu_0)}}{\alpha} \\
&+ \left(\sup_{\lambda>0} \lambda^{\frac{1}{2}} |F_\alpha(\lambda)\lambda - 1|\right)^2 \frac{\sum_{l=L+1}^{\infty} (\hat{x}, v_j)^2}{\alpha} \\
\leq & C_{\nu_0}^2 L \sigma_L^{2(1-\nu_0)} \|\hat{x}\|^2 \alpha^{\nu_0 - 1} + C_1^2 \sum_{l=L+1}^{\infty} (\hat{x}, v_j)^2 < 2\varepsilon
\end{aligned}$$

for all $\alpha < \left(\varepsilon^{-1} C_{\nu_0}^2 L \sigma_L^{2(1-\nu_0)} \|\hat{x}\|^2\right)^{-\frac{1}{\nu_0 - 1}}$, therefore $\|(KR_\alpha - Id)Kx\|/\sqrt{\alpha} = 0$ for $\alpha \to 0$.

□

We will first consider the case without emergency stop. We will treat the two cases $(\hat{y}, u_l) \neq 0$ for infinitely many $l \in \mathbb{N}$ and $(\hat{y}, u_l) = 0$ for all $l \in \mathbb{N}$ sufficiently large separately.

Proposition 1.6.3. *Assume that $(\hat{y}, u_l) \neq 0$ for infinitely many $l \in \mathbb{N}$, then there is a (deterministic) sequence $(q_n)_{n\in\mathbb{N}}$ with $q_n \to 0^+$ and*

$$\mathbb{P}\left(\alpha_n \leq q_n\right) \to 1$$

as $n \to \infty$

Proof.

It suffices to show that $\mathbb{P}\left(\alpha_n \leq \varepsilon\right) \to 1$ as $n \to \infty$ for arbitrary $\varepsilon > 0$. Let $\varepsilon > 0$. Then there is a $L \in \mathbb{N}$ such that $(\hat{y}, u_L) \neq 0$ and $\left(F_{q^k}(\sigma_L^2)\sigma_L^2 - 1\right)^2 > 1/2$ for all $k \in \mathbb{N}_0$ with $q^k \geq \varepsilon$ (because the F_{q^k} are bounded and $\sigma_l \to 0$ for $l \to \infty$). Set

$$\Omega_n := \left\{|\sqrt{n}\delta_n^{est} - \gamma| < \gamma \ , \ (\bar{Y}_n, u_L)^2 \geq (\hat{y}, u_L)^2/2\right\}. \tag{1.12}$$

Then for $n \geq 16\gamma^2/(\hat{y}, u_L)^2$,

$$\begin{aligned}\delta_n^{est}\chi_{\Omega_n} &\le \frac{2\gamma}{\sqrt{n}}\chi_{\Omega_n} < \sqrt{\frac{(\hat{y},u_L)^2}{4}}\chi_{\Omega_n} \le \sqrt{\left(F_{q^k}(\sigma_L^2)\sigma_L^2-1\right)^2(\bar{Y}_n,u_L)^2}\chi_{\Omega_n} \\ &\le \sqrt{\sum_{l=1}^{\infty}\left(F_{q^k}(\sigma_l^2)\sigma_l^2-1\right)^2\left(\bar{Y}_n,u_l\right)^2\chi_{\Omega_n}} = \|(KR_{q^k}-Id)\bar{Y}_n\|\chi_{\Omega_n}\end{aligned}$$

for all $k\in\mathbb{N}_0$ with $q^k\ge\varepsilon$. Thus for Ω_n given in (1.12)

$$\lim_{n\to\infty}\mathbb{P}\left(\alpha_n\le\varepsilon\right)\ge\lim_{n\to\infty}\mathbb{P}\left(\Omega_n\right)=1 \tag{1.13}$$

by Lemma 1.6.1 and since $(\bar{Y}_n,u_L)=\sum_{i=1}^n(Y_i,u_L)/n\to\mathbb{E}(Y_1,u_L)=(\hat{y},u_L)\ne 0$ almost surely for $n\to\infty$.

□

We are now ready for the central lemma.

Lemma 1.6.4. *Assume that $(\hat{y},u_l)\ne 0$ for infinitely many $l\in\mathbb{N}$, then there holds*

$$\mathbb{P}\left(\|(KR_{\alpha_n}-Id)(\bar{Y}_n-\hat{y})\|\le\delta_n^{est}/2,\ \|(KR_{\alpha_n/q}-Id)(\bar{Y}_n-\hat{y})\|\le\delta_n^{est}/2\right)\to 1$$

as $n\to\infty$.

Proof. By Proposition 1.6.3 there is a $(q_n)_{n\in\mathbb{N}}$ with $q_n\to 0$ and $\mathbb{P}\left(\alpha_n/q\le q_n\right)\to 1$ as $n\to\infty$. Therefore

$$\begin{aligned}&\mathbb{P}\left(\|(KR_{\alpha_n}-Id)(\bar{Y}_n-\hat{y})\|\le\frac{\delta_n^{est}}{2},\ \|(KR_{\alpha_n/q}-Id)(\bar{Y}_n-\hat{y})\|\le\frac{\delta_n^{est}}{2}\right)\\ \ge&\mathbb{P}\left(\sup_{0<\alpha\le q_n}\|(KR_\alpha-Id)(\bar{Y}_n-\hat{y})\|\le\frac{\gamma}{4\sqrt{n}},\ \alpha_n/q\le q_n,\ \delta_n^{est}>\frac{\gamma}{2\sqrt{n}}\right)\\ =&1-\mathbb{P}\left(\left(\sup_{0<\alpha\le q_n}\|(KR_\alpha-Id)(\bar{Y}_n-\hat{y})\|\le\frac{\gamma}{4\sqrt{n}},\ \alpha_n/q\le q_n,\ \delta_n^{est}>\frac{\gamma}{2\sqrt{n}}\right)^C\right)\\ \ge&1-\mathbb{P}\left(\sup_{0<\alpha\le q_n}\|(KR_\alpha-Id)(\bar{Y}_n-\hat{y})>\frac{\gamma}{4\sqrt{n}}\right)-\mathbb{P}\left(\alpha_n/q>q_n\right)\\ &-\mathbb{P}\left(\delta_n^{est}<\frac{\gamma}{2\sqrt{n}}\right)\to 1\end{aligned}$$

as $n\to\infty$ by Lemma 1.6.1. □

1.6.1.1 Proof of Theorem 1.2.4 without emergency stop

So let $\hat{x} = (K^*K)^{\nu/2}\xi$ with $\xi \in \mathcal{X}, \|\xi\| \leq \rho$. We first assume that $(\hat{y}, u_j) \neq 0$ for infinitely many $j \in \mathbb{N}$. Define

$$\Omega_n := \Big\{ \|(KR_{\alpha_n} - Id)(\bar{Y}_n - \hat{y})\| \leq \frac{\delta_n^{est}}{2},\ \|(KR_{\alpha_n/q} - Id)(\bar{Y}_n - \hat{y})\| \leq \frac{\delta_n^{est}}{2}, \tag{1.14}$$

$$|\sqrt{n}\delta_n^{est} - \gamma| \leq \frac{\gamma}{2},\ \alpha_n < 1 \Big\}. \tag{1.15}$$

There holds $\lim_{n\to\infty} \mathbb{P}(\Omega_n) = 1$. We decompose the total error in two parts

$$\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\| \leq \|R_{\alpha_n}(\bar{Y}_n - \hat{y})\| + \|R_{\alpha_n}\hat{y} - K^+\hat{y}\|$$

and restrict to Ω_n. For the approximation error, by (1.9), (1.7) and since K has dense image,

$$\begin{aligned}
&\|R_{\alpha_n}\hat{y} - K^+\hat{y}\|\chi_{\Omega_n} \\
\leq& \|KR_{\alpha_n}\hat{y} - KK^+\hat{y}\|^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} \rho^{\frac{1}{\nu+1}} \chi_{\Omega_n} \\
=& \|KR_{\alpha_n}\hat{y} - \hat{y}\|^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} \rho^{\frac{1}{\nu+1}} \chi_{\Omega_n} \\
\leq& \left(\|(KR_{\alpha_n} - Id)\bar{Y}_n\| + \|(KR_{\alpha_n} - Id)(\hat{y} - \bar{Y}_n)\|\right)^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} \rho^{\frac{1}{\nu+1}} \chi_{\Omega_n} \\
\leq& \left(\delta_n^{est} + \frac{\delta_n^{est}}{2}\right)^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} \rho^{\frac{1}{\nu+1}} \chi_{\Omega_n} \leq \left(\frac{9}{4}\gamma\right)^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} \rho^{\frac{1}{\nu+1}} \left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu}{\nu+1}}.
\end{aligned}$$

For the data propagation error we first bound α_n from below. By the defining relation of the discrepancy principle (note that Algorithm 1 does not terminate immediately on Ω_n),

$$\begin{aligned}
\delta_n^{est}\chi_{\Omega_n} &\leq \|(KR_{\alpha_n/q} - Id)\bar{Y}_n\|\chi_{\Omega_n} \\
&\leq \|(KR_{\alpha_n/q} - Id)\hat{y}\|\chi_{\Omega_n} + \|(KR_{\alpha_n/q} - Id)(\bar{Y}_n - \hat{y})\|\chi_{\Omega_n} \\
&\leq \rho C_{\nu+1}\left(\frac{\alpha_n}{q}\right)^{\frac{\nu+1}{2}} + \frac{\delta_n^{est}}{2} \\
\Longrightarrow & \alpha_n \geq q\left(\frac{\gamma}{\rho C_{\nu+1}4}\right)^{\frac{2}{\nu+1}} \chi_{\Omega_n} := b_n\chi_{\Omega_n}.
\end{aligned} \tag{1.16}$$

Set

$$L := \left(\frac{9}{4}\gamma\right)^{\frac{\nu}{\nu+1}} C_0^{\frac{1}{\nu+1}} + 1. \tag{1.17}$$

To finish the proof (for the case that $(\hat{y}, u_j) \neq 0$ for infinitely many $j \in \mathbb{N}$), it now suffices to show that there is a $C > 0$ such that

$$\mathbb{P}\left(\|R_{\alpha_n}\bar{Y}_n - R_{\alpha_n}\hat{y}\| \leq \rho^{\frac{1}{\nu+1}} \left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu}{\nu+1}} \right) \geq 1 - C\varepsilon \tag{1.18}$$

for all $\varepsilon > 0$, if $n \geq n(\varepsilon)$ large enough. We set $J_\varepsilon := \min\{j \in \mathbb{N} \;:\; \sum_{j' \geq j} \mathbb{E}(Y_1 - \hat{y}, u_j)^2 \leq \varepsilon$ and $Z := \sum_{j=J_\varepsilon+1}^{\infty}(Y_1 - \hat{y}, u_j)u_j$. Then

$$\begin{aligned}
&\mathbb{P}\left(\|R_{\alpha_n}(\bar{Y}_n - \hat{y})\| \leq \rho^{\frac{1}{\nu+1}} n^{-\frac{\nu}{2(\nu+1)}} \right)\\
=&\mathbb{P}\left(\sum_{j=1}^{\infty} \left(F_{\alpha_n}(\sigma_j^2)\sigma_j\right)^2 (\bar{Y}_n - \hat{y}, u_j)^2 \leq \rho^{\frac{2}{\nu+1}} n^{-\frac{\nu}{(\nu+1)}} \right)\\
=&\mathbb{P}\left(\sum_{j=1}^{J_\varepsilon} \left(F_{\alpha_n}(\sigma_j^2)\sigma_j\right)^2 (\bar{Y}_n - \hat{y}, u_j)^2 + \|R_{\alpha_n}Z\|^2 \leq \rho^{\frac{2}{\nu+1}} n^{-\frac{\nu}{(\nu+1)}} \right)\\
\geq&\mathbb{P}\left(\frac{C_R}{\sigma_{J_\varepsilon}^2} \sum_{j=1}^{J_\varepsilon} (\bar{Y}_n - \hat{y}, u_j)^2 + \frac{C_R C_F}{\alpha_n}\|Z\|^2 \leq \rho^{\frac{2}{\nu+1}} n^{-\frac{\nu}{(\nu+1)}} \right)\\
\geq&\mathbb{P}\left(\frac{C_R}{\sigma_{J_\varepsilon}^2} \sum_{j=1}^{J_\varepsilon} (\bar{Y}_n - \hat{y}, u_j)^2 + \frac{C_R C_F}{b_n}\|Z\|^2 \leq \rho^{\frac{2}{\nu+1}} n^{-\frac{\nu}{(\nu+1)}},\ \alpha_n \geq b_n \right)\\
\geq&1 - \mathbb{P}\left(\frac{C_R}{\sigma_{J_\varepsilon}^2} \sum_{j=1}^{J_\varepsilon} (\bar{Y}_n - \hat{y}, u_j)^2 + \frac{C_R C_F}{b_n}\|Z\|^2 > \rho^{\frac{2}{\nu+1}} n^{-\frac{\nu}{(\nu+1)}} \right) - \mathbb{P}\left(\alpha_n < b_n\right).
\end{aligned}$$

By Markov's inequality

$$
\begin{aligned}
&\mathbb{P}\left(\frac{C_R}{\sigma^2_{J_\varepsilon}}\sum_{j=1}^{J_\varepsilon}(\bar{Y}_n-\hat{y},u_j)^2+\frac{C_RC_F}{b_n}\|Z\|^2>\rho^{\frac{2}{\nu+1}}n^{-\frac{\nu}{(\nu+1)}}\right)\\
\le&\rho^{-\frac{2}{\nu+1}}n^{\frac{\nu}{\nu+1}}\mathbb{E}\left[\frac{C_R}{\sigma^2_{J_\varepsilon}}\sum_{j=1}^{J_\varepsilon}(\bar{Y}_n-\hat{y},u_j)^2+\frac{C_RC_F}{b_n}\|Z\|^2\right]\\
\le&\rho^{-\frac{2}{\nu+1}}n^{\frac{\nu}{\nu+1}}\left(\frac{C_R}{\sigma^2_{J_\varepsilon}n}\sum_{j=1}^{J_\varepsilon}\mathbb{E}(Y_1-\hat{y},u_j)^2\right.\\
&\qquad\left.+C_RC_F\left(\frac{4\sqrt{n}\rho C_{\nu+1}}{\gamma}\right)^{\frac{2}{\nu+1}}\frac{1}{n}\sum_{j=J_\varepsilon}^{\infty}\mathbb{E}(Y_1-\hat{y},u_j)^2\right)\\
\le&\rho^{-\frac{2}{\nu+1}}\frac{C_R}{\sigma^2_{J_\varepsilon}}n^{\frac{-1}{\nu+1}}+C_RC_F\left(\frac{4C_{\nu+1}}{\gamma}\right)^{\frac{2}{\nu+1}}\varepsilon\le\frac{C}{2}\varepsilon
\end{aligned}
$$

for n large enough and $C:=1+C_RC_F\left(\frac{2C_{\nu+1}}{\gamma}\right)^{\frac{2}{\nu+1}}$. Moreover by (1.16) and Lemma 1.6.1 $\mathbb{P}\left(\alpha_n\le b_n\right)\le\mathbb{P}\left(\Omega_n^C\right)\le C\varepsilon/2$ for n large enough which proves assertion (1.18).

Now we prove the assertion of Theorem 1.2.4 for the special case that $J:=\sup\left(j\in\mathbb{N}\,:\,(\hat{y},u_j)\neq 0\right)<\infty$. In this case we cannot expect a result similar to Lemma 1.6.4 (for example, α_n will not converge to 0 in probability for spectral cut-off), but the true solution $\hat{x}$ has arbitrarily large smoothness. Let $\varepsilon>0$ such that $\sigma_J^{-\varepsilon}\le 2$ and set $\nu'=\nu+\varepsilon$. Then

$$
\hat{x}=\sum_{j=1}^{J}\sigma_j^{\nu}(\xi,v_j)v_j=\sum_{j=1}^{J}\sigma_j^{\nu+\varepsilon}(\xi',v_j)v_j=(K^*K)^{\frac{\nu+\varepsilon}{2}}\xi'
$$

and

$$
\|\xi'\|=\sqrt{\sum_{j=1}^{J}(\xi,v_j)^2/\sigma_j^{2\varepsilon}}\le\sigma_J^{-\varepsilon}\rho\le 2\rho=\rho'.
$$

For the approximation error it is

$$
\begin{aligned}
&\|R_{\alpha_n}\hat{y} - K^+\hat{y}\| \\
=&\sqrt{\sum_{j=1}^{J}\left(F_{\alpha_n}(\sigma_j^2)\sigma_j - \frac{1}{\sigma_j}\right)^2 (\hat{y}, u_j)^2} = \sqrt{\sum_{j=1}^{J}\left(F_{\alpha_n}(\sigma_j^2)\sigma_j^2 - 1\right)^2 \frac{(\hat{y}, u_j)^2}{\sigma_j^2}} \\
\leq&\frac{1}{\sigma_J^2}\left(\sqrt{\sum_{j=1}^{J}\left(F_{\alpha_n}(\sigma_j^2)\sigma_j^2 - 1\right)^2 (\bar{Y}_n, u_j^2)} + \sqrt{\sum_{j=1}^{J}\left(F_{\alpha_n}(\sigma_j^2)\sigma_j^2 - 1\right)^2 (\bar{Y} - \hat{y}, u_j)^2}\right) \\
\leq&\frac{1}{\sigma_J^2}\left(\|(KR_{\alpha_n} - Id)\bar{Y}_n\| + \|(KR_{\alpha_n} - Id)(\bar{Y}_n - \hat{y})\|\right) \\
\leq&\frac{1}{\sigma_J^2}\left(\delta_n^{est} + C_0\delta_n^{true}\right).
\end{aligned}
$$

We now deduce a lower bound for the regularisation parameter and set $b_n := \left(\frac{1}{\rho'}\frac{\gamma}{4C_{\nu'+1}\sqrt{n}}\right)^{\frac{2}{\nu'+1}}$ with $\gamma = 1$ or $\gamma = \sqrt{\mathbb{E}\|Y_1 - \hat{y}\|^2}$, depending on if we used the sample variance or not. We claim that

$$
\mathbb{P}\left(\alpha_n \leq b_n\right) \to 1 \tag{1.19}
$$

as $n \to \infty$. Define

$$
\begin{aligned}
\Omega_n := \Big\{&|\sqrt{n}\delta_n^{est} - \gamma| < \gamma/2\ ,\quad \sup_{0<\alpha\leq b_n}\|(KR_\alpha - Id)(\bar{Y}_n - \hat{y})\| < \gamma/\sqrt{16n}, & (1.20)\\
&\delta_n^{est}, \delta_n^{true} \leq \sqrt{n}^{\varepsilon'-1}\Big\} & (1.21)
\end{aligned}
$$

for $\varepsilon' < \frac{\nu'}{\nu'+1} - \frac{\nu}{\nu+1}$. By (1.10)

$$
\begin{aligned}
\|(KR_\alpha - Id)\bar{Y}_n\|\chi_{\Omega_n} &\leq \|(KR_\alpha - Id)\hat{y}\|\chi_{\Omega_n} + \|(KR_\alpha - Id)(\bar{Y}_n - \hat{y})\|\chi_{\Omega_n} & (1.22)\\
&\leq C_{\nu+1}\rho b_n^{\frac{\nu+1}{2}}\chi_{\Omega_n} + \frac{\gamma}{4\sqrt{n}}\chi_{\Omega_n} = \frac{\gamma}{2\sqrt{n}}\chi_{\Omega_n} < \delta_n^{est}\chi_{\Omega_n},
\end{aligned}
$$

for all $\alpha \leq b_n$, so

$$
\alpha_n \geq q b_n \chi_{\Omega_n} \tag{1.23}
$$

for n large enough and the claim (1.19) follows with $\mathbb{P}\left(\Omega_n\right) \to 1$ as $n \to \infty$ (by Lemma 1.6.1). Finally,

$$
\begin{aligned}
&\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|\chi_{\Omega_n}\\
\leq&\|R_{\alpha_n}\|\|\bar{Y}_n - \hat{y}\|\chi_{\Omega_n'} + \|R_{\alpha_n}\hat{y} - K^+\hat{y}\|\chi_{\Omega_n}\\
\leq&\sqrt{C_R C_F}\frac{\delta_n^{true}}{\sqrt{\alpha_n}}\chi_{\Omega_n} + \frac{1}{\sigma_J^2}\left(\delta_n^{est} + C_0\delta_n^{true}\right)\chi_{\Omega_n}\\
\leq&\sqrt{C_R C_F}\left(\rho'\frac{4C_{\nu'+1}}{\gamma}\right)^{\frac{1}{\nu'+1}}\sqrt{n}^{\frac{1}{\nu'+1}+\varepsilon'-1} + \frac{1}{\sigma_J^2}\left(1+C_0\right)\sqrt{n}^{\varepsilon'-1}\\
\leq&\left(\sqrt{C_R C_F}\left(\frac{2C_{\nu'+1}}{\gamma}\right)^{\frac{1}{\nu'+1}}\rho^{\frac{1}{\nu'+1}} + \frac{1+C_0}{\sigma_J^2}\right)\left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu'}{\nu'+1}-\varepsilon'}\\
\leq&L\rho^{\frac{1}{\nu+1}}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu}{\nu+1}}
\end{aligned}
$$

for n large enough and L given in (1.17). The proof is finished with $\lim_{n\to\infty}\mathbb{P}\left(\Omega_n\right) = 1$.

1.6.1.2 Proof of Theorem 1.2.2 without emergency stop

W.l.o.g. we may assume that there are arbitrarily large $l \in \mathbb{N}$ with $(\hat{y}, u_l) \neq 0$, since otherwise we could apply Theorem 1.2.4 with any $\nu > 0$. Let $\varepsilon' > 0$ be arbitrary. Since $(R_\alpha)_{\alpha>0}$ is a regularisation and by (1.11) there is a $\varepsilon'' > 0$ such that

$$
\|R_\alpha\hat{y} - K^+\hat{y}\| \leq \varepsilon/2 \quad \text{ and } \quad \|(KR_{\alpha/q} - Id)\hat{y}\|/\sqrt{\alpha/q} \leq \sqrt{\frac{q}{8\gamma C_R C_F}}\varepsilon\varepsilon'
$$

for all $\alpha \leq \varepsilon''$. Set

$$
\begin{aligned}
\Omega_n :=& \Bigg\{|\sqrt{n}\delta_n^{est} - \gamma| < \gamma/2\ ,\ \|(KR_{\alpha_n/q} - Id)(\bar{Y}_n - \hat{y})\| \leq \frac{\delta_n^{est}}{2}\\
&\qquad \alpha_n \leq \varepsilon'',\ \delta_n^{true} \leq \frac{1}{\varepsilon'\sqrt{n}}\Bigg\}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
\delta_n^{est}\chi_{\Omega_n} &\leq \|(KR_{\alpha_n/q} - Id)\bar{Y}_n\|\chi_{\Omega_n}\\
&\leq \|(KR_{\alpha_n/q} - Id)\hat{y}\|\chi_{\Omega_n} + \|(KR_{\alpha_n/q} - Id)\left(\bar{Y}_n - \hat{y}\right)\|\chi_{\Omega_n}\\
&\leq \|(KR_{\alpha_n/q} - Id)\hat{y}\|\chi_{\Omega_n} + \frac{\delta_n^{est}}{2}.
\end{aligned}
$$

$$
\implies \frac{\gamma}{2}\sqrt{\frac{q}{n\alpha_n}}\chi_{\Omega_n} \leq \frac{\sqrt{q}\delta_n^{est}}{\sqrt{\alpha_n}}\chi_{\Omega_n} \leq \|(KR_{\alpha_n/q} - Id)\hat{y}\|/\sqrt{\alpha_n/q}\chi_{\Omega_n} \leq \sqrt{\frac{q\gamma^2}{8C_RC_F}}\varepsilon\varepsilon' \tag{1.24}
$$

by definition of Ω_n. Finally,

$$
\begin{aligned}
\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|\chi_{\Omega_n} &\leq \|R_{\alpha_n}\hat{y} - K^+\hat{y}\|\chi_{\Omega_n} + \|R_{\alpha_n}(\bar{Y}_n - \hat{y})\|\chi_{\Omega_n}\\
&\leq \frac{\varepsilon}{2} + \|R_{\alpha_n}\|\|\bar{Y}_n - \hat{y}\|\chi_{\Omega_n} \leq \frac{\varepsilon}{2} + \sqrt{\frac{C_RC_F}{\alpha_n}}\delta_n^{true}\chi_{\Omega_n}\\
&\leq \frac{\varepsilon}{2} + \sqrt{\frac{4C_RC_F}{\gamma^2 q}}\frac{\gamma}{2}\sqrt{\frac{q}{n\alpha_n}}\delta_n^{true}\sqrt{n}\chi_{\Omega_n}\\
&\leq \frac{\varepsilon}{2} + \sqrt{\frac{4C_RC_F}{\gamma^2 q}}\sqrt{\frac{q\gamma^2}{8C_RC_F}}\varepsilon\varepsilon'\frac{1}{\varepsilon'} \leq \varepsilon,
\end{aligned}
$$

where we used (1.24) and the definition of Ω_n in the fifth step. Thus $\mathbb{P}\left(\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\| \leq \varepsilon\right) \geq \mathbb{P}\left(\Omega_n\right) \geq 1 - \varepsilon'^2\mathbb{E}\|Y_1 - \hat{y}\|^2$ for $n \to \infty$ (by Lemmata 1.6.1 and 1.6.4, Proposition 1.6.3 and since $\mathbb{P}\left(\delta_n^{true} \geq \frac{1}{\varepsilon'\sqrt{n}}\right) \leq \varepsilon'^2\mathbb{E}\|Y_1 - \hat{y}\|^2$ by Tschebyscheff's inequality) and the claim follows with $\varepsilon' \to 0$.

1.6.1.3 Proofs for the emergency stop case

Again, denote by α_n the output of Algorithm 1 without the emergency stop. For the emergency stop, we have to consider $\|R_{\max\{\alpha_n,1/n\}}\bar{Y}_n - K^+\hat{y}\|$. It suffices to show that

$$
\mathbb{P}\left(\alpha_n \geq 1/n\right) \to 1 \tag{1.25}
$$

for $n \to \infty$. First assume that $K^+\hat{y} = (K^*K)^{\frac{\nu}{2}}\xi$ for some $\xi \in \mathcal{X}$ with $\|\xi\| \leq \rho$ and $0 < \nu \leq \nu_0 - 1$. With (1.16) or (1.23) and Lemma 1.6.1 it follows that

$$
\mathbb{P}\left(\alpha_n \geq q\left(\frac{\gamma}{4\rho C_{\nu+1}\sqrt{n}}\right)^{\frac{2}{\nu+1}}\right) \geq \mathbb{P}\left(\Omega_n\right) \to 1 \tag{1.26}
$$

for $n \to \infty$, with Ω_n given in (1.14) or (1.20)), thus we obtain (1.25). Otherwise, if there are no such ν, ρ and w, then (1.24) implies that for all $\varepsilon''' := \sqrt{\frac{q\gamma^2}{8C_RC_F}\varepsilon\varepsilon'}$ (where all quantities are given as in Section 1.6.1.2),

$$\mathbb{P}\left(\alpha_n \geq \frac{\delta_n^{est2}}{\varepsilon'''}\right) = \mathbb{P}\left(\frac{\alpha_n}{q\delta_n^{est2}} \geq \frac{1}{q\varepsilon}\right) \geq \mathbb{P}\left(\Omega_n\right) \tag{1.27}$$

for $n \to \infty$, with Ω_n given in (1.12) and we obtain the assertion with $\varepsilon, \varepsilon' \to 0$ and Lemma 1.6.1.

1.6.2 Proof of Corollary 1.2.5

Fix $\varepsilon > 0$. Denote by α_n the output of the discrepancy principle with emergency stop and set

$$\Omega_n := \{\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\| \leq \varepsilon\}. \tag{1.28}$$

It is

$$\|R_\alpha\hat{y} - K^+\hat{y}\| \leq \|R_\alpha K - Id\|\|\hat{x}\| \leq C \tag{1.29}$$

for all $\alpha > 0$. By the triangle inequality,

$$\begin{aligned}\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 &= 2\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - R_{\alpha_n}\hat{y}\|^2 + 2\mathbb{E}\|R_{\alpha_n}\hat{y} - K^+\hat{y}\|^2\\ &\leq 2\mathbb{E}\left[\|R_{\alpha_n}\|^2\delta_n^{true2}\right] + 2C^2 \leq 2C_RC_F\mathbb{E}\left[\delta_n^{true2}/\alpha_n\right] + 2C^2\\ &\leq 2nC_RC_F\mathbb{E}\delta_n^{true2} + 2C^2 = 2C_RC_F\mathbb{E}\|Y_1 - \hat{y}\|^2 + 2C^2 \leq C',\end{aligned}$$

where C' does not depend on n and where we used $\alpha_n \leq 1$ and (1.29) in the second step and $\alpha_n \geq 1/n$ in the fourth. By (1.28) there holds $\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|\chi_{\Omega_n} \leq \varepsilon$, so

$$\begin{aligned}\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 &= \mathbb{E}\left[\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2\chi_{\Omega_n}\right] + \mathbb{E}\left[\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2\chi_{\Omega_n^C}\right]\\ &\leq \varepsilon^2 + \mathbb{E}\left[\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2\chi_{\Omega_n^C}\right].\end{aligned}$$

We apply Cauchy-Schwartz to the second term

$$\begin{aligned}\mathbb{E}\left[\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2\chi_{\Omega_n^C}\right] &\leq \sqrt{\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^4\mathbb{E}\chi^2_{\Omega_n^C}}\\ &= \sqrt{\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^4\ \mathbb{P}\left(\Omega_n^C\right)}\end{aligned}$$

and we claim that there is a constant A with $\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^4 \leq A$ for all $n \in \mathbb{N}$.

$$
\begin{aligned}
&\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^4\\
\leq& 4\left(\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - R_{\alpha_n}\hat{y}\|^4 + 2\mathbb{E}\left[\|R_{\alpha_n}\bar{Y}_n - R_{\alpha_n}\hat{y}\|^2\|R_{\alpha_n}\hat{y} - K^+\hat{y}\|^2\right]\right.\\
&\qquad \left.+\mathbb{E}\|R_{\alpha_n}\hat{y} - K^+\hat{y}\|^4\right)\\
\leq& 4\left(\mathbb{E}\left[\|R_{\alpha_n}\|^4\delta_n^{true4}\right] + 2C^2\mathbb{E}\left[\|R_{\alpha_n}\|^2\delta_n^{true2}\right] + C^4\right)\\
\leq& B\left(\mathbb{E}\left[\delta_n^{true4}/\alpha_n^2\right] + \mathbb{E}\left[\delta_n^{true2}/\alpha_n\right] + 1\right)
\end{aligned}
$$

for some constant B, where we used (1.29) in the second step. First,

$$
\begin{aligned}
&\mathbb{E}\left[\delta_n^{true4}/\alpha_n^2\right]\\
\leq& n^2\mathbb{E}\|\bar{Y}_n - \hat{y}\|^4 = n^2\mathbb{E}\left[\sum_{j,j'\geq 1}\left(\bar{Y}_n - \hat{y}, u_j\right)^2\left(\bar{Y}_n - \hat{y}, u_{j'}\right)^2\right]\\
=& \frac{1}{n^2}\left(\sum_{j,j'\geq 1}\sum_{i,i',l,l'=1}^{n}\mathbb{E}\left[\left(Y_i - \hat{y}, u_j\right)\left(Y_l - \hat{y}, u_j\right)\left(Y_{i'} - \hat{y}, u_{j'}\right)\left(Y_{l'} - \hat{y}, u_{j'}\right)\right]\right)\\
\leq& \frac{1}{n^2}\sum_{j,j'\geq 1}\left(n\mathbb{E}\left[\left(Y_1 - \hat{y}, u_j\right)^2\left(Y_1 - \hat{y}, u_{j'}\right)^2\right] + n^2\mathbb{E}\left[\left(Y_1 - \hat{y}, u_j\right)^2\right]\mathbb{E}\left[\left(Y_1 - \hat{y}, u_{j'}\right)^2\right]\right.\\
&\qquad \left.+\ 2n^2\left(\mathbb{E}\left[\left(Y_1 - \hat{y}, u_j\right)\left(Y_1 - \hat{y}, u_{j'}\right)\right]\right)^2\right)\\
\leq& \frac{n+2n^2}{n^2}\mathbb{E}\left[\sum_{j,j'\geq 1}\left(Y_1 - \hat{y}, u_j\right)^2\left(Y_1 - \hat{y}, u_{j'}\right)^2\right]\\
&\qquad +\mathbb{E}\left[\sum_{j\geq 1}\left(Y_1 - \hat{y}, u_j\right)^2\right]\mathbb{E}\left[\sum_{j'\geq 1}\left(Y_1 - \hat{y}, u_{j'}\right)^2\right]\\
\leq& \frac{n+2n^2}{n^2}\mathbb{E}\left[\left(\sum_{j\geq 1}\left(Y_1 - \hat{y}, u_j\right)^2\right)^2\right] + \left(\mathbb{E}\left[\sum_{j\geq 1}\left(Y_1 - \hat{y}, u_j\right)^2\right]\right)^2\\
=& \frac{n+2n^2}{n^2}\mathbb{E}\|Y_1 - \hat{y}\|^4 + \left(\mathbb{E}\left[\|Y_1 - \hat{y}\|^2\right]\right)^2 \leq B_1
\end{aligned}
$$

for some constant B_1, where in the fourth step we used that the Y_i are i.i.d, that $\mathbb{E}\left(Y_1 - \hat{y}, u_j\right) = \left(\mathbb{E}[Y_1] - \hat{y}, u_j\right) = 0$ and that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for independent (and integrable) random variables (so the relevant cases are the ones where either all indices i, i', l, l' are equal or exactly pairwise two). Then we used Jensen's inequality in the fifth step. Moreover,
$\mathbb{E}\left[\delta_n^{true2}/\alpha_n\right] \leq n\mathbb{E}\left[\delta_n^{true2}\right] = \mathbb{E}\|Y_1 - \hat{y}\|^2 = B_2$, so the claim holds for $A = B(B_1 + B_2 + 1)$. By Theorem 3 it holds that $\mathbb{P}\left(\Omega_n\right) \to 1$ for $n \to \infty$, thus $\mathbb{P}\left(\Omega_n^C\right) \leq \varepsilon^2/A$

for n large enough and

$$\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^2 \leq \varepsilon\mathbb{E}[\chi_{\Omega_n}] + \sqrt{\mathbb{E}\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|^4 \ \mathbb{P}\left(\Omega_n^C\right)} \leq 2\varepsilon.$$

1.6.3 Proofs of Theorem 1.3.1, 1.3.2 and 1.3.3

Proof of Theorem 1.3.1

Note that $p > 1$ is necessary because of $\sum_{j=1}^{\infty} \mathbb{E}(Y_1 - \hat{y}, u_j)^2 = \mathbb{E}\|Y_1 - \hat{y}\|^2 \overset{!}{\leq} \infty$.

$$\begin{aligned}
\mathbb{E}\|R_\alpha \bar{Y}_n - K^+\hat{y}\|^2 &= \mathbb{E}\|R_\alpha(\bar{Y}_n - \hat{y})\|^2 + \|R_\alpha\hat{y} - K^+\hat{y}\|^2 \\
&= \sum_{\substack{j \\ \sigma_j^2 > \alpha}} \sigma_j^{-2}\mathbb{E}(\bar{Y}_n - \hat{y}, u_j)^2 + \sum_{\substack{j \\ \sigma_j^2 \leq \alpha}} \sigma_j^{-2}(\hat{y}, u_j)^2 \\
&= \frac{1}{n}\sum_{j=1}^{N} \sigma_j^{-2}\mathbb{E}(Y_1 - \hat{y}, u_j)^2 + \sum_{j=N+1}^{\infty} \sigma_j^{-2\nu}(\xi, v_j)^2 \\
&\asymp \frac{1}{n}\sum_{j=1}^{N} j^{q-p} + \rho\sum_{j=N+1}^{\infty} j^{-\nu q}(\xi, v_j)^2,
\end{aligned}$$

where $N = N(\alpha) = \max\{j \geq 1 \ : \ \sigma_j^2 > \alpha\}$. Therefore it holds that

$$\begin{aligned}
\sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_\alpha\bar{Y}_n - K^+\hat{y}\|^2 &\asymp \frac{1}{n}\int_{j=1}^{N} x^{q-p}dx + \rho N^{-\nu q} \\
&\asymp \begin{cases} \frac{1}{n} + \rho N^{-\nu q} & q-p < -1 \\ \frac{1}{n}\log(N) + \rho N^{-\nu q} & q-p = -1 \\ \frac{1}{n}N^{q-p+1} + \rho N^{-\nu q} & q-p > -1 \end{cases}.
\end{aligned}$$

The right hand side is minimised by the choices

$$N = N(n) \asymp \begin{cases} (\rho n)^{\frac{1}{\nu q}} & q-p \leq -1 \\ (\rho n)^{\frac{1}{(1+\nu)q+1-p}} & q-p > -1 \end{cases}.$$

Thus we obtain

$$\min_{\alpha>0} \sup_{\substack{\xi\in\mathcal{X},\ \|\xi\|\leq\rho \\ K^+\hat{y}=(K^*K)^{\nu/2}\xi}} \mathbb{E}\|R_\alpha\bar{Y}_n - K^+\hat{y}\|^2 \asymp \begin{cases} \frac{1}{n} & q-p < -1 \\ \frac{\log(\rho n)}{n} & q-p = -1 \\ \rho^{\frac{q+1-p}{(\nu+1)q+1-p}} \left(\frac{1}{n}\right)^{\frac{\nu}{\nu+1-\frac{p-1}{q}}} & q-p > -1 \end{cases}.$$

□

Proof of Theorem 1.3.2

The choice of r guarantees, that $SK : \mathcal{X} \to \mathcal{Y}$ is compact with singular values $\sigma_j(SK)^2 \asymp j^{r-q}$. Moreover by the choice of r,

$$\|S\hat{y}\|^2 = \sum_{j=1}^{\infty} j^r(\hat{y}, u_j)^2 \leq C \sum_{j=1}^{\infty} j^{r-q}(\hat{x}, v_j)^2 < \infty$$

and

$$\mathbb{E}\|S(Y_1 - \hat{y})\|^2 = \sum_{j=1}^{\infty} j^r \mathbb{E}(Y_1 - \hat{y}, u_j)^2 \leq C' \sum_{j=1}^{\infty} j^{r-p} < \infty,$$

where C, C' are constants. Thus $SY_1, SY_2, ...$ are i.i.d with $\mathbb{E}\left[SY_1\right] = S\hat{y}$ and $\mathbb{E}\|SY_1\|^2 < \infty$. By assumption, there exists $\xi \in \mathcal{X}$ with $\hat{x} = (K^*K)^{\nu/2}\xi$ and $\|\xi\| < \rho$ and a_j, with $\inf_{j\in\mathbb{N}} a_j = C'' > 0$ and $\sigma_j = j^{-q}a_j, \sigma_j(SK) = j^{r-q}a_j$. It follows that

$$\begin{aligned}\hat{x} = (K^*K)^{\nu/2}\xi &= \sum_{j=1}^{\infty} \sigma_j^{\nu}(\xi, v_j)v_j = \sum_{j=1}^{\infty} j^{-\nu q} a_j^{\nu}(\xi, v_j)v_j \\ &= \sum_{j=1}^{\infty} j^{-\frac{q}{q-r}\nu(q-r)} a_j^{\nu}(\xi, v_j)v_j = \sum_{j=1}^{\infty} \sigma_j(SK)^{\frac{q}{q-r}\nu} a_j^{-\frac{\nu r}{q-r}}(\xi, v_j)v_j \\ &= ((SK)^*SK)^{\frac{q}{2(q-r)}\nu}\,\xi',\end{aligned}$$

with $\xi' := \sum_{j=1}^{\infty} a_j^{-\frac{\nu r}{q-r}}(\xi, v_j)v_j \in \mathcal{X}$ and $\|\xi'\| \leq C''^{\frac{-r\nu}{q-r}}\rho =: \rho'$. Thus, for $\nu' := \frac{q}{q-r}\nu$ it holds that $\hat{x} = ((SK)^*SK)^{\nu'/2}\xi'$, so by Theorem 1.2.4 there exists $L' > 0$ with

$$\mathbb{P}\left(\|R'_{\alpha_n}S\bar{Y}_n - \hat{x}\| \leq L'\rho'^{\frac{1}{\nu'+1}}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu'}{\nu'+1}}\right) \to 1$$

as $n \to \infty$. It is

$$\frac{\nu'}{\nu'+1} = \frac{\nu}{\nu+1-\frac{r}{q}} \quad \text{and} \quad \frac{1}{\nu'+1} = \frac{q-r}{(\nu+1)q-r}.$$

Finally, considering the cases $C'' \leq 1$ and $C'' > 1$ separately yields

$$\rho'^{\frac{1}{\nu'+1}} = \left(C''^{-\frac{r\nu}{q-r}}\rho\right)^{\frac{1}{\frac{q}{q-r}\nu+1}} = C''^{\frac{-r\nu}{q\nu+q-r}}\rho^{\frac{1}{\nu'+1}} \leq \max\left(1, 1/C''\right)\rho^{\frac{1}{\nu'+1}}$$

and the claim holds with $L := L' \max(1, 1/C'')$.

