On a Dynamic Variant of the Iteratively Regularized Gauss-Newton Method with Sequential Data

Citation for published version:
Chada, N, Iglesias, M, Lu, S & Werner, F 2023, 'On a Dynamic Variant of the Iteratively Regularized Gauss-Newton Method with Sequential Data', SIAM Journal on Scientific Computing, vol. 45, no. 6, pp. A3020-A3046. https://doi.org/10.1137/22m1512442

Digital Object Identifier (DOI):
10.1137/22m1512442

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Peer reviewed version

Published In:
SIAM Journal on Scientific Computing

Publisher Rights Statement:
© 2023 Society for Industrial and Applied Mathematics. This author accepted manuscript is deposited under a CC BY license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

General rights
Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
ON A DYNAMIC VARIANT OF THE ITERATIVELY REGULARIZED
GAUSS-NEWTON METHOD WITH SEQUENTIAL DATA *

NEIL K. CHADA†, MARCO IGLESIAS‡, SHUAI LU§, AND FRANK WERNER¶

Abstract. For numerous parameter and state estimation problems, assimilating new data as they become available can help produce accurate and fast inference of unknown quantities. While most existing algorithms for solving those kind of ill-posed inverse problems can only be used with a single instance of the observed data, in this work we propose a new framework that enables existing algorithms to invert multiple instances of data in a sequential fashion. Specifically we will work with the well-known iteratively regularized Gauss–Newton method (IRGNM), a variational methodology for solving nonlinear inverse problems. We develop a theory of convergence analysis for a proposed dynamic IRGNM algorithm in the presence of Gaussian white noise. We combine this algorithm with the classical IRGNM to deliver a practical (blended) algorithm that can invert data sequentially while producing fast estimates. Our work includes the proof of well-definedness of the proposed iterative scheme, as well as various error bounds that rely on standard assumptions for nonlinear inverse problems. We use several numerical experiments to verify our theoretical findings, and to highlight the benefits of incorporating sequential data. The context of the numerical experiments comprises various parameter identification problems including a toy elliptic PDE example, and that of electrical impedance tomography.

Key words. Inverse problems, regularization theory, Gauss–Newton method, convergence rates

AMS subject classifications. 94A12, 86A22, 60G35, 62M99

1. Introduction. A common problem in numerous scientific disciplines is the estimation of some unknown function \(u^\dagger \in \mathcal{X}\), from observations of an exact quantity \(y^\dagger \in \mathcal{Y}\) of the form

\[
y^\dagger = F(u^\dagger),
\]

where we assume that \(F : D(F) \subset \mathcal{X} \to \mathcal{Y}\) is a nonlinear continuous operator acting between two Hilbert spaces \(\mathcal{X}\) and \(\mathcal{Y}\) with domain of definition \(D(F)\). Due to the unavoidable presence of observational noise in real applications, the idealised equation (1.1) must be replaced by

\[
y^\delta = F(u^\dagger) + \sigma \xi,
\]

where \(\xi\) could be deterministic, random uniformly bounded, or some other (unbounded) sort of noise. Problems associated with (1.1) or (1.2) are commonly referred to as inverse problems [47, 48], concerned with the estimation of some unobservable parameter or quantity of interest. Such examples of particular applications include,
but not limited to, geophysical sciences, medical imaging and numerical weather prediction [8, 43, 44].

In the classical deterministic literature, it is assumed that the noise $\xi$ is an element of $\mathcal{Y}$ with small norm, $\|\xi\|_\mathcal{Y} \leq \delta$. In this case, one has $y^\delta \in \mathcal{Y}$, and a well-known regularization method to recover $u^1$ from $y^\delta$ is the iteratively regularized Gauss–Newton method (IRGNM) [35], proposed originally by Bakushinskii [2]. At each iteration, the IRGNM solves a variational problem of the form

$$
\hat{u}_{n+1} := \arg \min_{u \in \mathcal{X}} \left[ \|F(\hat{u}_n) + F'[\hat{u}_n](u - \hat{u}_n) - y^\delta\|_\mathcal{Y}^2 + \alpha_n \|u - \hat{u}_0\|_\mathcal{X}^2 \right],
$$