□

Proof of Theorem 1.3.3 For $d_j := j^{p-1-\varepsilon_2}$ it holds that

$$\begin{aligned}\hat{x} = (K^*K)^{\frac{\nu}{2}}\xi &= \sum_{j=1}^{\infty} \sigma_j^{\nu}(\xi, v_j)v_j = \sum_{j=1}^{\infty} j^{-q\nu}(\xi, v_j)v_j \qquad (1.30)\\ &= \sum_{j=1}^{\infty} \left(j^{-(q+1+\varepsilon_2-p)}\right)^{\frac{q\nu}{q+1+\varepsilon_2-p}} (\xi, v_j)v_j = \sum_{j=1}^{\infty} (d_j\sigma_j)^{\frac{q\nu}{q+1+\varepsilon_2-p}} (\xi, v_j)v_j\\ &= \sum_{j=1}^{\infty} (d_j\sigma_j)^{\nu'}(\xi, v_j)v_j\end{aligned}$$

with $\nu' := q\nu/(q+1+\varepsilon_2-p)$. By Theorem 2 of [Ang12] it holds that $\mathbb{E}[|s_{n,j}^2 - \mathbb{E}(Y_1 - \hat{y}, u_j)^2|^2] \leq 4\mathbb{E}(Y_1 - \hat{y}, u_j)^4$, thus

$$\begin{aligned}&\mathbb{P}\left(|s_{n,j}^2 - \mathbb{E}(Y_1 - \hat{y}, u_j)^2| \leq \frac{\mathbb{E}(Y_1 - \hat{y}, u_j)^2}{2},\ \forall j \leq m_n\right)\\ \geq& 1 - \sum_{j=1}^{m_n} \mathbb{P}\left(|s_{n,j}^2 - \mathbb{E}(Y_1 - \hat{y}, u_j)^2| > \frac{\mathbb{E}(Y_1 - \hat{y}, u_j)^2}{2}\right)\\ \geq& 1 - \sum_{j=1}^{m_n} \frac{\mathbb{E}[|s_{n,j}^2 - \mathbb{E}(Y_1 - \hat{y}, u_j)^2|^2]}{(\mathbb{E}[(Y_1 - \hat{y}, u_j)^2]/2)^2}\\ =& 1 - \frac{16m_n}{n} \sup_{j=1,\ldots,m_n} \frac{\mathbb{E}[(Y_1 - \hat{y}, u_j)^4]}{(\mathbb{E}[(Y_1 - \hat{y}, u_j)^2])^2} \geq 1 - \frac{16m_nC_p}{n}\\ \geq& 1 - 16C_pn^{-\varepsilon_1} \to 1\end{aligned}$$

as $n \to \infty$. From that directly follows

$$\mathbb{P}\left(\frac{d_j^2}{2} \leq d_{j,n}^2 \leq 2d_j^2,\ \forall j = 1, ..., m_n\right) \to 1, \qquad (1.31)$$

$$\mathbb{P}\left(|\sqrt{n}\delta_n^{est} - \gamma| \leq \frac{\gamma}{2}\right) \to 1 \qquad (1.32)$$

for $\gamma := \sqrt{\mathbb{E}\|Y_1 - \hat{y}\|^2 \sum_{j=1}^{\infty} j^{-(1+\varepsilon_2)}}$ as $n \to \infty$, because $q > p-1$ implies $\min(j^{p-1-\varepsilon_2}, j^q) = j^{(p-1-\varepsilon_2)}$. We prove a modulation of Lemma 1.6.4.

Lemma 1.6.5. *It holds that*

$$\mathbb{P}\left(\sqrt{\sum_{j=k_n}^{m_n} d_{j,n}^2(\bar{Y}_n - \hat{y}, u_j)^2} \le \frac{\delta_n^{est}}{2}\right) \to 1$$

as $n \to \infty$.

Proof of Lemma 1.6.5 we first show that there is $(q_n)_{n\in\mathbb{N}} \subset \mathbb{N}$

$$\mathbb{P}\left(k_n \ge q_n\right) \to 1 \quad \text{and} \quad q_n \to \infty \tag{1.33}$$

as $n \to \infty$. For that it suffices to show that $\lim_{n\to\infty} \mathbb{P}\left(k_n \ge k\right) = 0$ for all $k \in \mathbb{N}$. To see this set

$$\Omega_n := \left\{\|\sqrt{n}\delta_n^{est} - \gamma\| \le \frac{\gamma}{2},\ (\bar{Y}_n, u_L)^2 \ge (\hat{y}, u_L)^2/2,\ d_L^2 \le 2d_{L,n}^2\right\}. \tag{1.34}$$

Then for $n \ge \max(k, 16\gamma^2/(d_L(\hat{y}, u_L))^2)$,

$$\begin{aligned}\delta_n^{est}\chi_{\Omega_n} &\le \frac{2\gamma}{\sqrt{n}}\chi_{\Omega_n} < \sqrt{\frac{d_L^2(\hat{y}, u_L)^2}{4}}\chi_{\Omega_n} \le \sqrt{\frac{d_L^2(\bar{Y}_n, u_L)^2}{2}}\chi_{\Omega_n} \\ &\le \sqrt{d_{L,n}^2(\bar{Y}_n, u_L)^2} \le \sqrt{\sum_{j=k}^{m_n} d_{j,n}^2(\bar{Y}_n, u_j)^2}.\end{aligned}$$

Thus $k_n\chi_{\Omega_n} > k\chi_{\Omega_n}$ by Algorithm 2 and (1.33) follows with $\lim_{n\to\infty} \mathbb{P}\left(\Omega_n\right) = 1$ (because of (1.31), (1.32) and the law of large numbers). For $\varepsilon > 0$

$$\begin{aligned}&\mathbb{P}\left(\sqrt{\sum_{j=k_n}^{m_n} d_{j,n}^2(\bar{Y}_n - \hat{y}, u_j)^2} \le \frac{\delta_n^{est}}{2}\right) \\ \ge&\mathbb{P}\left(\sqrt{\sum_{j=q_n}^{m_n} d_j^2(\bar{Y}_n - \hat{y}, u_j)^2} \le \frac{\gamma}{4\sqrt{n}},\ \delta_n^{est} \ge \frac{\gamma}{2\sqrt{n}},\ d_j^2 \ge \frac{d_{j,n}^2}{2}\ \forall j \le m_n,\ k_n \ge q_n\right) \\ \ge&1 - \mathbb{P}\left(\sqrt{\sum_{j=q_n}^{m_n} d_j^2(\bar{Y}_n - \hat{y}, u_j)^2} > \frac{\gamma}{4\sqrt{n}}\right) - \mathbb{P}\left(\delta_n^{est} < \frac{\gamma}{2\sqrt{n}}\right) \\ &\quad - \mathbb{P}\left(d_j^2 < \frac{d_{j,n}^2}{2},\ \forall j \le m_n\right) - \mathbb{P}\left(k_n < q_n\right) \\ \ge&1 - \varepsilon - \mathbb{P}\left(\sqrt{\sum_{j=q_n}^{m_n} d_j^2(\bar{Y}_n - \hat{y}, u_j)^2} > \frac{\gamma}{4\sqrt{n}}\right)\end{aligned}$$

for n large enough because of (1.31), (1.32) and (1.33). Now

$$\mathbb{P}\left(\sqrt{\sum_{j=q_n}^{m_n} d_j^2(\bar{Y}_n-\hat{y},u_j)^2} > \frac{\gamma}{4\sqrt{n}}\right) \leq \frac{16n}{\gamma^2}\sum_{j=q_n}^{m_n} d_j^2 \frac{\mathbb{E}(Y_1-\hat{y},u_j)^2}{n}$$
$$\leq \frac{16}{\gamma^2}\sum_{j=q_n}^{m_n} j^{-(1+\varepsilon_2)} \leq \varepsilon$$

for n large enough, because $\lim_{n\to\infty}\sum_{j=q_n}^{m_n} j^{-(1+\varepsilon_2)} \leq \lim_{n\to\infty}\sum_{j=q_n}^{\infty} j^{-(1+\varepsilon_2)} = 0$. Since ε was arbitrary, the proof of Lemma 1.6.5 is concluded.

□

We start the main proof and decompose as usual

$$\|\bar{X}_n - K^+\hat{x}\| \leq \sqrt{\sum_{j=1}^{k_n}\frac{(\bar{Y}_n-\hat{y},u_j)^2}{\sigma_j^2}} + \sqrt{\sum_{j=k_n}^{\infty}(\hat{x},u_j)^2}$$

and first consider the approximation error. With the convention $\sum_{j=s}^{t} = 0$ for $s > t$, a standard application of Hölder's inequality for $p = \frac{\nu'+1}{\nu'}$ and $q = \nu'+1$, (1.30) and the triangle inequality yield

$$\sqrt{\sum_{j=k_n+1}^{m_n}(\hat{x},u_j)^2} = \sqrt{\sum_{j=k_n+1}^{m_n}(d_j\sigma_j)^{2\nu'}(\xi,v_j)^2}$$
$$\leq \sqrt{\left(\sum_{j=k_n+1}^{m_n}(d_j\sigma_j)^{2(\nu'+1)}(\xi,v_j)^2\right)^{\frac{\nu'}{\nu'+1}}\left(\sum_{j=k_n+1}^{m_n}(\xi,v_j)^2\right)^{\frac{1}{\nu'+1}}}$$
$$\leq \rho^{\frac{1}{\nu'+1}}\left(\sqrt{\sum_{j=k_n+1}^{m_n}(d_j\sigma_j)^2(\hat{x},v_j)^2}\right)^{\frac{\nu'}{\nu'+1}} = \rho^{\frac{1}{\nu'+1}}\left(\sqrt{\sum_{j=k_n+1}^{m_n}d_j^2(\hat{y},v_j)^2}\right)^{\frac{\nu'}{\nu'+1}}$$
$$\leq \rho^{\frac{1}{\nu'+1}}\left(\sqrt{\sum_{j=k_n+1}^{m_n}d_j^2(\bar{Y}_n,u_j)^2} + \sqrt{\sum_{j=k_n+1}^{m_n}d_j^2(\bar{Y}_n-\hat{y},u_j)^2}\right)^{\frac{\nu'}{\nu'+1}}.$$

Thus for

$$\Omega_n := \Bigg\{ \sqrt{\sum_{j=k_n}^{m_n} d_{j,n}^2 (\bar{Y}_n - \hat{y}, u_j)^2} \leq \frac{\delta_n^{est}}{2}, \ |\delta_n^{est} - \frac{\gamma}{\sqrt{n}}| \leq \frac{\gamma}{2\sqrt{n}}, \\ \frac{d_j^2}{2} \leq d_{j,n}^2 \leq 2 d_j^2 \ \forall j \leq m_n \Bigg\} \tag{1.35}$$

there holds

$$\begin{aligned} &\left(\sqrt{\sum_{j=k_n+1}^{m_n} d_j^2 (\bar{Y}_n, u_j)^2} + \sqrt{\sum_{j=k_n+1}^{m_n} d_j^2 (\bar{Y}_n - \hat{y}, u_j)^2} \right)^{\frac{\nu'}{\nu'+1}} \chi_{\Omega_n} \\ \leq& 4^{\frac{\nu'}{\nu'+1}} \left(\sqrt{\sum_{j=k_n+1}^{m_n} d_{j,n}^2 (\bar{Y}_n, u_j)^2} + \sqrt{\sum_{j=k_n+1}^{m_n} d_{j,n}^2 (\bar{Y}_n - \hat{y}, u_j)^2} \right)^{\frac{\nu'}{\nu'+1}} \chi_{\Omega_n} \\ \leq& 4^{\frac{\nu'}{\nu'+1}} \left(\delta_n^{est} + \delta_n^{est} \right) \chi_{\Omega_n} \leq \left(\frac{16\gamma}{\sqrt{n}} \right)^{\frac{\nu'}{\nu'+1}}. \end{aligned}$$

Consequently,

$$\begin{aligned} \sqrt{\sum_{j=k_n+1}^{\infty} (\hat{x}, u_j)^2 \chi_{\Omega_n}} &\leq \sqrt{\sum_{j=k_n+1}^{m_n} (\hat{x}, u_j)^2 \chi_{\Omega_n}} + \sqrt{\sum_{j=m_n}^{\infty} (\hat{x}, u_j)^2} \\ &\leq \rho^{\frac{1}{\nu'+1}} \left(\frac{16\gamma}{\sqrt{n}} \right)^{\frac{\nu'}{\nu'+1}} + \sqrt{\sum_{j=m_n+1}^{\infty} \sigma_j^{2\nu} (\xi, v_j)^2} \\ &\leq \frac{L}{2} \max \left(\rho^{\frac{q+1-p}{\nu q+q+1-p}} \left(\frac{1}{\sqrt{n}} \right)^{\frac{\nu}{\nu + \frac{q+1+\varepsilon_2-p}{q}}}, \rho \left(\frac{1}{\sqrt{n}} \right)^{2(1-\varepsilon_1)q\nu} \right) \end{aligned} \tag{1.36}$$

for $L = (32\gamma)^{\frac{\nu'}{\nu'+1}}$. To finish the proof we need to verify a similar bound for the data propagation error. By definition of the discrepancy principle (Algorithm 2) and Ω_n

$$
\begin{aligned}
\delta_n^{est}\chi_{\Omega_n} &< \sqrt{\sum_{j=k_n}^{m_n} d_{j,n}^2(\bar{Y}_n, u_j)^2\chi_{\Omega_n}} \\
&\leq \sqrt{\sum_{j=k_n}^{m_n} d_{j,n}^2(\hat{y}, u_j)^2\chi_{\Omega_n}} + \sqrt{\sum_{j=k_n}^{m_n} d_{j,n}^2(\bar{Y}_n - \hat{y}, u_j)^2\chi_{\Omega_n}} \\
&< 2\sqrt{\sum_{j=k_n}^{m_n} d_j^2(\hat{y}, u_j)^2} + \frac{\delta_n^{est}}{2} = 2\sqrt{\sum_{j=k_n}^{m_n} (d_j\sigma_j)^{2(1+\nu')}(\xi, v_j)^2} + \frac{\delta_n^{est}}{2}\chi_{\Omega_n} \\
&\leq 2\rho k_n^{-\frac{(q+1+\varepsilon_2-p)(1+\nu')}{2}} + \frac{\delta_n^{est}}{2}\chi_{\Omega_n} \\
\Longrightarrow k_n\chi_{\Omega_n} &\leq \left(\frac{4\rho}{\delta_n^{est}}\right)^{\frac{2}{(q+1+\varepsilon_2-p)(1+\nu')}} \chi_{\Omega_n} \leq \left(\frac{16\rho^2 n}{\gamma^2}\right)^{\frac{1}{(q(\nu+1)+1+\varepsilon_2-p}} =: b_n'.
\end{aligned}
$$

We set $b_n := \min(b_n', m_n)$, so $k_n\chi_{\Omega_n} \leq b_n$ and

$$
\begin{aligned}
&\mathbb{P}\left(\sqrt{\sum_{j=1}^{k_n} \frac{(\bar{Y}_n - \hat{y}, u_j)^2}{\sigma_j^2}} \leq \frac{L}{2}\max\left(\rho^{\frac{1}{\nu'+1}}\left(\frac{1}{\sqrt{n}}\right)^{\frac{\nu'}{\nu'+1}}, \rho\left(\frac{1}{\sqrt{n}}\right)^{2(1-\varepsilon_1)q\nu}\right)\right) \\
&\geq \mathbb{P}\left(\sqrt{\sum_{j=1}^{b_n} \frac{(\bar{Y}_n - \hat{y}, u_j)^2}{\sigma_j^2}} \leq A\left(\frac{1}{\sqrt{n}}\right)^{\min\left(\frac{\nu'}{\nu'+1}, 2(1-\varepsilon_1)q\nu\right)}, \Omega_n\right) \\
&\geq 1 - \mathbb{P}\left(\sqrt{\sum_{j=1}^{b_n} \frac{(\bar{Y}_n - \hat{y}, u_j)^2}{\sigma_j^2}} > A\left(\frac{1}{\sqrt{n}}\right)^{\min\left(\frac{\nu'}{\nu'+1}, 2(1-\varepsilon_1)q\nu\right)}\right) - \mathbb{P}\left(\Omega_n^C\right)
\end{aligned}
$$

with $A := \frac{L}{2}\max\left(\rho^{\frac{1}{\nu'+1}}, \rho\right)$. Finally, with constants A', A''

$$
\begin{aligned}
&\mathbb{P}\left(\sqrt{\sum_{j=1}^{b_n}\frac{(\bar{Y}_n-\hat{y},u_j)^2}{\sigma_j^2}}>A\left(\frac{1}{\sqrt{n}}\right)^{\min\left(\frac{\nu'}{\nu'+1},2(1-\varepsilon_1)q\nu\right)}\right)\\
\leq& An^{\min\left(\frac{\nu'}{\nu'+1},2(1-\varepsilon_1)q\nu\right)}\sum_{j=1}^{b_n}j^q\mathbb{E}(\bar{Y}_n-\hat{y},u_j)^2\\
\leq& A'n^{\min\left(\frac{\nu'}{\nu'+1},2(1-\varepsilon_1)q\nu\right)-1}\int_1^{b_n}x^{q-p}dx\leq A'n^{\min\left(\frac{\nu'}{\nu'+1},2(1-\varepsilon_1)q\nu\right)-1}b_n^{q+1-p}\\
\leq& A''n^{\min\left(\frac{\nu'}{\nu'+1},2(1-\varepsilon_1)q\nu\right)-1+(q+1-p)\min\left(\frac{1}{q(\nu+1)+1+\varepsilon_2-p},(1-\varepsilon_1)\right)}\\
\leq& A''n^{\min\left(\frac{1}{q(\nu+1)+1+\varepsilon_2-p},2(1-\varepsilon_1)\right)q\nu-1+(q+1-p)\min\left(\frac{1}{q(\nu+1)+1+\varepsilon_2-p},2(1-\varepsilon_1)\right)}\\
\leq& A''n^{\min\left(\frac{q(\nu+1)+1-p}{q(\nu+1)+1+\varepsilon_2-p},2(1-\varepsilon_1)(q(\nu+1)+1-p)\right)-1}\to 0
\end{aligned}
$$

as $n\to\infty$, since $\varepsilon_2>0$. With $\mathbb{P}\left(\Omega_n^C\right)\to 0$ this concludes the proof of Theorem 1.3.3.

□

1.7 Numerical demonstration

We conclude with some numerical results.

1.7.1 Differentiation of binary option prices

A natural example is given if the data is acquired by a Monte-Carlo simulation, here we consider an example from mathematical finance. The buyer of a binary call option receives after T days a payoff Q, if then a certain stock price S_T is higher then the strike value K. Otherwise he gets nothing. Thus the value V of the binary option depends on the expected evolution of the stock price. We denote by r the riskfree rate, for which we could have invested the buying price of the option until the expiry rate T. If we already knew today for sure, that the stock price will hit the strike (insider information), we would pay $V=e^{-rT}Q$ for the binary option (e^{-rT} is called discount factor). Otherwise, if we believed that the stock price will hit the strike with probability p, we would pay $V=e^{-rT}Qp$. In the Black Scholes model one assumes, that the relative change of the stock price in a short time interval is normally distributed, that is

$$S_{t+\delta t}-S_t\sim\mathcal{N}(\mu\delta t,\sigma^2\delta t).$$

Under this assumption one can show that (see [HB16])

$$S_T = S_0 e^{sT},$$

where S_0 is the initial stock price and $s \sim \mathcal{N}\left(\mu - \sigma^2/2, \sigma^2/T\right)$. Under this assumptions one has $V = e^{-rT} Q\Phi(d)$, with

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}} d\xi, \qquad d = \frac{\log \frac{S_0}{K} + T\left(\mu - \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T}}.$$

Ultimately we are interested in the sensitivity of V with respect to the starting stock price S_0, that is $\partial V(S_0)/\partial S_0$. We formulate this as the inverse problem of differentiation. Set $\mathcal{X} = \mathcal{Y} = L^2([0,1] =$ and define

$$\begin{aligned} K :& L^2([0,1]) \to L^2([0,1]) \\ f \mapsto & Af = g : x \mapsto \int_0^x f(y)dy. \end{aligned}$$

Then our true data is $\hat{y} = V = e^{-rT} Q\Phi(d)$. To demonstrate our results we now approximate $V : S_0 \mapsto e^{-rT} Qp(S_0)$ through a Monte-Carlo approach. That is we generate independent gaussian random variables $Z_1, Z_2, ...$ identically distributed to s and set $Y_i := e^{-rT} Q\chi_{\{S_0 e^{TZ_i} \geq K\}}$. Then we have $\mathbb{E}Y_i = e^{-rT} Q\mathbb{P}(S_0 e^{TZ_i}) = e^{-rT} Qp(S_0) = V(S_0)$ and $\mathbb{E}\|Y_i\|^2 \leq e^{-rT} Q < \infty$. We replace $L^2([0,1])$ with piecewise continuous linear splines on a homogeneous grid with $m = 50000$ elements (we can calculate Kg exactly for such a spline g). We use in total $n = 10000$ random variables for each simulation. As parameters we chose $r = 0.0001, T = 30, K = 0.5, Q = 1, \mu = 0.01, \sigma = 0.1$. It is easy to see that $\hat{x} = K^+\hat{y} \in \mathcal{X}_\nu$ for all $\nu > 0$ using the transformation $z(\xi) = 0,5e^{\sqrt{0,3}\xi - 0,15}$. Since the qualification of the Tikhonov regularisation is 2, Theorem 1.2.4 gives an error bound which is asymptotically proportional to $(1/\sqrt{n})^{\frac{1}{2}}$. In Figure 1 we plot the L^2 average of 100 simulations (with the sample variance as error bars) of the discrepancy principle together with the (translated) optimal error bound. In this case the emergency stop did not trigger once - this is plausible, since the true solution is very smooth, which yields comparably higher values of the regularisation parameter and also, the error distribution is Gaussian.

Let us stress that this is only an academic example to demonstrate the possibility of using our new generic approach in the context of Monte Carlo simulations. Explicit solution formulas for standard binary options are well-known, and for more complex financial derivatives with discontinuous payoff profiles (such as autocallables or Coco-bonds) one would rather resort to stably differentiable Monte Carlo methods ([AHHK13] or [GHR20]) or use specific regularization methods for numerical differentiation [HS01].

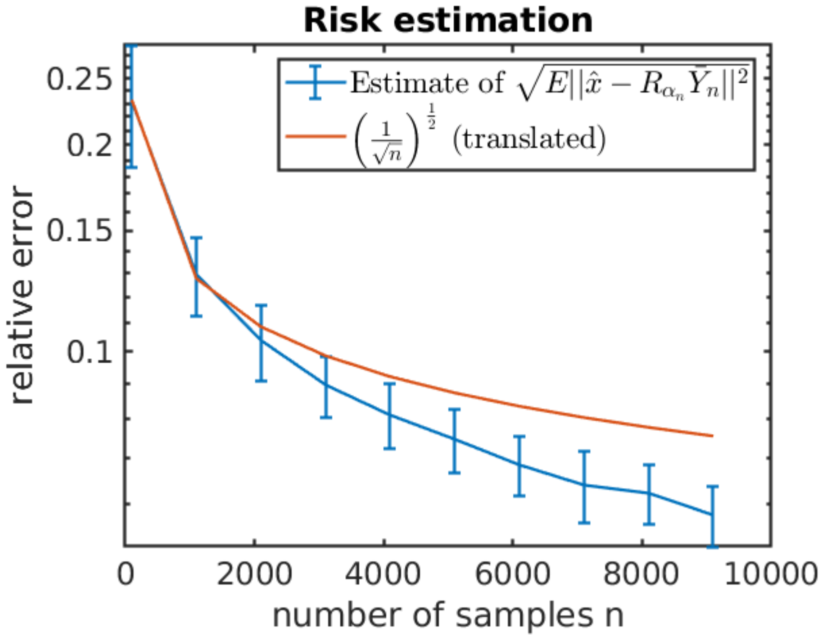

Figure 1.1: Estimated Risk of a binary option.

1.7.1.1 Rescaling of the measurements and the operator

We now numerically investigate the idea of rescaling the measurements to improve the relative smoothness and hence the convergence speed, cf. section 1.3. We reduce the discretisation dimension to $m = 5000$ and calculate the singular value decomposition with the function 'csvd' from the regularisation toolbox [Han10]. We apply Algorithm 2 for $\varepsilon_1 = 0.5, \varepsilon_2 = 0.1$ and $n = 5 * [10^1\ ,\ 3.3 * 10^1\ ,\ 10^2 ... 3.3 * 10^4\ ,\ 10^5]$ and compare the result to Algorithm 1 with the same n. In Figure 1.2 we plot the L^2 average of the relative errors for 100 simulations of the discrepancy principle together with the sample variance of the relative errors. We clearly see, that the convergence is faster in the rescaled case.

1.7.2 Inverse heat equation

We consider the toy problem 'heat' from [Han10]. We chose the discretisation level $m = 100$ and set $\sigma = 0.7$. Under this choice, the last seven singular values (calculated with the function 'csvd') fall below the machine precision of 10^{-16}. The discretised large systems of linear equations are solved iteratively using the conjugate gradient method ('pcg' from MATLAB) with a tolerance of 10^{-8}. As a regularisation method we chose Tikhonov regularisation and we compared the a priori choice $\alpha_n = 1/\sqrt{n}$, the discrepancy principle (dp) and the discrepancy principle with emergency stop (dp+es), as implemented in Algorithm 1 with $q = 0.7$ and estimated sample variance. The unbiased i.i.d measurements fulfill $\sqrt{\mathbb{E}\|Y_i - \hat{y}\|^2} \approx 1.16$ and $\mathbb{E}\|Y_i - \mathbb{E}Y_i\|^k = \infty$

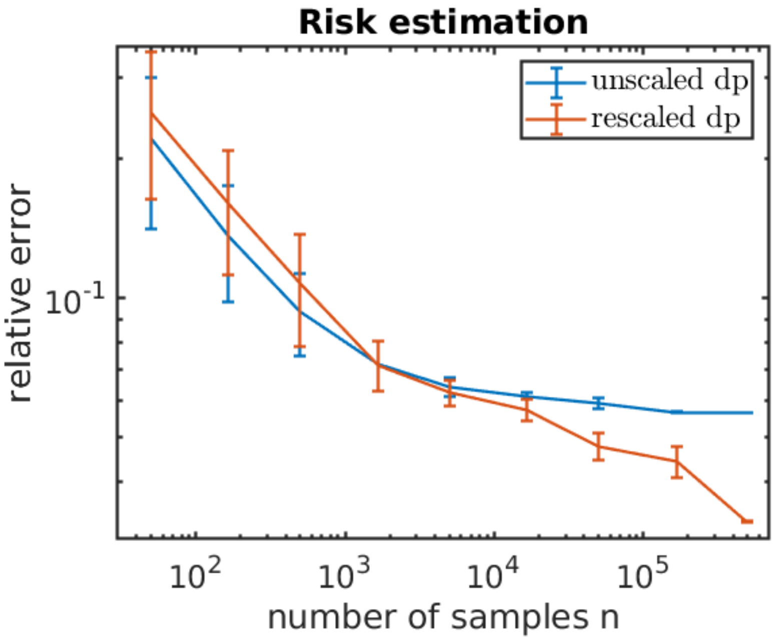

Figure 1.2: Estimated risk of a binary option for unscaled and rescaled operator and measurements.

for $k \geq 3$. Concretely, we chose $Y_i := \hat{y} + E_i$ with $E_i := U_i * Z_i * v$, where the U_i are independent and uniformly on $[-1/2, 1/2]$ distributed, the Z_i are independent Pareto distributed (MATLAB function 'gprnd' with parameters $1/3$, $1/2$ and $3/2$), and v is a uniform permutation of $1, 1/2^{\frac{3}{4}}, ..., 1/m^{\frac{3}{4}}$. Thus we chose a rather ill-posed problem together with a heavy-tailed error distribution. We considered three different sample sizes $n = 10^3, 10^4, 10^5$ with 200 simulations for each one. The results are presented as boxplots in Figure 3. It is visible, that the results are much more concentrated for a priori regularisation and discrepancy principle with emergency stop, indicating the L^2 convergence (strictly speaking we do not know if the discrepancy principle with emergency stop converges in L^2, since the additional assumption of Corollary 1.2.5 is violated here). Moreover the statistics of the discrepancy principle with and without emergency stop become more similiar with increasing sample size - with the crucial difference, that the outliers as such we denote the red crosses above the blue box, thus the cases where the method performed badly) are only present in case of the discrepancy principle without emergency stop, causing non-convergence in L^2, see Figure 2. Thus here the discrepancy principle with emergency stop is superior to the discrepancy principle without emergency stop, in particular for large sample sizes. Beside that, the error is falling slower in case of the a priori parameter choice. The number of outliers falls with increasing sample size from 37 for $n = 10^3$ to 18 for $n = 10^5$, indicating the (slow) convergence in probability of the discrepancy principle. Note that $\delta_n^{true}/\delta_n^{est} \approx 1.9$ (in average), if we only consider the runs yielding outliers. This illustrates, that the lack of convergence in L^2 is caused by the occasional underestimation of the data error.

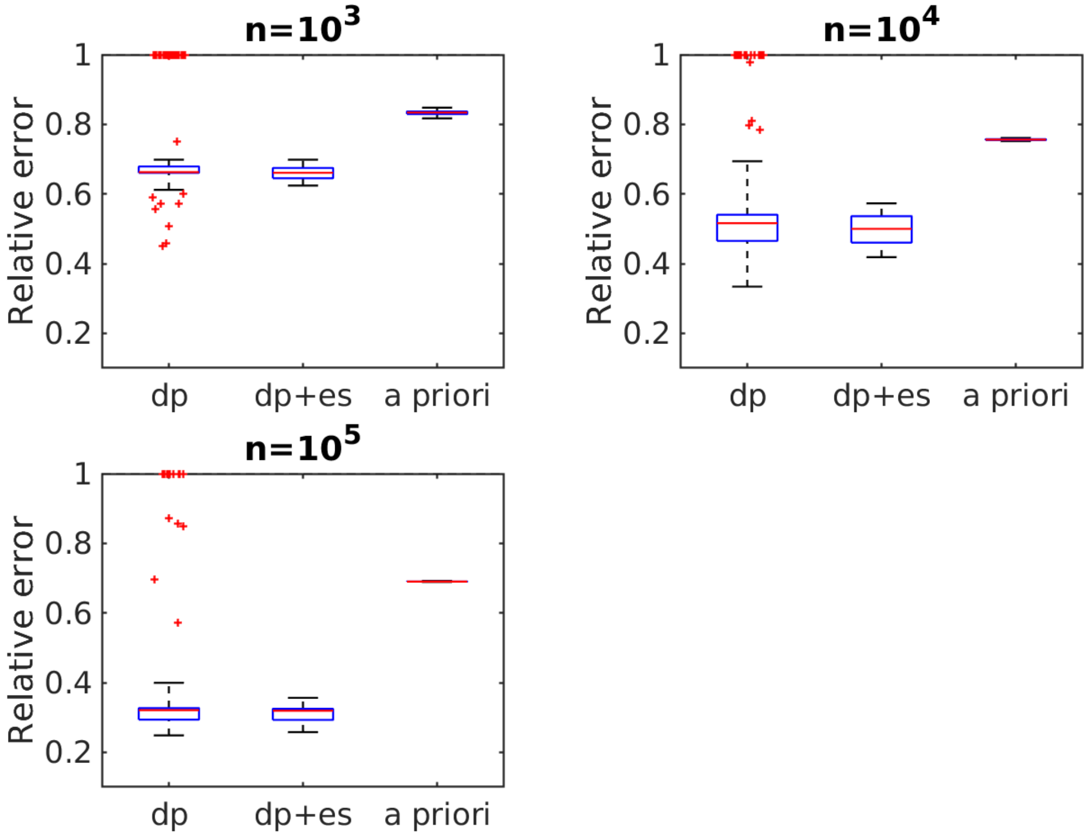

Figure 1.3: Comparison of Tikhonov regularisation with discrepancy principle (dp, Algorithm 1), discrepancy principle with emergency stop (dp+es, Algorithm 1 (optional)) and a priori choice for 'heat'. Boxplots of the relative errors $\|R_{\alpha_n}\bar{Y}_n - K^+\hat{y}\|/\|K^+\hat{y}\|$ for 200 simulations with three different sample sizes.

Table 1.1: Estimated relative L^2 error for 'heat'

n	$e_{\mathrm{d}p}$	$e_{\mathrm{d}p+es}$	$e_{apriori}$
1e3	572.49	0.66	0.83
1e4	79.45	0.49	0.76
1e5	107.19	0.31	0.69

1.8 Concluding remarks

In this chapter, we have shown how to solve a linear inverse problem under arbitrary stochastic noise of unknown error distribution (unbiased and with finite variance), if we are able to repeat the measurement independently. Important further investigations could deal with a rigorous analysis of the approach presented numerically in section 1.7.1.1 on the rescaling of the measurements, or on the relation to heuristic parameter choice rules with their noise conditions, as touched in section 1.4. It could also be worth to take a closer look at settings involving Monte-Carlo simulations, c.f. the numerical example in 1.7.1. Hereby one should probably also take errors of the forward operator into account. Finally, we did not touch nonlinear problems or problems in Banach spaces, see [EHN96],[KNS08], [SKHK12] and [IJ15] for introducing monographs. The approach of using averaged data is first of all independent of the linear Hilbert space setting, and therefore should be investigated in these scenarios as well.

Chapter 2

The white noise case

The sections 2.1 to 2.5 are, up to minor changes, submitted for publication [HJP20b].

We again consider the ill-posed equation $K\hat{x} = \hat{y}$ for a given $\hat{y} \in \mathcal{D}(K^+)$, where K^+ is the generalised inverse of the compact and linear operator K and the right hand side $\hat{y}$ is ad hoc unknown and has to be reconstructed from measurements. In the preceeding chapter we presented an approach, how to solve the problem for multiple unbiased and independent measurements of $\hat{y}$, if the arbitrary unknown error distribution has finite variance. Now we take discretisation into account and extend the results to arbitrary (unknown) white noise settings.

As an arbitrary element of an infinite-dimensional space, $\hat{y}$ cannot be measured directly, but we may measure $l(\hat{y})$ for various linear functionals $l \in \mathcal{L}(\mathcal{Y}, \mathbb{R})$. If the unknown $\hat{y}$ is for example a continuous function, one may think of performing point evaluations or measuring the integrals of that function over small parts of the domain. We will refer to these linear functionals as measurement channels in the following. We assume that we have multiple and unbiased samples on each measurement channel, corrupted randomly by additive noise. So,

$$Y_{ij} := l_j(\hat{y}) + \delta_{ij} \tag{2.1}$$

is the i-th sample on the j-th measurement channel, with $\|l_1\| = \|l_2\| = ...$ and unbiased and independent measurement errors δ_{ij}, $i, j \in \mathbb{N}$ with arbitrary unknown distribution. Thus $\big(Y_{i1} - l_1(\hat{y}) \quad ... \quad Y_{im} - l_m(\hat{y})\big)^T \in \mathbb{R}^m, i \in \mathbb{N}$ are i.i.d white noise vectors with unknown distribution. We assume, that $(l_j)_{j\in\mathbb{N}}$ is complete and square-summable, i.e. for all $y \in \mathcal{Y} \setminus \{0\}$ there is a l_j with $l_j(y) \neq 0$ and $\sum_{j=1}^{\infty} l_j(y)^2 < \infty$. For a fixed number m of measurement channels and a large number n of repetitions we obtain an approximation

$$\bar{Y}_n^{(m)} := \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n Y_{i1} \\ ... \\ \frac{1}{n}\sum_{i=1}^n Y_{im} \end{pmatrix} \approx \begin{pmatrix} l_1(\hat{y}) \\ ... \\ l_m(\hat{y}) \end{pmatrix}.$$

As a first approach we apply the ideas of Tikhonov (0.2) and minimise the following functional with finite-dimensional residuum (fdr)

$$\arg\min_{x\in\mathcal{X}} \left\| \begin{pmatrix} l_1(Kx) \\ \dots \\ l_m(Kx) \end{pmatrix} - \bar{Y}_n^{(m)} \right\|_{\mathbb{R}^m}^2 + \alpha\|x\|_{\mathcal{X}}^2. \tag{2.2}$$

The main question of this chapter is, whether the unique minimiser of (2.2), denoted by $R_\alpha^{(m)}\bar{Y}_n^{(m)}$, converges to $\hat{x}$ for $m,n\to\infty$, for adequately chosen $\alpha=\alpha(m,n)$. With the Bakushinskii veto (Theorem 0.0.1) in mind, this choice must depend on the measurement error $\|\bar{Y}_n^{(m)} - \begin{pmatrix} l_1(\hat{y}) & \dots & l_m(\hat{y})\end{pmatrix}^T\|$. Here, the i.i.d assumption yields a natural estimator

$$\delta_{m,n}^{est} := \sqrt{\frac{m}{n}s_{m,n}^2}, \tag{2.3}$$

where $s_{m,n}^2 := \frac{1}{m}\sum_{j=1}^m \frac{1}{n-1}\sum_{i=1}^n \left(Y_{ij} - \frac{1}{n}\sum_{l=1}^n Y_{lj}\right)^2$ is the mean of the sample variances. Clearly, $\delta_{m,n}^{est}$ converges to 0 in probability (and a.s. and in root mean square), iff $m/n\to 0$ (see Proposition 2.3.3 below). For the determination of the regularisation parameter α in (2.2) we once more consider the discrepancy principle (0.6) and solve

$$\left\| \begin{pmatrix} l_1(KR_\alpha^{(m)}\bar{Y}_n^{(m)}) \\ \dots \\ l_m(KR_\alpha^{(m)}\bar{Y}_n^{(m)}) \end{pmatrix} - \bar{Y}_n^{(m)} \right\| \approx \delta_{m,n}^{est} \tag{2.4}$$

(see Algorithm 3 with $C_0 = 1$ for the numerical implementation). We obtain the following convergence result for the discrepancy principle.

Corollary 2.0.1. *Assume that K is injective with dense image and that $(\delta_{ij})_{i,j\in\mathbb{N}}$ are independent and identically distributed with zero mean and bounded variance. Moreover assume, that $(l_j)_{j\in\mathbb{N}}$ is complete and square-summable. Then, with $\alpha_{m,n}$ determined by the discrepancy principle* (2.4), *we have that*

$$\lim_{\substack{m\to\infty\\ n\to\infty\\ m/n\to 0}} \mathbb{P}\left(\|R_{\alpha_{m,n}}^{(m)}\bar{Y}_n^{(m)} - K^+\hat{y}\| \ge \varepsilon\right) = 0$$

for all $\varepsilon > 0$.

All the details to this result can be found in Section 2.1, where we also more generally treat filter based regularisations, as well as a priori parameter choice rules and discretisations $l_j^{(m)}$, $j = 1,...,m$. Let us stress, that Corollary 2.0.1 guarantees convergence without any quantitative knowledge of the quality of the discretisation (error), for an arbitrary unknown error distribution. In view of the Bakushinskii veto it might be surprising, that no knowledge of some kind of discretisation error is required. Corollary 2.0.1 does not give a convergence rate, however, the numerical experiments in Section 2.4 indicate, that there might hold order optimality in various settings.

In order to obtain convergence rates we consider a second approach, which is about to first construct from the measured data in $\mathbb{R}^m$ continuous measurements in the Hilbert space $\mathcal{Y}$, see f.e. the recent preprint [GH20]. For that we solve the optimisation problem

$$Z_n^{(m)} := \arg\min_{y\in\mathcal{Y}} \left\| \begin{pmatrix} l_1(y) \\ \dots \\ l_m(y) \end{pmatrix} - \bar{Y}_n^{(m)} \right\|. \tag{2.5}$$

We restrict to discretisations, for which (2.5) is well-conditioned, see Assumption 2.2.1. For general discretisations one would need to add an additional regularisation term. Then, instead of (2.2), we solve the following functional with infinite-dimensional residuum (idr)

$$\arg\min_{x\in\mathcal{X}} \left\|Kx - Z_n^{(m)}\right\|_{\mathcal{Y}}^2 + \alpha \left\|x\right\|_{\mathcal{X}}^2 \tag{2.6}$$

and the regularisation parameter α has to be chosen accordingly to $\|Z_n^{(m)} - \hat{y}\|$. With

$$y^{(m)} := \arg\min_{y\in\mathcal{Y}} \left\| \begin{pmatrix} l_1(y) \\ \dots \\ l_m(y) \end{pmatrix} - \begin{pmatrix} l_1(\hat{y}) \\ \dots \\ l_m(\hat{y}) \end{pmatrix} \right\|.$$

we may decompose this term into a measurement error and a discretisation error

$$\|Z_n^{(m)} - \hat{y}\| \leq \|Z_n^{(m)} - y^{(m)}\| + \|y^{(m)} - \hat{y}\|.$$

Assume that we know an asymptotic bound δ_m^{disc} for the discretisation error $\|\hat{y} - y^{(m)}\|$ (which is natural in various settings, see Section 2.2). One may estimate $\|Z_n^{(m)} - y^{(m)}\|$ (see Algorithm 2), and should use that many repetitions $n(m, \delta_m^{disc})$, such that this estimator approximately equals δ_m^{disc}. The regularisation parameter α is then again determined via the discrepancy principle

$$\|KR_\alpha Z_{n(m,\delta_m^{disc})}^{(m)} - Z_{n(m,\delta_m^{disc})}^{(m)}\| \approx 2\delta_m^{disc}, \tag{2.7}$$

with $R_\alpha Z_{n(m,\delta_m^{disc})}^{(m)}$ the unique solution of (2.6) (see Algorithm 4 with $C_0 = 1$ for the numerical implementation). We obtain the following result on the convergence and the order optimality.

Corollary 2.0.2. *Assume that K is injective with dense image and that $(\delta_{ij})_{i,j\in\mathbb{N}}$ are independent with zero mean and finite variance. Moreover, the discretisation is complete and well-conditioned (see Propposition 2.2.3). Let $(\delta_m^{disc})_{m\in\mathbb{N}}$ be an known upper bound for the discretisation error converging to 0 and determine α_m with the discrepancy principle* (2.7)*. Then*

$$\lim_{m\to\infty}\mathbb{P}\left(\|R_{\alpha_m}Z^{(m)}_{n(m,\delta_m^{disc})}-K^+\hat{y}\|\geq\varepsilon\right)=0$$

for all $\varepsilon>0$. If moreover there is a $0<\nu\leq1$ and a $\rho>0$ such that $K^+\hat{y}=(K^*K)^{\nu/2}\xi$ for some $\xi\in\mathcal{X}$ with $\|\xi\|\leq\rho$, then

$$\mathbb{P}\left(\|R_{\alpha_m}Z^{(m)}_{n(m,\delta_m^{disc})}-K^+\hat{y}\|\leq L'\rho^{\frac{1}{\nu+1}}\delta_m^{\frac{\nu}{\nu+1}}\right)\to1$$

for $m\to\infty$ and some constant L'.

The rest of the chapter is organised as follows. In Section 2.1 and Section 2.2 we will show the L^2 convergence (a.k.a. convergence of the mean squared error) of a priori parameter choice rules and the convergence in probability of the discrepancy principle for the both approaches respectively. The proofs are deferred to Section 2.3 and we conclude with a numerical study in Section 2.4 and some final remarks in Section 2.5.

2.1 Approach with finite-dimensional residuum

We start with a precise and more general definition of our discretisation scheme. Therefore we introduce as follows the discretisation (operators)

$$P_m:\mathcal{Y}\to\mathbb{R}^m,\ y\mapsto\begin{pmatrix}l_1^{(m)}(y)\\ \dots\\ l_m^{(m)}(y)\end{pmatrix},\tag{2.8}$$

with the corresponding measurements

$$Y_{ij}^{(m)}:=l_j^{(m)}(\hat{y})+\delta_{ij}^{(m)}.$$

That is, the measurement channels and also the error distribution may depend on the number m of measurement channels now. We will often use, that by the Riesz representation theorem there are unique $(\eta_j^{(m)})_{j\leq m,m\in\mathbb{N}}$ such that $l_j^{(m)}(y)=(\eta_j^{(m)},y)$

for all $y \in \mathcal{Y}$. From now on we consider more generally filter-based regularisations $R_\alpha^{(m)} := F_\alpha\left((P_mK)^* P_mK\right)(P_mK)^*$, where $(F_\alpha)_\alpha$ fulfills Assumption 2.1.1 below.

Assumption 2.1.1 (Filter). *$(F_\alpha)_{\alpha>0}$ is a family of piecewise continuous real valued functions on $[0,\|K\|]^2$, with*

$$\lim_{\alpha\to 0} \sup_{\varepsilon\le\lambda\le\|K\|^2} |F_\alpha(\lambda) - 1/\lambda| = 0 \tag{2.9}$$

for all $\varepsilon > 0$ and $\lambda|F_\alpha(\lambda)| \le C_R \in \mathbb{R}$ for all $\lambda \in (0,\|K\|^2]$ and $\alpha > 0$. Moreover it has qualification $\nu_0 \ge 0$, i.e. ν_0 is maximal such that for all $0 \le \nu \le \nu_0$ there is a constant $C_\nu \in \mathbb{R}$ such that

$$\sup_{\lambda\in(0,\|K\|^2]} \lambda^{\frac{\nu}{2}} |1 - F_\alpha(\lambda)\lambda| \le C_\nu \alpha^{\frac{\nu}{2}}.$$

Hereby, for $\nu = 0$ the constant C_0 is assumed to be known. Finally, there is a constant $C_F \in \mathbb{R}$ with $|F_\alpha(\lambda)| \le C_F/\alpha$ for all $\alpha > 0$ and $\lambda \in (0,\|K\|^2]$.

Remark 2.1.2. Assumption 2.1.1 coincides with the classical ones as given in Assumption 1.1.2 in the preceding chapter up to (2.9), which is replaced by the weaker condition $\lim_{\alpha\to 0} F_\alpha(\lambda) = 1/\lambda$, for all $\lambda \in (0,\|K\|^2]$ there. However, it is easy to verify that the generating filter of all popular methods, e.g. truncated singular value, (iterated) Tikhonov or Landweber regularisation, fulfill Assumption 2.1.1. In all this cases it holds that $C_0 = 1$.

We require the discretisation to converge in the following sense.

Assumption 2.1.3 (Disretisation for finite-dimensional residuum). *There exists an injective operator $A \in \mathcal{L}(\mathcal{Y})$ such that $\lim_{m\to\infty} P_m^* P_m y = Ay$ for all $y \in \mathcal{Y}$.*

We list some popular discretisation schemes which fulfill Assumption 2.1.3, starting with the one from the introduction.

Proposition 2.1.4. *Assume that $l_j^{(m)} = l_j$ for all $j = 1, ..., m$ and $m \in \mathbb{N}$ with $(l_j)_{j\in\mathbb{N}} \subset \mathcal{L}(\mathcal{Y}, \mathbb{R})$, where $(l_j)_{j\in\mathbb{N}}$ is complete and square-summable, i.e. for all $y \in \mathcal{Y} \setminus \{0\}$ there is a l_j such that $l_j(y) \ne 0$ and there holds $\sum_{j=1}^\infty l_j(y)^2 < \infty$. Then Assumption 2.1.3 is fulfilled.*

Often the limit operator A will be the identity $Id = Id_\mathcal{Y}$, e.g. in the case when we discretise by box or hat functions.

Proposition 2.1.5. *Assume that $\mathcal{Y} = L^2([0,1])$ and we discretise by box functions, i.e. $l_j^{(m)} = (\eta_j^{(m)}, \cdot)$, with $\eta_j^{(m)} = \sqrt{m}\chi_{[\frac{j-1}{m},\frac{j}{m})}$ for $j = 1, ..., m$ and $m \ge 2$. Then Assumption 2.1.3 is fulfilled with $A = Id$.*

Proposition 2.1.6. *Assume that $\mathcal{Y} = L^2([0,1])$ and we discretise by hat functions, i.e. $l_j^{(m)} = (\eta_j^{(m)}, \cdot)$, with*

1. $\frac{\eta_j^{(m)}}{\sqrt{m-1}} = (1 - j + (m-1)x)\chi_{[\frac{j-1}{m-1}, \frac{j}{m-1})} + (j + 1 - (m-1)x)\chi_{[\frac{j}{m-1}, \frac{j+1}{m-1})}$ *for* $j = 2, ..., m-1$,

2. $\eta_1^{(m)} = \sqrt{2(m-1)}(1 + j - (m-1)x)\chi_{[\frac{j}{m-1}, \frac{j+1}{m-1})}$,

3. $\eta_m^{(m)} = \sqrt{2(m-1)}((m-1)x - j + 1)\chi_{[\frac{j-1}{m-1}, \frac{j}{m-1}]}$.

Then Assumption 2.1.3 is fulfilled with $A = Id$.

2.1.1 A priori regularisation with finite-dimensional residuum

We start with a priori regularisations and impose the following assumption on the error, which is weaker than the one in the introduction. Basically, solely independence on each measurement channel and a uniform boundedness of the variances are required.

Assumption 2.1.7 (Error for a priori regularisation)**.** *For all $m, j \in \mathbb{N}$, the random variables $\left(\delta_{ij}^{(m)}\right)_{i \in \mathbb{N}}$ are independent with zero mean and there exists $C_d \in \mathbb{R}$ with*

$$\sup_{\substack{m,i,j \in \mathbb{N} \\ j \le m}} \mathbb{E}[\delta_{ij}^{(m)^2}] \le C_d.$$

Since the sample variance depends on the data, we set $s_{m,n}^2 = 1$ here, such that $\delta_{m,n}^{est} = \sqrt{m/n}$. This has the advantage, that the regularisation parameter α is independent of the measurements $Y_{ij}^{(m)}$. We obtain convergence in L^2 for a priori regularisation.

Theorem 2.1.8. *Assume that the discretisation fulfills Assumption 2.1.3 and that the error is accordingly to Assumption 2.1.7 and $(F_\alpha)_{\alpha>0}$ fulfills Assumption 2.1.1. Take an a priori parameter choice rule with $\alpha(\delta) \overset{\delta \to 0}{\longrightarrow} 0$ and $\delta / \sqrt{\alpha(\delta)} \overset{\delta \to 0}{\longrightarrow} 0$. Then there holds*

$$\lim_{\substack{m,n \to \infty \\ m/n \to 0}} \mathbb{E} \| R_{\alpha(\delta_{m,n}^{est})}^{(m)} \bar{Y}_n^{(m)} - K^+ \hat{y} \|^2 = 0.$$

2.1.2 A posteriori regularisation with finite dimensional residuum

We turn our attention to the discrepancy principle. The regularisation parameter is determined through

$$\|P_m K R_\alpha^{(m)} \bar{Y}_n^{(m)} - \bar{Y}_n^{(m)}\| \approx \delta_{m,n}^{est} \tag{2.10}$$

and in the definition of $\delta_{m,n}^{est} = \sqrt{s_{m,n}^2 m/n}$ we choose the mean of the sample variances

$$s_{m,n}^2 := \frac{1}{m}\sum_{j=1}^{m} \frac{1}{n-1}\sum_{i=1}^{n} \left(Y_{ij}^{(m)} - \frac{1}{n}\sum_{l=1}^{n} Y_{lj}^{(m)} \right)^2 ,$$

since we will need a sharp estimation of the right hand side. We implement the discrepancy principle with Algorithm 3.

Algorithm 3 Discrepancy principle with fdr approach

```
Choose $\tau > C_0$ (from Assumption 2.1.1) and $q \in (0,1)$;
Input: Measurements $Y_{ij}^{(m)} = l_j^{(m)}(\hat{y}) + \delta_{ij}^{(m)}$ with $i \le n$ and $j \le m$;
Set $\bar{Y}_n^{(m)} = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} Y_{i1}^{(m)} \\ \dots \\ Y_{im}^{(m)} \end{pmatrix}$;
Set $\delta_{m,n}^{est} = \sqrt{\frac{m}{n}\frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{ij}^{(m)} - \frac{1}{n}\sum_{l=1}^{n} Y_{lj}^{(m)}\right)^2}$;
$k = 0$;
while $\left\| \begin{pmatrix} l_1^{(m)}(K R_{q^k}^{(m)} \bar{Y}_n^{(m)}) \\ \dots \\ l_m^{(m)}(K R_{q^k}^{(m)} \bar{Y}_n^{(m)}) \end{pmatrix} - \bar{Y}_n^{(m)} \right\| > \tau \delta_{m,n}^{est}$ do
    $k = k + 1$;
end while
$\alpha_{m,n} = q^k$;
```

Algorithm 3 terminates (with probability tending to 1 for $m \to \infty$), if K has dense image and (for m large enough) $\mathbb{E}(Y_{11}^{(m)} - \mathbb{E}Y_{11}^{(m)})^2 > 0$, for details see section 1.2 of the preceding chapter. We extend the assumptions of the error in the introduction.

Assumption 2.1.9 (Error for a posteriori regularisation)**.** *It holds that either*

1. *the random variables* $\left(\delta_{ij}^{(m)}\right)_{i,j,m\in\mathbb{N}}$ *are i.i.d. with zero mean and bounded variance, or,*

2. *there are $C_d \in \mathbb{R}$ and $p > 1$ such that for all $m \in \mathbb{N}$ the random variables $\left(\delta_{ij}^{(m)}\right)_{i,j\in\mathbb{N}}$ are i.i.d with zero mean and $\frac{\mathbb{E}|\delta_{ij}^{(m)}|^{2p}}{(\mathbb{E}\delta_{ij}^{(m)^2})^p} \le C_d$.*

The main difference between Assumption 2.1.9.1 and 2.1.9.2 is, that for the latter the error distribution may vary with m, to the cost of a uniform moment condition.

Remark 2.1.10. Assumption 2.1.9.2 guarantees, that the error distribution does not degenerate too much. It is trivially fulfilled, if f.e. $\delta_{ij}^{(m)} \overset{d}{=} c_m X$, with $\mathbb{E}|X|^{2p} < \infty, (c_m)_{m\in\mathbb{N}} \subset \mathbb{R}\setminus\{0\}$.

Now we are ready to prove convergence of the discrepancy principle. In contrast to the previous section, where we showed convergence in L^2 for a priori regularisation methods, the result will now be on convergence in probability, as convergence in L^2 is not expected (compare this to the counterexample in section 1.2.1 of the preceding chapter).

Theorem 2.1.11. *Assume that K is injective with dense image and that the discretisation fulfills Assumption 2.1.3 and that the error is accordingly to Assumption 2.1.9 and $(F_\alpha)_{\alpha>0}$ fulfills Assumption 2.1.1 with a qualification $\nu_0 > 1$. Then, with $\alpha_{m,n}$ the output of Algorithm 3,*

$$\lim_{\substack{m,n\to\infty \\ m/n\to 0}} \mathbb{P}\left(\|R_{\alpha_{m,n}}^{(m)}\bar{Y}_n^{(m)} - K^+\hat{y}\| \ge \varepsilon\right) = 0$$

for all $\varepsilon > 0$.

Corollary 2.0.1 is an easy consequence of Theorem 2.1.11 and Proposition 2.1.4. We conclude the section with some remarks regarding Assumptions 2.1.3 and 2.1.9.

Remark 2.1.12. A natural generalisation of the assumption in the introduction, that $(l_j)_{j\in\mathbb{N}}$ is complete, would be the following: For all $y \in \mathcal{Y}\setminus\{0\}$ there is a $\varepsilon > 0$ such that $\|P_m y\| \ge \varepsilon$ for m large enough. However, the following counter example shows that this condition is not sufficient to guarantee, that the discretisation error tends to 0: Let $(e_j)_{j\in\mathbb{N}}$ be an orthonormal basis of $\mathcal{Y}$. Set $l_j^{(m)}(y) = (y, e_j)$ for $j = 2, ..., m$ and $l_1^{(m)}(y) = (y, e_1/\sqrt{2} + e_{m+1}/\sqrt{2})$. For $y \ne 0$ we set $\varepsilon = |(y, e_j)|/2$ with $j = \min\{j\ : (y, e_j) \ne 0\}$. Then clearly, $\|P_m y\| \ge \varepsilon$ for m large enough. But, it is $\mathcal{N}(P_m) =< e_1/\sqrt{2} - e_{m+1}/\sqrt{2}, e_{m+2}, e_{m+3}, ..., >$, thus $P_{\mathcal{N}(P_m)}e_1 = e_1/\sqrt{2} \not\to 0$ for $m \to \infty$.

Remark 2.1.13. As already mentioned, Assumption 2.1.9 excludes distributions which are too degenerated and guarantees, that $\mathbb{E}\delta_{11}^{(m)^2}$ is in some sense uniformly estimatable. We quickly sketch what can go wrong, if the distributions degenerate too much. Assume that $(\delta_{ij}^{(m)})_{ij}$ are i.i.d. for all $m \in \mathbb{N}$, with

$$\mathbb{P}(\delta_{ij}^{(m)} = x) = \begin{cases} \frac{1}{m^4} & \text{for } x = -\sqrt{m^4-1} \\ \frac{m^4-1}{m^4} & \text{for } x = 1/\sqrt{m^4-1} \end{cases}.$$

Thus $\mathbb{E}\delta_{11}^{(m)} = 0$ and $\mathbb{E}\delta_{11}^{(m)2} = 1$, but, for any $p > 1$,

$$\frac{\mathbb{E}|\delta_{11}^{(m)}|^{2p}}{(\mathbb{E}\delta_{11}^{(m)2})^p} \geq \frac{1}{m^4}|\sqrt{m^4-1}|^{2p} = \left(1 - \frac{1}{m^4}\right)|m^4-1|^p \to \infty$$

as $m \to \infty$. Thus Assumption 2.1.9 is violated and with the choice $n(m) = m^2$ it holds that $\lim_{m\to\infty} \frac{m}{n(m)} = 0$, but we have that

$$\begin{aligned}\mathbb{P}\left(\delta_{m,n(m)}^{est} = 0\right) &= \mathbb{P}\left(s_{m,n(m)}^2 = 0\right) \\ &= \mathbb{P}\left(\delta_{ij}^{(m)} = 1/\sqrt{m^4-1}\ ,\ i = 1,...,m^2, j = 1,...,m\right) \\ &= \left(1 - \frac{1}{m^4}\right)^{m^3} = \left(\left(1 - \frac{1}{m^4}\right)^{m^4}\right)^{\frac{1}{m}} \to 1\end{aligned}$$

as $m \to \infty$. Thus with asymptotic probability 1 the discrepancy principle cannot even be applied for this choice of n. The number of repetitions $n(m) = m$ is simply too small to estimate the variance of $\delta_{11}^{(m)}$ adequately.

2.2 Approach with infinite-dimensional residuum

We turn our attention to the second approach (2.6). The strategy is to use the measured data to construct virtual measurements in the infinite-dimensional Hilbert space $\mathcal{Y}$ and then to regularise the infinite-dimensional problem using classical methods. For the regularisation we will need in the following an upper bound for the discretisation error, which we denote by $\delta_m^{disc} \geq \|\hat{y} - P_m^+ P_m \hat{y}\|$. Decomposing the true data error yields

$$\|\hat{y} - P_m^+ \bar{Y}_n^{(m)}\| \leq \|\hat{y} - P_m^+ P_m \hat{y}\| + \|P_m^+ P_m \hat{y} - P_m^+ \bar{Y}_n^{(m)}\|.$$

As in the approach with a finite-dimensional residuum, there is a generic way to estimate the (projected) measurement error $\|P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - P_m^+ P_m \hat{y}\|$. So that it is natural to choose the number of repetitions $n(m, \delta_m^{disc})$ in such a way, that this estimator approximately equals the discretisation error δ_m^{disc}. After that one may use any deterministic regularisation together with total estimated noise level

$$2\delta_m^{disc} \approx \|\hat{y} - P_m^+ P_m \hat{y}\| + \|P_m^+ P_m \hat{y} - P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)}\| \geq \|\hat{y} - P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)}\|. \quad (2.11)$$

We again consider regularisations $R_\alpha := F_\alpha(K^*K)K^*$ induced by a regularising filter fulfilling Assumption 2.1.1 and impose the following restrictions on the discretisation and our a priori knowledge of it.

Assumption 2.2.1 (Discretisation for infinite dimensional residuum)**.** *We assume that we know an asymptotic upper bound $(\delta_m^{disc})_{m\in\mathbb{N}}$ for the discretisation error and asymptotic upper and lower bounds $(c_m)_{m\in\mathbb{N}}, (C_m)_{m\in\mathbb{N}}$ for the singular values $(\sigma_j^{(m)})_{j\leq m, m\in\mathbb{N}}$ of $(P_m)_{m\in\mathbb{N}}$. More precisely, these bounds have to fulfill $\|\hat{y} - P_m^+ P_m \hat{y}\| \leq \delta_m^{disc}, c_m \leq \sigma_j^{(m)} \leq C_m$ for all $j = 1, .., m$ and m large enough, and $\delta_m^{disc} \to 0$ as $m \to \infty$ and*

$$\limsup_{m\in\mathbb{N}} \kappa(P_m) := \limsup_{m} \|P_m\|\|P_m^+\| = \limsup_{m\in\mathbb{N}} \frac{\max_{j=1,...,m} \sigma_j^{(m)}}{\min_{j=1,...,m} \sigma_j^{(m)}} \leq \limsup_{m\in\mathbb{N}} \frac{C_m}{c_m} < \infty. \tag{2.12}$$

Often the stability assumption (2.12) can be guaranteed by an angle condition for the unique $\eta_j^{(m)} \in \mathcal{Y}$, which fulfill $l_j^{(m)}(y) = (\eta_j, y)$ for all $y \in \mathcal{Y}$.

Proposition 2.2.2. *Assume that*

$$\sup_{m\in\mathbb{N}} \sup_{j\leq m} \sum_{i\neq j} \frac{|(\eta_i^{(m)}, \eta_j^{(m)})|}{\|\eta_1^{(m)}\|^2} \leq c < 1.$$

Then $c_m := \|\eta_1^{(m)}\|^2(1-c) \leq \sigma_j^{(m)} \leq \|\eta_1^{(m)}\|^2(1+c) =: C_m$ for $j = 1, .., m$ and m large enough and thus $\kappa(P_m) \leq \frac{1+c}{1-c}$.

Clearly, the angle condition is always satisfied for orthogonal discretisations. We now show that Assumption 2.2.1 is fulfilled for various popular discretisation schemes. We start with the example from the introduction.

Proposition 2.2.3. *Assume that $l_j^{(m)} = l_j = (\eta_j, \cdot)$ for all $j = 1, ..., m$ and $m \in \mathbb{N}$, with $(l_j)_{j\in\mathbb{N}} \subset \mathcal{L}(\mathcal{Y}, \mathbb{R})$ and $(\eta_j)_{j\in\mathbb{N}} \subset \mathcal{Y}$, and that we know c and δ_m^{disc} such that $\delta_m^{disc} \geq \|\hat{y} - P_m^+ P_m \hat{y}\|$ and $(l_j)_{j\in\mathbb{N}}$ is complete, i.e. for all $y \in \mathcal{Y} \setminus \{0\}$ there is a l_j such that $l_j(y) \neq 0$, and well-conditioned, that is*

$$\sup_{j\in\mathbb{N}} \sum_{\substack{i=1 \\ i\neq j}}^{\infty} |(\eta_i, \eta_j)|/\|\eta_1\|^2 \leq c < 1.$$

Then Assumption 2.2.1 is fulfilled for δ_m^{disc} and $c_m = 1 - c, C_m = 1 + c$.

Next we consider discretisation along the singular directions of K.

Proposition 2.2.4. *Assume that the singular value decomposition $(\sigma_l, v_l, u_l)_{l\in\mathbb{N}}$ of K is known. Then for the discretisation $l_j^{(m)} = (u_j, \cdot)$ Assumption 2.2.1 is (asymptotically) fulfilled, with the bounds $\delta_m^{disc} = f_m\sigma_{m+1}$ (where f_m is any sequence with $f_m \to \infty$ as $m \to \infty$) and $c_m = C_m = 1$.*

In many important cases, for example if K is a Fredholm integral equation with sufficient smoothing kernel, Assumption 2.2.1 is also fulfilled for discretisation with box or hat functions.

Proposition 2.2.5. *Consider $\mathcal{X} = \mathcal{Y} = L^2(0,1)$ and $\eta_j^{(m)}$ the box functions from Proposition. If $\hat{y}$ is continuously differentiable, then Assumption 2.2.1 is fulfilled with bounds $\delta_m = f_m/m$ and $c_m = C_m = 1$, where $(f_m)_m$ is arbitrary with $\lim_m f_m = \infty$.*

Proposition 2.2.6. *Consider $\mathcal{X} = \mathcal{Y} = L^2(0,1)$ and $\eta_j^{(m)}$ the hat functions from Proposition. If $\hat{y}$ is continuously differentiable, then Assumption 2.2.1 is fulfilled with bounds $\delta_m = f_m/m$ and $c_m = 1/6$ and $C_m = 7/6$, where $\lim_m f_m = \infty$. If $\hat{y}$ is twice continuously differentiable, then Assumption 2.2.1 is fulfilled with bounds $\delta_m = f_m/m^2$ and $c_m = 1/6$ and $C_m = 7/6$, with $\lim_m f_m = \infty$.*

It remains to determine the number of repetitions $n(m, \delta_m^{disc})$, such that the (back projected) measurement error fulfills $\|P_m^+ P_m \hat{y} - P_m^+ \bar{Y}^{(m)}_{n(m,\delta_m^{disc})}\| \approx \delta_m^{disc}$. This number depends on the singular value composition of P_m and the variance $\mathbb{E}{\delta_{11}^{(m)}}^2$. More precisely, with $(\sigma_j^{(m)}, v_j^{(m)}, u_j^{(m)})_{j\leq m}$ the singular value decomposition of P_m and $e_1, ...e_m$ is the standard basis of $\mathbb{R}^m$, it is

$$\begin{aligned}\|P_m^+\bar{Y}_n^{(m)} - P_m^+P_m\hat{y}\|^2 &= \sum_{j=1}^m \frac{1}{{\sigma_j^{(m)}}^2}\left(\sum_{l=1}^m\sum_i \delta_{ij}^{(m)}/n(v_j, e_l)\right)^2\\ \Longrightarrow \mathbb{E}\|P_m^+\bar{Y}_n^{(m)} - P_m^+P_m\hat{y}\|^2 &= \frac{\mathbb{E}{\delta_{11}^{(m)}}^2}{n}\sum_{j=1}^m \frac{1}{{\sigma_j^{(m)}}^2}.\end{aligned}$$

Thus with our lower bound $c_m \leq \sigma_j^{(m)}$ we determine

$$n(m, \delta_m^{disc}) := \min\left\{n \geq 2 \ : \ \frac{ms_{m,n}^2}{nc_m^2} \leq {\delta_m^{disc}}^2\right\},$$

with $s_{m,n}^2 = 1$ or $s_{m,n}^2 = \frac{1}{m}\sum_{j=1}^m \frac{1}{n-1}\sum_{i=1}^n \left(Y_{ij}^{(m)} - \frac{1}{n}\sum_{l=1}^n Y_{lj}^{(m)}\right)^2$.

2.2.1 A priori regularisation for infinite-dimensional residuum

For a priori regularisations we set $s_{m,n}^2 = 1$, so that $n(m, \delta)$ and the measurements $Y_{ij}^{(m)}$ are independent. The convergence result holds true with the same assumption for the error as in Section 2.1.1.

Theorem 2.2.7. *Assume that the discretisation fulfills Assumption 2.2.1 and that the error is accordingly to Assumption 2.1.7 and $(F_\alpha)_{\alpha>0}$ fulfills Assumption 2.1.1. Take an a priori parameter choice rule with $\alpha(\delta) \overset{\delta\to 0}{\longrightarrow} 0$ and $\delta/\sqrt{\alpha(\delta)} \overset{\delta\to 0}{\longrightarrow} 0$. Then there holds*

$$\lim_{m\to\infty} \mathbb{E}\|R_{\alpha(\delta_m^{disc})} P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - K^+\hat{y}\|^2 = 0$$

for $n(m, \delta_m^{disc}) = \lceil \frac{m}{c_m^2 \delta_m^{disc^2}} \rceil$.

Remark 2.2.8. Note that for a priori regularisation one can relax the condition on δ_m^{disc} in Assumption 2.2.1 to $\lim_{m\to\infty} \delta_m^{disc} = 0$ and $\limsup_{m\to\infty} \frac{\delta_m^{disc}}{\|\hat{y} - P_m^+ P_m \hat{y}\|} > 0$.

2.2.2 A posteriori regularisation for infinite-dimensional residuum

Now we determine the stopping index $n(m, \delta_m^{disc})$ more accurately with the sample variance and set $s_{m,n}^2 := \frac{1}{m}\sum_{j=1}^m \frac{1}{n-1}\sum_{i=1}^n \left(Y_{ij}^{(m)} - \frac{1}{n}\sum_{l=1}^n Y_{lj}^{(m)}\right)^2$. We implement the discrepancy principle in Algorithm 2.

Algorithm 4 Discrepancy principle with idr approach

```
Choose $\tau > C_0$ (from Assumption 2.1.1) and $q \in (0,1)$;
Input: Number of measurement channels $m$, measurements $Y_{ij}^{(m)}$, $j \leq m, i \in \mathbb{N}$, upper bound $\delta_m^{disc}$ for discretisation error, lower bound $c_m$ for singular values of $P_m$;
Determine $n(m, \delta_m^{disc}) := \min \left\{ n' \geq 1 \ : \ \frac{m s_{m,n'}^2}{n' c_m^2} \leq \delta_m^{disc 2} \right\}$ from measurements $Y_{ij}^{(m)}$.
Set $\bar{Y}_{n(m,\delta_m^{disc})}^{(m)} = \frac{1}{n(m,\delta_m^{disc})} \sum_{i=1}^{n(m,\delta_m^{disc})} \begin{pmatrix} Y_{i1}^{(m)} & \ldots & Y_{in}^{(m)} \end{pmatrix}^T$;
$k = 0$;
while $\|(K R_{q^k} P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)}\| > 2\tau\delta_m^{disc}$ do
    $k = k + 1$;
end while
$\alpha_m = q^k$;
```

Algorithm 4 terminates under the same conditions as Algorithm 3. The back propagating of the measurements induces correlations, which forces us to impose slightly stricter conditions on the error distribution than in the setting before. On the other hand, the regularisation is now done in $\mathcal{Y}$ (no matter which m), which allows to use classical results to obtain a convergence rate.

Theorem 2.2.9. *Assume that K is injective with dense image and that the discretisation fulfills Assumption 2.2.1 and that the error is accordingly to Assumption 2.1.9, with $p \geq 2$ in the case of 2.1.9.2 and $(F_\alpha)_{\alpha>0}$ fulfills Assumption 2.1.1 with a qualification $\nu_0 > 1$. For $\tau > C_0$, let α_m and $\bar{Y}_{n(m,\delta_m^{disc})}^{(m)}$ be the output of the discrepancy principle as implemented in Algorithm 4. Then*

$$\lim_{m\to\infty} \mathbb{P}\left(\|R_{\alpha_m} P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - K^+\hat{y}\| \geq \varepsilon \right) = 0.$$

If moreover, there is a $0 < \nu \leq \nu_0 - 1$ and a $\rho > 0$ such that $K^+\hat{y} = (K^*K)^{\nu/2}\xi$ for some $\xi \in \mathcal{X}$ with $\|\xi\| \leq \rho$, then

$$\mathbb{P}\left(\|R_{\alpha_m} P_m{}^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - K^+\hat{y}\| \leq L' \rho^{\frac{1}{\nu+1}} \left(\delta_m^{disc}\right)^{\frac{\nu}{\nu+1}} \right) \to 1$$

for $m \to \infty$ and some constant L'.

Now Corollary 2.0.2 in the introduction is an easy consequence of Theorem 2.2.9 and Proposition 2.2.3.

2.3 Proofs

In this section we collect the proofs. We will need the singular value decomposition of an injective compact operator A (see [Cav11]): there exists a monotone sequence $\|A\| = \sigma_1 \geq \sigma_2 \geq \ldots > 0$. Moreover there are families of orthonormal vectors $(u_l)_{l\leq \dim(\mathcal{R}(A))}$ and $(v_l)_{l\leq \dim(\mathcal{R}(A))}$ with $\overline{span(u_l : l \leq \dim(\mathcal{R}(A))} = \overline{\mathcal{R}(A)}$, $\overline{span(v_l : l \leq \dim(\mathcal{R}(A))} = \mathcal{N}(A)^\perp$ such that $Av_l = \sigma_l v_l$ and $A^* u_l = \sigma_l v_l$.

2.3.1 Proofs for finite-dimensional residuum

The assumptions for the discretisation when using the first approach (with finite-dimensional residuum) are such, that the discretised operators $K^*P_m^*P_mK$ converge uniformly to a compact and injective operator K^*AK. The uniform convergence guarantees, that the eigenvalues and spaces of the former converge pointwise to the ones of the latter, and the injectivity of the limit operator assures, that the unknown $\hat{x}$ is determined arbitrarily precisely by finitely many eigenvectors of the latter. We make this precise with the following lemma.

Lemma 2.3.1. *Assume that K is injective and that Assumption 2.1.3 holds true. Then*

$$\|K^*P_m^*P_mK - K^*AK\| \to 0$$

for $m \to \infty$ and K^*AK is injective, compact, self-adjoint and positive semidefinite. Denote by $(\lambda_j^{(m)})_{j\leq m}$ and $(\lambda_j^{(\infty)})_{j\in\mathbb{N}}$ the nonzero eigenvalues with corresponding orthonormal eigenvectors $(v_j^{(m)})_{j\leq m}$ of
$K^*P_m^*P_mK$ and K^*AK respectively, ordered decreasingly. Then

1. *$\lim_{m\to\infty} \lambda_j^{(m)} = \lambda_j^{(\infty)}$ for all $j \in \mathbb{N}$, and*
2. *for all $x \in \mathcal{X}$ and $\varepsilon > 0$, there is a $M = M(x,\varepsilon) \in \mathbb{N}$, such that*

$$\limsup_{m\to\infty} \sum_{j=M+1}^{m} (x, v_j^{(m)})^2 \leq \varepsilon.$$

Proof.

Denote by (σ_j, u_j, v_j) the singular value decomposition of K and set $A_m = P_m^*P_m$ and
$C := \max\{\|A\|, \sup_m \|A_m\|\} < \infty$ (uniform boundedness principle). For $\varepsilon > 0$ arbitrary define $M \in \mathbb{N}$ implicitly through $2C\sigma_{M+1} \leq \varepsilon/2$. Then

$$
\begin{aligned}
&\|A_mK - AK\| \\
&= \sup_{\substack{x\in\mathcal{X}\\ \|x\|=1}} \|A_mKx - AKx\| = \sup_{\substack{\sum\alpha_j^2=1\\ x=\sum\alpha_j u_j}} \left\|\sum_{j=1}^{\infty} \alpha_j (A_mK - AK)u_j\right\| \\
&\le \sup_{\substack{\sum\alpha_j^2=1\\ x=\sum\alpha_j u_j}} \sum_{j=1}^{M} \sigma_j|\alpha_j|\,\|(A_m - A)v_j\| + \sup_{\substack{\sum\alpha_j^2=1\\ x=\sum\alpha_j u_j}} \left\|(A_m - A)\sum_{j=M+1}^{\infty} \sigma_j\alpha_j v_j\right\| \\
&\le \sigma_1 \sum_{j=1}^{M} \|(A_m - A)v_j\| + \|A_m - A\| \sup_{\substack{\sum\alpha_j^2=1\\ x=\sum\alpha_j u_j}} \left\|\sum_{j=M+1}^{\infty} \sigma_j\alpha_j v_j\right\| \\
&\le \sigma_1 \sum_{j=1}^{M} \|(A_m - A)v_j\| + 2C\sigma_{M+1} \le \sigma_1 \sum_{j=1}^{M} \|(A_m - A)v_j\| + \varepsilon/2
\end{aligned}
$$

Because $A_m \to A$ pointwise there is an $m_0 \in \mathbb{N}$, such that $\sigma_1 \sum_{j=1}^{M} \|(A_m - A)v_j\| \le \varepsilon/2$ for all $m \ge m_0$, thus $A_mK \to AK$ and therefore $K^*A_mK \to K^*AK$ for $m \to \infty$ uniformly. Since $K^*P_m^*P_mK$ is compact, self-adjoint and positive semidefinite, so is K^*AK as its uniform limit. Then (1.) holds by Section 6 of [BO91]. We define iteratively $I_1 := \{j \,:\, \lambda_j^{(\infty)} = \lambda_1^{(\infty)}\}$, $I_i := \{j \,:\, \lambda_j^{(\infty)} = \lambda_{\max(I_{i-1})+1}\}$. So the cardinality of I_i is the algebraic multiplicity of the i-th largest eigenvalue of K^*AK. We define the corresponding eigenspaces $E_i := \mathit{span}\left(v_j^{(\infty)}\ ,\ j \in I_i\right)$, $E_i^m := \mathit{span}\left(v_j^{(m)}\ ,\ j \in I_i\right)$. With $P_{E_i}, P_{E_i^m}$ the orthogonal projections onto E_i and E_i^m, by Theorem 7.1 of [BO91] there is a constant C_i such that $\|P_{E_i^m} - P_{E_i}\| \le C_i\|K^*P_m^*P_mK - K^*AK\|$ (for m sufficiently large). Thus there is a $M \in \mathbb{N}$ with $M = \sum_{i=1}^{i^*} |I_i|$ for some $i^* \in \mathbb{N}$ such that

$$
\begin{aligned}
\left|\sum_{j=1}^{M}\left(\hat{x}, v_j^{(m)}\right)^2 - \sum_{j=1}^{M}\left(\hat{x}, v_j^{(\infty)}\right)^2\right| &\le \sum_{i=1}^{i^*} \left|\|P_{E_i^m}\hat{x}\|^2 - \|P_{E_i}\hat{x}\|^2\right| \\
&\le \sum_{i=1}^{i^*} \left(|\|P_{E_i^m}\hat{x}\| + \|P_{E_i}\hat{x}\|\right) \left|\|P_{E_i^m}\hat{x}\| - \|P_{E_i}\hat{x}\|\right| \\
&\le 2\|\hat{x}\| \sum_{i=1}^{i^*} \|P_{E_i}^m\hat{x} - P_{E_i}\hat{x}\| \\
&\le 2\|\hat{x}\|^2 \|K^*P_m^*P_mK - K^*AK\| \sum_{i=1}^{i^*} C_i \le \varepsilon/2
\end{aligned}
$$

for m sufficiently large and

$$\left| \|\hat{x}\|^2 - \sum_{j=1}^{M} \left(\hat{x}, v_j^{(\infty)}\right)^2 \right| = \sum_{j=M+1}^{\infty} (\hat{x}, v_j^{(\infty)})^2 \leq \varepsilon/2,$$

where the second assertion followed from the injectivity of K^*AK. Thus

$$\begin{aligned}\sum_{j=M+1}^{m} \left(\hat{x}, v_j^{(m)}\right)^2 &\leq \left\|P_{(P_mK)^\perp}\hat{x}\right\|^2 - \sum_{j=1}^{M} \left(\hat{x}, v_j^{(m)}\right)^2 \\ &\leq \|\hat{x}\|^2 - \sum_{j=1}^{M} \left(\hat{x}, v_j^{(\infty)}\right)^2 + \sum_{j=1}^{M} \left(\hat{x}, v_j^{(m)}\right)^2 - \sum_{j=1}^{M} \left(\hat{x}, v_j^{(\infty)}\right)^2 \leq \varepsilon\end{aligned}$$

for m sufficiently large. □

2.3.1.1 Proof of Proposition 2.1.4

It is $\sup_{m\in\mathbb{N}} \|P_my\| = \sup_{m\in\mathbb{N}} \sum_{j=1}^{m} l_j(y)^2 < \infty$, thus $\sup_m \|P_m\| < \infty$ and with the embedding $\mathbb{R}^m \subset l^2(\mathbb{N})$ it follows that $\lim P_my = P_\infty y$, with $P_\infty y = \begin{pmatrix} l_1(y) & l_2(y) & ... \end{pmatrix}$. Thus $P_m^*P_my \to Ay$ with $A = P_\infty^*P_\infty$ and A is injective because of the completeness condition.

2.3.1.2 Proof of Proposition 2.1.5

Since smooth functions are dense in L^2, it suffices to consider the case where y is smooth. We have that $P_m^*P_m = P_m^+P_m = P_{\mathcal{N}(P_m)^\perp}$ and $\mathcal{N}(P_m)^\perp := \{\sum_{j=1}^{m} \alpha_j\Lambda_j^{(m)}\}$ is the set of all functions constant on a homogeneous grid with m elements. Since the set of all functions constant on a homogeneous grid is dense in the set of smooth functions, the claim follows.