where $\hat{u}_0 \in \mathcal{X}$ is some initial guess, $F'[u]$ is the Fréchet (or some other) derivative of $F$ at $u$, and $\{\alpha_n\}_{n=1}^N$ is a sequence of regularization parameters chosen such that

$$
a_0 \leq 1, \quad \alpha_n \searrow 0, \quad 1 \leq \frac{\alpha_n}{\alpha_{n+1}} \leq C_{\text{dec}}, \quad \text{for all } n \in \mathbb{N},
$$

for some constant $C_{\text{dec}}$. Typically one uses $\alpha_n = \alpha_0 C_{\text{dec}}^{-n}$. Alternatively, we can express the minimization procedure of (1.3) in terms of the first order optimality condition as

$$
\hat{u}_{n+1} = \hat{u}_n - (F'[\hat{u}_n] + \alpha_n \text{Id} \mathcal{X})^{-1} (F'[\hat{u}_n] (F(\hat{u}_n) - y^\delta) + \alpha_n (\hat{u}_n - \hat{u}_0)),
$$

with the adjoint $F'[u] : \mathcal{Y} \to \mathcal{X}$ of $F'[u] : \mathcal{X} \to \mathcal{Y}$. Convergence (rate) analysis for the classical IRGNM can be found in [7, 31, 34] and extension towards the random noise or the Banach space setting can be found in [3, 32, 33] and references therein.

In contrast to the case of deterministic noise, here we consider a common stochastic model which assumes (unbounded) random noise $\xi$. More precisely, we require $\xi : \mathcal{Y} \to L^2(\Omega, \mathcal{A}, \mathbb{P})$ to be a Hilbert space process with an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$ satisfying $\mathbb{E}[\xi] = 0$ and $\|\text{Cov}[\xi]\|_{\mathcal{Y} \to \mathcal{Y}} \leq 1$. This means that $\xi$ can be interpreted as a (random) element of the algebraic dual space $\mathcal{Y}^*$, but in general not as an element of the topological dual space $\mathcal{Y}'$ of $\mathcal{Y}$. In this case, expression (1.2) has to be understood in a weak sense because the interpretation $\xi \in \mathcal{Y}^*$ does not allow for a Riesz representation $\xi \in \mathcal{Y}$; hence, in general, $y^\delta \notin \mathcal{Y}$. This implies that for every $g \in \mathcal{Y}$, the quantity

$$
\langle \xi, g \rangle := \langle \xi, g \rangle_{\mathcal{Y}' \times \mathcal{Y}} = \xi(g)
$$

is a real-valued random variable with $\mathbb{E}[\xi, g] = \langle \mathbb{E}[\xi] , g \rangle = 0$, and finite second moment $\mathbb{E}[\xi, g] = (\text{Cov}[\xi] g, g) \leq \|g\|_{\mathcal{Y}}^2$. Note that the notation in (1.5) is compatible with the deterministic case $y^\delta \in \mathcal{Y}$, and thus with the case where $\xi \in Y$. We refer to [14] for more details on Hilbert space processes and infinite-dimensional random models.

In this study we employ a observational model with sequential noisy observations of the form

$$
Y_n = F(u^1) + \sigma \xi_n, \quad n = 1, 2, \ldots,
$$

where $\xi_n : \mathcal{Y} \to L^2(\Omega, \mathcal{A}, \mathbb{P})$ are i.i.d. (independent, identically distributed) Hilbert space processes again satisfying $\mathbb{E}[\xi_n] = 0$, $\|\text{Cov}[\xi_n]\|_{\mathcal{Y} \to \mathcal{Y}} \leq 1$. Observations of the form (1.6) are available in nearly all practical applications, but usually not treated as such. Instead, sequential observations of the form (1.6) are used to generate a final