2.3.1.3 Proof of Proposition 2.1.6

As above w.l.o.g. y is assumed to be smooth. We denote by $A_m \in \mathbb{R}^{m\times m}$ the matrix representing $P_m : \mathcal{N}(P_m)^\perp \to \mathbb{R}^m$ with respect to the bases $(\eta_j^{(m)})_{j=1,...,m} \subset \mathcal{N}(P_m)^\perp$ and $(e_j)_{j=1,...,m} \subset \mathbb{R}^m$, where the latter is the canonical basis of $\mathbb{R}^m$. So

$$P_m^*P_m\eta_j^{(m)} = \sum_{i=1}^{m} (A_m^*A_m)_{ij}\, \eta_i^{(m)},$$

and

$$(A_m)_{ij} = \left(P_m \eta_i^{(m)}, e_j\right)_{\mathbb{R}^m} = l_j^{(m)}(\eta_i^{(m)}) = (\eta_j^{(m)}, \eta_i^{(m)})_{\mathcal{Y}}$$

with

$$(\eta_j^{(m)}, \eta_i^{(m)}) = \begin{cases} 2/3 & , i = j \\ 1/3 & , |i-j| = 1, \min(i,j) = 1 \text{ or } \max(i,j) = m \\ 1/6 & , |i-j| = 1, \min(i,j) > 1 \text{ and } \max(i,j) < m \\ 0 & , else. \end{cases}$$

So it is

$$\|P_m\| \leq \sqrt{\|P_m\|_1 \|P_m\|_\infty} = \max_{j=1,\dots,m} \sum_{i=1}^m |(A_m)_{ij}| = \frac{7}{6},$$

where $\|.\|$, $\|.\|_1$ and $\|.\|_\infty$ are the spectral, the maximum absolute column and row norm respectively, and

$$P_m^* P_m \eta_j^{(m)} = \frac{\eta_{j-2}^{(m)}}{36} + \frac{2\eta_{j-1}^{(m)}}{9} + \frac{\eta_j^{(m)}}{2} + \frac{2\eta_{j+1}^{(m)}}{9} + \frac{\eta_{j+2}^{(m)}}{36},$$

for $j = 4, ..., m-3$. Denote by $y_m = \sum_{j=1}^m y\left(\frac{j-1}{m-1}\right) \eta_j^{(m)} \sqrt{\frac{3}{2(m-1)}}$ the interpolating spline of y, then

$$
\begin{aligned}
&\|y - P_m^* P_m y\| \\
&\leq \|y_m - P_m^* P_m y_m\| + \|(I - P_m^* P_m)(y - y_m)\| \\
&\leq \left\| \sum_{j=1}^{m} y\left(\frac{j-1}{m-1}\right) \sqrt{\frac{3}{2(m-1)}} \left(I_m - P_m^* P_m\right) \eta_j^{(m)} \right\| + 2\|y - y_m\| \\
&\leq 2\|yg - y_m\| + 6 \sup_t |y(t)| \sqrt{\frac{3}{2(m-1)}} \left(1 + \frac{7^2}{6^2}\right) + \\
&\left\| \sum_{j=4}^{m-3} \left(\frac{y\left(\frac{j}{m-1}\right)}{2} - \frac{2y\left(\frac{j+1}{m-1}\right)}{9} - \frac{2y\left(\frac{j-1}{m-1}\right)}{9} - \frac{y\left(\frac{j+2}{m-1}\right)}{36} - \frac{y\left(\frac{j-2}{m-1}\right)}{36} \right) \frac{\sqrt{3}\eta_j^{(m)}}{\sqrt{2(m-1)}} \right\| \\
&\leq 2\|y - y_m\| + 30 \sup_t |y(t)| \frac{1}{\sqrt{m-1}} \\
&+ \sup_{j \leq m} \left| \frac{y\left(\frac{j}{m-1}\right)}{2} - \frac{2y\left(\frac{j+1}{m-1}\right)}{9} - \frac{2y\left(\frac{j-1}{m-1}\right)}{9} - \frac{y\left(\frac{j+2}{m-1}\right)}{36} - \frac{y\left(\frac{j-2}{m-1}\right)}{36} \right| \\
&* \left\| \sum_{j=4}^{m-3} \frac{\sqrt{3}\eta_j^{(m)}}{\sqrt{2(m-1)}} \right\| \\
&\quad \leq 2\|y - y_m\| + 30 \sup_{t \in (0,1)} |y(t)| \frac{1}{\sqrt{m-1}} + \sup_{t \in (0,1)} |y'(t)| \frac{3}{m} \to 0
\end{aligned}
$$

as $m \to \infty$.

2.3.1.4 Proof of Theorem 2.1.8

We will need the following proposition for the convergence proofs.

Proposition 2.3.2. *Assume that Assumption 2.1.3 is fulfilled. Then, $P_{\mathcal{N}(P_m K)} x \to 0$ as $m \to \infty$, for all $x \in \mathcal{X}$.*

Proof. We assume w.l.o.g. that $x_m := P_{\mathcal{N}(P_m K)} x \rightharpoonup z \in \mathcal{X}$ for $m \to \infty$ (weakly). Then $\lim_{m\to\infty} K x_m = Kz$. Thus

$$
\begin{aligned}
\|AKz\| &= \limsup_{m\to\infty} \|P_m^* P_m K z\| = \limsup_{m\to\infty} \|P_m^* P_m (Kz - Kx_m)\| \\
&\leq \limsup_{m\to\infty} \|P_m\|^2 \|Kz - Kx_m\| = 0,
\end{aligned}
$$

so $AKz = 0$ hence by injectivity $z = 0$. In particular, $\left(P_{\mathcal{N}(P_m K)} v_i, v_i\right) \to (0, v_i) = 0$ for $m \to \infty$ and $i \in \mathbb{N}$ (set $x = v_i$ the i-th singular vector of K), so

$$1 \geq \|P_{\mathcal{N}(P_mK)}v_i - v_i\|^2 = \|P_{\mathcal{N}(P_mK)}v_i\|^2 - 2(P_{\mathcal{N}(P_mK)}v_i, v_i) + 1$$

and therefore

$$\limsup_{m\to\infty} \|P_{\mathcal{N}(P_mK)}v_i\| = 0.$$

Finally, by injectivity of K, for $\varepsilon > 0$ there is a $M \in \mathbb{N}$ with $\sum_{j=M+1}^{\infty}(x, v_j)^2 \leq \varepsilon$, so

$$\limsup_{m\to\infty} \|P_{\mathcal{N}(P_mK)}x\|^2 \leq \sum_{j=1}^{M}(x, v_j)^2 \limsup_{m\to\infty} \|P_{\mathcal{N}(P_mK)}v_j\|^2 + \varepsilon = \varepsilon$$

and the claim follows with $\varepsilon \to 0$. □

We come to the main proof and split

$$\begin{aligned}
&\mathbb{E}\left\|R^{(m)}_{\alpha(\delta^{est}_{m,n})}\bar{Y}^{(m)}_n - K^+\hat{y}\right\|^2 \\
&\leq \left\|K^+\hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})}P_m\hat{y}\right\|^2 + \mathbb{E}\left\|R^{(m)}_{\alpha(\delta^{est}_{m,n})}P_m\hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})}\bar{Y}^{(m)}_n\right\|^2 \\
&\leq \left\|K^+\hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})}P_m\hat{y}\right\|^2 + \left\|R^{(m)}_{\alpha(\delta^{est}_{m,n})}\right\|^2 \mathbb{E}\left\|\bar{Y}^{(m)}_n - P_m\hat{y}\right\|^2
\end{aligned}$$

and because of independence,

$$\mathbb{E}\left\|\bar{Y}^{(m)}_n - P_m\hat{y}\right\|^2 = \mathbb{E}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\delta^{(m)}_{ij}\right)^2 = \frac{1}{n}\sum_{j=1}^{m}\mathbb{E}\delta^{(m)2}_{1j} \leq \frac{m}{n}C_d = \left(\delta^{est}_{m,n}\right)^2 C_d.$$

Assumption 2.1.1 implies, that

$$\|R_\alpha\| \leq \sqrt{C_R C_F/\alpha}, \tag{2.13}$$

see f.e. [EHN96] or Proposition 1 of [HJP20a]. Therefore it follows that

$$\left\|R^{(m)}_{\alpha(\delta^{est}_{m,n})}\right\|^2 \mathbb{E}\delta^{meas2}_m \leq \left(\left\|R^{(m)}_{\alpha(\delta^{est}_{m,n})}\right\|\delta^{est}_{m,n}\right)^2 C_d \leq C_d C_R C_F \frac{\delta^{est\ 2}_{m,n}}{\alpha(\delta^{est}_{m,n})} \to 0 \tag{2.14}$$

for $m, n \to \infty, m/n \to 0$. Now

$$
\begin{aligned}
&\left\| K^+\hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})} P_m \hat{y} \right\| \\
\leq &\left\| K^+\hat{y} - (P_m K)^+ P_m \hat{y} \right\| + \left\| (P_m K)^+ P_m \hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})} P_m \hat{y} \right\| \\
= &\left\| K^+ K \hat{x} - (P_m K)^+ P_m K \hat{x} \right\| + \left\| (P_m K)^+ P_m \hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})} P_m \hat{y} \right\| \\
= &\left\| \hat{x} - P_{\mathcal{N}(P_m K)^\perp} \hat{x} \right\| + \left\| (P_m K)^+ P_m \hat{y} - R^{(m)}_{\alpha(\delta^{est}_{m,n})} P_m \hat{y} \right\|
\end{aligned}
$$

and

$$
\lim_{m\to\infty} \| \hat{x} - P_{\mathcal{N}(P_m K)^\perp} \hat{x} \| = \lim_{m\to\infty} \| P_{\mathcal{N}(P_m K)} \hat{x} \| = 0 \tag{2.15}
$$

by Proposition 2.3.2. Finally, for $\varepsilon > 0$, by Lemma 2.3.1.2 there is a $M \in \mathbb{N}$ such that $\sum_{j=M+1}^{m} \left(\hat{x}, v_j^{(m)} \right)^2 \leq \varepsilon$ for m large enough and therefore

$$
\begin{aligned}
&\left\| (P_m K)^+ P_m K \hat{x} - R^{(m)}_{\alpha(\delta^{est}_{m,n})} P_m K \hat{x} \right\|^2 \\
= &\sum_{j=1}^{m} \left| 1 - F_{\alpha(\delta^{est}_{m,n})}(\sigma_j^{(m)^2}) \sigma_j^{(m)^2} \right|^2 \left(\hat{x}, v_j^{(m)} \right)^2 \\
\leq &\sum_{j=1}^{M} \left| 1 - F_{\alpha(\delta^{est}_{m,n})}(\sigma_j^{(m)^2}) \sigma_j^{(m)^2} \right|^2 \left(\hat{x}, v_j^{(m)} \right)^2 + \sum_{j=M+1}^{m} \left(\hat{x}, v_j^{(m)} \right)^2 \\
\leq &\|\hat{x}\|^2 \sup_{j=1,\dots,M} \left| 1 - F_{\alpha(\delta^{est}_{m,n})}(\sigma_j^{(m)^2}) \sigma_j^{(m)^2} \right|^2 + \varepsilon.
\end{aligned}
$$

By Lemma 2.3.1.1, (2.9) and since $\alpha(\delta^{est}_{m,n}) \to 0$ for $m, n \to \infty, m/n \to 0$,

$$
\sup_{j=1,\dots,M} \left| 1 - F_{\alpha(\delta^{est}_{m,n})}(\sigma_j^{(m)^2}) \sigma_j^{(m)^2} \right| \leq \sup_{\frac{\sigma_M^{(\infty)^2}}{2} \leq \lambda \leq \|K\|^2} \left| 1 - F_{\alpha(\delta^{est}_{m,n})}(\lambda)\lambda \right| \leq \frac{\sqrt{\varepsilon}}{\|\hat{x}\|}
$$

for all m, n sufficiently large and m/n sufficiently small. Thus with $\varepsilon \to 0$ it follows that

$$
\lim_{\substack{m,n\to\infty \\ m/n\to 0}} \left\| (P_m K)^+ P_m K \hat{x} - R^{(m)}_{\alpha(\delta^{est}_{m,n})} P_m K \hat{x} \right\| = 0,
$$

which concludes the proof together with (2.14) and (2.15).

2.3.1.5 Proof of Theorem 2.1.11

By the nature of white noise we cannot expect the error to concentrate along a certain direction, in contrast to the setting in chapter 1. However, the independence between the measurement channels implies, that its amplitude is highly concentrated. First, the following Proposition affirms that we are estimating the variance correctly.

Proposition 2.3.3. *Assume that the error fulfills Assumption 2.1.9. Then for the sample variance*

$$s_{m,n}^2 = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{ij}^{(m)} - \frac{1}{n}\sum_{l=1}^{m} Y_{lj}^{(m)}\right)^2$$

there holds

$$\lim_{m\to\infty}\mathbb{P}\left(\sup_{n\geq 2}\left|s_{m,n}^2 - \mathbb{E}\delta_{11}^{(m)^2}\right| \geq \varepsilon\mathbb{E}\delta_{11}^{(m)^2}\right) = 0$$

for all $\varepsilon > 0$.

Proof. As a sum of m reversed martingales, $\left(s_{m,-n}^2 - \mathbb{E}\delta_{11}^{(m)^2}\right)_{n\leq -2}$ is a reversed martingale adapted to the filtration

$$\mathcal{F}_{-n} = \sigma\left(\sum_{i=1}^{n}(\delta_{i1}^{(m)} - \overline{\delta_{i1}^{(m)}})^2, ..., \sum_{i=1}^{n}(\delta_{im}^{(m)} - \overline{\delta_{im}^{(m)}})^2\right), n \geq 2.$$

Under Assumption 2.1.9.2, by the Kolmogorov-Doob-inequalities there holds

$$\mathbb{P}\left(\sup_{n\geq 2}\left|s_{m,n}^2 - \mathbb{E}\delta_{11}^{(m)^2}\right| \geq \varepsilon\mathbb{E}\delta_{11}^{(m)^2}\right) \leq \frac{\mathbb{E}\left|s_{m,2}^2 - \mathbb{E}\delta_{11}^{(m)^2}\right|^p}{\left(\varepsilon\mathbb{E}\delta_{11}^{(m)^2}\right)^p}.$$

By Marcinkiewicz-Zygmund inequality [Gut13] there exists C_p such that

$$\begin{aligned}
\mathbb{E}\left|s_{m,2}^2 - \mathbb{E}\delta_{11}^{(m)^2}\right|^p &= \mathbb{E}\left|\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{2}(\delta_{ij}^{(m)} - \overline{\delta_{ij}^{(m)}})^2 - \mathbb{E}\delta_{11}^{(m)^2}\right|^p \\
&\leq \frac{C_p}{m^{p-1}}\mathbb{E}\left|\sum_{i=1}^{2}(\delta_{i1}^{(m)} - \overline{\delta_{i1}}^{(m)})^2 - \mathbb{E}\delta_{11}^{(m)^2}\right|^p \\
&\leq \frac{2^{p-1}(4^p+1)C_p}{m^{p-1}}\mathbb{E}|\delta_{11}^{(m)}|^{2p},
\end{aligned}$$

so

$$\mathbb{P}\left(\sup_{n\geq 2}\left|s_{m,n}^2-\mathbb{E}\delta_{11}^{(m)^2}\right|\geq\varepsilon\mathbb{E}\delta_{11}^{(m)^2}\right)\leq\frac{\mathbb{E}\left|s_{m,2}^2-\mathbb{E}\delta_{11}^{(m)^2}\right|^p}{(\mathbb{E}\delta_{11}^{(m)^2})^p}\leq\frac{2^{p-1}(4^p+1)C_pC_d}{\varepsilon^p m^{p-1}}\to 0$$

as $m\to\infty$. Under Assumption 2.1.9.1, by the Kolmogorov-Doob-inequality,

$$\mathbb{P}\left(\sup_{n\geq 2}\left|s_{m,n}^2-\mathbb{E}\delta_{11}^{(m)^2}\right|\geq\varepsilon\mathbb{E}\delta_{11}^{(m)^2}\right)\leq\frac{\mathbb{E}\left|s_{m,2}^2-\mathbb{E}\delta_{11}^{(1)^2}\right|}{\varepsilon\mathbb{E}\delta_{11}^{(1)^2}}.$$

It is

$$s_{m,2}^2-\mathbb{E}\delta_{11}^{(1)^2}=\frac{1}{m}\sum_{j=1}^m\sum_{i=1}^2\left(\delta_{ij}^{(m)}-\overline{\delta_{ij}^{(m)}}\right)^2-\mathbb{E}\delta_{11}^{(1)^2}=:\frac{1}{m}\sum_{j=1}^m X_j^{(m)}$$

with $X_j^{(m)}, j=1,...,m, m\in\mathbb{N}$ are i.i.d and $\mathbb{E}X_j^{(m)}=0, \mathbb{E}|X_j^{(m)}|<\infty$. To finish the proof we need to show that $\mathbb{E}|\sum_j X_m/m|\to 0$ as $m\to\infty$. Let $\varepsilon'>0$. By dominated convergence and integrability of $X_j^{(m)}$, there is $M>0$ large enough such that for $Y_j^{(m)}:=X_j^{(m)}\chi_{\left\{|X_j^{(m)}|\leq M\right\}}$ and $Z_j^{(m)}:=X_j^{(m)}\chi_{\left\{|X_j^{(m)}|>M\right\}}$ it holds that $\mathbb{E}|Z_1^{(1)}|\leq\varepsilon$. So, since $X_j^{(m)}$ are i.i.d,

$$\begin{aligned}\mathbb{E}\left|\sum_{j=1}^m X_j^{(m)}\right| &\leq \mathbb{E}\left|\sum_{j=1}^m Y_j^{(m)}-\mathbb{E}Y_j^{(m)}\right|+\mathbb{E}\left|\sum_{j=1}^m Z_j^{(m)}-\mathbb{E}Z_j^{(m)}\right| \qquad (2.16)\\ &\leq\sqrt{\mathbb{E}\left|\sum_{j=1}^m Y_j^{(m)}-\mathbb{E}Y_j^{(m)}\right|^2}+\sum_{j=1}^m\mathbb{E}\left|Z_j^{(m)}-\mathbb{E}Z_j^{(m)}\right|\\ &\leq\sqrt{m\mathbb{E}\left|Y_1^{(1)}-\mathbb{E}Y_1^{(1)}\right|^2}+2m\mathbb{E}|Z_1^{(1)}|\leq\sqrt{m2M\mathbb{E}|X_1^{(1)}|}+2m\varepsilon,\end{aligned}$$

thus $\mathbb{E}\left|\sum_{j=1}^m X_j^{(m)}/m\right|\leq 3\varepsilon$ for m large enough. □

Now we need the following Lemma.

Lemma 2.3.4. *Assume that the error model is accordingly to Assumption 2.1.9. Then there holds*

$$\lim_{m,n\to\infty}\mathbb{P}\left(\left|\frac{\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|-\delta_{m,n}^{est}}{\delta_{m,n}^{est}}\right|\geq\varepsilon\right)=0.$$

Proof. It is

$$\begin{aligned}&\frac{\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|-\delta^{est}_{m,n}}{\delta^{est}_{m,n}}\\ =&\sqrt{\frac{\mathbb{E}\delta_{11}^{(m)^2}}{s^2_{m,n}}}\left(\frac{\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|-\sqrt{m\mathbb{E}\delta_{11}^{(m)^2}/n}}{\sqrt{m\mathbb{E}\delta_{11}^{(m)^2}/n}}+1-\sqrt{\frac{s^2_{m,n}}{\mathbb{E}\delta_{11}^{(m)^2}}}\right).\end{aligned}$$

Thus by Proposition 2.3.3 it suffices to show that

$$\lim_{m,n\to\infty}\mathbb{P}\left(\left|\frac{\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|^2-\frac{m}{n}\mathbb{E}\delta_{11}^{(m)^2}}{\frac{m}{n}\mathbb{E}\delta_{11}^{(m)^2}}\right|\geq\varepsilon\right)=0.$$

Let us first assume that Assumption 2.1.9.1 holds true. Then, by Markov's inequality

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|^2-\frac{m}{n}\mathbb{E}\delta_{11}^{(m)^2}}{\frac{m}{n}\mathbb{E}\delta_{11}^{(m)^2}}\right|\geq\varepsilon\right)&\leq\frac{\mathbb{E}\left|\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|^2-\frac{m}{n}\mathbb{E}\delta_{11}^{(m)^2}\right|}{\varepsilon\frac{m}{n}\mathbb{E}\delta_{11}^{(m)^2}}\\ &=\frac{1}{m\varepsilon}\mathbb{E}\left|\sum_{j=1}^m\left(\left(\frac{\sum_{i=1}^n\delta_{ij}^{(m)}}{\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}}\right)^2-1\right)\right|.\end{aligned}$$

Now with

$$X_{jn}^{(m)}:=\left(\frac{\sum_{i=1}^n\delta_{ij}^{(m)}}{\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}}\right)^2-1,$$

it holds that $(X_{jn}^{(m)})_{j=1}, j=1,...,m, m\in\mathbb{N}$ are i.i.d (for each fixed n) and $\mathbb{E}X_{jn}^{(m)}=0, \mathbb{E}|X_{jn}^{(m)}|=2<\infty$. We proceed similiar as at the end of the proof of Proposition 2.3.3, with the additional technical difficulty due to the dependence on n. Let $\varepsilon>0$ and Z be a standard Gaussian (thus $\mathbb{E}Z^2=1$ in particular). Then for M large enough, it holds that

$$\mathbb{E}\left[\chi_{\{|Z^2-1|\geq M\}}\right]\leq\frac{\varepsilon}{4} \tag{2.17}$$

$$\mathbb{E}\left[Z^2\chi_{\{|Z^2-1|<M\}}\right]\geq\mathbb{E}[Z^2]-\frac{\varepsilon}{4}=1-\frac{\varepsilon}{4}. \tag{2.18}$$

By the standard central limit theorem for real valued random variables, it holds that

$$\frac{\sum_{i=1}^n \delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[\delta_{11}^{(1)^2}\right]}} \to Z$$

weakly, as $n \to \infty$. Since

$$\begin{aligned} f_1 &: \mathbb{R} \to \mathbb{R},\ x \mapsto \chi_{\{|x^2-1|\geq M\}}, \\ f_2 &: \mathbb{R} \to \mathbb{R},\ x \mapsto x^2\chi_{\{|x^2-1|<M\}} \end{aligned}$$

are bounded functions whose set of discontinuities has Lebegue measure 0, it holds that

$$\mathbb{E}\left[f_p\left(\frac{\sum_{i=1}^n \delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[\delta_{11}^{(1)^2}\right]}}\right)\right] \to \mathbb{E}\left[f_p(Z)\right]$$

as $n \to \infty$ for $p = 1, 2$ by Portemanteaus lemma (see e.g. [Kle13]). Thus by (2.17) there is a n^* such that

$$\begin{aligned} \mathbb{E}\left[f_1\left(\frac{\sum_{i=1}^n \delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[\delta_{11}^{(1)^2}\right]}}\right)\right] &\leq \mathbb{E}\left[f_1(Z)\right] + \left|\mathbb{E}\left[f_1\left(\frac{\sum_{i=1}^n \delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[\delta_{11}^{(1)^2}\right]}}\right) - f_1(Z)\right]\right| \leq \frac{\varepsilon}{2} \\ \mathbb{E}\left[f_2\left(\frac{\sum_{i=1}^n \delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[\delta_{11}^{(1)^2}\right]}}\right)\right] &\geq \mathbb{E}\left[f_2(Z)\right] - \left|\mathbb{E}\left[f_2\left(\frac{\sum_{i=1}^n \delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[\delta_{11}^{(1)^2}\right]}}\right) - f_2(Z)\right]\right| \geq 1 - \frac{\varepsilon}{2} \end{aligned}$$

for all $n \geq n^*$ and $p = 1, 2$. We again set $Y_{jn}^{(m)} := X_{jn}^{(m)}\chi_{\left\{|X_{jn}^{(m)}|\leq M\right\}}$ and $Z_{jn}^{(m)} := X_{jn}^{(m)}\chi_{\left\{|X_{jn}^{(m)}|>M\right\}}$ and define

$$f_3 := \mathbb{R} \to \mathbb{R},\ x \mapsto x^2\chi_{\{|x^2-1|\geq M|}.$$

Then

$$
\begin{aligned}
\mathbb{E}|Z_{1n}^{(n)}| &\leq \mathbb{E}\left[f_3\left(\frac{\sum_{i=1}^{n}\delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[{\delta_{11}^{(1)}}^2\right]}}\right)\right] + \mathbb{E}\left[f_1\left(\frac{\sum_{i=1}^{n}\delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[{\delta_{11}^{(1)}}^2\right]}}\right)\right] \qquad (2.19)\\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^{n}\delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[{\delta_{11}^{(1)}}^2\right]}}\right)^2\right] - \mathbb{E}\left[f_2\left(\frac{\sum_{i=1}^{n}\delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[{\delta_{11}^{(1)}}^2\right]}}\right)\right]\\
&\quad + \mathbb{E}\left[f_1\left(\frac{\sum_{i=1}^{n}\delta_{i1}^{(1)}}{\sqrt{n\mathbb{E}\left[{\delta_{11}^{(1)}}^2\right]}}\right)\right]\\
&\leq 1-(1-\frac{\varepsilon}{2})+\frac{\varepsilon}{2}=\varepsilon,
\end{aligned}
$$

for all $n \geq n^*$, where we used that $f_2(x) = f_3(x) = x^2$ in the second step. With the same argumentation as in (2.16),

$$
\begin{aligned}
\mathbb{E}\left|\sum_{j=1}^{m} X_{jn}^{(m)}\right| &\leq \mathbb{E}\left|\sum_{j=1}^{m} Y_{jn}^{(m)} - \mathbb{E}Y_{jn}^{(m)}\right| + \mathbb{E}\left|\sum_{j=1}^{m} Z_{jn}^{(m)} - \mathbb{E}Z_{jn}^{(m)}\right|\\
&\leq \sqrt{\mathbb{E}\left|\sum_{j=1}^{m} Y_{jn}^{(m)} - \mathbb{E}Y_{jn}^{(m)}\right|^2} + \sum_{j=1}^{m}\mathbb{E}\left|Z_{jn}^{(m)} - \mathbb{E}Z_{jn}^{(m)}\right|\\
&\leq \sqrt{m\mathbb{E}\left|Y_{1n}^{(1)} - \mathbb{E}Y_{1n}^{(1)}\right|^2} + 2m\mathbb{E}|Z_{1n}^{(1)}| \leq \sqrt{m2M\mathbb{E}|X_{1n}^{(1)}|} + 2m\varepsilon\\
&\leq \sqrt{4mM} + 2m\varepsilon,
\end{aligned}
$$

for all $n \geq n^*$, where we used $\mathbb{E}|X_{1n}^{(1)}| \leq 2$ and (2.19) in the last step. Thus $\mathbb{E}|\sum_{j=1}^{m} X_{jm}^{(m)}/m| \leq 3\varepsilon$ for m, n large enough. The claim is proved.

Now assume that Assumption 2.1.9.2 holds true. Then, by Markov's inequality,

$$
\mathbb{P}\left(\left|\frac{\left\|\bar{Y}_n^{(m)} - P_m\hat{y}\right\|^2 - \frac{m}{n}\mathbb{E}{\delta_{11}^{(m)}}^2}{\frac{m}{n}\mathbb{E}{\delta_{11}^{(m)}}^2}\right| \geq \varepsilon\right) \leq \frac{\mathbb{E}\left|\frac{n}{m\mathbb{E}{\delta_{11}^{(m)}}^2}\left\|\bar{Y}_n^{(m)} - P_m\hat{y}\right\|^2 - 1\right|^p}{\varepsilon^p}
$$

and using further twice the Marcinkiewicz-Zygmund inequality, one obtains

$$\begin{aligned}
&\mathbb{E}\left|\frac{n}{m\mathbb{E}\delta_{11}^{(m)^2}}\left\|\bar{Y}_n^{(m)}-P_m\hat{y}\right\|^2-1\right|^p\\
=&\frac{1}{m^p}\mathbb{E}\left|\sum_{j=1}^{m}\left(\left(\sum_{i=1}^{n}\delta_{ij}^{(m)}/\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}\right)^2-1\right)\right|^p\\
\leq&\frac{B_p m^{\max(1,p/2)}}{m^p}\mathbb{E}\left|\left(\sum_{i=1}^{n}\delta_{i1}^{(m)}/\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}\right)^2-1\right|^p\\
\leq&\frac{2^{p-1}B_p}{m^{\min(p-1,p/2)}}\left(\mathbb{E}\left|\sum_{i=1}^{n}\delta_{i1}^{(m)}/\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}\right|^{2p}+1^p\right)\\
\leq&\frac{2^{p-1}B_p}{m^{\min(p-1,p/2)}}\left(B_{2p}\mathbb{E}\left|\delta_{11}^{(m)}\right|^{2p}/\left(\mathbb{E}\delta_{11}^{(m)^2}\right)^p+1\right)\leq\frac{C}{m^{\min(p-1,p/2)}}\to 0
\end{aligned}$$

as $m\to\infty$, where we have used independence and $\mathbb{E}\left(\sum_{i=1}^{n}\delta_{ij}^{(m)}/\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}\right)^2=1$ in the second step.

□

Before we will start with the main proof, we need one last proposition.

Proposition 2.3.5. *For all $\varepsilon>0$, there are $m_0\in\mathbb{N}$ and $\alpha_0>0$ such that*

$$\lim_{m\to\infty}\left\|P_mKR_\alpha^{(m)}P_mK\hat{x}-P_mK\hat{x}\right\|/\sqrt{\alpha}\leq\varepsilon$$

for all $m\geq m_0$ and $\alpha\leq\alpha_0$.

Proof. Lemma 2.3.1.2 guarantees the existence of $M\in\mathbb{N}$, such that

$$C_1^2\sum_{j=M+1}^{m}(\hat{x},v_j^{(m)})^2\leq\varepsilon/2$$

for m sufficiently large. Then

$$\begin{aligned}
&\left\|(P_mKR_\alpha^{(m)} - Id)P_mK\hat{x}\right\|^2/\alpha = \sum_{j=1}^{m}\left(F_\alpha(\sigma_j^{(m)^2})\sigma_j^{(m)^2} - 1\right)^2 \frac{\sigma_j^{(m)^2}}{\alpha}(\hat{x}, v_j^{(m)})^2 \\
&\leq \left(\sup_{\lambda>0}\lambda^{\frac{\nu_0}{2}}|F_\alpha(\lambda)\lambda - 1|\right)^2 \|\hat{x}\|^2 \sum_{j=1}^{M}\frac{\sigma_j^{(m)^{2(1-\nu_0)}}}{\alpha} \\
&\qquad + \left(\sup_{\lambda>0}\lambda^{\frac{1}{2}}|F_\alpha(\lambda)\lambda - 1|\right)^2 \frac{\sum_{j=M+1}^{m}(\hat{x}, v_j^{(m)})^2}{\alpha} \\
&\leq C_{\nu_0}^2 M\sigma_M^{(m)^{2(1-\nu_0)}}\|\hat{x}\|^2\alpha^{\nu_0-1} + C_1^2\sum_{l=M+1}^{m}(\hat{x}, v_j^{(m)})^2 \\
&\leq 2C_{\nu_0}^2 M\sigma_M^{(\infty)^{2(1-\nu_0)}}\|\hat{x}\|^2\alpha^{\nu_0-1} + \varepsilon/2 \leq \varepsilon
\end{aligned}$$

for m sufficiently large and α sufficiently small, where we have used that the qualification of $(F_\alpha)_{\alpha>0}$ is bigger than one in the third and Lemma 2.3.1.1 in the fourth step. □

We start with the main proof. We define

$$\Omega_{m,n} := \left\{\left\|\bar{Y}_n^{(m)} - P_m\hat{y}\right\| \leq \frac{\tau + C_0}{2C_0}\delta_{m,n}^{est}\ ,\ \delta_{m,n}^{est} \leq c\varepsilon\right\},$$

with $c \leq \frac{1}{2}\max\left\{\frac{C_0+3\tau}{\sigma_M^{(\infty)^2}}, \frac{(\tau+C_0))\sqrt{C_RC_F}}{\sqrt{\varepsilon'}}\right\}^{-1}$, where ε' is given below.

By Proposition 2.3.2,

$$\left\|(P_mK)^+P_m\hat{y} - K^+\hat{y}\right\| = \left\|P_{\mathcal{N}(P_mK)}\hat{x}\right\| \leq \varepsilon$$

for m large enough, and by Lemma 2.3.1.2,

$$\begin{aligned}
&\left\| R^{(m)}_{\alpha_{m,n}} P_m \hat{y} - K^+ \hat{y} \right\|^2 \\
\leq& \sum_{j=1}^{M} \left| F_{\alpha_{m,n}}(\sigma_j^{(m)^2})\sigma_j^{(m)^2} - 1\right|^2 (\hat{x}, v_j^{(m)})^2 + \sum_{j=M+1}^{m} (\hat{x}, v_j^{(m)})^2 \\
\leq& \frac{1}{\sigma_M^{(m)^2}} \sum_{j=1}^{M} \left| F_{\alpha_{m,n}}(\sigma_j^{(m)^2})\sigma_j^{(m)^2} - 1\right|^2 \sigma_j^{(m)^2}(\hat{x}, v_j^{(m)})^2 + \varepsilon/2 \\
=& \frac{1}{\sigma_M^{(m)^2}} \left\|(P_m K R^{(m)}_{\alpha_{m,n}} - Id) P_m \hat{y}\right\| + \varepsilon/2 \\
\leq& \frac{1}{\sigma_M^{(m)^2}} \left(\left\|(P_m K R^{(m)}_{\alpha_{m,n}} - Id)\bar{Y}_n^{(m)}\right\| + \left\|(P_m K R^{(m)}_{\alpha_{m,n}} - Id)(P_m \hat{y} - \bar{Y}_n^{(m)})\right\|\right) + \varepsilon/2
\end{aligned}$$

for m sufficiently large. So Lemma 2.3.1.1 and the defining relation of the discrepancy principle and of $\Omega_{m,n}$ ensure that

$$\left\| R^{(m)}_{\alpha_{m,n}} P_m \hat{y} - K^+ \hat{y} \right\| \chi_{\Omega_{m,n}} \leq \frac{2}{\sigma_M^{(\infty)^2}} \left(\tau \delta^{est}_{m,n} + C_0 \frac{\tau + C_0}{2C_0} \delta^{est}_{m,n} \right) \chi_{\Omega_{m,n}} + \varepsilon/2 \leq \varepsilon$$

for m sufficiently large. Moreover,

$$\begin{aligned}
&\tau \delta^{est}_{m,n} \chi_{\Omega_{m,n}} \\
\leq& \left\|(P_m K R^{(m)}_{\alpha_{m,n}/q} - Id)\bar{Y}_n^{(m)}\right\| \chi_{\Omega_{m,n}} \\
\leq& \left\|(P_m K R_{\alpha_{m,n}/q} - Id) P_m \hat{y}\right\| + \left\|(P_m K R^{(m)}_{\alpha_{m,n}/q} - Id)(\bar{Y}_n^{(m)} - P_m \hat{y})\right\| \chi_{\Omega_{m,n}} \\
\leq& \left\|(P_m K R_{\alpha_{m,n}/q} - Id) P_m \hat{y}\right\| + C_0 \frac{\tau + C_0}{2C_0} \delta^{est}_{m,n} \chi_{\Omega_{m,n}}, \\
\Longrightarrow& \delta^{est}_{m,n} \chi_{\Omega_{m,n}} \leq \frac{2}{\tau - C_0} \left\|(P_m K R_{\alpha_{m,n}/q} - Id) P_m \hat{y}\right\|
\end{aligned}$$

Proposition 2.3.5 guarantees the existence of ε' such that for m large enough

$$\left\| P_m K R^{(m)}_{\alpha} P_m \hat{y} - P_m \hat{y} \right\| / \sqrt{\alpha} \leq \frac{(\tau - C_0) q C_0}{(\tau + C_0)\sqrt{C_R C_F}} \frac{\varepsilon}{2}$$

for all $\alpha \leq \varepsilon'/q$. So with (2.13),

$$\begin{aligned}
&\|R^{(m)}_{\alpha_{m,n}}(\bar{Y}^{(m)}_n - P_m\hat{y})\|\chi_{\Omega_{m,n}}\\
&\le \|R_{\alpha_{m,n}}\|\|\bar{Y}^{(m)}_n - P_m\hat{y}\|\chi_{\Omega_{m,n}} \le \sqrt{\frac{C_R C_R}{\alpha_{m,n}}}\frac{\tau + C_0}{2C_0}\delta^{est}_{m,n}\chi_{\Omega_{m,n}}\\
\le&\frac{(\tau + C_0)\sqrt{C_R C_F}}{2C_0}\left(\frac{\delta^{est}_{m,n}}{\sqrt{\alpha_{m,n}}}\chi_{\Omega_{m,n}\cap\{\alpha_{m,n}\le\varepsilon'\}} + \frac{\delta^{est}_{m,n}}{\sqrt{\alpha_{m,n}}}\chi_{\Omega_{m,n}\cap\{\alpha_{m,n}\ge\varepsilon'\}}\right)\\
\le&\frac{(\tau + C_0)\sqrt{C_R C_F}}{2C_0}\left(\frac{2}{(\tau - C_0)q}\frac{\left\|(P_m K R_{\frac{\alpha_{m,n}}{q}} - Id)P_m\hat{y}\right\|}{\sqrt{\alpha_{m,n}/q}}\chi_{\{\alpha_{m,n}\le\varepsilon'\}} + \frac{\delta^{est}_{m,n}}{\sqrt{\varepsilon'}}\chi_{\Omega_{m,n}}\right)\\
\le&\frac{(\tau + C_0)\sqrt{C_R C_F}}{2C_0}\left(\frac{2}{(\tau - C_0)q}\frac{(\tau - C_0)qC_0}{(\tau + C_0)\sqrt{C_R C_F}}\frac{\varepsilon}{2} + \frac{c\varepsilon}{\sqrt{\varepsilon'}}\right) \le \varepsilon/2 + \varepsilon/2
\end{aligned}$$

for m large enough. Putting it all together,

$$\begin{aligned}
&\left\|R^{(m)}_{\alpha_{m,n}}\bar{Y}^{(m)}_n - K^+\hat{y}\right\|\chi_{\Omega_{m,n}}\\
\le&\left\|R^{(m)}_{\alpha_{m,n}}(\bar{Y}^{(m)}_n - P_m\hat{y})\right\|\chi_{\Omega_{m,n}} + \left\|R^{(m)}_{\alpha_{m,n}}P_m\hat{y} - (P_m K)^+ P_m\hat{y}\right\|\chi_{\Omega_{m,n}}\\
&\qquad + \left\|(P_m K)^+ P_m\hat{y} - K^+\hat{y}\right\|\chi_{\Omega_{m,n}}\\
\le&3\varepsilon
\end{aligned}$$

for m sufficiently large, which together with $\lim_{\substack{m,n\to\infty\\ m/n\to 0}}\mathbb{P}\left(\Omega_{m,n}\right) = 1$ finishes the proof.

2.3.2 Proofs for infinite-dimensional residuum

For the second approach (with infinite-dimensional residuum), we need to guarantee stable inversion of the discretisation operator P_m. Afterwards we will show strong concentration of the back projected measurements in $\mathcal{Y}$ in order to use classical results from deterministic regularisation theory.

2.3.2.1 Proof of Proposition 2.2.2

It is $\kappa(P_m) = \kappa(P_m|_{\mathcal{N}(P_m)^\perp})$. We again denote by $A_m \in \mathbb{R}^{m\times m}$ the matrix representing $P_m : \mathcal{N}(P_m)^\perp \to \mathbb{R}^m$ with respect to the bases $(\eta^{(m)}_j)_{j=1,\dots,m} \subset \mathcal{N}(P_m)^\perp$ and $(e_j)_{j=1,\dots,m} \subset \mathbb{R}^m$, where the latter is the canonical basis of $\mathbb{R}^m$. Thus

$$(A_m)_{ij} = \left(P_m\eta^{(m)}_i, e_j\right)_{\mathbb{R}^m} = l^{(m)}_j(\eta^{(m)}_i) = (\eta^{(m)}_j, \eta^{(m)}_i)_{\mathcal{Y}}.$$

By assumption, we have that

$$\left\| \frac{A_m}{\|\eta_1^{(m)}\|^2} - I_m \right\| \leq \sqrt{\left\| \frac{A_m}{\|\eta_1^{(m)}\|^2} - I_m \right\|_1 \left\| \frac{A_m}{\|\eta_1^{(m)}\|^2} - I_m \right\|_\infty}$$
$$= \max_{j=1,\dots,m} \sum_{i \neq j} \frac{|(\eta_j^{(m)}, \eta_i^{(m)})|}{\|\eta_1^{(m)}\|^2} =: c < 1,$$

where $I_m \in \mathbb{R}^{m \times m}$ is the identity and $\|.\|, \|.\|_1, \|.\|_\infty$ are the spectral and the maximum absolute column or row norm. So by (2.3) in [Rum11], it is

$$1 - c \leq \sigma_j \left(\frac{A_m}{\|\eta_1^{(m)}\|^2} \right) \leq 1 + c, \tag{2.20}$$

for $j = 1, ..., m$, where $\sigma_1(A), ..., \sigma_m(A)$ denote the singular values of $A \in \mathbb{R}^{m \times m}$. This proves the proposition.

2.3.2.2 Proof of Proposition 2.2.3

The bounds c_m, C_m follow directly from Proposition 2.2.2. It remains to show that $\|\hat{y} - P_m^+ P_m \hat{y}\| \to 0$ as $m \to \infty$. It holds that $\mathcal{N}(P_1) \supseteq \mathcal{N}(P_2) \supseteq ....$ In particular, there is an orthonormal basis $(w_i)_{i \in \mathbb{N}}$ such that $\mathcal{N}(P_m) = span(w_{m+1}, w_{m+2}, ...)$. Thus, $\delta_m^{disc} = \|P_{\mathcal{N}(P_m)} y\| = \sqrt{\sum_{j=m+1}^\infty (y, w_j)^2} \to 0$ as $m \to \infty$.

2.3.2.3 Proof of Proposition 2.2.4

The bound for the discretisation error follows from

$$\|\hat{y} - {P_m}^+ P_m \hat{y}\|^2 = \sum_{j > m} (\hat{y}, u_j)^2 = \sum_{j > m} \sigma_j^{2+2\nu} (w, v_j)^2 \leq \sigma_{m+1}^{2(1+\nu)} \|w\|^2.$$

Since $(\eta_j^{(m)}, \eta_i^{(m)}) = (v_j, v_i)$ and $(v_j)_{j \in \mathbb{N}}$ is an orthonormal basis, the claim follows.

2.3.2.4 Proof of Proposition 2.2.5

The choice $c_m = C_m = 1$ follows from Proposition 2.2.2, since $(\eta_j^{(m)})_{j=1,...,m}$ are orthonormal for all $m \in \mathbb{N}$. Denote by $y_m = \sum_{j=1}^m \hat{y}((j-1)/m) \chi_{(\frac{j-1}{m}, \frac{j}{m})} \in \mathcal{R}(P_m^*) =$

$\mathcal{N}(P_m)^\perp$ the piecewise constant interpolating spline of the continuously differentiable function $\hat{y}$. Then there holds

$$\begin{aligned}\|\hat{y} - P_m^+ P_m \hat{y}\| = \|\hat{y} - P_{\mathcal{N}(P_m)^\perp}\hat{y}\| \leq \|\hat{y} - y_m\| &\leq \sqrt{\int_0^1 (\hat{y}(t) - y_m(t))^2 dt}\\ &= \sqrt{\sum_{j=1}^m \int_{\frac{j-1}{m}}^{\frac{j}{m}} \left(\hat{y}(t) - \hat{y}\left((\frac{j-1}{m}\right)\right)^2 dt}\\ &= \sqrt{\sum_{j=1}^m \int_{\frac{j-1}{m}}^{\frac{j}{m}} y'(\xi_t)\left(t - \frac{j-1}{m}\right)^2 dt} \leq \frac{\sup_{t' \in (0,1)} |\hat{y}'(t')|}{m},\end{aligned}$$

with $\xi_t \in [\frac{j-1}{m}, \frac{j}{m})$.

2.3.2.5 Proof of Proposition 2.2.6

It is

$$(\eta_j^{(m)}, \eta_i^{(m)}) = \begin{cases} 2/3 & , i = j \\ 1/3 & , |i-j| = 1, \min(i,j) = 1 \text{ or } \max(i,j) = m \\ 1/6 & , |i-j| = 1, \min(i,j) > 1 \text{ and } \max(i,j) < m \\ 0 & , else \end{cases}$$

Therefore

$$\sup_{m \in \mathbb{N}} \max_{j \leq m} \frac{\sum_{j \neq i} |(\eta_j^{(m)}, \eta_i^{(m)})|}{\|\eta_1^{(m)}\|^2} = \frac{1/2}{2/3} = \frac{3}{4},$$

so that the bounds c_m, C_m follow with Proposition 2.2.2. Let $y_m \in \mathcal{N}(P_m)^\perp$ be the interpolating spline of continuously differentiable $\hat{y}$. By the mean value theorem there are $\xi_t, \zeta_t \in [\frac{j-1}{m-1}, \frac{j}{m-1})$ such that

$$\begin{aligned}&\hat{y}(t) - y_m(t)\\ =&\hat{y}\left(\frac{j-1}{m-1}\right) + \hat{y}'(\xi_t)\left(t - \frac{j-1}{m-1}\right)\\ &\quad - \left(\hat{y}\left(\frac{j-1}{m-1}\right) + \left((\hat{y}\left(\frac{j}{m-1}\right) - \hat{y}\left(\frac{j-1}{m-1}\right)\right)((m-1)t - (j-1))\right)\\ =&(y'(\xi_t) - y'(\zeta_t)\left(t - \frac{j-1}{m-1}\right)\end{aligned}$$

for $t \in [\frac{j-1}{m-1}, \frac{j}{m-1})$. Thus

$$\begin{aligned}\|\hat{y} - P_m^+ P_m \hat{y}\| \leq \|\hat{y} - y_m\| &\leq \sqrt{\sum_{j=1}^{m} \int_{\frac{j-1}{m-1}}^{\frac{j}{m-1}} (\hat{y}'(\xi_t) - \hat{y}'(\zeta_t))^2 \left(t - \frac{j-1}{m-1}\right)^2 dt} \\ &\leq \frac{2\sqrt{m}\sup_{t\in(0,1)} |\hat{y}'(t)|}{(m-1)^{3/2}} \leq \frac{2^{5/2}\sup_{t'\in(0,1)} |\hat{y}'(t')|}{m}\end{aligned}$$

If $\hat{y}$ is twice continuously differentiable, then there are $\xi'_t, \zeta'_t \in (\frac{j-1}{m-1}, \frac{j}{m-1}]$ such that

$$\begin{aligned}|\hat{y}'(\xi_t) - \hat{y}'(\zeta_t)| &= \left|\hat{y}''(\xi'_t)\left(\xi_t - \frac{j-1}{m-1}\right) - \hat{y}''(\zeta'_t)\left(\zeta_t - \frac{j-1}{m-1}\right)\right| \\ &\leq \frac{2\sup_{t'\in(0,1)} |\hat{y}''(t')|}{m-1}\end{aligned}$$

for $t \in [\frac{j-1}{m-1}, \frac{j}{m-1})$, so that

$$\begin{aligned}\|\hat{y} - P_m^+ P_m \hat{y}\| \leq \|\hat{y} - y_m\| &\leq \sqrt{\sum_{j=1}^{m} \int_{\frac{j-1}{m-1}}^{\frac{j}{m-1}} \left(\frac{2\sup_{t'\in(0,1)} |\hat{y}''(t')|}{m-1}\right)^2 \left(t - \frac{j-1}{m-1}\right)^2 dt} \\ &\leq \frac{2\sqrt{m}\sup_{t'\in(0,1)} |\hat{y}''(t')|}{(m-1)^{5/2}} \leq \frac{2^{7/2}\sup_{t'\in(0,1)} |\hat{y}''(t')|}{m^2}.\end{aligned}$$

2.3.2.6 Proof of Theorem 2.2.7

We use the bias-variance decomposition

$$
\begin{aligned}
&\mathbb{E}\left\|R_{\alpha(\delta_m^{disc})}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-K^+\hat{y}\right\|^2\\
=&\mathbb{E}\left\|R_{\alpha(\delta_m^{disc})}P_m^+(\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-P_m\hat{y})\right\|^2+\left\|R_{\alpha(\delta_m^{disc})}P_m^+P_m\hat{y}-K^+\hat{y}\right\|^2\\
\leq&\mathbb{E}\left\|R_{\alpha(\delta_m^{disc})}P_m^+(\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-P_m\hat{y})\right\|^2+2\left\|R_{\alpha(\delta_m^{disc})}P_m^+P_m\hat{y}-R_{\alpha(\delta_m^{disc})}\hat{y}\right\|^2\\
&\qquad+2\left\|R_{\alpha(\delta_m^{disc})}\hat{y}-K^+\hat{y}\right\|^2\\
\leq&\left\|R_{\alpha(\delta_m^{disc})}\right\|^2\left(\|P_m^+\|^2\mathbb{E}\left\|P_m\hat{y}-\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}\right\|^2+2\left\|P_m^+P_m\hat{y}-\hat{y}\right\|^2\right)\\
&\qquad+2\left\|R_{\alpha(\delta_m^{disc})}\hat{y}-K^+\hat{y}\right\|^2\\
\leq&\frac{C_RC_F}{\alpha(\delta_m^{disc})}\left(\frac{\mathbb{E}\delta_{11}^{(m)^2}m}{c_m^2n(m,\delta_m^{disc})}+2\delta_m^{disc^2}\right)+2\left\|R_{\alpha(\delta_m^{disc})}\hat{y}-K^+\hat{y}\right\|^2\\
\leq&\left(C_RC_F(C_d+2)\right)\frac{\delta_m^{disc^2}}{\alpha(\delta_m^{disc})}+2\left\|R_{\alpha(\delta_m^{disc})}\hat{y}-K^+\hat{y}\right\|^2\to 0
\end{aligned}
$$

as $m\to\infty$.

2.3.2.7 Proof of Theorem 2.2.9

The proof of Theorem 2.2.9 is more technical than the one of Theorem 2.1.11, due to correlations coming from the back projecting of the measurements and the data-dependent determination of the stopping index $n(m,\delta_m^{disc})$. However, under slightly stronger conditions we obtain a similar concentration property of the measurement error.

Lemma 2.3.6. *Assume that the discretisation fulfills Assumption 2.1.3 and the error is accordingly to Assumption 2.1.9, with $p\geq 2$ in the case of Assumption 2.1.9.2. For $m\in\mathbb{N},\delta_0,\delta>0$ and the sample variance*

$$
s_{m,n}^2:=\frac{1}{m}\sum_{j=1}^m\frac{1}{n-1}\sum_{i=1}^n\left(Y_{ij}^{(m)}-\frac{1}{n}\sum_{l=1}^nY_{lj}^{(m)}\right)^2,
$$

consider the (random) choice

$$
n(m,\delta)=\min\left\{n'\geq 1:\frac{ms_m^2(n')}{c_m^2n'}\leq\delta^2\right\}
$$

with $\sigma_1^{(m)},...,\sigma_m^{(m)}$ the singular values of P_m. Then for any $\varepsilon>0$ there holds

$$\lim_{m\to\infty}\sup_{0<\delta\le\delta_0}\mathbb{P}\left(\left|\frac{\left\|P_m^+\bar{Y}^{(m)}_{n(m,\delta)}-P_m^+P_m\hat{y}\right\|-\delta_m}{\delta_m}\right|\ge\varepsilon\right)=0$$

with $\bar{Y}^{(m)}_{n(m,\delta)}=\frac{1}{n(m,\delta)}\sum_{i=1}^{n(m,\delta)}\begin{pmatrix}Y^{(m)}_{i1} & \dots & Y^{(m)}_{im}\end{pmatrix}^T$ *and* $\delta_m:=\delta\sqrt{\sum_{j=1}^m\frac{c_m^2}{m\sigma_j^{m2}}}$.

Proof. The auxiliary parameter δ_m has to be introduced due to the fact, that with the choice of $n(m,\delta)$ we are actually overestimating $\mathbb{E}\left\|P_m^+\bar{Y}^{(m)}_{n(m,\delta)}-P_m^+P_m\hat{y}\right\|^2$, since $c_m\le\sigma_j^{(m)}$. We define

$$\mu_m^\delta:=\frac{m\mathbb{E}[\delta_{11}^{(m)^2}]}{c_m^2\delta^2}$$

$$I_\varepsilon(m,\delta):=\left[(1-\varepsilon)\mu_m^\delta,(1+\varepsilon)\mu_m^\delta\right].$$

$$\delta_{m,n}^{meas}:=\|P_m^+\bar{Y}_n^{(m)}-P_m^+P_m\hat{y}\|=\sqrt{\sum_{j=1}^m\lambda_j^{(m)}\left(\sum_{l=1}^m\sum_{i=1}^n\frac{\delta_{ij}^{(m)}}{n}(u_j^{(m)},e_l^{(m)})\right)^2}$$

where $\lambda_j^{(m)}=\sigma_j^{(m)^{-2}}$ and $(u_j^{(m)})_{j\le m},(e_j^{(m)})_{j\le m}\subset\mathbb{R}^m$ are the singular basis of P_m (fulfilling
$P_mP_m{}^*u_j^{(m)}=\sigma_j^{(m)^2}u_j^{(m)}$) and the canonical basis of $\mathbb{R}^m$ respectively. So

$$\mathbb{E}\delta_{m,n}^{meas2}=\sum_{j=1}^m\lambda_j\mathbb{E}\left(\sum_{l=1}^m\sum_{i=1}^n\frac{\delta_{il}^{(m)}}{n}(u_j^{(m)},e_l^{(m)})\right)^2=\frac{\mathbb{E}\delta_{11}^{(m)^2}}{n}\sum_{j=1}^m\lambda_j$$

and

$$\begin{aligned}\mathbb{P}\left(\left|\frac{{\delta_{m,n(m,\delta)}^{meas}}^2-\delta_m^2}{\delta_m^2}\right|\le\varepsilon\right)&\ge\mathbb{P}\left(\left|\frac{{\delta_{m,n(m,\delta)}^{meas}}^2-\delta_m^2}{\delta_m^2}\right|\le\varepsilon,n(m,\delta)\in I_{\varepsilon'}\right)\\&\ge\mathbb{P}\left(\sup_{n\in I_{\varepsilon'}}\left|\frac{\delta_{m,n}^{meas2}-\delta_m^2}{\delta_m^2}\right|\le\varepsilon,n(m,\delta)\in I_{\varepsilon'}\right)\\&\ge1-\mathbb{P}\left(\sup_{n\in I_{\varepsilon'}}\left|\frac{\delta_{m,n}^{meas2}-\delta_m^2}{\delta_m^2}\right|>\varepsilon\right)-\mathbb{P}\left(n(m,\delta)\notin I_{\varepsilon'}\right).\end{aligned}$$

Since

$$\left|\frac{\delta_{m,n}^{meas2}-\delta_m^2}{\delta_m^2}\right| \le \left(\left|\frac{\delta_{m,n}^{meas2}-\mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}}\right| + \left|\frac{\mathbb{E}\delta_{m,n}^{meas2}-\delta_m^2}{\mathbb{E}\delta_{m,n}^{meas2}}\right|\right)\frac{\mathbb{E}\delta_{m,n}^{meas2}}{\delta_m^2}$$

and

$$\sup_{n\in I_{\varepsilon'}}\left|\frac{\mathbb{E}\delta_{m,n}^{meas2}-\delta_m^2}{\delta_m^2}\right| = \frac{\varepsilon'}{1-\varepsilon'}, \quad \sup_{n\in I_{\varepsilon'}}\frac{\mathbb{E}\delta_{m,n}^{meas2}}{\delta_m^2} = \frac{1}{1-\varepsilon'},$$

we conclude that for $\varepsilon' = \frac{3}{16}\varepsilon \le 1/4$

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{{\delta_{m,n(m,\delta)}^{meas}}^2-\delta_m^2}{\delta_m^2}\right| \le \varepsilon\right)\\
\ge& 1-\mathbb{P}\left(\sup_{n\in I_{\varepsilon'}}\left|\frac{\delta_{m,n}^{meas2}-\mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}}\right| > \varepsilon(1-\varepsilon')-\frac{\varepsilon'}{1-\varepsilon'}\right) - \mathbb{P}\left(n(m,\delta)\notin I_{\varepsilon'}\right)\\
\ge& 1-\mathbb{P}\left(\sup_{n\in I_{\frac{3}{16}\varepsilon}}\left|\frac{\delta_{m,n}^{meas2}-\mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}}\right| > \varepsilon/2\right) - \mathbb{P}\left(n(m,\delta)\notin I_{\frac{3}{16}\varepsilon}\right).
\end{aligned} \tag{2.21}$$

Thus it remains to show that the both terms with negative sign tend to zero.

Proposition 2.3.7. *For every $\varepsilon > 0$ there holds*

$$\sup_{\delta_0\ge\delta>0}\mathbb{P}\left(n(m,\delta)\in I_\varepsilon(m,\delta)\right)\to 1$$

for $m\to\infty$.

Proof. For m large enough it is $\lfloor(1+\varepsilon)\mu_m^\delta\rfloor \ge (1+\varepsilon/2)\mu_m^\delta$ and

$$\begin{aligned}
&\{n(m,\delta)\in I_\varepsilon(m,\delta)\} = \left\{\left|n(m,\delta)-\mu_m^\delta\right| \le \varepsilon\mu_m^\delta\right\}\\
\supseteq&\left\{\frac{ms_{m,n}^2}{c_m^2 n} > \delta^2\ ,\ \forall\ n<(1-\varepsilon)\mu_m^\delta\right\}\\
&\quad\cap\left\{\frac{ms_{m,n}^2}{c_m^2 n} \le \delta^2\ ,\ \text{for } n=\lfloor(1+\varepsilon)\mu_m^\delta\rfloor\right\}\\
=&\left\{ms_{n,m}^2 > \frac{n}{\mu_m^\delta}\ ,\ \forall n<(1-\varepsilon)\mu_m^\delta\right\}\cap\left\{s_{n,m}^2 \le \frac{n}{\mu_m^\delta}\ ,\ \text{for } n=\lfloor(1+\varepsilon)\mu_m^\delta\rfloor\right\}\\
\supseteq&\left\{|s_{n,m}^2-\mathbb{E}[\delta_{11}^{(m)^2}]| \le \varepsilon/2\mathbb{E}[\delta_{11}^{(m)^2}]\ ,\ \forall n\ge 2\right\},
\end{aligned}$$

and the claim follows by Proposition 2.3.3.

□

For the first term in (2.21) we will need the following proposition.

Proposition 2.3.8. *For (X_l) i.i.d, $l = 1, ..., m$, with $\mathbb{E}X_l = 0$, $\mathbb{E}X_l^2 = 1$ and $\mathbb{E}X_l^4 < \infty$ and*
$(u_j)_{j\leq m}, (e_j)_{j\leq m} \subset \mathbb{R}^m$ orthonormal bases and $(\lambda_j)_{j\leq m} \in \mathbb{R}^+$, it holds that

$$\mathbb{E}\left|\sum_{j=1}^m \lambda_j \left(\left(\sum_{l=1}^m X_l(u_j, e_l)\right)^2 - 1\right)\right|^2 \leq \max_{j\leq m} \lambda_j^2(\mathbb{E}X_1^4 + 5)m.$$

Proof. By Jensen's inequality,

$$\begin{aligned}
&\left(\mathbb{E}\left[\left|\sum_{j=1}^m \lambda_j \left(\left(\sum_{l=1}^m X_l(u_j, e_l)\right)^2 - 1\right)\right|\right]\right)^2 \\
\leq &\mathbb{E}\left[\left|\sum_{j=1}^m \lambda_j \left(\left(\sum_{l=1}^m X_l(u_j, e_l)\right)^2 - 1\right)\right|^2\right] \\
= &\sum_{j,j'=1}^m \lambda_j\lambda_{j'} \left(\mathbb{E}\left[\left(\sum_{l=1}^m X_l(u_j, e_l)\right)^2 \left(\sum_{l'=1}^m X_{l'}(u_{j'}, e_{l'})\right)^2\right]\right. \\
&\quad \left. -2\mathbb{E}\left[\left(\sum_{l=1}^m X_l(u_j, e_l)\right)^2\right] + 1\right) \\
= &\sum_{j,j'=1}^m \lambda_j\lambda_{j'} \left(\sum_{l,l',l'',l'''=1}^m \mathbb{E}\left[X_l X_{l'} X_{l''} X_{l'''}\right](u_j, e_l)(u_j, e_{l'})(u_{j'}, e_{l''})(u_{j'}, e_{l'''})\right. \\
&\quad \left. +2\left(\mathbb{E}[X_1]^2\right)^2 - 1\right) \\
= &\sum_{j,j'=1}^m \lambda_j\lambda_{j'} \left(\mathbb{E}X_1^4 \sum_{l=1}^m (u_j, e_l)^2(u_{j'}, e_l)^2 + \left(\mathbb{E}[X_1^2]\right)^2 \sum_{\substack{l,l'=1 \\ l\neq l'}}^m (u_j, e_l)^2(u_{j'}, e_{l'})^2\right. \\
&\quad \left. +2\left(\mathbb{E}[X_1^2]\right)^2 \sum_{\substack{l,l'=1 \\ l\neq l'}}^m (u_j, e_l)(u_j, e_{l'})(u_{j'}, e_l)(u_{j'}, e_{l'}) - 1\right).
\end{aligned}$$

With

$$\sum_{\substack{l'=1\\l'\neq l}}^{m}(u_{j'},e_{l'})^2 = 1-(u_{j'},e_l)^2$$

and

$$\sum_{\substack{l'=1\\l'\neq l}}^{m}(u_j,e_{l'})(u_{j'},e_{l'}) = (u_j,u_{j'})-(u_j,e_l)(u_{j'},e_l)$$

we further deduce that

$$\begin{aligned}
&\left(\mathbb{E}\left[\left|\sum_{j=1}^{m}\lambda_j\left(\left(\sum_{l=1}^{m}X_l(u_j,e_l)\right)^2-1\right)\right|\right]\right)^2\\
&=\sum_{j,j'=1}^{m}\lambda_j\lambda_{j'}\left(\mathbb{E}X_1^4\sum_{l=1}^{m}(u_j,e_l)^2(u_{j'},e_l)^2+\sum_{l=1}^{m}(u_j,e_l)^2(1-(u_{j'},e_l)^2)\right.\\
&\qquad\left.+2\sum_{l=1}^{m}(u_j,e_l)(u_{j'},e_l)\left((u_j,u_{j'})-(u_j,e_l)(u_{j'},e_l)\right)-1\right)\\
&=\sum_{j,j'=1}^{m}\lambda_j\lambda_{j'}\left(\mathbb{E}X_1^4\sum_{l=1}^{m}(u_j,e_l)^2(u_{j'},e_l)^2+1-\sum_{l=1}^{m}(u_j,e_l)^2(u_{j'},e_l)^2)\right.\\
&\qquad\left.+2\left((u_j,u_{j'})^2-\sum_{l}(u_j,e_l)^2(u_{j'},e_l)^2\right)-1\right)\\
&\le\max_{j\le m}\lambda_j^2\left(\sum_{l=1}^{m}\sum_{j,j'=1}^{m}|\mathbb{E}X_1^4-3|(u_j,e_l)^2(u_{j'},e_l)^2+2\sum_{j,j'=1}^{m}(u_j,u_{j'})^2\right)\\
&\le\max_{j\le m}\lambda_j^2(\mathbb{E}X_1^4+5)m,
\end{aligned}$$

□

Finally, it is

$$\begin{aligned} M_n^{(m)} &:= n\frac{\delta_{m,n}^{meas2} - \mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}} \\ &= n\frac{\sum_{j=1}^m \lambda_j \left(\sum_{l=1}^m \sum_{i=1}^n \frac{\delta_{il}^{(m)}}{\sqrt{n}}(u_j^{(m)}, e_l^{(m)})\right)^2 - \mathbb{E}\delta_{11}^{(m)^2} \sum_{j=1}^m \lambda_j}{\mathbb{E}\delta_{11}^{(m)^2} \sum_{j=1}^m \lambda_j} \\ &= \frac{n}{\sum_{j'=1}^m \lambda_{j'}} \sum_{j=1}^m \lambda_j \left(\left(\sum_{l=1}^m \sum_{i=1}^n \frac{\delta_{il}^{(m)}}{\sqrt{n\mathbb{E}\delta_{11}^{(m)^2}}}(u_j^{(m)}, e_l^{(m)}) \right)^2 - 1 \right). \end{aligned}$$

It is easy to verify that $(M_n^{(m)})_{n\in\mathbb{N}}$ is a martingale adapted to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ generated by the measurement errors, $\mathcal{F}_n := \sigma\left(\delta_{ij}^{(m)}\ ,\ i \le n, j \le m\right)$ for every fixed $m \in \mathbb{N}$. Now assume that Assumption 2.1.9.2 with $p \ge 2$ holds true. With $n_- := (1 + \frac{3}{16}\varepsilon)\mu_m^\delta, n_+ := (1 + \frac{3}{16}\varepsilon)\mu_m^\delta$, we obtain via the Kolmogorov-Doob-inequality

$$\begin{aligned} \mathbb{P}\left(\sup_{n \in I_{\frac{3}{16}\varepsilon}} \left| \frac{\delta_{m,n}^{meas2} - \mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}} \right| \ge \frac{\varepsilon}{2} \right) &= \mathbb{P}\left(n_- \sup_{n \in I_{\frac{3}{16}\varepsilon}} \left| \frac{\delta_{m,n}^{meas2} - \mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}} \right| \ge \frac{n_-\varepsilon}{2} \right) \\ &\le \mathbb{P}\left(\sup_{n \in I_{\frac{3}{16}\varepsilon}} |M_n^{(m)}| \ge \frac{n_-\varepsilon}{2} \right) \le \frac{4\mathbb{E}\left[M_{n_+}^{(m)^2}\right]}{\varepsilon^2 n_-^2}. \end{aligned}$$

With $X_l := \sum_i \delta_{ij}^{(m)} / \sqrt{n\mathbb{E}\delta_{ij}^{(m)^2}}$ Proposition 2.3.8 yields

$$\begin{aligned} \frac{4\mathbb{E}\left[M_{n_+}^{(m)^2}\right]}{n_-^2\varepsilon^2} &= \frac{4n_+^2}{n_-^2\varepsilon^2} \frac{\max_{j\le m} \lambda_j^2 (\mathbb{E}X_1^4 + 5)m}{(\sum_j \lambda_j)^2} \\ &= \frac{4n_+^2}{\varepsilon^2 n_-^2} \frac{\max_{j\le m}\sigma_j^{-4}}{\min_{j\le m}\sigma_j^{-4}} \left(\frac{\mathbb{E}\delta_{11}^{(m)^4}}{n_+(\mathbb{E}\delta_{11}^{(m)^2})^2} + 3\frac{n_+ - 1}{n_+} + 5 \right) \frac{1}{m} \\ &= \frac{n_+^2}{\varepsilon^2 n_-^2} \kappa(P_m)^4 \left(\frac{C_d}{n_+} + 3\frac{n_+ - 1}{n_+} + 5 \right) \frac{1}{m} \to 0 \end{aligned}$$

as $m \to \infty$. In the following we write u_j and e_j for $u_j^{(m)}$ and $e_j^{(m)}$. Under Assumption 2.1.9.1, the Kolmogorov-Doob-inequality yields

$$\mathbb{P}\left(\sup_{n \in I_{\frac{3}{16}\varepsilon}} \left| \frac{\delta_{m,n}^{meas2} - \mathbb{E}\delta_{m,n}^{meas2}}{\mathbb{E}\delta_{m,n}^{meas2}} \right| \ge \frac{\varepsilon}{2} \right) \le \frac{\mathbb{E}\left|M_{n_+}^{(m)}\right|}{\varepsilon n_-}.$$

We set $S_m := \frac{M^{(m)}_{n_+}}{n_+}\sum_{j=1}^m \lambda_j$ and $Z_l^{(m)} := \sum_{i=1}^n \delta_{il}^{(m)}/\sqrt{n_+\mathbb{E}\delta_{11}^{(m)^2}}$. So $Z_l^{(m)}, j = 1, ..., m, m \in \mathbb{N}$ are i.i.d.. For $K > 0$ we truncate

$$\begin{aligned} V_l^{(m)} &:= Z_l^{(m)}\chi_{\{|Z_l^{(m)}|\leq K\}} - \mathbb{E}\left[Z_l^{(m)}\chi_{\{|Z_l^{(m)}|\leq K\}}\right] \\ W_l^{(m)} &:= Z_l^{(m)}\chi_{\{|Z_l^{(m)}|> K\}} - \mathbb{E}\left[Z_l^{(m)}\chi_{\{|Z_l^{(m)}|> K\}}\right]. \end{aligned}$$

Then $\mathbb{E}V_l^{(m)} = \mathbb{E}W_l^{(m)} = 0 = V_l^{(m)}W_l^{(m)}$ and therefore

$$\begin{aligned} &\mathbb{E}|S_m| \\ =&\mathbb{E}\left|\sum_{j=1}^m \lambda_j\left(\left(\sum_{l=1}^m Z_l^{(m)}(u_j,e_l)\right)^2 - 1\right)\right| \\ \leq&\mathbb{E}\left|\sum_{j=1}^m \lambda_j\left(\left(\sum_{l=1}^m V_l^{(m)}(u_j,e_l)\right)^2 - \mathbb{E}\left[V_1^{(1)^2}\right]\right)\right| + \mathbb{E}\left|\sum_{j=1}^m \lambda_j\left(\sum_{l=1}^m W_l^{(m)}(u_j,e_l)\right)^2\right| \\ &\quad + 2\mathbb{E}\left|\sum_{j=1}^m \lambda_j \sum_{\substack{l,l'=1\\ l\neq l'}}^m V_l^{(m)}W_{l'}^{(m)}(u_j,e_l)(u_j,e_{l'})\right| + \left|1-\mathbb{E}\left[V_1^{(1)}\right]^2\right|\sum_{j=1}^m \lambda_j. \end{aligned}$$

Since $\mathbb{E}\left[V_1^{(1)^4}\right] < \infty$, by Proposition 2.3.8 above and Jensen's inequality,

$$\begin{aligned} &\mathbb{E}\left|\sum_{j=1}^m \lambda_j\left(\left(\sum_{l=1}^m V_l^{(m)}(u_j,e_l)\right)^2 - \mathbb{E}\left[V_1^{(1)^2}\right]\right)\right| \\ \leq&\sqrt{\mathbb{E}\left|\sum_{j=1}^m \lambda_j\left(\left(\sum_{l=1}^m V_l^{(m)}(u_j,e_l)\right)^2 - \mathbb{E}\left[V_1^{(1)^2}\right]\right)\right|^2} \\ \leq&\|P_m^+\|^2\sqrt{\mathbb{E}\left[V_1^{(1)^4}\right] + 5}\sqrt{m}. \end{aligned}$$

For the second term,

$$\begin{aligned}\mathbb{E}\left|\sum_{j=1}^{m}\lambda_j\left(\sum_{l=1}^{m}W_l^{(m)}(u_j,e_l)\right)^2\right| &\leq \mathbb{E}\left|\|P_m^+\|^2\sum_{l,l'=1}^{m}W_l^{(m)}W_{l'}^{(m)}\sum_{j=1}^{m}(u_j,e_l)(u_j,e_{l'})\right|\\ &=\|P_m^+\|^2\mathbb{E}\left|\sum_{l,l'=1}^{m}W_l^{(m)}W_{l'}^{(m)}(e_l,e_{l'})\right|\\ &=m\|P_m^+\|^2\mathbb{E}\left[{W_1^{(1)}}^2\right].\end{aligned}$$

For the third term we calculate the variance,

$$\begin{aligned}&\left(\mathbb{E}\left|\sum_{j=1}^{m}\lambda_j\sum_{\substack{l,l'=1\\ l\neq l'}}^{m}V_l^{(m)}W_{l'}^{(m)}(u_j,e_l)(u_j,e_{l'})\right|\right)^2\\ \leq&\mathbb{E}\left|\sum_{j=1}^{m}\lambda_j\sum_{\substack{l,l'=1\\ l\neq l'}}^{m}V_l^{(m)}W_{l'}^{(m)}(u_j,e_l)(u_j,e_{l'})\right|^2\\ \leq&\mathbb{E}\sum_{j,j'=1}^{m}\lambda_j\lambda_{j'}\sum_{\substack{l,l'=1\\ l\neq l'}}^{m}\sum_{\substack{l'',l'''=1\\ l''\neq l'''}}^{m}V_l^{(m)}W_{l'}^{(m)}V_{l''}^{(m)}W_{l'''}^{(m)}(u_j,e_l)(u_j,e_{l'})(u_{j'},e_{l''})(u_{j'},e_{l'''})\\ =&\mathbb{E}\left[{V_1^{(1)}}^2\right]\sum_{j,j'=1}^{m}\lambda_j\lambda_j'\sum_{\substack{l,l'=1\\ l\neq l'}}^{m}(u_j,e_l)(u_j,e_{l'})(u_{j'},e_l)(u_{j'},e_{l'})\\ =&\mathbb{E}\left[{V_1^{(1)}}^2\right]\mathbb{E}\left[{W_1^{(1)}}^2\right]\sum_{j,j'=1}^{m}\lambda_j\lambda_{j'}\sum_{l=1}^{m}(u_j,e_l)(u_{j'},e_l)\left((u_j,u_{j'})-(u_j,e_l)(u_{j'},e_{l'})\right)\\ =&\mathbb{E}\left[{V_1^{(1)}}^2\right]\mathbb{E}\left[{W_1^{(1)}}^2\right]\sum_{j,j'=1}^{m}\lambda_j\lambda_{j'}\left((u_j,u_{j'})^2-\sum_{l=1}^{m}(u_j,e_l)^2(u_{j'},e_l)^2\right)\\ \leq&\mathbb{E}\left[{V_1^{(1)}}^2\right]\mathbb{E}\left[{W_1^{(1)}}^2\right]\sum_{j,j'=1}^{m}\lambda_j\lambda_{j'}(u_j,u_{j'})^2=\mathbb{E}\left[{V_1^{(1)}}^2\right]\mathbb{E}\left[{W_1^{(1)}}^2\right]\sum_{j=1}^{m}\lambda_j^2\\ \leq&\mathbb{E}\left[{V_1^{(1)}}^2\right]\mathbb{E}\left[{W_1^{(1)}}^2\right]\|P_m^+\|^4m.\end{aligned}$$

Altogether,

$$\begin{aligned}
&\frac{2\mathbb{E}|M_{n_+}^{(m)}|}{\varepsilon n_-}\\
\le&\frac{2n_+}{\varepsilon n_-}\frac{1}{\sum_j \lambda_j}\mathbb{E}|S_m|\\
\le&\frac{2n_+}{\varepsilon n_-}\frac{1}{\sum_j \lambda_j}\left(\|P_m^+\|^2\sqrt{\mathbb{E}\left[V_1^{(1)^4}\right]}+5\sqrt{m}+m\|P_m^+\|^2\mathbb{E}\left[W_1^{(1)^2}\right]\right.\\
&\qquad\left.+\sqrt{\mathbb{E}\left[V_1^{(1)^2}\right]\mathbb{E}\left[W_1^{(1)^2}\right]}\|P_m^+\|^2\sqrt{m}|1-\mathbb{E}V^2|m\|P_m^+\|^2\right)\\
\le&\frac{2n_+\kappa(P_m)^2}{\varepsilon n_-\sqrt{m}}\left(\sqrt{\mathbb{E}\left[V_1^{(1)^4}\right]}+5+\sqrt{\mathbb{E}\left[V_1^{(1)^2}\right]\mathbb{E}\left[W_1^{(1)^2}\right]}\right)\\
&\qquad+\frac{2n_+\kappa(P_m)^2}{\varepsilon n_-}\left(\mathbb{E}\left[W_1^{(1)^2}\right]+\left|1-\mathbb{E}\left[V_1^{(1)^2}\right]\right|\right).
\end{aligned}$$

The claim follows with $\lim_{K\to\infty}\mathbb{E}\left[V_1^{(1)^2}\right]=1, \lim_{K\to\infty}\mathbb{E}\left[W_1^{(1)^2}\right]=0$ and $\sup_m \kappa(P_m)^2<\infty$.

□

We come to the main proof

Proof. We set

$$\Omega_m := \left\{\left\|P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-P_m^+P_m\hat{y}\right\|\le\frac{\tau+C_0}{2C_0}\delta_m^{disc}\right\}.$$

Then,

$$\begin{aligned}
\left\|P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-\hat{y}\right\|\chi_{\Omega_m}\le&\left\|P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-P_m^+P_m\hat{y}\right\|\chi_{\Omega_m}+\left\|P_m^+P_m\hat{y}-\hat{y}\right\|\chi_{\Omega_m}\\
\le&\frac{\tau+3C_0}{2C_0}\delta_m^{disc}.
\end{aligned}\tag{2.22}$$

By Algorithm 2 it is

$$\begin{aligned}
&\alpha_m\\
:=&\left\{q^k\ ,\ k\in\mathbb{N}_0\ ,\ \left\|KR_{\alpha_m}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}\right\|\le2\tau\delta_m^{disc}\right\}\\
=&\left\{q^k\ ,\ k\in\mathbb{N}_0\ ,\ \left\|KR_{\alpha_m}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}-P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})}\right\|\le\frac{4\tau C_0}{\tau+3C_0}\frac{\tau+3C_0}{2C_0}\delta_m^{disc}\right\}
\end{aligned}$$

and because of $\frac{4\tau C_0}{\tau+3C_0} > C_0$,(2.22) and $\lim_{m\to\infty}\delta_m^{disc} = 0$, it follows that

$$\lim_{m\to\infty}\left\|R_{\alpha_m}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})} - K^+\hat{y}\right\|\chi_{\Omega_m} = 0$$

by Theorem 4.17 and Remark 4.18 from [EHN96]. With the same reasoning it follows that there is a $L' \in \mathbb{R}$ such that

$$\left\|R_{\alpha_m}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})} - K^+\hat{y}\right\|\chi_{\Omega_m} \leq L'\rho^{\frac{1}{\nu+1}}\delta_m^{disc\,\frac{\nu}{\nu+1}},$$

if there are $0 < \nu \leq \nu_0 - 1$ and $\xi \in \mathcal{X}$ with $K^+\hat{y} = (K^*K)^{\nu/2}\xi$ and $\|\xi\| \leq \rho$. Lemma 2.3.6 implies that $\lim_{m\to\infty}\mathbb{P}\left(\Omega_m\right) = 1$, which concludes the proof.

□

2.4 Numerical Demonstration

We provide numerical experiments to complement the theoretical analysis. Three model examples, i.e. `phillips` (mildly ill-posed, smooth), `gravity` (severely ill-posed, medium smooth) and `shaw` (severely ill-posed, non smooth), are taken from the open source `MATLAB` package Regutools [Han94].The problems cover a variety of setting, e.g., different solution smoothness and degree of ill-posedness. These examples are discretisations of Fredholm/Volterra integral equations of the first kind, by means of either the Galerkin approximation with piecewise constant basis functions or quadrature rules. We approximate our infinite-dimensional K with one of the above examples with dimension $m_\infty \gg 1$. The number of measurements channels m is then always chosen such that $m \ll m_\infty$. In most of the examples we use discretisation by box functions as follows, compare to Lemma 2.1.5. With $k = m_\infty/m$ we set

$$\begin{aligned} P_m &: \mathbb{R}^{m_\infty} \to \mathbb{R}^m \\ &\begin{pmatrix} y_{(i-1)k+1} \\ \dots \\ y_{(i-1)k+k} \end{pmatrix} \mapsto \frac{1}{\sqrt{k}}\left(y_{(i-1)k+1} + \dots + y_{(i-1)k+k}\right)e_i \end{aligned}$$

where $i = 1, ..., m$ and $e_1, ..., e_m$ is the canonical basis of $\mathbb{R}^m$. In Subsection 2.4.3 we will also consider discretisation by hat functions to give an example with nonorthogonal discretisation. We chose a shifted generalised Pareto distribution for the distribution of the measurement error, e.g. $\delta_{ij}^{(m)} = Z_{ij}^{(m)} - EZ_{ij}^{(m)}$, where $Z_{ij}^{(m)}$ are i.i.d and follow a generalised Pareto distribution (gprnd(l,σ,θ,m,n) in Matlab, with $l = 1/3$, $\sigma = \sqrt{(1-l)^2(1-2l)\|\hat{y}\|}$ and $\theta = 0$). This distribution is highly non symmetric with a heavy tail. The above choices for the parameters imply that $\mathbb{E}{\delta_{ij}^{(m)}}^2 = \|\hat{y}\|$ and $\mathbb{E}|\delta_{ij}^{(m)}|^3 = \infty$. Thus the error fulfills Assumption 2.1.9.1 in all the examples. The parameter τ in the definition of the discrepancy principle is set to $\tau = 1.2$. All the statistical quantities are computed for 100 independent runs, and the results are presented as box plots.

2.4.1 Convergence of finite-dimensional residuum approach

First we visualise the convergence of the discrepancy principle with the finite-dimensional
residuum approach, as stated in Corollary 2.0.1. We use discretisation by box functions as presented above and set $m_\infty = 4000$ and $m = 5, 10, 20$. For each m we plot in Figure 2.1 the resulting relative errors $\|R_{\alpha_{m,n}}^{(m)} \bar{Y}_n^{(m)} - \hat{x}\| / \|\hat{x}\|$ for $n = 10, ..., 10^9$ repetitions. For m fix, the relative errors first decrease steadily, and then saturate (at $\|\hat{x} - (P_m K)^+ P_m K \hat{x}\|$), as the number of repetitions n grows. The saturation level decreases rapidly while m grows, confirming the convergence of the approach. It is notable, that for all examples a fairly small number of measurement channels is sufficient to yield good approximations.