(1.7) \[ Z_N = N^{-1} \sum_{n=1}^{N} Y_n = F (u^\dagger) + \sigma N \sum_{n=1}^{N} \xi_n, \]

as the average of the (first) \( N \) sequential observations. The rationale behind is that

the covariance operator of \( Z_N \) satisfies

\[ \text{Cov} [Z_N] = \text{Cov} \left[ \frac{\sigma}{N} \sum_{n=1}^{N} \xi_n \right] = \frac{\sigma^2}{N^2} \sum_{n=1}^{N} \text{Cov} [\xi_n], \]

and hence the noise level of \( Z_N \) is \( \frac{\sigma}{\sqrt{N}} \) instead of \( \sigma \) for each of the observations \( Y_n \) in

(1.6). In our situation, where we assume that all the available data (i.e. \( Y_n \) in (1.6) or

\( Z_N \) in (1.7) are a.s. not elements in \( \mathcal{Y} \)), the above classic IRGNM (cIRGNM) is not
directly available. However, noticing that \( \|y - y\|_{\mathcal{Y}}^2 \) in (1.3) is in finite dimensions just
the negative log-likelihood functional of the normal distribution, it seems reasonable
to replace \( \|y - y\|_{\mathcal{Y}}^2 \) by

\[ \mathcal{S}(g; Z_N) := \frac{1}{2} \|g\|_{\mathcal{Y}}^2 - \langle g, Z_N \rangle, \quad g \in \mathcal{Y}, \]

as this is the infinite-dimensional negative log-likelihood in the Cameron-Martin-
Girsanov sense, cf. [50]. This leads to the following method modification of the
cIRGNM in case of random noise:

\[ \hat{u}_{n+1} := \arg \min_{u \in \mathcal{X}} \left[ \mathcal{S} (F (\hat{u}_n) + F' [\hat{u}_n] (u - \hat{u}_n); Z_N) + \alpha_n \|u - \hat{u}_0\|_{\mathcal{Y}}^2 \right]. \]

Note that all terms in (1.9) - especially the term \( \langle Z_N, F (\hat{u}_n) + F' [\hat{u}_n] (u - \hat{u}_n) \rangle \) -
are well-defined, since we have \( F (\hat{u}_n) + F' [\hat{u}_n] (u - \hat{u}_n) \in \mathcal{Y} \) for all \( u \in \mathcal{X} \). This
method has been proposed and analyzed in [26], and allows for further generalizations
including different noise models or general convex penalty terms.

In this work we take a different focus motivated by many practical problems, for
which one does not want to begin the reconstruction procedure until the (final) mea-

urement \( Y_N \) has been collected (so that \( Z_N \) can be computed). Instead, it would be
preferable to start the reconstruction immediately after obtaining \( Y_1 \) and update our
estimate of the unknown as the new observations \( Y_2, Y_3, ... Y_N \) become available. This
motivation aligns with the aim of online algorithms for (linear) inverse problems which
have been recently attracted much attention to solve filtering and data assimilation
problems [15, 28, 29].

In the context of the IRGNM, we propose to sequentially incorporate data by
the following scheme which we call the dynamic iteratively regularized Gauss-
Newton method (dIRGNM):

\[ \hat{u}_{n+1} := \arg \min_{u \in \mathcal{X}} \left[ \mathcal{S} (F (\hat{u}_n) + F' [\hat{u}_n] (u - \hat{u}_n); Z_n) + \alpha_n \|u - \hat{u}_0\|_{\mathcal{Y}}^2 \right]. \]