2.4.2 (Semi-)Convergence of infinite-dimensional residuum approach

Now we come to the discrepancy principle with the infinite-dimensional residuum approach, as stated in Corollary 2.0.2. Again we chose discretisation by box functions for the measurements with $m_\infty = 4000$ and this time we set $m = 20, 50, 100$. For each m we plot in Figure 2.1 the resulting relative errors $\|R_{\alpha_m} P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - \hat{x}\| / \|\hat{x}\|$ for varying upper bound δ_m^{disc} from Assumption 2.2.1. More precisely we chose the latter in relation to the exact discretisation error $d_m := \|\hat{y} - P_m^+ P_m \hat{y}\|$. In particular we also consider $\delta_m^{disc} < d_m$ and we exhibit a semi-convergence. Strictly speaking, the last two choices ($d_m/2$ and $d_m/4$) for δ_m^{disc} violate Assumption 2.2.1 and we thus illustrate the sensitiveness to underestimation of the true discretisation error. It is notable that for the choice $\delta_m^{disc} = d_m/2$ (e.g. underestimation of the discretisation error by a factor 1/2) the relative errors are still decreasing. This is explained by the fact, that the estimation in (2.11) is quite coarse. Together with the choice $\tau = 1.2$ this yields, that it still holds that the true unknown error $\|P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - \hat{y}\|$ fulfills $\|P_m^+ \bar{Y}_{n(m,\delta_m^{disc})}^{(m)} - \hat{y}\| < 2\tau\delta_m^{disc}$. For the choice $\delta_m^{disc} = d_m/4$ the errors then diverge. The semi-convergence is in contrast to the saturation observed in the left column of Figure 2.1 and illustrates the fundamental difference, that for the finite-dimensional approach no quantitative knowledge of the discretisation error is required, while for the infinite-dimensional approach it is.

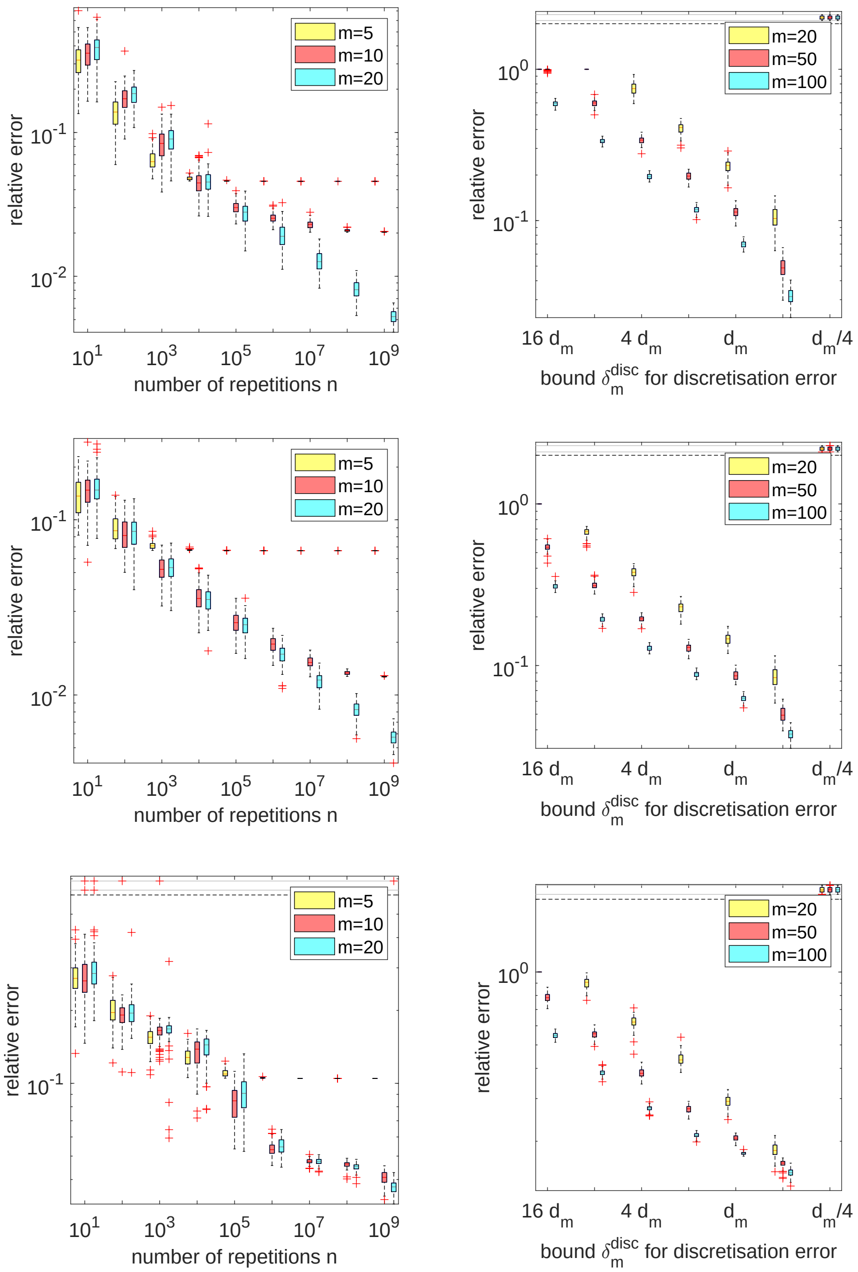

Figure 2.1: Results of approach (2.2) and (2.6) with the discrepancy principle as implemented in Algorithm 1 (left column) or 2 (right column) respectively, for 'phillips' (first row), 'gravity' (second row) and 'shaw' (third row), visualised as boxplots for 100 independent runs. *Left column*: Relative errors $\|R^{(m)}_{\alpha_{m,n}}\bar{Y}^{(m)}_n - \hat{x}\|/\|\hat{x}\|$ against number of repetitions n for several numbers of measurement channels m. *Right column*: Relative errors $\|R_{\alpha_m}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})} - \hat{x}\|/\|\hat{x}\|$ against bound for the discretisation error δ_m^{disc} for several numbers of measurement channels m. δ_m^{disc} is chosen in relation to the exact discretisation error $d_m := \|\hat{y} - P_m^+P_m\hat{y}\|$.

2.4.3 Comparison of the both approaches

We now compare the both approaches directly. We consider discretisation by box functions with $m_\infty = 4000$ and $m = 50, 100, 200$ and discretisation by hat functions (compare to Proposition 2.1.6). The latter is precisely implemented as follows. With $k = \frac{m_\infty - 1}{m-1}$ we set

$$\begin{aligned} P_m &: \mathbb{R}^{m_\infty} \to \mathbb{R}^m \\ &\begin{pmatrix} y_{(i-1)k+1} \\ \dots \\ y_{(i+1)k+1} \end{pmatrix} \mapsto \frac{1}{\sqrt{\sum_{j=1}^{2k+1} a_j^2}} \left(a_1 y_{(i-1)k+1} + \dots + a_{2k+1} y_{(i+1)k+1} \right) e_i \end{aligned}$$

where $i = 2, ..., m-1$ and

$$a_i := \begin{cases} (i-1)/k & i \leq k+1, \\ 1 - (i-k-1)/k & i \geq k+1. \end{cases}$$

For the boundaries we set,

$$\begin{pmatrix} y_1 \\ \dots \\ y_{k+1} \end{pmatrix} \mapsto \frac{1}{\sqrt{\sum_{i=k+1}^{k=2k+1} a_i^2}} \left(a_{k+1} y_1 + \dots + a_{2k+1} y_{k+1} \right) e_1$$

and

$$\begin{pmatrix} y_{m_\infty - (k+1)} \\ \dots \\ y_{m_\infty} \end{pmatrix} \mapsto \frac{1}{\sqrt{\sum_{i=1}^{k=k+1} a_i^2}} \left(a_1 y_{m_\infty - (k+1)} + \dots + a_{k+1} y_{m_\infty} \right) e_m.$$

Here we use $m_\infty = 4132$ and $m = 18, 28, 52$. We first applied Algorithm 2 with exact upper bound $\delta_m^{disc} = \|\hat{y} - P_m^+ P_m \hat{y}\|$. The (random) stopping index $n(m, \delta_m^{disc})$ from Algorithm 2 is then used as the number of repetitions n in Algorithm 1. We plot in Figure 2.2 the relative errors of the both approaches for growing number of measurement channels m. We observe the stated convergence as m grows. Moreover, the errors of the approach with finite-dimensional residuum are even slightly better than the ones of the approach with infinite-dimensional approach in all the examples. This gives numerical evidence, that also the first approach is order optimal in various settings.

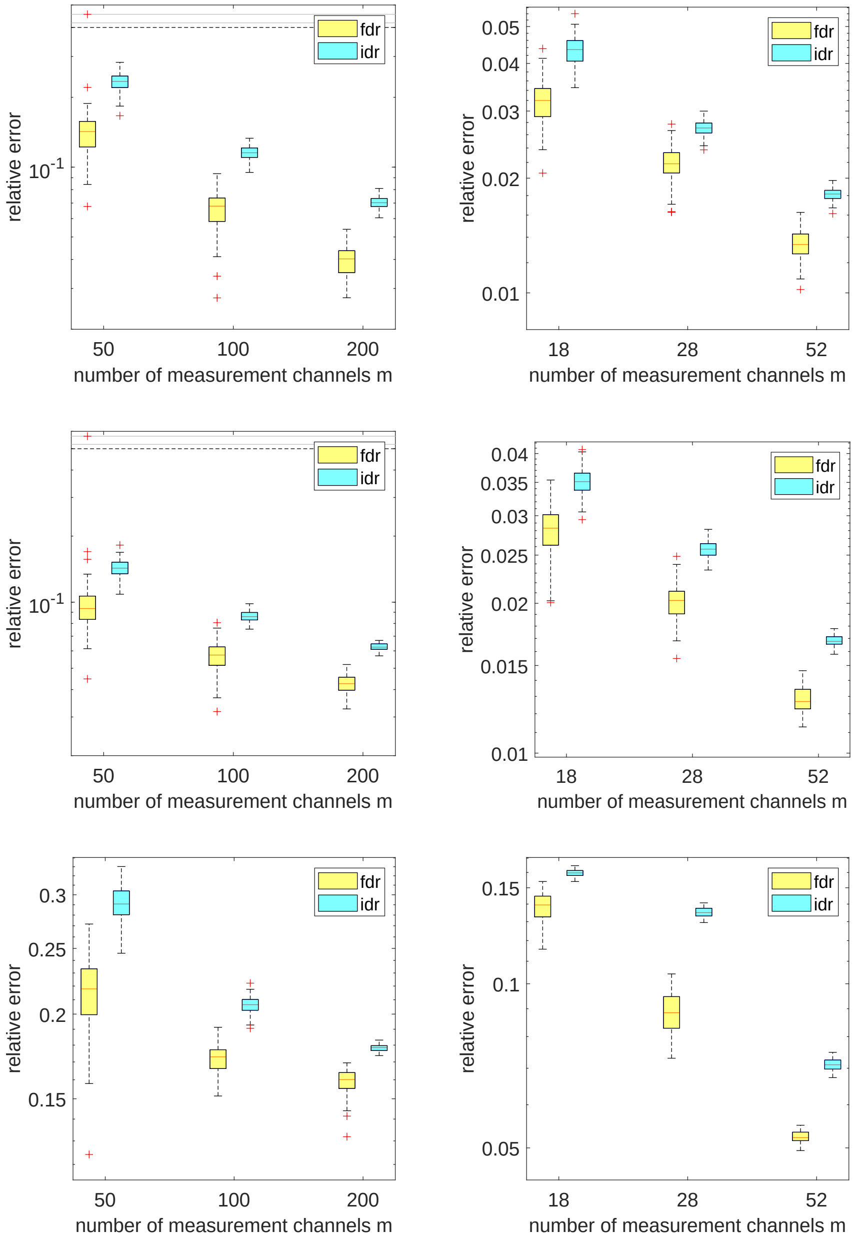

Figure 2.2: Direct comparison of both approaches (2.2) (fdr) and (2.6) (idr) with discrepancy principle as implemented in Algorithm 1 and 2 for 'phillips' (first line), 'gravity' (second line) and 'shaw' (third line). For the discretisation of the measurements either box functions (first column) or hat functions (second column) are used. Concretely, the relative errors $\|R_{\alpha_{m,n(m,\delta_m^{disc})}}\bar{Y}^{(m)}_{n(m,\delta_m^{disc})} - \hat{x}\|/\|\hat{x}\|$ (fdr) and $\|R_{\alpha_m}P_m^+\bar{Y}^{(m)}_{n(m,\delta_m^{disc})} - \hat{x}\|/\|\hat{x}\|$ (idr) are plotted against the number of measurement channels m, where δ_m^{disc} is chosen to be the exact discretisation error $\|\hat{y} - P_m^+P_m\hat{y}\|$ and $n(m,\delta_m^{disc})$ is calculated with Algorithm 2.

2.5 Concluding remarks

In this chapter, we have analysed linear inverse problems under unknown white noise. We presented two approaches for the solution. In both cases, we used multiple discretised measurements to prove convergence in probability against the true solution, as the number of repetitions and the number of measurement channels tend to infinity. The first approach neither required knowledge of the arbitrary error distribution, nor quantitative knowledge of the quality of the discretisation to obtain convergence. For the second approach we also proved an optimal convergence rate, under additional knowledge of the discretisation error.

We want to pronounce two important outstanding questions. Firstly, the discretisation considered in this article entered the problem through discretised measurements. In particular, this is determined by the practical problem and the way the data is measured or acquired. In order to solve the problem numerically, as in the preceding section, one also has to discretise the true unknown $\hat{x}$. In contrast to the measurements, here there is more freedom to choose the numerical discretisation, since one is basically only limited by computational power. It therefore is of high interest to find an optimal choice for that. Secondly, it might come as a surprise that in all the numerical examples the approach with finite-dimensional residuum (fdr) gives slightly better results than the one with infinite-dimensional residuum (idr), even though the theoretical results do only guarantee the optimality of the latter one. Thus an important open question is to derive natural and verifiable conditions, which rigorously guarantee optimality of the first approach.

Chapter 3

The discrepancy principle for stochastic gradient descent

Sections 3.1, 3.2 and 3.4 are, up to minor changes, published in [JJ20]. Section 3.3 contains yet unpublished results.

In chapter 1 and 2 we mainly focused on classical filter-based regularisation methods. Relatively novel methods used heavily in machine learning do not fit into this framework, and its application and convergence properties remain largely unexplored, in particular in the context of regularisation theory of inverse problems. Here we focus on a seemingly simple method, the stochastic gradient descent. It is classically formulated in a finite-dimensional setting, and we hence study the following finite-dimensional (though possibly the dimension may be extremely large) linear inverse problem:

$$Ax = \hat{y}, \tag{3.1}$$

where $x \in \mathbb{R}^{m'}$ is the unknown signal of interest, $\hat{y} \in \mathbb{R}^m$ is the exact data and $A \in \mathbb{R}^{m \times m'}$ is the system matrix. In practice, we have access only to a corrupted version y^δ of the exact data $\hat{y} = A\hat{x}$ (with the reference solution $\hat{x}$ being any exact solution). In order to isolate the difficulties arising intrinsically from the usage of stochastic gradient descent, we first restrict to classical deterministic noise. The case of i.i.d. measurements and estimated noise level is discussed afterwards in Section 3.3. So the measurement is

$$y^\delta = \hat{y} + \xi$$

where $\xi \in \mathbb{R}^m$ denotes the noise, with a noise level $\delta = \|\xi\|$. When the size of the problem (3.1) is massive, the classical methods from the previous chapters may become infeasible due to computational complexity. Especially computationally cheap and thus attractive is a simple stochastic gradient descent (SGD) [RM51, BCN18]. In its simplest form, it reads as follows: given an initial guess $x_1^\delta = x_1 \in \mathbb{R}^{m'}$, let

$$x_{k+1}^\delta := x_k^\delta - \eta_k((a_{i_k}, x_k^\delta) - y_{i_k}^\delta)a_{i_k}, \quad k = 1, 2, \ldots, \tag{3.2}$$

where $\eta_k > 0$ is a decreasing stepsize, a_i is the i-th row of the matrix A (as a column vector), $(\cdot, \cdot)$ denotes Euclidean inner product on $\mathbb{R}^{m'}$, and the row index i_k at the

kth SGD iteration is chosen uniformly (with replacement) from the set $\{1, ..., m\}$. It can be derived by applying stochastic gradient descent to the quadratic functional:

$$J(x) = \frac{1}{2m}\|Ax - y^\delta\|^2 = \frac{1}{m}\sum_{i=1}^{m} f_i(x), \quad \text{with } f_i(x) = \frac{1}{2}((a_i, x) - y_i^\delta)^2.$$

Distinctly, the method (3.2) operates only on one single data pair (a_{i_k}, y_{i_k}) each time, and thus it is directly scalable to the data size m of problem (3.1). This feature makes it especially attractive in the context of massive data.

As already stated in the introduction, a central open problem is the verification of adaptive stopping rules for stochastic gradient descent. In this chapter we once more take the focus on the discrpancy principle 0.6. Specifically in this context, with x_k^δ being the kth iterate constructed by an iterative regularization method, the principle determines the stopping index $k(\delta)$ by

$$k(\delta) := \min\left\{k \in \mathbb{N} : \|Ax_k^\delta - y^\delta\| \leq \tau\delta\right\}, \tag{3.3}$$

where the constant $\tau > 1$ is fixed. Note that the stopping index $k(\delta)$ depends on the random iterate x_k^δ, and thus it is also a random variable, which poses the main challenge in the theoretical analysis. The use of the discrepancy principle in the context of stochastic iterative methods has not been explored so far, to the best of our knowledge. The goal of this chapter is to study the basic properties of the discrepancy principle for SGD. It is worth noting that a direct computation of the residual $\|Ax_k^\delta - y^\delta\|$ at every SGD iteration is demanding. However, one may compute it not at every SGD iteration but only with a given frequency (e.g., per epoch, see Section 3.4), as done by the popular stochastic variance reduced gradient [JZ13], for which residual evaluation is a part of gradient computation. Also there are efficient methods to compute the residual $\|Ax_k^\delta - y^\delta\|$ using randomized SVD [KJ19], by exploiting the intrinsic low-rank nature for many practical inverse problems.

3.1 Convergence and a finite termination property

Now we specify the algorithmic parameters for SGD, and state the main results of the work. Throughout, we make the following assumption on the stepsizes and the regularity condition on the ground truth solution $\hat{x}$, i.e., the minimum-norm solution defined by

$$\hat{x} = \arg\min_{x:Ax=\hat{y}} \|x\|. \tag{3.4}$$

The stepsize schedule in (i) is commonly known as the polynomially decaying stepsize schedule, and (ii) is the classical power type source condition, where $B = m^{-1}(A^TA)$ (with m being the data size, i.e., the number of rows in A), imposing a type of smoothness on the solution $\hat{x}$ (relative to the system matrix A and the initial guess x_1). In the analysis and computation below, x_1 is fixed at 0. Generally, in classical regularization theory for infinite-dimensional inverse problems, the source element w plays the role of a Lagrangian multiplier of the constrained problem in (3.4), whose existence is not ensured for an operator with a nonclosed range and has to be assumed [EHN96, IJ15]. In the finite-dimensional case, the existence of a source element w for the case $\frac{\nu}{2} \leq 1$ is ensured, but the norm of the source element w can be arbitrarily large.

Assumption 3.1.1. *The following conditions hold.*

(i) The stepsizes η_j satisfy $\eta_j = c_0 j^{-\alpha}$, with $\alpha \in (0,1)$ and $c_0 \leq (\max_{j=1,\ldots,n} \|a_j\|^2)^{-1}$.

(ii) There is a $\nu > 0$ and a $w \in \mathbb{R}^{m'}$ such that $\hat{x} - x_1 = B^{\frac{\nu}{2}} w$.

The first theorem gives a finite-iteration termination property of the discrepancy principle, where $\mathbb{P}$ is with respect to the filtration generated by the random index $(i_k)_{k=1}^{\infty}$. It can also be viewed as a partial result on the optimality. It implies in particular that for $\nu < 1$, the data propagation error is of optimal order. The proof relies crucially on the observation that the variance component of the mean squared residual contributes only marginally for sufficiently large k.

Theorem 3.1.2. *Let Assumption 3.1.1 be fulfilled, and $k(\delta)$ be determined by the discrepancy principle* (3.3). *Then for all $0 < r < 1$ and $\tau > \tau^* > 1$, with $c = \left(\frac{\tau^*-1}{\sqrt{m} c_\nu}\right)^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}} + 2$, there holds*

$$\mathbb{P}\left(k(\delta) \leq c\delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}\right) \to 1 \quad \text{as } \delta \to 0^+,$$

with the constant $c_\nu = \left(\frac{(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)}{c_0 e (2^{1-\alpha}-1)}\right)^{\frac{\nu}{2}+\frac{1}{2}} \|w\|$.

Remark 3.1.3. The condition $r < 1$ is related to an apparent saturation phenomenon with SGD: for any $\nu > 1$, the SGD iterate x_k^δ with *a priori* stopping can only achieve a convergence rate comparable with that for $\nu = 1$ in the setting of Assumption 3.1.1, at least for the analysis in [JL19]. However, in the very recent preprint [JZZ20b] a refined convergence analysis is presented, showing that this saturation actually does not occur, if the initial step size c_0 is sufficiently small.

The second contribution of this chapter is on the convergence in probability of the SGD iterate $x_{k(\delta)}^\delta$ with the stopping index $k(\delta)$ determined by (3.3). This result has one drawback. In the proof, we have to assume that the stopping index $k(\delta)$ is independent of the iterates $x_{k(\delta)}^\delta$. In practice, this can be achieved by running SGD twice with the same data (y^δ, δ): the first round is for the determination of $k(\delta)$, then the second (independent) round is stopped using $k(\delta)$. This increases the computational expense by a factor of 2. However, the numerical results in Section 3.4 show that one can use the iterate from the first run without compromising the accuracy.

Theorem 3.1.4. *Let Assumption 3.1.1 be fulfilled, and $k(\delta)$ be determined by the discrepancy principle* (3.3)*. Then for all $\varepsilon > 0$ there holds*

$$\mathbb{P}\left(\|x_{k(\delta)}^\delta - \hat{x}\| \geq \varepsilon\right) \to 0 \quad \text{as } \delta \to 0^+,$$

where $(x_k^\delta)_{k\in\mathbb{N}}$ are SGD iterates independent of $k(\delta)$, with the same data (y^δ, δ).

In sum, Theorems 3.1.2 and 3.1.4 confirm that the discrepancy principle is a valid *a posteriori* stopping rule for SGD. However, they do not give a rate of convergence, which remains an open problem. Numerically, we observe that the convergence rate obtained by the discrepancy principle is nearly order-optimal for low-regularity solutions, as the *a priori* rule in the regime in [JL19], and the performance is competitive with the standard Landweber method. Thus, the method is especially attractive for finding a low-accuracy solution. However, for very smooth solutions (i.e., large ν), it manifested as an undesirable saturation phenomenon, due to the presence of the significant variance component (when compared with the approximation error), under the setting of Assumption 3.1.1. The rest of the chapter is organized as follows. In Sections 3.2.1 and 3.2.2, we prove Theorems 3.1.2 and 3.1.4, respectively. Several auxiliary results needed for the proof of Theorem 3.1.2 are given in Section 3.2.3. The setup with repeatedly i.i.d. measurements is discussed in Section 3.3. Finally, several numerical experiments are presented in Section 3.4 to complement the theoretical analysis. We conclude with some useful notation. We denote the SGD iterate for exact data $\hat{y}$ by x_k, and that for noisy data y^δ by x_k^δ. The expectation $\mathbb{E}[\cdot]$ is with respect to the filtration $\mathcal{F}_k$, generated by the random indices $\{i_1, \ldots, i_k\}$.

3.2 Proofs

In this section we gather the proofs.

3.2.1 The proof of Theorem 3.1.2

In this section, we give the proof of Theorem 3.1.2. First, we give several preliminary facts. By the construction in (3.2), since x_k^δ is measurable with respect to $\mathcal{F}_{k-1}$,

$$\begin{aligned}\mathbb{E}[x_{k+1}^\delta|\mathcal{F}_{k-1}] &= x_k^\delta - \eta_k m^{-1}\sum_{i=1}^n((a_i, x_k^\delta) - y_i^\delta)a_i \\ &= x_k^\delta - \eta_k m^{-1}(A^tAx_k^\delta - A^ty^\delta).\end{aligned}$$

Thus, by the law of total expectation, the sequence $(\mathbb{E}[x_k^\delta])_{k\in\mathbb{N}}$ satisfies the following recursion:

$$\mathbb{E}[x_{k+1}^\delta] = \mathbb{E}[x_k^\delta] - \eta_k(\bar{A}^t\bar{A}\mathbb{E}[x_k^\delta] - \bar{A}^t\bar{y}^\delta) \tag{3.5}$$

with $\bar{A} = m^{-\frac{1}{2}}A$ and $\bar{y}^\delta = m^{-\frac{1}{2}}y^\delta$. This is exactly the classical Landweber method [Lan51] (but with diminishing stepsizes) applied to the rescaled linear system $\bar{A}x = \bar{y}^\delta$. For the Landweber method, the discrepancy principle (3.3), e.g., regularizing property and optimal convergence rates, has been thoroughly studied for both linear and nonlinear inverse problems (see, e.g., [EHN96, Chapter 6] and [KNS08]). The key insight for the analysis below is the following empirical observation: for a suitably large k, typically the variance component $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2] \ll \delta^2$, as confirmed by the numerical experiments in Section 3.4.2. This fact allows us to transfer the results for the Landweber method to SGD.

The proof of Theorem 3.1.2 employs two preliminary results, whose lengthy proofs are deferred to Section 3.2.3. The first result gives an upper bound of the following stopping index $k^*(\delta)$, for any $\tau^* > 1$, defined by

$$k^*(\delta) := \min\{k \in \mathbb{N} \ : \ \|A\mathbb{E}[x_k^\delta] - y^\delta\| \le \tau^*\delta\}. \tag{3.6}$$

Clearly, $k^*(\delta)$ is the stopping index by the classical discrepancy principle, when applied to the sequence $(\mathbb{E}[x_k^\delta])_{k\in\mathbb{N}}$, which is exactly the Landweber method, in view of the relation (3.5).

Proposition 3.2.1. *Let Assumption 3.1.1 be fulfilled. Then for $k^*(\delta)$ defined in* (3.6)*, there holds*

$$k^*(\delta) \le \Big(\frac{\tau^* - 1}{\sqrt{m}c_\nu}\delta\Big)^{-\frac{2}{(1-\alpha)(\nu+1)}} + 2, \tag{3.7}$$

with $c_\nu = \big(\frac{(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)}{c_0 e(2^{1-\alpha}-1)}\big)^{\frac{\nu}{2}+\frac{1}{2}}\|w\|$.

The second result gives an upper bound on the variance component $\mathbb{E}[\|A(x_{\kappa(\delta)}^\delta - \mathbb{E}[x_{\kappa(\delta)}^\delta])\|^2]$ of the mean squared residual $\mathbb{E}[\|Ax_k - y^\delta\|^2]$. It indicates that the variance $\mathbb{E}[\|A(x_{k(\delta)}^\delta - \mathbb{E}[x_{k(\delta)}^\delta])\|^2]$ contributes only marginally to the mean squared residual $\mathbb{E}[\|Ax_{k(\delta)}^\delta - y^\delta\|^2]$, and consequently the squared residual $\|Ax_{k(\delta)}^\delta - y^\delta\|^2$ of individual realizations of SGD may be used instead for determining an appropriate stopping index.

Proposition 3.2.2. *Under Assumption 3.1.1 with $\kappa(\delta) \ge \delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}$ and $0 < r < 1$, there holds*

$$\mathbb{E}[\|A(x_{\kappa(\delta)}^\delta - \mathbb{E}[x_{\kappa(\delta)}^\delta])\|^2] = o(\delta^2), \quad \text{as } \delta \to 0^+.$$

Now we can present the proof of Theorem 3.1.2.

Proof. Set $1 < \tau^* < \tau$ and $\bar{k}(\delta) = [c\delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}] + 2$ ($[\cdot]$ denotes taking the integral part of a real number), with $c = \big(\frac{\tau^*-1}{\sqrt{m}c_\nu}\big)^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}$. By the definition of $k(\delta)$ in (3.3), the event $\mathcal{E} = \{k(\delta) \le \bar{k}(\delta)\}$ is given by

$$\mathcal{E} = \{\exists i \in \{1, \ldots, \bar{k}(\delta)\} \text{ such that } \|Ax_i^\delta - y^\delta\| \le \tau\delta\}.$$

Thus, $\mathcal{E} \supset \{\|Ax^\delta_{\bar{k}(\delta)} - y^\delta\| \leq \tau\delta\}$. Consequently,

$$\begin{aligned}\mathbb{P}(k(\delta) \leq \bar{k}(\delta)) &\geq \mathbb{P}(\|Ax^\delta_{\bar{k}(\delta)} - y^\delta\| \leq \tau\delta)\\ &\geq \mathbb{P}(\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\| \leq (\tau-\tau^*)\delta,\ \|A\mathbb{E}[x^\delta_{\bar{k}(\delta)}] - y^\delta\| \leq \tau^*\delta).\end{aligned}$$

By the choice of $\bar{k}(\delta)$, Proposition 3.2.1 implies

$$\|A\mathbb{E}[x^\delta_{\bar{k}(\delta)}] - y^\delta\| \leq \tau^*\delta.$$

Consequently,

$$\begin{aligned}\mathbb{P}(k(\delta) \leq \bar{k}(\delta)) &\geq \mathbb{P}(\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\| \leq (\tau-\tau^*)\delta)\\ &= 1 - \mathbb{P}(\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\| > (\tau-\tau^*)\delta).\end{aligned}$$

Meanwhile, by Tschebyscheff's inequality [Fel68, p. 233], we have

$$\mathbb{P}(\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\| > (\tau-\tau^*)\delta) \leq \frac{\mathbb{E}\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\|^2}{(\tau-\tau^*)^2\delta^2}.$$

Therefore,

$$\mathbb{P}(k(\delta) \leq \bar{k}(\delta)) \geq 1 - \frac{\mathbb{E}\|A(x^\delta_{\bar{k}(\delta)} - \mathbb{E}[x^\delta_{\bar{k}(\delta)}])\|^2}{(\tau-\tau^*)^2\delta^2},$$

which together with Proposition 3.2.2 directly implies

$$\mathbb{P}(k(\delta) \leq \bar{k}(\delta)) \to 1 \quad \text{as } \delta \to 0^+.$$

This completes the proof of the theorem. □

Remark 3.2.3. The condition $r < 1$ is related to an apparent saturation phenomenon with SGD: for any $\nu > 1$, the SGD iterate x_k^δ with *a priori* stopping can only achieve a convergence rate comparable with that for $\nu = 1$ in the setting of Assumption 3.1.1, at least for the current analysis [JL19]. It remains unclear whether this is an intrinsic drawback of SGD or due to limitations of the proof technique.

Remark 3.2.4. In practice, we prefer computing the residual with a frequency $\omega m \in \mathbb{N}$:

$$k_\omega(\delta) := \min\{\omega m k \ : \ k \in \mathbb{N}\ ,\ \|Ax^\delta_{\omega m k} - y^\delta\| \leq \tau\delta\}.$$

Since one of the numbers $[c\delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}] + 2, ..., [c\delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}] + \omega m + 1$ is of the form $\omega m k$, with $k \in \mathbb{N}$, there holds

$$\mathbb{P}\left(k_\omega(\delta) \leq c\delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}} + \omega m + 1\right) \to 1 \quad \text{as } \delta \to 0^+.$$

That is, the upper bound on the stopping index remains largely valid for a variant of the discrepancy principle (3.3) evaluated with a given frequency.

Remark 3.2.5. The finite-iteration termination property in Theorem 3.1.2 relies heavily on the assumption $\alpha < 1$ in the definition of the stepsize schedule. Without this condition, Theorem 3.1.2 (and thus also the convergence in probability) generally do not hold. Indeed, if $\operatorname{rank}(A) \geq 2$, $\hat{y} \neq 0$ and $\alpha > 1$, then there holds

$$\liminf_{\delta \to 0^+} \mathbb{P}(k(\delta) = \infty) > 0. \tag{3.8}$$

To prove this assertion, let $k^* \in \mathbb{N}$ be such that $\eta_k \|A\|^2 \leq \frac{1}{2}$ for all $k \geq k^*$. Since $\operatorname{rank}(A) \geq 2$ and $\hat{y} \neq 0$, there exists an index $j \in \{1, \ldots, m\}$ such that $\hat{y} \notin \operatorname{span}(Aa_j)$. In view of the fact $Ax_k \chi_{\{i_1 = \ldots = i_{k^*-1} = j\}} \in \operatorname{span}(Aa_j)$, for $k \in \{1, ..., k^*\}$, there exists an $\eta > 0$ with

$$\mathbb{P}\left(\|Ax_k - \hat{y}\| \geq \eta,\ \forall k \leq k^*\right) \geq \mathbb{P}\left(i_1 = ... = i_{k^*-1} = j\right) > 0.$$

Meanwhile for $k > k^*$, similiar to (3.10) below, there holds

$$\begin{aligned}\|Ax_k - \hat{y}\| &\geq \|Ax_{k-1} - \hat{y}\| - \eta_{k-1} |(Ax_{k-1} - \hat{y}, e_{i_{k-1}})| \|AA^t e_{i_{k-1}}\| \\ &\geq ... \geq \|Ax_{k^*} - \hat{y}\| \prod_{i=k^*}^{k-1} (1 - \eta_i \|A\|^2).\end{aligned}$$

Using the elementary inequalities $1 + x \leq e^x$ for all $x \in \mathbb{R}$ and $1 + x \geq e^{x - x^2}$ for all $x \in [-\frac{1}{2}, 0]$ and the estimate (3.10) below, we deduce

$$\begin{aligned}&\|Ax_k^\delta - y^\delta\| \\ \geq& \|Ax_k - \hat{y}\| - \|A(x_k - x_k^\delta) - (\hat{y} - y^\delta)\| \\ \geq& \|Ax_{k^*} - \hat{y}\| \prod_{i=k^*}^{k-1} (1 - \|A\|^2 \eta_i) - \delta \prod_{i=1}^{k-1} (1 + \|A\|^2 \eta_i) \\ \geq& \|Ax_{k^*} - \hat{y}\| \exp\Big(- c_0 \|A\|^2 \sum_{i=k^*}^{k-1} i^{-\alpha} - c_0^2 \|A\|^4 \sum_{i=k^*}^{k-1} i^{-2\alpha} \Big) - \delta \exp\Big(c_0 \|A\|^2 \sum_{i=1}^{k-1} i^{-\alpha} \Big) \\ \geq& c' \|Ax_{k^*} - \hat{y}\| - c'' \delta,\end{aligned}$$

with

$$c' := e^{-c_0 \|A\|^2 \sum_{i=1}^\infty i^{-\alpha} - c_0^2 \|A\|^4 \sum_{i=1}^\infty i^{-2\alpha}} > 0 \quad \text{and} \quad c'' := e^{c_0 \|A\|^2 \sum_{i=1}^\infty i^{-\alpha}} < \infty.$$

So for small enough $\delta > 0$, there holds

$$\|Ax_k^\delta - y^\delta\| \chi_{\{\|Ax_i - \hat{y}\| \geq \eta,\ \forall i \leq k^*\}} \geq c' \eta - c'' \delta > \tau \delta.$$

Consequently,

$$\liminf_{\delta > 0} \mathbb{P}(k(\delta) = \infty) \geq \mathbb{P}\left(\|Ax_i - \hat{y}\| \geq \eta,\ \forall i \leq k^*\right) > 0.$$

This shows the assertion (3.8).

3.2.2 The proof of Theorem 3.1.4

In this section, we prove Theorem 3.1.4. It employs the following proposition, which states that potential early stopping actually does not cause any problem.

Proposition 3.2.6. *For all $\varepsilon > 0$, there is a sequence $(k_\delta^-)_\delta$ with $k_\delta^- \to \infty$ for $\delta \to 0^+$, such that*

$$\|x^\delta_{k(\delta)} - \hat{x}\|\chi_{\{k(\delta)\leq k_\delta^-\}} \leq \varepsilon$$

for $\delta > 0$ small enough.

Proof. It suffices to show that for all $K \in \mathbb{N}$

$$\|x_{k(\delta)} - \hat{x}\|\chi_{\{k(\delta)\leq K\}} \to 0 \quad \text{as } \delta \to 0^+. \tag{3.9}$$

In order to show this, we need the following two estimates for the iterated noise:

$$\|A(x_k^\delta - x_k) - (y^\delta - \hat{y})\| \leq \delta \prod_{j=1}^{k-1}(1 + \eta_j\|A\|^2), \tag{3.10}$$

$$\|x_k^\delta - x_k\| \leq \delta\|A\| \sum_{j=1}^{k-1} \eta_j \prod_{i=1}^{j-1}(1 + \eta_i\|A\|^2), \tag{3.11}$$

with the conventions $\sum_{j=1}^{0} = 0$ and $\prod_{j=1}^{0} = 1$. We prove the estimates (3.10) and (3.11) by mathematical induction. Note that $a_i = A^t e_i$. For the estimate (3.10), by the triangle inequality and the defining relation (3.2) of SGD iteration,

$$\begin{aligned}
&\|A(x_{k+1}^\delta - x_{k+1}) - (y^\delta - \hat{y})\| \\
&\leq \|A(x_k^\delta - x_k) - (y^\delta - \hat{y})\| + \eta_k\| \left(A(x_k^\delta - x_k) - (y^\delta - \hat{y}), e_{i_k}\right) AA^t e_{i_k}\| \\
&\leq \|A(x_k^\delta - x_k) - (y^\delta - \hat{y})\| \left(1 + \eta_k\|A\|^2\right),
\end{aligned}$$

and since $x_1 = x_1^\delta$, $\|A(x_1^\delta - x_1) - (y^\delta - \hat{y})\| = \|y^\delta - \hat{y}\| \leq \delta$. For the estimate (3.11), we have $\|x_1^\delta - x_1\| = 0$ and

$$\|x_{k+1}^\delta - x_{k+1}\| \leq \|x_k^\delta - x_k\| + \eta_k\|A\|\|A(x_k^\delta - x_k) - (y^\delta - \hat{y})\|,$$

so the claim follows using the estimate (3.10). Now, for each fixed K, since there are only finitely many different realisations of the first K SGD iterates, there is a (deterministic) $\eta > 0$, which depends on K, such that

$$\min_{k=1,\ldots,K} \left(\|Ax_k - \hat{y}\| - \eta\right) \chi_{\{\|Ax_k - \hat{y}\|>0\}} \geq 0, \tag{3.12}$$

where without loss of generality, we have assumed $\hat{y} \neq 0$. Therefore, using estimates (3.10) and (3.12),

$$\begin{aligned}&\|Ax_k^\delta - y^\delta\|\chi_{\{\|Ax_k-\hat{y}\|>0\}}\\ &\geq \|Ax_k - \hat{y}\|\chi_{\{\|Ax_k-\hat{y}\|>0\}} - \|A(x_k - x_k^\delta) - (\hat{y} - y^\delta)\|\chi_{\{\|Ax_k-\hat{y}\|>0\}}\\ &\geq \Big(\eta - \delta\prod_{j=1}^{k-1}(1+\eta_j\|A\|^2)\Big)\chi_{\|Ax_k-\hat{y}\|>0\}} > \tau\delta\chi_{\{\|Ax_k-\hat{y}\|>0\}},\end{aligned}$$

for any $\delta < \frac{\eta}{\tau+\prod_{j=1}^{K-1}(1+\eta_j\|A\|^2)}$. Then by the definition of the discrepancy principle in (3.3), this implies

$$\{k(\delta) \leq K\} \subset \{\|Ax_{k(\delta)} - \hat{y}\| = 0\}$$

for $\delta > 0$ small enough. Meanwhile, since by construction $x_{k(\delta)} \in \mathcal{R}(A^t) = \mathcal{N}(A)^\perp$, $\|Ax_{k(\delta)} - \hat{y}\| = 0$ implies $x_{k(\delta)} = \hat{x}$, the minimum norm solution. The proof of (3.9) is concluded by

$$\begin{aligned}\|x_{k(\delta)}^\delta - \hat{x}\|\chi_{\{k(\delta)\leq K\}} &= \|x_{k(\delta)}^\delta - x_{k(\delta)}\|\chi_{\{k(\delta)\leq K\}}\\ &\leq \delta\|A\|\sum_{j=1}^{K-1}\eta_j\prod_{i=1}^{j-1}(1+\eta_i\|A\|^2) \to 0\end{aligned}$$

for $\delta \to 0^+$, where we have used estimate (3.11). This completes the proof of the proposition. □

Now we can state the proof of Theorem 3.1.4.

Proof of Theorem 3.1.4 Fix $\varepsilon > 0$. Proposition 3.2.6 and Theorem 3.1.2 guarantee the existence of two sequences $(k_\delta^-)_\delta, (k_\delta^+)_\delta$, with $k_\delta^- \leq k_\delta^+ \leq c\delta^{-\frac{2}{(1-\alpha)(\min(\nu,r)+1)}}$, $k_\delta^- \to \infty$ for $\delta \to 0^+$ and

$$\|x_{k(\delta)}^\delta - \hat{x}\|\chi_{\{k(\delta)\leq k_\delta^-\}} \leq \varepsilon \quad \text{for } \delta \text{ small enough}$$

and

$$\mathbb{P}\left(k(\delta) \leq k_\delta^+\right) \to 1 \quad \text{for } \delta \to 0^+.$$

Consequently, for $\delta > 0$ small enough, there holds

$$\begin{aligned}&\mathbb{P}(\|x_{k(\delta)}^\delta - \hat{x}\| > \varepsilon)\\ &= \mathbb{P}(\|x_{k(\delta)}^\delta - \hat{x}\| > \varepsilon, k(\delta) \leq k_\delta^-) + \mathbb{P}(\|x_{k(\delta)}^\delta - \hat{x}\| > \varepsilon, k(\delta) > k_\delta^-)\\ &= \mathbb{P}(\|x_{k(\delta)} - \hat{x}\| > \varepsilon, k(\delta) > k_\delta^-)\\ &= \mathbb{P}(\|x_{k(\delta)} - \hat{x}\| > \varepsilon, k_\delta^- < k(\delta) \leq k_\delta^+) + \mathbb{P}(\|x_{k(\delta)} - \hat{x}\| > \varepsilon, k(\delta) > k_\delta^+)\\ &\leq \mathbb{P}(\|x_{k(\delta)} - \hat{x}\| > \varepsilon, k_\delta^- < k(\delta) \leq k_\delta^+) + \mathbb{P}(k(\delta) > k_\delta^+).\end{aligned}$$

In view of Theorem 3.1.2, it remains to show that

$$\mathbb{P}(\|x_{k(\delta)} - \hat{x}\| > \varepsilon, k_\delta^- < k(\delta) \leq k_\delta^+) \to 0 \qquad \text{for } \delta \to 0^+.$$

To this end, let $\Omega_\delta := \{k_\delta^- \le k(\delta) \le k_\delta^+\}$ and we split the error into three parts in a customary way: approximation error, data propagation error and stochastic error. Specifically, by the triangle inequality, there are constants c_1 and c_2 such that

$$\begin{aligned}
&\|x_{k(\delta)}^\delta - \hat{x}\|\chi_{\Omega_\delta}\\
=&\sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \hat{x}\|\chi_{\{k(\delta)=k\}}\\
\le&\sum_{k=k_\delta^-}^{k_\delta^+} \left(\|\mathbb{E}[x_k] - \hat{x}\| + \|\mathbb{E}[x_k] - \mathbb{E}[x_k^\delta]\| + \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\right)\chi_{\{k(\delta)=k\}}\\
\le&\sum_{k=k_\delta^-}^{k_\delta^+} \left(c_1(k-1)^{-(1-\alpha)\frac{\nu}{2}} + c_2\delta(k-1)^{\frac{1-\alpha}{2}} + \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\right)\chi_{\{k(\delta)=k\}}\\
\le& c_1\left(k_\delta^- - 1\right)^{-(1-\alpha)\frac{\nu}{2}} + c_2\delta\left(k_\delta^+ - 1\right)^{\frac{1-\alpha}{2}} + \sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}},
\end{aligned}$$

where we have used [JL19, Theorem 3.2] and Lemma 3.2.8 below in the third line. The first two terms clearly tend to 0 for $\delta \to 0^+$ (since $k_\delta^- \to \infty$, and $\delta(k_\delta^+)^{\frac{1-\alpha}{2}} \to 0$, in view of Theorem 3.1.2). By Markov's inequality [Fel68, p. 242] and the independence assumption between $k(\delta)$ and $x_{k(\delta)}^\delta$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}} > \varepsilon'\right) &\le \frac{\sum_{k=k_\delta^-}^{k_\delta^+} \mathbb{E}\left[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}}\right]}{\varepsilon'}\\
&= \frac{\sum_{k=k_\delta^-}^{k_\delta^+} \mathbb{E}\left[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|\right]\mathbb{P}\left(k(\delta)=k\right)}{\varepsilon'}.
\end{aligned}$$

Now Jensen's inequality and Proposition 3.2.12 below (with $s = 0$, $\gamma < \min(\alpha, 1-\alpha)$ and $\beta < 1-\alpha$) give

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=k_\delta^-}^{k_\delta^+} \|x_k^\delta - \mathbb{E}[x_k^\delta]\|\chi_{\{k(\delta)=k\}} > \varepsilon'\right) &\le \frac{\sum_{k=k_\delta^-}^{k_\delta^+} \sqrt{\mathbb{E}\left[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2\right]}\mathbb{P}\left(k(\delta)=k\right)}{\varepsilon'}\\
&\le \frac{\sqrt{c((k_\delta^-)^{-\beta} + \delta^2(k_\delta^-)^{-\gamma})}\sum_{k=k_\delta^-}^{k_\delta^+} \mathbb{P}\left(k(\delta)=k\right)}{\varepsilon'}\\
&= \frac{\sqrt{c((k_\delta^-)^{-\beta} + \delta(k_\delta^-)^{-\gamma})}\mathbb{P}\left(\Omega_\delta\right)}{\varepsilon'} \to 0
\end{aligned}$$

as $\delta \to 0^+$. Thus it follows that

$$\mathbb{P}\left(\|x_k(\delta) - \hat{x}\| > \varepsilon, k_\delta^- < k(\delta) \le k_\delta^+\right) \to 0$$

as $\delta \to 0^+$. This completes the proof of the theorem. □

Remark 3.2.7. Clearly, with $k_\omega(\delta)$ given as in Remark 3.2.4, there holds

$$\mathbb{P}\left(\|x_{k_\omega(\delta)} - \hat{x}\| \geq \varepsilon\right) \to 0$$

for $\delta \to 0^+$. That is, the convergence remains valid for the variant of the discrepancy principle (3.3) evaluated with a frequency.

3.2.3 The proofs of Propositions 3.2.1 and 3.2.2

In this part, we prove Propositions 3.2.1 and 3.2.2, which are used in the proof of the Theorems 3.1.2 and 3.1.4. We shall use the following result from [JL19, Theorem 3.1] frequently. Note that $\|B^{\frac{1}{2}}(x_k - \hat{x})\| = \|Ax_k - \hat{y}\|/\sqrt{m}$.

Lemma 3.2.8. *Let Assumption 3.1.1 be fulfilled, then for $s \in \{0, \frac{1}{2}\}$ and $c_{\nu,s} := \left(\frac{(\frac{\nu}{2}+s)(1-\alpha)}{c_0 e(2^{1-\alpha}-1)}\right)^{\frac{\nu}{2}+s} \|w\|$, there holds*

$$\|B^s(\mathbb{E}[x_{k+1}] - \hat{x})\| \leq c_{\nu,s} k^{-(\frac{\nu}{2}+s)(1-\alpha)}.$$

3.2.3.1 The proof of Proposition 3.2.1

Proof. We may assume $k^* > 2$. By the definition of $k^*(\delta)$ and the triangle inequality

$$\begin{aligned}\tau^*\delta &\leq \|A\mathbb{E}[x^\delta_{k^*-1}] - y^\delta\| \\ &\leq \|A\mathbb{E}[x_{k^*-1}] - \hat{y}\| + \|A\mathbb{E}[x^\delta_{k^*-1} - x_{k^*-1}] + (\hat{y} - y^\delta)\|.\end{aligned}$$

By Lemma 3.2.8, the term $\|A\mathbb{E}[x_{k^*-1}] - \hat{y}\|$ is bounded by

$$\|A\mathbb{E}[x_{k^*-1}] - \hat{y}\| \leq c_\nu (k^* - 2)^{-(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)}, \quad \text{with } c_\nu = \sqrt{m} c_{\nu,\frac{1}{2}}. \tag{3.13}$$

Next we claim

$$\|A\mathbb{E}[x^\delta_{k^*-1} - x_{k^*-1}] + (\hat{y} - y^\delta)\| \leq \delta. \tag{3.14}$$

Combining (3.13) with (3.14) immediately implies the desired assertion. It remains to show the claim (3.14). To this end, we employ the filter of the Landweber method. The relation (3.5) implies that $\mathbb{E}[x^\delta_k]$ satisfies the following recursion

$$A\mathbb{E}[x^\delta_{k+1}] - y^\delta = \left(I - \frac{\eta_k}{m} AA^t\right)\left(A\mathbb{E}[x^\delta_k] - y^\delta\right).$$

Using this yields

$$A\mathbb{E}[x^\delta_k] - y^\delta = \prod_{j=1}^{k-1}\left(I - \frac{\eta_j}{m} AA^t\right)\left(Ax_1 - y^\delta\right), \tag{3.15}$$

and consequently, by the choice of c_0,

$$\|A\mathbb{E}[x_k^\delta - x_k] + (\hat{y} - y^\delta)\| = \left\| \prod_{j=1}^{k-1} \left(I - \frac{\eta_j}{m} AA^t\right) (\hat{y} - y^\delta) \right\| \leq \delta. \tag{3.16}$$

This completes the proof of the proposition. □

3.2.3.2 Proof of Proposition 3.2.2

The proof of Proposition 3.2.2 employs several technical estimates [JL19].

Lemma 3.2.9. *For any $j < k$, and any symmetric and positive semidefinite operator S and stepsizes $\eta_j \in (0, \|S\|^{-1}]$ and $p \geq 0$, there holds*

$$\| \prod_{i=j}^{k} (I - \eta_i S) S^p \| \leq \frac{p^p}{e^p (\sum_{i=j}^k \eta_i)^p}.$$

Next we recall two useful estimates taken from [JL19].

Lemma 3.2.10. *For $\eta_j = \eta_0 j^{-\alpha}$ with $\alpha \in (0,1)$, $\beta \in [0,1]$ and $r \geq 0$, there hold*

$$\sum_{j=1}^{[\frac{k}{2}]} \frac{\eta_j^2}{(\sum_{\ell=j+1}^k \eta_\ell)^r} j^{-\beta} \leq c_{\alpha,\beta,r} k^{-r(1-\alpha)+\max(0,1-2\alpha-\beta)},$$

$$\sum_{j=[\frac{k}{2}]+1}^{k-1} \frac{\eta_j^2}{(\sum_{\ell=j+1}^k \eta_\ell)^r} j^{-\beta} \leq c'_{\alpha,\beta,r} k^{-((2-r)\alpha+\beta)+\max(0,1-r)},$$

where we slightly abuse the notation $k^{-\max(0,0)}$ for $\ln k$, and $c_{\alpha,\beta,r}$ and $c'_{\alpha,\beta,r}$ are given by

$$c_{\alpha,\beta,r} = 2^r \eta_0^{2-r} \begin{cases} \frac{2\alpha+\beta}{2\alpha+\beta-1}, & 2\alpha+\beta > 1, \\ 2, & 2\alpha+\beta = 1, \\ \frac{2^{2\alpha+\beta-1}}{1-2\alpha-\beta}, & 2\alpha+\beta < 1, \end{cases} \quad \text{and} \quad c'_{\alpha,\beta,r} = 2^{2\alpha+\beta} \eta_0^{2-r} \begin{cases} \frac{r}{r-1}, & r > 1, \\ 2, & r = 1, \\ \frac{2^{r-1}}{1-r}, & r < 1. \end{cases}$$

The next result gives an important recursion between the variance estimate.

Lemma 3.2.11. *Let Assumption 3.1.1 be fulfilled. Then for the SGD iterate x_k^δ, with $\phi_j^s = \|B^{\frac{1}{2}+s}\Pi_{j+1}^k(B)\|$, there holds*

$$\begin{aligned} &\mathbb{E}[\|B^s(x_{k+1}^\delta - \mathbb{E}[x_{k+1}^\delta])\|^2] \\ \leq &\sum_{j=1}^k \eta_j^2 (\phi_j^s)^2 \left(c_s \mathbb{E}[\|B^s \left(x_j^\delta - \mathbb{E}[x_j^\delta]\right) \|^2] + 2c_\nu j^{-2(1-\alpha)(\frac{\nu}{2}+\frac{1}{2})} + 2\delta^2\right), \end{aligned}$$

with $s \in \{0, \frac{1}{2}\}$ and c_s, c_ν given below.

Proof. By [JL19, Theorem 3.3] and the bias variance decomposition, the left hand side (LHS) is bounded by

$$\begin{aligned}\text{LHS} \leq& \sum_{j=1}^{k} \eta_j^2 (\phi_j^s)^2 \mathbb{E}[\|Ax_j^\delta - y^\delta\|^2 \\ =& \sum_{j=1}^{k} \eta_j^2 (\phi_j^s)^2 \left(\mathbb{E}[\|A\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2] + \|A\mathbb{E}[x_j^\delta] - y^\delta\|^2\right).\end{aligned}$$

Now by the triangle inequality and (3.16),

$$\begin{aligned}\text{LHS} \leq& \sum_{j=1}^{k} \eta_j^2 (\phi_j^s)^2 \Big(\mathbb{E}[\|A\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2] \\ &+ \left(\|A\mathbb{E}[x_j] - \hat{y}\| + \|A\left(\mathbb{E}[x_j^\delta] - \mathbb{E}[x_j]\right) - \left(y^\delta - \hat{y}\right)\|\right)^2\Big)\end{aligned}$$

Since $\|A\mathbb{E}[x_1] - \hat{y}\| = \|\hat{y}\|$, and

$$\|A\mathbb{E}[x_j] - \hat{y}\| \leq \sqrt{m}c_{\nu,\frac{1}{2}}(j-1)^{-(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)} \leq \sqrt{m}c_{\nu,\frac{1}{2}} 2^{(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)} j^{-(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)}$$

for $j \geq 2$ by Lemma 3.2.8. Thus, with $c_\nu := \left(\max\{\|\hat{y}\|, \sqrt{m}c_{\nu,\frac{1}{2}} 2^{(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)}\}\right)^2$,

$$\text{LHS} \leq \sum_{j=1}^{k} \eta_j^2 (\phi_j^s)^2 \left(n^{2s}\|A\|^{4(\frac{1}{2}-s)}\mathbb{E}[\|B^s\left(x_j^\delta - \mathbb{E}[x_j^\delta]\right)\|^2] + 2c_\nu j^{-2(1-\alpha)(\frac{\nu}{2}+\frac{1}{2})} + 2\delta^2\right)$$

which completes the proof of the lemma with $c_s = m^{2s}\|A\|^{4(\frac{1}{2}-s)}$. □

The next result gives a sharp estimate on $\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$.

Proposition 3.2.12. *Let Assumption 3.1.1 be fulfilled. Then for the SGD iterate* x_k^δ, *the mean squared error* $\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ *with* $s \in \{0, \frac{1}{2}\}$ *satisfies*

$$\mathbb{E}[\|B^s(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2] \leq c(\alpha, \nu, m, s, \beta, \gamma)(k^{-\beta} + \delta^2 k^{-\gamma})$$

for $\beta < \min\left((1+2s)(1-\alpha), (1+\nu)(1-\alpha)+\alpha\right)$ *and* $\gamma < \min(\alpha, 1-\alpha)$.

Proof. Lemma 3.2.11 implies that the weighted mean squares error $d_j^s = \mathbb{E}[\|B^s(x_k^\delta - \hat{x})\|^2]$ satisfies the following recursion

$$d_{k+1}^s \leq \sum_{j=1}^{k} \eta_j^2 (\phi_j^s)^2 \left(c_s d_j^s + 2c_\nu j^{-2(1-\alpha)(\frac{\nu}{2}+s)} + 2\delta^2\right) \tag{3.17}$$

Now we prove the desired assertion by mathematical induction (with $\beta = (\nu+1)(1-\alpha)$):

$$d_k^s \leq c(k^{-\beta} + \delta^2 k^{-\gamma}),$$

where the constant $c \geq 1$ is to be determined. This assertion holds trivially for all finite k, up to k^*, provided that c is sufficiently large. Now suppose the assertion holds for $k \geq k^*$, and we prove the assertion for $k+1$. Indeed, it follows from the recursion (3.17), the induction hypothesis and since $\beta < 2(1-\alpha)(\frac{\nu}{2}+\frac{1}{2})$, that

$$\begin{aligned} d_{k+1}^s &\leq \sum_{j=1}^k \eta_j^2(\phi_j^s)^2(c_s c(j^{-\beta}+j^{-\gamma}\delta^2) + 2c_\nu j^{2(1-\alpha)(\frac{\nu}{2}+s)} + 2\delta^2) \\ &\leq c_s c\sum_{j=1}^k \eta_j^2(\phi_j^s)^2 j^{-\beta} + (c_s c+2)\delta^2\sum_{j=1}^k \eta_j^2(\phi_j^s)^2 + 2c_\nu\sum_{j=1}^k \eta_j^2(\phi_j^s)^2 j^{-2(1-\alpha)(\frac{\nu}{2}+\frac{1}{2})} \\ &\leq (c_s c+2c_\nu)\sum_{j=1}^k \eta_j^2(\phi_j^s)^2 j^{-\beta'} + (c_s c+2)\delta^2\sum_{j=1}^k \eta_j^2(\phi_j^s)^2. \end{aligned}$$

with $\beta' = \min(\beta,(1+\nu)(1-\alpha))$. Without loss of generality, we may assume that $\beta' \geq 1-2\alpha$. By Lemmas 3.2.9 and 3.2.10, the first sum is bounded by

$$\begin{aligned} \sum_{j=1}^k \eta_j^2(\phi_j^s)^2 j^{-\beta'} \leq & e^{-2}c_{\alpha,\beta',1+2s}k^{-(1+2s)(1-\alpha)+\max(0,1-2\alpha-\beta')} \\ & + e^{-1}c'_{\alpha,\beta',1}\|B\|k^{-(\alpha+\beta')}\ln k + c_0^2\|B\|^2k^{-(2\alpha+\beta')}. \end{aligned} \tag{3.18}$$

Since $\beta'+\alpha > \beta$ and $\max(0,1-2\alpha-\beta') = 0$, thus,

$$\begin{aligned} &\sum_{j=1}^k \eta_j^2\phi_j^2 j^{-\beta'} \\ \leq&(e^{-2}c_{\alpha,\beta',1+2s}k^{-(1+2s)(1-\alpha)+\beta}\ln k + e^{-1}c'_{\alpha,\beta',1}\|B\|k^{-(\alpha+\beta')+\beta}\ln k + c_0^2\|B\|^2k^{-\alpha})k^{-\beta}. \end{aligned}$$

Meanwhile, with $-(1+2s)(1-\alpha)+\max(0,1-2\alpha) = -\min((1+2s)(1-\alpha),\alpha+2s(1-\alpha))$, we obtain

$$\begin{aligned} &\sum_{j=1}^k \eta_j^2(\phi_j)^2 \\ \leq&e^{-2}c_{\alpha,0,2}k^{-\min((1+2s)(1-\alpha),\alpha+2s(1-\alpha))} + e^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha}\ln k + c_0^2\|B\|^2k^{-2\alpha} \\ \leq&(e^{-2}c_{\alpha,0,1+2s}k^{-\min((1-\alpha),\alpha)+\gamma} + e^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha+\gamma}\ln k + c_0^2\|B\|^2k^{-2\alpha+\gamma})k^{-\gamma} \end{aligned}$$

Combining the preceding estimates yields

$$\begin{aligned} &d_{k+1} \\ \leq&(cc_s+2c_\nu)\Big(e^{-2}c_{\alpha,\beta',1+2s}k^{-(1+2s)(1-\alpha)+\beta}\ln k + e^{-1}c'_{\alpha,\beta',1}\|B\|k^{-(\alpha+\beta')+\beta}\ln k \\ &\quad +c_0^2\|B\|^2k^{-\alpha}\Big)k^{-\beta} + (c_s c+2)\delta^2\left(e^{-2}c_{\alpha,0,1+2s}k^{-\min((1-\alpha),\alpha)+\gamma}\right. \\ &\quad \left.+e^{-1}c'_{\alpha,0,1}\|B\|k^{-\alpha+\gamma}\ln k + c_0^2\|B\|^2k^{-2\alpha+\gamma}\right)k^{-\gamma}. \end{aligned}$$

Since by assumption, $\beta < (1+2s)(1-\alpha)$, $\beta < \alpha + \beta'$ and $\gamma < \min(\alpha, 1-\alpha)$, there exists k^* such that for all $k \geq k^*$

$$\begin{aligned}
\frac{1}{4} &> (c_s + 2c_\nu)\Big(e^{-2} c_{\alpha,\beta',1+2s} k^{-(1+2s)(1-\alpha)+\beta} \ln k + e^{-1} c'_{\alpha,\beta',1} \|B\| k^{-(\alpha+\beta')+\beta} \ln k \\
&\qquad + c_0^2 \|B\|^2 k^{-2\alpha}\Big)\,, \\
\frac{1}{4} &> (c_s + 2)\delta^2 \left(e^{-2} c_{\alpha,0,1+2s} k^{-\min((1-\alpha),\alpha)+\gamma} + e^{-1} c'_{\alpha,0,1} \|B\| k^{-\alpha+\gamma} \ln k \right. \\
&\qquad \left. + c_0^2 \|B\|^2 k^{-2\alpha+\gamma}\right).
\end{aligned}$$

Thus, with this choice of k^* and $k \geq k^*$,

$$\begin{aligned}
d_{k+1} &\leq \tfrac{c}{4}\left(k^{-\beta} + \delta^2 k^{-\gamma}\right) \leq c\tfrac{(1+k^{-1})^\beta}{4}\left((k+1)^{-\beta} + \delta^2 (k+1)^{-\gamma}\right) \\
&< c\left((k+1)^{-\beta} + \delta^2 (k+1)^{-\gamma}\right)
\end{aligned}$$

and we obtain the desired assertion. □

Remark 3.2.13. The m factor in the estimate is due to the variance inflation of using stochastic gradients in place of gradient in SGD. This factor can be reduced by suitable variance reduction techniques, e.g., mini-batching and stochastic variance reduced gradient [JZ13]. Note that with [JL19, Theorems 3.1 and 3.2] and $s = 0$, Proposition 3.2.12 gives an improved (regarding the exponents) a priori bound for the mean squared error $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$.

Last, using Lemma 3.2.11 and Proposition 3.2.12, we can prove Proposition 3.2.2.

Proof of Proposition 3.2.2 Using Lemma 3.2.11 and Proposition 3.2.12 with $s = \frac{1}{2}$ and $c = c(\alpha, \nu, m, s, \beta, \gamma)$, we deduce

$$\mathbb{E}[\|A(x_{\kappa(\delta)}^\delta - \mathbb{E}[x_{\kappa(\delta)}^\delta])\|^2] \leq mc\left(\kappa(\delta)^{-\beta} + \delta^2 \kappa(\delta)^{-\gamma}\right).$$

We choose $\gamma > 0$. If $\nu < 1$ and $r > 2\nu$, then we can choose $\beta > (1-\alpha)(\nu+1)$, so with the choice $\kappa(\delta) = \delta^{-\frac{2}{(1-\alpha)(\nu+1)}}$, the claim follows. Otherwise, if $\nu \geq 1$, then we can choose $\beta > (1-\alpha)(r+1)$, so with the choice $\kappa(\delta) = \delta^{-\frac{2}{(1-\alpha)(r+1)}}$ the claim again follows. This completes the proof of the proposition. □

3.3 Stochastic error

In contrast to Chapter 1 and 2, we here assumed that we know and upper bound $\delta \geq \|\hat{y} - y^\delta\|$ of the data error. We now discuss the case, when we have multiple unbiased measurements of the data $\hat{y}$.

3.3.1 The case with finite variance

Assume that we have unbiased i.i.d. measurements $Y_1, ..., Y_n$ of $\hat{y}$ with finite variance (i.e $\mathbb{E}\|Y_1 - \hat{y}\|^2$ is of smaller order than the discretisation dimension m). We take the mean $\bar{Y}_n$ as our approximation of $\hat{y}$ with estimated data error $\delta_n^{est} = s_n^2/\sqrt{n}$. We are facing now exactly the same problem as in Chapter 1, i.e. it will occasionaly hold that $\delta_n^{est} < \|\bar{Y}_n - \hat{y}\| := \delta_n^{true}$. For simplicity we restrict to the case $s_n^2 = 1$, the case $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n \|Y_i - \bar{Y}_n\|^2$ can be treated almost the same way, see Chapter 1. We show in the following, that this will in essence not change the results. We assume that the measurements are independent of the random sampling of the row index. We denote by $\mathbb{E}_{sgd}$ the expectation with respect to the random sampling of the row index, and with $\mathbb{E}$ the total expectation. We denote by X_k^n the SGD iterates for noisy random data $\bar{Y}_n$ and with x_k the ones for exact data $\hat{y}$. First note, that δ_n^{est} and δ_n^{true} are of the same order, i.e. for arbitrary sequences $(c_n)_{n\in\mathbb{N}}, (C_n)_{n\in\mathbb{N}}$ with $c_n \to 0, C_n \to \infty$ it holds that

$$\mathbb{P}\left(c_n\delta_n^{est} \leq \delta_n^{true} \leq C_n\delta_n^{est}\right) \to 1$$

as $n \to \infty$. Thus the $\delta = \delta_n^{true}$ in Proposition 3.2.2 may be replaced with δ_n^{est} and in order to reproduce Theorem 3.1.2 and 3.1.4 it is sufficient to just rework Proposition 3.2.1. However, $\mathbb{E}_{sgd}[X_k^n]$ are the iterates of the filter-based Landweber method (applied to noisy random data $\bar{Y}_n$) and hence can be treated as in Chapter 1.

Proposition 3.3.1. *Let Assumption 3.1.1 be fulfilled. Then, for*

$$k^*(n) := \min\left\{k \in \mathbb{N} \ : \ \|A\mathbb{E}_{sgd}[X_k^n] - \bar{Y}_n\| \leq \tau^*\delta_n^{est}\right\}$$

it holds that

$$\mathbb{P}\left(k^*(n) \leq \bar{k}(n) := \left(\frac{\tau^* - 1}{\sqrt{m}c_\nu}\delta_n^{est}\right)^{-\frac{2}{(1-\alpha)(\nu+1)}} + 1\right) \geq 1 - f(n) \to 1,$$

as $n \to \infty$, with $c_\nu = \left(\frac{(\frac{\nu}{2}+\frac{1}{2})(1-\alpha)}{c_0 e(2^{1-\alpha}-1)}\right)^{\frac{\nu}{2}+\frac{1}{2}}\|w\|$ and

$$f(n) := \sum_{l=1}^r \prod_{j=1}^{\bar{k}(n)-1}\left(1 - \frac{\eta_j}{m}\sigma_l^2\right)^2 \mathbb{E}\left(Y_1 - \hat{y}, u_l\right)^2,$$

with $(\sigma_l, u_l, v_l)_{l=1}^r$ the singular value decomposition of A.

Proof.

Decomposition of the residual and Lemma 3.2.9 (with $\mathbb{E}_{sgd}$ instead of $\mathbb{E}$) give

$$\begin{aligned}
&\|A\mathbb{E}_{sgd}[X^n_{\bar{k}(n)}] - \bar{Y}_n\| \\
\leq &\|A\mathbb{E}_{sgd}[X_{\bar{k}(n)}] - \hat{y}\| + \|A\mathbb{E}_{sgd}\left[X^n_{\bar{k}(n)} - x_{\bar{k}(n)}\right] - (\bar{Y}_n - \hat{y})\| \\
\leq &\sqrt{m}c_\nu(\bar{k}(n) - 1)^{-(\frac{\nu+1)(1-\alpha)}{2}} + \|A\mathbb{E}_{sgd}\left[X^n_{\bar{k}(n)} - x_{\bar{k}(n)}\right] - (\bar{Y}_n - \hat{y})\| \\
= &(\tau^* - 1)\delta^{est}_n + \|A\mathbb{E}_{sgd}\left[X^n_{\bar{k}(n)} - x_{\bar{k}(n)}\right] - (\bar{Y}_n - \hat{y})\|.
\end{aligned}$$

Therefore,

$$\Big\{\|A\mathbb{E}_{sgd}[X^n_{\bar{k}(n)}] - \bar{Y}_n\| \leq \tau^*\delta^{est}_n\} \supset \{\|A\mathbb{E}_{sgd}\left[X^n_{\bar{k}(n)} - x_{\bar{k}(n)}\right] - (\bar{Y}_n - \hat{y})\| \leq \delta^{est}_n\Big\}$$

and by definition of $k^*(n)$ and Tschebyscheff's inequality we deduce that

$$\begin{aligned}
\mathbb{P}\left(k^*(n) \leq \bar{k}(n)\right) \geq \mathbb{P}&\left(\|A\mathbb{E}_{sgd}[X^n_{\bar{k}(n)}] - \bar{Y}_n\| \leq \tau^*\delta^{est}_n\right) \\
\geq \mathbb{P}&\left(\|A\mathbb{E}_{sgd}\left[X^n_{\bar{k}(n)} - x_{\bar{k}(n)}\right] - (\bar{Y}_n - \hat{y})\| \leq \delta^{est}_n\right) \\
\geq &1 - \frac{\mathbb{E}\|A\mathbb{E}_{sgd}\left[X^n_{\bar{k}(n)} - x_{\bar{k}(n)}\right] - (\bar{Y}_n - \hat{y})\|^2}{\delta^{est2}_n} \\
= &1 - n\mathbb{E}\left[\left\|\sum_{i=1}^{n}\prod_{j=1}^{\bar{k}(n)-1}\left(I - \frac{\eta_j}{m}AA^t\right)(Y_i - \hat{y})\right\|^2\right] \\
= &1 - \sum_{l=1}^{r}\prod_{j=1}^{\bar{k}(n)-1}\left(1 - \frac{\eta_j}{m}\sigma_l^2\right)^2 \mathbb{E}\left(Y_1 - \hat{y}, u_l\right)^2 = 1 - f(n).
\end{aligned}$$

We show that $f(n) \to 0$ as $n \to \infty$. Let $\varepsilon > 0$ and $L \leq r$ such that

$$\sum_{l=L+1}^{r}\mathbb{E}\left(Y_1 - \hat{y}, u_l\right)^2 < \varepsilon/2.$$

Then,

$$\begin{aligned}
f(n) :=& \sum_{l=1}^{r} \prod_{j=1}^{\bar{k}(n)-1} \left(1 - \frac{\eta_j}{m} AA^t\right)^2 \mathbb{E}\left(Y_1 - \hat{y}, u_l\right)^2 \\
\leq& \sum_{l=1}^{L} \prod_{j=1}^{\bar{k}(n)-1} \left(1 - \frac{\eta_j}{m} AA^t\right)^2 \mathbb{E}\left(Y_1 - \hat{y}, u_l\right)^2 + \sum_{l=L+1}^{r} \mathbb{E}\left(Y_1 - \hat{y}, u_l\right) \\
\leq& \, L e^{-\frac{2\sigma_L^2}{m} \sum_{j=1}^{\bar{k}(n)-1} \eta_j} + \varepsilon/2 \leq \varepsilon
\end{aligned}$$

for n large enough (since $\sum_{j=1}^{\infty} \eta_j = \infty$). Note that the proof worked also for $m \to \infty$, given $c_0 \asymp m$. □

3.3.2 The white noise case

We now consider the white noise scenario from Chapter 2. So assume that Y_{ij}, $i \leq n, j \leq m$ are unbiased and i.i.d measurements of $\hat{y}_j, j = 1, ..., m$ (so that $\delta_{ij} := Y_{ij} - \hat{y}_j$ are i.i.d for $j \leq m, i \in \mathbb{N}$). We replace y^δ and δ by the mean and the estimated data error

$$\bar{Y}_n^{(m)} := \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} Y_{i1} \\ \ldots \\ Y_{in} \end{pmatrix} \qquad \delta_{m,n}^{est} = \sqrt{s_{m,n}^2 \frac{m}{n}},$$

where

$$s_{m,n}^2 := \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_{ij} - \frac{1}{n} \sum_{l=1}^{n} Y_{lj}\right)^2 = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=1}^{n} \left(\delta_{ij} - \frac{1}{n} \sum_{l=1}^{n} \delta_{lj}\right)^2$$

is the mean of the sample variances. Note that the true error

$$\delta_{m,n}^{true} := \|\bar{Y}_n^m - \hat{y}\| = \sqrt{\sum_{j=1}^{m} \left(\sum_{i=1}^{n} Y_{ij}/n - \hat{y}_j\right)^2} = \sqrt{\sum_{j=1}^{m} \left(\sum_{i=1}^{n} \delta_{ij}/n\right)^2}$$

and $\delta_{m,n}^{est}$ in fact only depend on the dimension m of the inverse problem and are otherwise independent of $A, \hat{x}$ and $\hat{y}$. The following lemma states, that for a fixed error distribution all the results remain true with high probability (for sufficiently large dimension m and number of measurements n), if we replace y^δ and δ with $\bar{Y}_n^{(m)}$ and $\delta_{m,n}^{est}$.

Lemma 3.3.2. *Let $\varepsilon > 0$. Then,*

$$\mathbb{P}\left(\left|\frac{\delta^{true}_{m,n} - \delta^{est}_{m,n}}{\delta^{est}_{m,n}}\right| \leq \varepsilon\right) \to 1$$

as $m, n \to \infty$, where the rate depend only on ε and the distribution of δ_{11}.

Proof. This follows directly with Lemma 2.3.4. □

3.4 Numerical experiments and discussions

Now we provide numerical experiments to complement the theoretical analysis. Three model examples, i.e., `phillips` (mildly ill-posed, smooth), `gravity` (severely ill-posed, medium smooth) and `shaw` (severely ill-posed, nonsmooth), are taken from the open source `MATLAB` package Regutools [Han07], available at `http://people.compute.dtu.dk/pcha/Regutools/` (last accessed on April 14, 2020). The problems cover a variety of setting, e.g., different solution smoothness and degree of ill-posedness. These examples are discretizations of Fredholm/Volterra integral equations of the first kind, by means of either the Galerkin approximation with piecewise constant basis functions or quadrature rules. All the examples are discretized into a linear system of size $m = m' = 1000$. In addition, we generate a synthetic example, termed `smoothed-phillips`, whose exact solution $\hat{x}$ is first generated by $\bar{x} = A^t A A^t \bar{y}$ and then normalized to have unit maximum, i.e., $\hat{x} = \bar{x}/\|\bar{x}\|_{\ell^\infty}$, where A is the system matrix and $\bar{y}$ the exact data from `phillips`, and the corresponding exact data is formed by $\hat{y} = A\hat{x}$. By its very construction, the solution $\hat{x}$ satisfies Assumption 3.1.1(ii) with an exponent $\nu > 4$, and thus it is very smooth in some sense. Throughout, the noisy data y^δ is generated according to

$$y_i^\delta := y_i^\dagger + \delta \max_j(|\hat{y}_j|)\xi_i, \quad i = 1, \ldots, n,$$

where the i.i.d. random variables ξ_i follow the standard Gaussian distribution (with zero mean and unit variance), and $\delta > 0$ denotes the relative noise level (by slightly abusing the notation). The parameter c_0 in the stepsize schedule in Assumption 3.1.1(i) is set to $(\max_i \|a_i\|^2)^{-1}$, the exponent α is taken from the set $\{0.1, 0.3, 0.5\}$, and unless otherwise stated, the stopping criterion is tested every 100 SGD iterations (see Remarks 3.2.4 and 3.2.7). SGD is always initialized with $x_1 = 0$, and the maximum number of epochs is fixed at 5000, where one epoch refers to n SGD iterations. The parameter τ in the discrepancy principle (3.3) is fixed at $\tau = 1.2$. All the statistical quantities presented below are computed from 100 independent runs.

3.4.1 Optimality

First, we verify the optimality of the discrepancy principle (3.3), against an order optimal regularization method. There are many possible choices, e.g., Landweber method and conjugate gradient method [EHN96, Chapters 6 and 7]. In this work, we employ the Landweber method as the benchmark. The Landweber method generally converges steadily although often slowly. However, it is known to be an order optimal regularization method with infinite qualification [EHN96, Theorem 6.5, p. 159], when terminated by the discrepancy principle (3.6), and further, it is the population version of SGD (the expected iterates $\big(\mathbb{E}[x_k^\delta]\big)_{k\in\mathbb{N}}$ are exactly the Landweber iterates; see (3.5)), and thus it serves a good benchmark for performance comparison in terms of the convergence rate. For the comparison, the Landweber method is initialized with $x_1 = 0$, with a constant stepsize $1/\|A\|^2$, and it is terminated with the discrepancy principle (3.6) with $\tau^* = 1.2$ (i.e., the same as for SGD) with the maximum number of iterations being fixed at 5000. The numerical results for the examples are summarized in Tables 3.1–3.4. In the tables, e_{sgd} and $\mathrm{std}(e_{sgd})$ denote the (sample) mean and the (sample) standard deviation of the (squared) error $\|x_{k_\delta}^\delta - \hat{x}\|^2$, respectively, i.e.,

$$e_{\rm sgd} = \mathbb{E}[\|x_{k_\delta}^\delta - \hat{x}\|^2] \quad \text{and} \quad std(e_{\rm sgd}) = \mathbb{E}[(\|x_{k^\delta}^\delta - \hat{x}\|^2 - e_{\rm sgd})^2]^{\frac{1}{2}},$$

and $k_{sgd} = \mathbb{E}[k_\delta]$ is the mean stopping index for SGD, in terms of the number of epochs. Likewise $e_{\mathrm{l}m}$ and $k_{\mathrm{l}m}$ denote the squared reconstruction error and stopping index, respectively, of the Landweber method, terminated according to the discrepancy principle (3.6).

Table 3.1: Comparison between SGD and LM for `phillips`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	$\mathrm{std}(e_{sgd})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{sgd})$	k_{sgd}	e_{sgd}	$\mathrm{std}(e_{sgd})$	k_{sgd}	$e_{\mathrm{l}m}$	$k_{\mathrm{l}m}$
1e-3	8.6e-3	4.5e-3	1.42	8.5e-3	4.4e-3	4.18	8.3e-3	4.6e-3	52.2	5.7e-3	361
5e-3	1.7e-2	8.4e-3	0.45	2.3e-2	8.8e-3	0.97	2.4e-2	7.3e-3	6.03	2.2e-2	128
1e-2	2.8e-2	1.6e-2	0.28	4.7e-2	2.0e-2	0.43	5.7e-2	2.0e-2	1.64	5.7e-2	51
5e-2	1.4e-1	9.7e-2	0.15	1.4e-1	9.0e-2	0.11	2.1e-1	9.6e-2	0.17	2.1e-1	15

The numerical results allow drawing a number of interesting observations. First, the exponent α in the stepsize schedule exerts a strong influence on the (expected) stopping index k_{sgd}. At low noise levels (i.e., small δ), k_{sgd} increases dramatically with the value of α. Meanwhile, for any fixed α, the error e_{sgd} increases steadily with the noise level δ, exhibiting the convergence behavior indicated in Theorem 3.1.4. Further, for each fixed δ, the error e_{sgd} is largely comparable for all different α values, although k_{sgd} increases with α. This behavior is qualitatively in good agreement with Theorem 3.1.2: the upper bound scales as $O(\delta^{-\frac{2}{(1-\alpha)(\min(2p,r)+1)}})$. Thus, in practice, in order to obtain relatively efficient SGD, one prefers small α values. Second, in

Table 3.2: Comparison between SGD and LM for `gravity`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{lm}	k_{lm}
1e-3	6.7e-1	2.6e-1	1.96	7.4e-1	2.7e-1	9.31	7.7e-1	2.4e-1	198	7.2e-1	640
5e-3	2.0e0	8.9e-1	0.45	2.5e0	1.1e0	0.88	2.7e0	1.1e0	6.21	2.4e0	95
1e-2	3.1e0	1.5e0	0.25	4.3e0	1.9e0	0.36	4.7e0	2.0e0	1.36	4.0e0	50
5e-2	9.0e0	5.3e0	0.14	1.1e1	6.6e0	0.10	1.5e1	7.4e0	0.13	1.6e1	9

Table 3.3: Comparison between SGD and LM for `shaw`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{lm}	k_{lm}
1e-3	8.2e0	9.3e-2	57.7	8.4e0	5.5e-2	891	2.0e1	5.6e-1	5000	1.2e1	5000
5e-3	2.7e1	1.2e0	0.94	2.8e1	1.1e0	3.81	2.8e1	1.0e0	51.69	2.8e1	189
1e-2	2.9e1	1.6e0	0.59	3.1e1	1.1e0	1.93	3.1e1	1.0e0	19.71	3.1e1	117
5e-2	5.0e1	1.0e1	0.15	6.0e1	8.0e0	0.25	6.7e1	7.4e0	0.818	6.8e1	22

Table 3.4: Comparison between SGD and LM for `smoothed-phillips`.

	$\alpha = 0.1$			$\alpha = 0.3$			$\alpha = 0.5$			LM	
δ	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{sgd}	std(e_{sgd})	k_{sgd}	e_{lm}	k_{lm}
1e-3	1.6e-1	6.8e-2	1.34	1.5e-1	5.8e-2	4.03	1.5e-1	6.0e-2	48	1.5e-3	29
5e-3	3.9e-1	2.0e-1	0.36	5.0e-1	2.0e-1	0.59	4.9e-1	1.9e-1	2.68	1.3e-2	18
1e-2	5.9e-1	2.6e-1	0.24	8.5e-1	3.7e-1	0.30	9.4e-1	3.9e-1	0.77	4.0e-2	15
5e-2	2.9e0	1.4e0	0.16	3.2e0	1.5e0	0.10	4.3e0	2.1e0	0.13	7.1e-1	9

terms of accuracy (measured by the mean squared error), SGD is competitive with the classical Landweber method for `phillips`, `gravity` and `shaw`: e_{sgd} and $e_{\mathrm{l}m}$ are fairly close to each other in most cases, and e_{sgd} can be smaller than $e_{\mathrm{l}m}$, which fully confirms the order-optimality of the discrepancy principle (3.3) for SGD for low regularity solutions, and also confirming the convergence in Theorem 3.1.4. In fact, empirically, the error seems to converge not only in probability, but also in L^2. A close inspection on the stopping index k_{sgd} is very telling: when the noise level δ is medium to large, the stopping index k_{sgd} of SGD, determined by (3.3), is ten-fold smaller than that for the Landweber method in terms of epoch count. In particular, when the noise level δ is relatively high, SGD can actually deliver an accurate solution within less than one epoch, i.e., going through only a fraction of all the available data points. Thus, in this regime, SGD is much more efficient than the Landweber method. These observations are valid for all the examples, despite their dramatic difference in degree of ill-posedness and solution smoothness. However, for `smoothed-phillips`, the achieved accuracy by SGD is far below than that by the

Landweber method for all three exponents α. This suboptimality in convergence rate is attributed to the saturation phenomenon for SGD, due to the dominance of the computational variance, when the true solution $\hat{x}$ is very smooth. The effect of the variance component will be examined more closely below in Section 3.4.2.

The example `shaw` is challenging for numerical recovery, since the solution is far less smooth, and at low noise level δ =1e-3, the discrepancy principle (3.6) cannot be reached even after 5000 Landweber iterations, see Table 3.3. A similar behavior is also observed for SGD with $\alpha = 0.3$ and $\alpha = 0.5$. Nonetheless, with $\alpha = 0.1$, the discrepancy principle (3.3) can be reached by SGD after a few hundred epochs, clearly showing the surprisingly beneficial effect of SGD noise for low-regularity solutions.

Next we examine more closely the performance of individual samples. The boxplots are shown in Fig. 3.1 for the examples at two different scenarios, i.e., fixed α and fixed δ. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively; The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. It is observed that for a fixed α, on average the error $\|x_{k(\delta)}^\delta - \hat{x}\|^2$ increases with the noise level δ samplewise, and also its distribution broadens. However, the required number of iterations to fulfill the discrepancy principle (3.3) decreases dramatically, as the noise level δ increases, concurring with the preceding observation that SGD is especially efficient for data with high noise levels. Meanwhile, with the noise level δ fixed, the value of α does not change the results much overall. However, a larger α can potentially make the percentile box larger and also more outliers, as shown by the results for `gravity` in Fig. 3.1, and thus give less accurate results. This observation is counter-intuitive in that smaller variance does not immediately lead to better accuracy. This might be related to the delicate interplay between the total error and various problem / algorithmic parameters, e.g., α and p. Further, the outliers in the boxplots mostly lie above the box. These observations are typical for all the examples.

3.4.2 How influential is the variance?

Now we examine more closely the dynamics of the SGD iteration via the bias-variance decomposition of the error $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$ and residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$:

$$\mathbb{E}[\|x_k^\delta - \hat{x}\|^2] = \|\mathbb{E}[x_k^\delta] - \hat{x}\|^2 + \mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2],$$
$$\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2] = \|A\mathbb{E}[x_k^\delta] - y^\delta\|^2 + \mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2].$$

In Fig. 3.2, we display the dynamics of mean squared error $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$ and the mean squared residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ together with their variance components for the examples at two different relative noise levels, i.e., δ =5e-3 and δ =5e-2. At each time, SGD is run for 100 epochs (i.e., 1e5 SGD iterations), and the results are recorded every 50 SGD iterations, starting from the 50th SGD iterations.

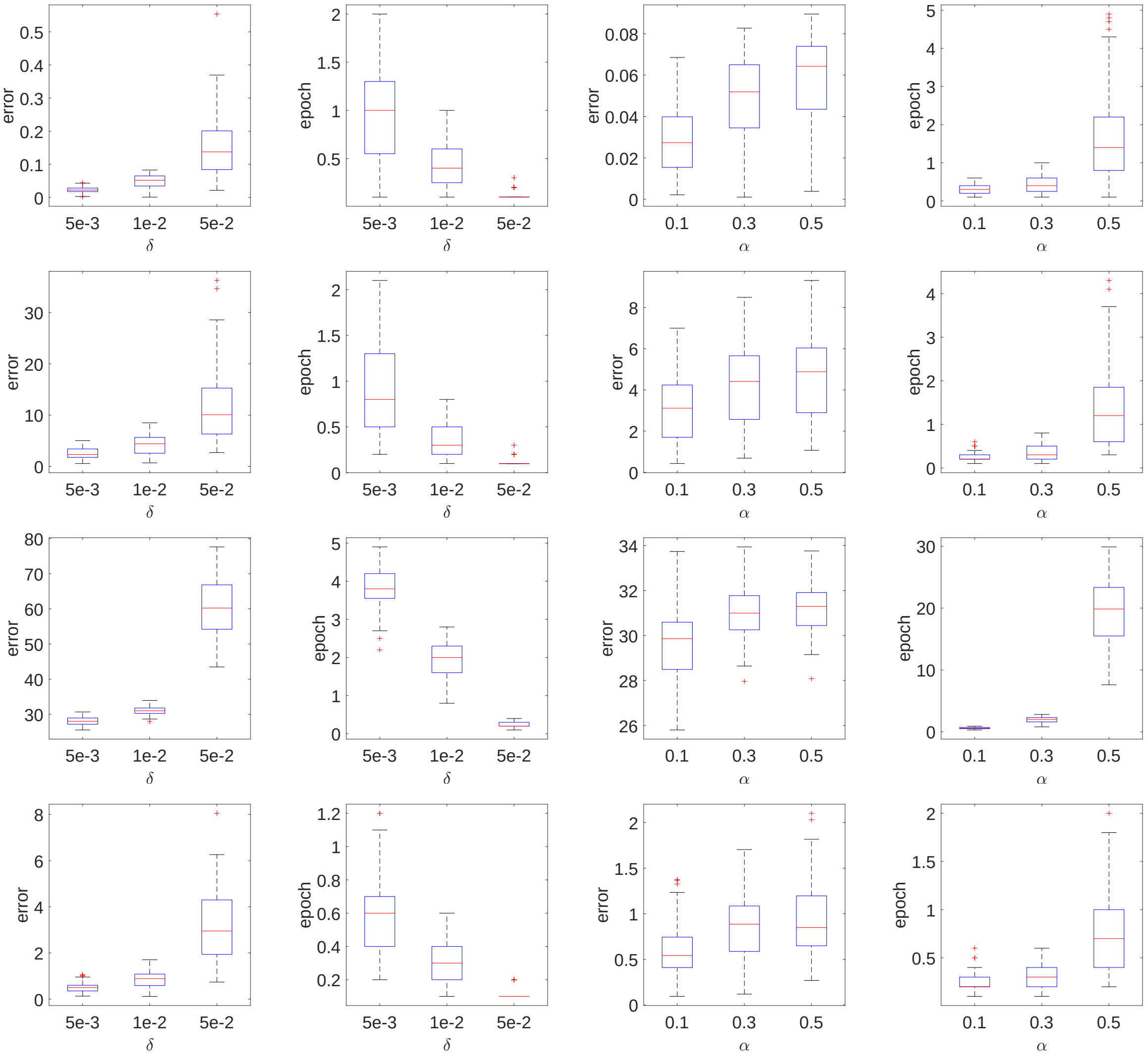

Figure 3.1: Box plots for the error $\|x_{k_\delta}^\delta - \hat{x}\|^2$ and the stopping index k_δ by SGD. The first two columns are obtained by SGD with $\alpha = 0.3$, whereas the last two columns are for the noise level δ =1e-2. The rows from top to bottom refer to `phillips`, `gravity`, `shaw` and `smoothed-phillips`, respectively.

In the plots, we have indicated the true noise $\|y^\delta - \hat{y}\|^2$, also denoted by δ^2. It is observed that both $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$ and $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ decay steadily at an algebraic rate up to a value comparable to the stopping index $k^*(\delta)$ for the Landweber method (by the discrepancy principle (3.6)). Beyond the critical threshold $k^*(\delta)$, the error $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$ exhibits a semiconvergence behavior in that it starts to increase, whereas the residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ nearly levels off at a value comparable with the noise level δ^2 (actually it oscillates slightly, since the SGD iterate is only descent for the residual on average). This is typical for iterative regularization methods for inverse problems, since for the later iterates, the noise becomes the dominating driving force. Proposition 3.2.12 with $s = \frac{1}{2}$ indicates that a similar behavior holds also for their variance components (up to slightly beyond $k^*(\delta)$). Actually, the residual variance

$\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ first decays as $O(k^{-2(1-\alpha)})$ (upon ignoring the δ term), which matches well the empirical rate in the plot. For the later iterates, as suggested by the δ term in Proposition 3.2.12, the decay is roughly $O(k^{-\alpha})$. Likewise, the error variance $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ decays slower at a rate $O(k^{-(1-\alpha)})$. Interestingly, the decay rates of $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ and $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ in the first and last columns are largely comparable, despite their drastic difference in the smoothness of the exact solution $\hat{x}$. Thus, the decay estimate in Proposition 3.2.12 is actually quite sharp, partially explaining the saturation phenomenon observed earlier. This behavior is consistently observed for all three α values. It is worth noting that for `smoothed-phillips`, the curves for $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$ nearly overlay each other, i.e., the bias component is negligible after the initial 50 iterations, due to high smoothness of the true solution, clearly indicating the saturation. For the other three examples, empirically, the variance components are of smaller order right after the initial 50 iterations. In particular, as stated in Proposition 3.2.2, $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ contributes very little to the mean squared residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ in the neighborhood of $k^*(\delta)$. This occurs for all three values of the exponent α in the stepsize schedule. The observations hold also for individual realizations; see Fig. 3.3 for the corresponding plots. The overall behavior of the curves in Fig. 3.3 is fairly similar to that in Fig. 3.2, except that the residual and error curves exhibit pronounced oscillations due to the randomness of the row index selection. Nonetheless, in the neighborhood of $k^*(\delta)$, the variance components remain much smaller in magnitude. This observation provides the key insight for the analysis in Section 3.2.1.

3.4.3 Independent run

The convergence analysis in Theorem 3.1.4 requires a SGD iterate $x_{k(\delta)}^\delta$ independent of the stopping index $k(\delta)$ determined by the discrepancy principle (3.3). In practice this can be achieved by an independent run of SGD, at the expense of slightly increasing the computational effort. Now we examine the impact of this choice, and we denote by DP and i-DP the SGD iterate used in (3.3) and that by an independent SGD run, respectively. The relevant numerical results are presented in Tables 3.5–3.8, where the numbers outside and inside the bracket denote e_{sgd} and $\mathrm{std}(e_{sgd})$, respectively. It is observed that DP gives only slightly better results in terms of the mean, but its standard deviation $std(e_{sgd})$ is generally much smaller than that by i-DP. Nonetheless, both the mean e_{sgd} and the standard deviation $std(e_{sgd})$ of i-DP are decreasing steadily as the noise level δ decreases to 0, confirming the convergence result in Theorem 3.1.4.

The difference is more clearly visualised in the boxplots in Fig. 3.4 (for `phillips` with two noise levels). A close look shows that the mean and percentile are fairly close to each other, but the i-DP result tends to have far more outliers lying above the box (marked by red cross in the plots). This is attributed to the fact that

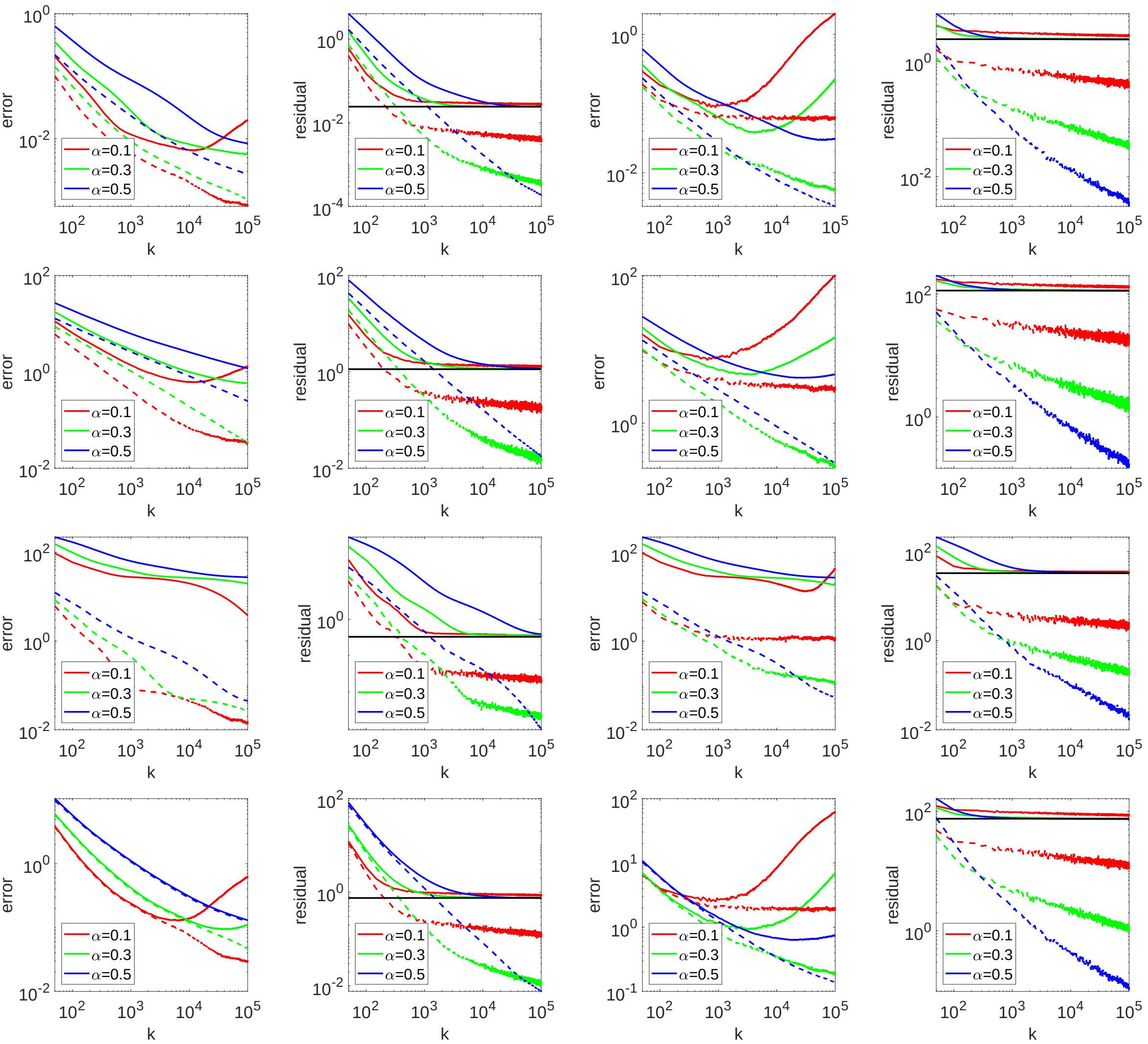

Figure 3.2: The decay of the mean squared error $\mathbb{E}[\|x_k^\delta - \hat{x}\|^2]$ and residual $\mathbb{E}[\|Ax_k^\delta - y^\delta\|^2]$ and their variance components $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ versus the SGD iteration number k. The solid and dashed curves denote the mean squared quantity and the variance component, respectively, and the black curve indicates the discrepancy $\delta^2 = \|y^\delta - \hat{y}\|^2$. The first two columns are for the noise level $\delta =$ 5e-3 and the last two columns are for the noise level $\delta =$ 5e-2. The rows from top to bottom refer to `phillips`, `gravity`, `shaw` and `smoothed-phillips`, respectively.

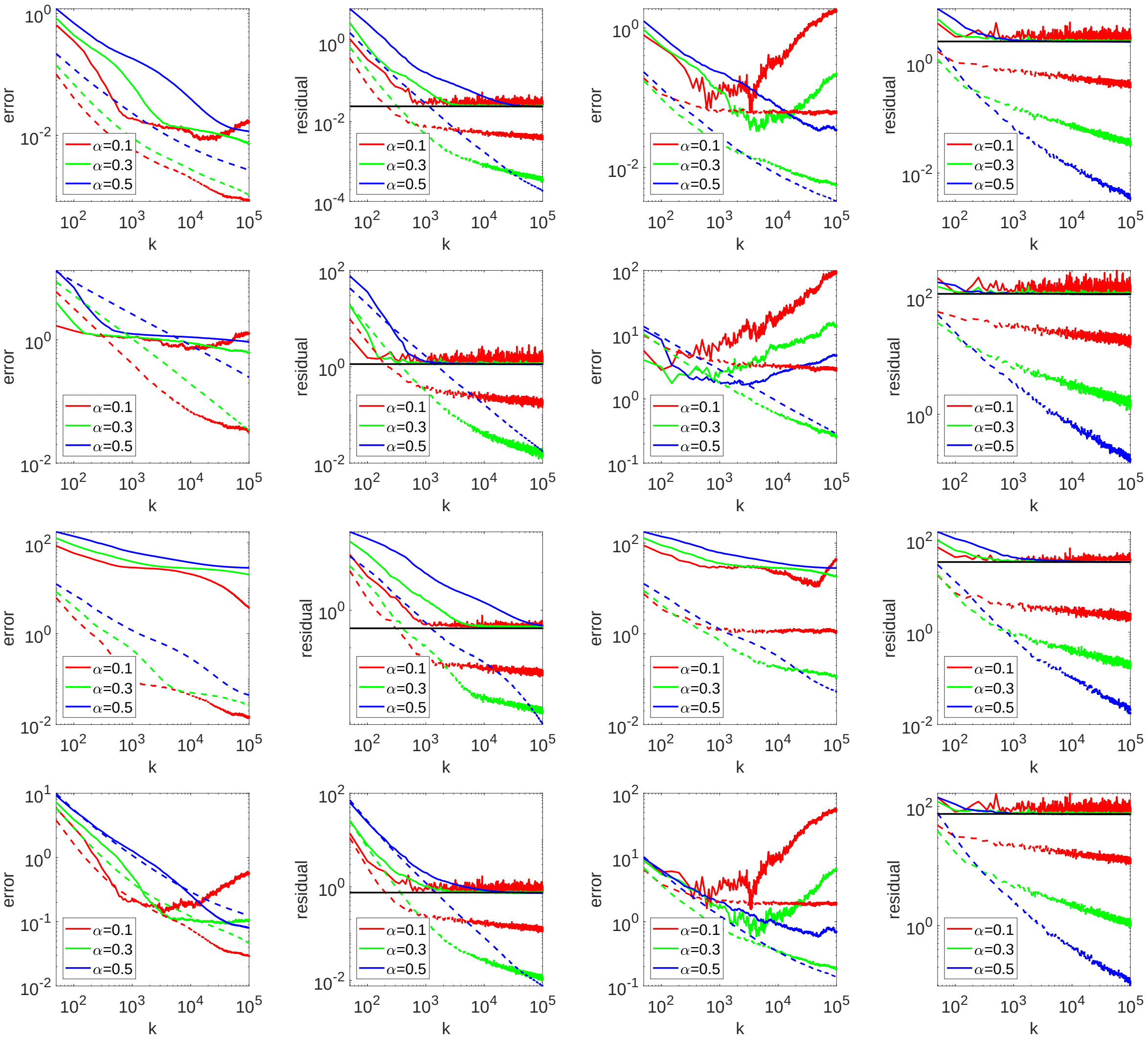

Figure 3.3: The decay of the squared error $\|x_k^\delta - \hat{x}\|^2$ and residual $\|Ax_k^\delta - y^\delta\|^2$ and their variance components $\mathbb{E}[\|x_k^\delta - \mathbb{E}[x_k^\delta]\|^2]$ and $\mathbb{E}[\|A(x_k^\delta - \mathbb{E}[x_k^\delta])\|^2]$ versus the SGD iteration number k. The solid and dashed curves denote the squared quantity and the variance components, respectively, and the black curve indicates the discrepancy $\delta^2 = \|y^\delta - \hat{y}\|^2$. The first two columns are for the noise level $\delta =$ 5e-3 and the last two columns are for the noise level $\delta =$ 5e-2. The rows from top to bottom refer to `phillips`, `gravity`, `shaw` and `smoothed-phillips`, respectively.

$k(\delta)$ determined by the discrepancy principle (3.3) is occasionally too small for an independent SGD run, and thus the corresponding residual is far above the target noise level in the discrepancy principle (3.3); see the boxplots in the last column of Fig. 3.4. That is, the outliers are due to stopping too early. This agrees with the observation that one iteration step of SGD has only a small effect on the high frequency components (because of the scaling with the corresponding small singular values). Thus, small $\|Ax_k^\delta - \hat{y}\|$ for $k \ll k^*(\delta)$ implies that also $\|x_k^\delta - \hat{x}\|$ is small. Although not presented, we note that this behavior is observed for all the examples at different noise levels. Thus, in practice, using the SGD iterate directly from the path for (3.3) is preferred, taking into account both accuracy and computational efficiency. It is an interesting theoretical question to analyze the convergence (and convergence rates) of the SGD iterate by (3.3).

Table 3.5: Comparison between DP and i-DP for `phillips`.

	$\alpha = 0.1$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP
1e-3	8.60e-3 (4.53e-3)	1.12e-2 (1.18e-2)	8.34e-3 (4.60e-3)	1.28e-2 (1.55e-2)
5e-3	1.70e-2 (8.41e-3)	2.31e-2 (2.43e-2)	2.48e-2 (7.38e-3)	4.17e-2 (3.63e-2)
1e-2	2.82e-2 (1.62e-2)	4.35e-2 (4.44e-2)	5.78e-2 (2.04e-2)	6.85e-2 (5.66e-2)
5e-2	1.41e-1 (9.70e-2)	1.53e-1 (8.97e-2)	2.11e-1 (9.69e-2)	2.47e-1 (1.93e-1)

Table 3.6: Comparison between DP and i-DP for `gravity`.

	$\alpha = 0.1$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP
1e-3	6.71e-1 (2.61e-1)	9.30e-1 (7.45e-1)	7.46e-1 (2.73e-1)	1.03e0 (8.04e-1)
5e-3	2.00e0 (8.91e-1)	2.43e0 (1.39e0)	2.53e0 (1.12e0)	3.74e0 (2.62e0)
1e-2	3.12e0 (1.57e0)	4.03e0 (2.54e0)	4.33e0 (1.92e0)	5.24e0 (3.13e0)
5e-2	9.07e0 (5.31e0)	1.01e1 (5.49e0)	1.15e1 (6.61e0)	1.19e1 (8.16e0)

Table 3.7: Comparison between DP and i-DP for `shaw`.

	$\alpha = 0.1$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP
1e-3	8.29e0 (9.35e-2)	8.30e0 (3.29e-1)	2.01e1 (5.64e-1)	2.00e1 (5.25e-1)
5e-3	2.77e1 (1.24e0)	2.77e1 (1.27e0)	2.82e1 (1.02e0)	2.80e1 (1.22e0)
1e-2	2.96e1 (1.65e0)	3.03e1 (2.58e0)	3.12e1 (1.08e0)	3.16e1 (2.44e0)
5e-2	5.02e1 (1.08e1)	5.34e1 (1.53e1)	6.70e1 (7.41e0)	7.04e1 (1.35e1)

Table 3.8: Comparison between DP and i-DP for `smoothed-phillips`.

	$\alpha = 0.1$		$\alpha = 0.5$	
δ	DP	i-DP	DP	i-DP
1e-3	1.63e-1 (6.87e-2)	1.92e-1 (1.27e-1)	1.55e-1 (6.09e-2)	1.93e-1 (1.88e-1)
5e-3	3.92e-1 (2.08e-1)	4.68e-1 (3.47e-1)	4.92e-1 (1.99e-1)	7.51e-1 (5.73e-1)
1e-2	5.95e-1 (2.64e-1)	8.12e-1 (5.04e-1)	9.46e-1 (3.93e-1)	1.46e0 (1.13e0)
5e-2	2.98e0 (1.44e0)	3.25e0 (1.52e0)	4.35e0 (2.13e0)	4.59e0 (3.29e0)

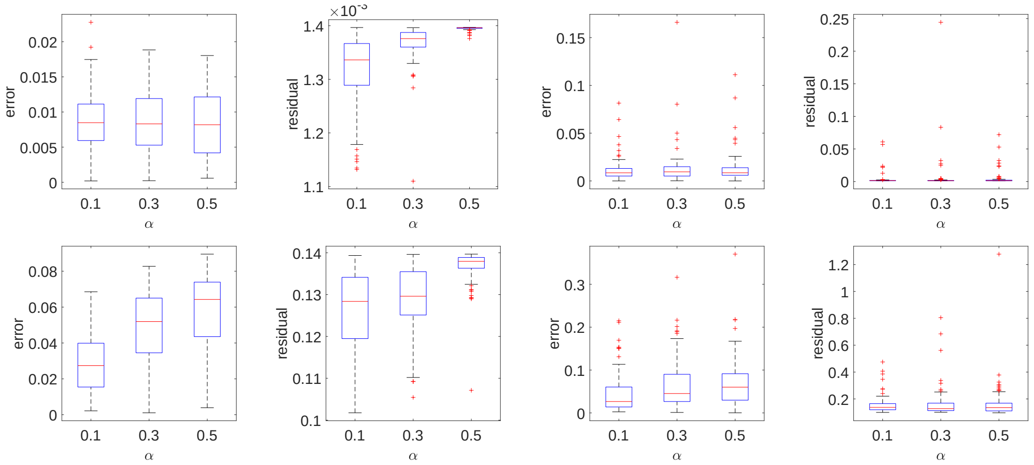

Figure 3.4: Boxplots for the error $\|x^\delta_{k(\delta)} - \hat{x}\|^2$ and the residual $\|Ax^\delta_{k(\delta)} - y^\delta\|^2$ for DP (the first two columns) and i-DP (the last two columns), for `phillips` at two noise levels, i.e., $\delta =$ 1e-3 (top) and $\delta =$ 1e-2 (bottom).

3.5 Concluding remarks

In this work, we have presented a preliminary study on the discrepancy principle as an *a posteriori* stopping rule for the popular stochastic gradient descent for solving linear inverse problems. We proved a finite-iteration termination property of the principle, and a consistency result in high probability for an independent version of discrepancy principle. Several numerical experiments indicate the feasibility of the rule as a stopping criterion.

There are several outstanding questions that deserve further research. First, one important question is the convergence of the dependent version of the discrepancy principle, and convergence rates (and also optimality, if possible!). This would put the discrepancy principle on a firm mathematical basis. Second, the analysis so far does not cover the critical case $\alpha = 1$ in the stepsize schedule. This choice is often adopted in the context of stochastic approximation [KY03] for optimal asymptotic behaviour, but it is unclear whether the discrepancy principle can be applied then.

Bibliography

[AAA$^+$19] Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. First M87 Event Horizon Telescope Results. III. Data Processing and Calibration. *The Astrophysical Journal Letters*, 875(1):L3, 2019.

[AHHK13] Thomas Alm, Bastian Harrach, Daphne Harrach, and Marco Keller. A Monte Carlo pricing algorithm for autocallables that allows for stable differentiation. *Journal of Computational Finance*, 17(1), 2013.

[AM90] Miguel A Ariño and Benjamin Muckenhoupt. Maximal functions on classical Lorentz spaces and Hardy's inequality with weights for nonincreasing functions. *Transactions of the American Mathematical Society*, 320(2):727–735, 1990.

[Ang12] Jordanka A Angelova. On moments of sample mean and variance. *Int. J. Pure Appl. Math*, 79(1):67–85, 2012.

[Bak84] AB Bakushinskii. Remarks on the choice of regularization parameter from quasioptimality and relation tests. *Zh. Vychisl. Mat. i Mat. Fiz.*, 24(8):1258–1259, 1984.

[Bar18] Johnathan M Bardsley. *Computational uncertainty quantification for inverse problems*, volume 19. SIAM, 2018.

[BCN18] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Rev.*, 60(2):223–311, 2018.

[Bec11] SMA Becker. Regularization of statistical inverse problems and the Bakushinskiĭ veto. *Inverse Problems*, 27(11):115010, 2011.

[BHMR07] Nicolai Bissantz, Thorsten Hohage, Axel Munk, and Frits Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.

[BHR18] Gilles Blanchard, Marc Hoffmann, and Markus Reiß. Optimal adaptation for early stopping in statistical inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1043–1075, 2018.

[BLMT09] Toni Buades, Yifei Lou, Jean-Michel Morel, and Zhongwei Tang. A note on multi-image denoising. In *2009 International Workshop on Local and*

Non-Local Approximation in Image Processing, pages 1–15. IEEE, 2009.

[BM12] Gilles Blanchard and Peter Mathé. Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse problems*, 28(11):115011, 2012.

[BO91] Ivo Babuška and John Osborn. Eigenvalue problems. 1991.

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag/Springer, Heidelberg, 2010.

[BR08] Frank Bauer and Markus Reiß. Regularization independent of the noise level: an analysis of quasi-optimality. *Inverse Problems*, 24(5):055009, 2008.

[BZAJ20] Riccardo Barbano, Chen Zhang, Simon Arridge, and Bangti Jin. Quantifying Model Uncertainty in Inverse Problems via Bayesian Deep Gradient Descent. *arXiv preprint arXiv:2007.09971*, 2020.

[Can06] Emmanuel J Candes. Modern statistical estimation via oracle inequalities. *Acta numerica*, 15:257, 2006.

[Cav11] Laurent Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, pages 3–96. Springer, 2011.

[CG$^+$06] Laurent Cavalier, Yu Golubev, et al. Risk hull method and regularization by projections of ill-posed inverse problems. *The Annals of Statistics*, 34(4):1653–1677, 2006.

[CGP$^+$02] Laurent Cavalier, GK Golubev, Dominique Picard, AB Tsybakov, et al. Oracle inequalities for inverse problems. *The Annals of Statistics*, 30(3):843–874, 2002.

[CT02] Laurent Cavalier and Alexandre Tsybakov. Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, 123(3):323–354, 2002.

[DJ94] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

[Don95] David L Donoho. Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Applied and computational harmonic analysis*, 2(2):101–126, 1995.

[EHN96] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems.* Kluwer Academic Publishers Group, Dordrecht, 1996.

[Fel68] William Feller. *An Introduction to Probability Theory and its Applications. Volume I.* John Wiley & Sons, Inc., New York-London-Sydney, third edition, 1968.

[G$^+$11] Yuri Golubev et al. Adaptive spectral regularizations of high dimensional linear models. *Electronic Journal of Statistics*, 5:1588–1617, 2011.

[GBH70] Richard Gordon, Robert Bender, and Gabor T. Herman. Algebraic Reconstruction Techniques (art) for three-dimensional electron microscopy and X-ray photography. *J. Theor. Biol.*, 29(3):471–476, IN1–IN2, 477–481, 1970.

[GH20] Henrik Garde and Nuutti Hyvönen. Mimicking relative continuum measurements by electrode data in two-dimensional electrical impedance tomography. *arXiv preprint arXiv:2001.10604*, 2020.

[GHR17] Daniel Gerth, Andreas Hofinger, and Ronny Ramlau. On the lifting of deterministic convergence rates for inverse problems with stochastic noise. *Inverse Problems & Imaging*, 11(4):663–687, 2017.

[GHR20] Thomas Gerstner, Bastian Harrach, and Daniel Roth. Monte Carlo pathwise sensitivities for barrier options. *Journal of Computational Finance*, 23(5), 2020.

[GSS14] Emmanuel S Garcia, David T Sandwell, and Walter HF Smith. Retracking Cryosat-2, Envisat and Jason-1 radar altimetry waveforms for improved gravity field recovery. *Geophysical Journal International*, 196(3):1402–1422, 2014.

[Gut13] Allan Gut. *Probability: a graduate course*, volume 75. Springer Science & Business Media, 2013.

[HA10] Umer Hassan and Muhammad Sabieh Anwar. Reducing noise by repetition: introduction to signal averaging. *European Journal of Physics*, 31(3):453, 2010.

[Han94] Per Christian Hansen. Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numerical algorithms*, 6(1):1–35, 1994.

[Han07] Per Christian Hansen. Regularization Tools version 4.0 for Matlab 7.3. *Numer. Algorithms*, 46(2):189–194, 2007.

[Han10] Per Christian Hansen. *Discrete inverse problems: insight and algorithms*, volume 7. Siam, 2010.

[HB16] John C Hull and Sankarshan Basu. *Options, futures, and other derivatives.* Pearson Education India, 2016.

[Heg92] Markus Hegland. An optimal order regularization method which does not use additional smoothness assumptions. *SIAM journal on numerical analysis*, 29(5):1446–1461, 1992.

[HJP20a] Bastian Harrach, Tim Jahn, and Roland Potthast. Beyond the Bakushinskii veto: Regularising linear inverse problems without knowing the noise distribution. *Numerische Mathematik*, 145(3):581–603, 2020.

[HJP20b] Bastian Harrach, Tim Jahn, and Roland Potthast. Regularising linear inverse problems under unknown non-Gaussian white noise. *arXiv preprint arXiv:2010.04519*, 2020.

[HM07] Bernd Hofmann and Peter Mathé. Analysis of profile functions for general linear regularization methods. *SIAM Journal on Numerical Analysis*, 45(3):1122–1141, 2007.

[Hof06] Andreas Hofinger. *Ill-posed problems: Extending the deterministic theory to a stochastic setup.* Trauner, 2006.

[HS01] Martin Hanke and Otmar Scherzer. Inverse problems light: numerical differentiation. *The American Mathematical Monthly*, 108(6):512–521, 2001.

[IJ15] Kazufumi Ito and Bangti Jin. *Inverse Problems: Tikhonov Theory and Algorithms.* World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2015.

[JJ20] Tim Jahn and Bangti Jin. On the discrepancy principle for stochastic gradient descent. *Inverse Problems*, 36(9):095009, sep 2020.

[JL19] Bangti Jin and Xiliang Lu. On the regularizing property of stochastic gradient descent. *Inverse Problems*, 35(1):015004, 27, 2019.

[JZ13] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *NIPS'13*, pages 315–323, Lake Tahoe, Nevada, 2013.

[JZZ20a] Bangti Jin, Zehui Zhou, and Jun Zou. On the convergence of stochastic gradient descent for nonlinear ill-posed problems. *SIAM J. Optim.*, 30(2):1421–1450, 2020.

[JZZ20b] Bangti Jin, Zehui Zhou, and Jun Zou. On the Saturation Phenomenon of Stochastic Gradient Descent for Linear Inverse Problems. *arXiv preprint arXiv:2010.10916*, 2020.

[KJ19] Tobias Kluth and Bangti Jin. Enhanced reconstruction in magnetic particle imaging by whitening and randomized SVD approximation. *Phys. Med. Biol.*, 64(12):125026, 2019.

[Kle13] Achim Klenke. *Probability theory: a comprehensive course.* Springer Science & Business Media, 2013.

[KN08] Stefan Kindermann and Andreas Neubauer. On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. *Inverse Problems & Imaging*, 2(2):291, 2008.

[KNS08] Barbara Kaltenbacher, Andreas Neubauer, and Otmar Scherzer. *Iterative Regularization Methods for Nonlinear Ill-posed Problems.* Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[KPJP18] Stefan Kindermann, Sergiy Pereverzyev Jr, and Andrey Pilipenko. The quasi-optimality criterion in the linear functional strategy. *Inverse Problems*, 34(7):075001, 2018.

[KS06] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media, 2006.

[KS07] Jari Kaipio and Erkki Somersalo. Statistical inverse problems: discretization, model reduction and inverse crimes. *Journal of computational and applied mathematics*, 198(2):493–504, 2007.

[KY03] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer-Verlag, New York, second edition, 2003.

[Lan51] L. Landweber. An iteration formula for Fredholm integral equations of the first kind. *Amer. J. Math.*, 73:615–624, 1951.

[Lav62] MM Lavrentiev. O nekotorykh nekorrektnykh zadachakh matematicheskoi fiziki. *Izdat. Sibirsk. Otdel. Akad. Nauk SSSR*, 1962.

[LM14] Shuai Lu and Peter Mathé. Discrepancy based model selection in statistical inverse problems. *Journal of Complexity*, 30(3):290–308, 2014.

[LPB$^+$18] Felix Lucka, Katharina Proksch, Christoph Brune, Nicolai Bissantz, Martin Burger, Holger Dette, and Frank Wübbeling. Risk estimators for choosing regularization parameters in ill-posed problems-properties and limitations. *Inverse Problems & Imaging*, 12(5):1121–1155, 2018.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 1991.

[Lyo04] Richard G Lyons. *Understanding digital signal processing, 3/E.* Pearson Education India, 2004.

[MBLW04] Craig D Mackay, John Baldwin, Nicholas Law, and Peter Warner. High-resolution imaging in the visible from the ground without adaptive optics: new techniques and results. In *Ground-based Instrumentation for Astronomy*, volume 5492, pages 128–136. International Society for Optics and Photonics, 2004.

[MH08] Peter Mathé and Bernd Hofmann. How general are general source conditions? *Inverse Problems*, 24(1):015009, 2008.

[Mor68] Vladimir Alekseevich Morozov. The error principle in the solution of operational equations by the regularization method. *Zhurnal Vychisli-*

tel'noi Matematiki i Matematicheskoi Fiziki, 8(2):295–309, 1968.

[MP01] Peter Mathé and Sergei V Pereverzev. Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods. *SIAM Journal on Numerical Analysis*, 38(6):1999–2021, 2001.

[MP03a] Peter Mathé and Sergei V Pereverzev. Discretization strategy for linear ill-posed problems in variable Hilbert scales. *Inverse Problems*, 19(6):1263, 2003.

[MP03b] Peter Mathé and Sergei V Pereverzev. Geometry of linear ill-posed problems in variable Hilbert scales. *Inverse problems*, 19(3):789, 2003.

[Nat86] Frank Natterer. *The Mathematics of Computerized Tomography.* B. G. Teubner, Stuttgart; John Wiley & Sons, Ltd., Chichester, 1986.

[Neu08] Andreas Neubauer. The convergence of a new heuristic parameter selection criterion for general regularization methods. *Inverse Problems*, 24(5):055005, 2008.

[NP15] Gen Nakamura and Roland Potthast. *Inverse Modeling.* 2053-2563. IOP Publishing, 2015.

[Phi62] David L Phillips. A technique for the numerical solution of certain integral equations of the first kind. *Journal of the ACM (JACM)*, 9(1):84–97, 1962.

[Rie13] Andreas Rieder. *Keine Probleme mit inversen Problemen: eine Einführung in ihre stabile Lösung.* Springer-Verlag, 2013.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[Rum11] Siegfried M Rump. Verified bounds for singular values, in particular for the spectral norm of a matrix and its inverse. *BIT Numerical Mathematics*, 51(2):367–384, 2011.

[SKHK12] Thomas Schuster, Barbara Kaltenbacher, Bernd Hofmann, and Kamil S Kazimierski. *Regularization methods in Banach spaces*, volume 10. Walter de Gruyter, 2012.

[TA77] AN Tikhonov and Vasili Ya Arsenin. *Methods for solving ill-posed problems.* John Wiley and Sons, Inc, 1977.

[Tau98] Ulrich Tautenhahn. Optimality for ill-posed problems under general source conditions. *Numerical Functional Analysis and Optimization*, 19(3-4):377–398, 1998.

[Ten17] Luis Tenorio. *An introduction to data analysis and uncertainty quantification for inverse problems.* SIAM, 2017.

[Tik63] Andrei Nikolaevich Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.

[Wah77] Grace Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667, 1977.

[Wer18] Frank Werner. Adaptivity and Oracle Inequalities in Linear Statistical Inverse Problems: A (Numerical) Survey. In *New Trends in Parameter Identification for Mathematical Models*, pages 291–316. Springer, 2018.

Lebenslauf

Tim Nikolas Jahn

Persönliche Daten

Geboren	am 19. März 1992 in Bensheim
Familienstand	verheiratet mit Sabrina Jahn (geborene Amrein), drei Kinder, geboren 2015, 2017 und 2018
Staatsangehörigkeit	Deutsch

Ausbildung und berufliche Tätigkeit

1998-2001	Grundschule in Nieder-Beerbach
2001-2010	Schuldorf Bergstraße in Seeheim-Jugenheim Abschluss: Abitur, Note: 1.1
2010-2011	Wehrdienst als Musiker im HMK 300 in Koblenz
2011-2016	Studium der Mathematik, Physik und Meteorologie an der Goethe-Universität Frankfurt
August 2015	Bachelorabschluss Physik, 'Neuron controlled robots in simulated physical environment', Betreuer: Prof. Dr. Claudius Gros, Abschlussnote: 1.1
September 2016	Masterabschluss Mathematik, 'The high temperature regime of a multi-species mean field spin glass', Betreuer: Prof. Dr. Nicola Kistler, Abschlussnote: 1.0
2014	Auslandssemester an der Universität Stockholm
2012 - 2016	Studentische Hilfskraft an den Fachbereichen 10 und 12 der Goethe-Universität Frankfurt
Seit Oktober 2016	Wissenschaftlicher Mitarbeiter von Prof. Dr. Bastian von Harrach am Institut für Mathematik der Goethe-Universität Frankfurt