Note that this algorithm can be started as soon as \( Y_1 \) (and hence \( Z_1 \)) is available,
i.e. right after the first set of observations are collected. We emphasise that the
main difference between (1.9) and (1.10) is the index \( n \) in the used data \( Z_n \) (com-
pared to \( Z_N \) in (1.9)). However, this ensures that the data \( Y_n \) (and hence \( Z_n \)) that
is currently available are assimilated sequentially via (1.7) into the algorithm (1.10). We develop theory which predicts that the performance of the proposed dIRGNM algorithm is similar to that of the cIRGNM applied to the full datum $Z_N$. However, the dIRGNM offers the advantage of an online algorithm since it can be employed as soon as new measurements become available. Consequently, we expect the overall inversion process to be faster than the classical approach which, in turn, requires us to first collect all measurements prior to running the inversion algorithm. There are further practical advantages of the proposed framework including that the dIRGNM provides reconstructions of the unknown during the measurement process. These reconstructions could be useful to assess whether additional measurements are needed. Our numerical results confirm our theoretical insights and demonstrate that the performance of the dIRGNM (and the corresponding blended version) leads to faster overall computational time without compromising the accuracy of the reconstruction.

1.1. Literature overview. In many real-world application areas, it is common to have experimental settings that allow us to sequentially acquire multiple observations of the physical process under consideration (e.g. by repeating the experiment). The classical approach for solving this kind of inverse problem is to first produce the average of those observations, and use this average with a standard regularisation method to infer the unknown quantity/parameter of interest. A class of methods for solving ill-posed inverse problems is the so-called variational regularization, which includes the well-known Tikhonov regularization, as well as various other methods such as Landweber iteration, steepest descent and $\nu$-methods [17, 35, 42].

The analysis of the convergence of most existing iterative methods, including those cited above, assume that observed data remain the same throughout the iterative procedure. However, exploring sequential variants of these methods in which data are updated as they become available can bring substantial benefits in practical settings. The focus on the IRGNM is particularly relevant since, for data assimilation problems, the Gauss-Newton method has been shown to have striking similarities with Kalman filtering methodologies that sequentially update parameters and states of dynamical processes [5, 6, 10, 12, 20].

The extensive and successful use of Kalman filter methods for large-scale data assimilation applications such as ocean and weather forecasting [9, 40, 43], has prompted a body of work aimed at importing and adapting those methodologies for solving ill-posed inverse problems. In [15, 29], for example, regularization theory was used to analyze convergence of data assimilation algorithms, such as the Kalman filter, 3DVAR and 4DVAR in the context of solving linear inverse problems. These works have shown that using multiple instances of noisy observations lead to more robust and stable algorithms when a scaling regularization parameter is appropriately tuned. In the nonlinear case, however, whether the convergence of filtering methods, such as the ensemble Kalman filter [18, 19] and extended Kalman filter, can be improved by using multiple instances of data is still an open problem. Our work on the dIRGNM, in addition to providing practical algorithms that can invert data sequentially, will also pave the way towards understanding the dynamic behaviour of data assimilation algorithms for nonlinear inverse problems.

1.2. Aim of the paper & outline. Our primary focus and contribution from this work is the development and understanding of the dIRGNM, which, as stated earlier, is a modified version of the IRGNM that enable us to sequentially invert observed data. We propose two particular forms of a dynamic IRGNM, the first is given above in (1.10) which is intended for our analysis with infinitely many observations.
The second form, which we refer to as, the blended iterated regularized Gauss-Newton method (bIRGNM) combines the cIRGNM with the dIRGNM in the practical case when finitely many observations are available. The motivation behind the blended scheme is to obtain improved performance by initially running the dIRGNM for various but finitely many observations, followed by running the cIRGNM with the average of all acquired observations. Based on standard assumptions we prove well-definedness for both algorithms. In addition, we derive appropriate error bounds and convergence rates. As typical for Newton-type methods, our results are local in the sense that they require the initial guess \( \hat{u}_0 \) to be sufficiently close to the true solution \( u^\dagger \). In order to prevent from data over-fitting, our analysis also includes recommended choices for the regularization parameter. We employ two PDE-constrained parameter identification problems in order to demonstrate their computational advantages over the cIRGNM.

The outline of this paper is as follows. In Section 2 we provide the necessary background and material related to the dIRGNM. The main results pertaining the convergence analysis of the dIRGNM are presented in Section 3 with the proposed algorithms. The numerical experiments are discussed in Section 4. We include tests on two PDE-constrained parameter identification problems. Some auxiliary analysis and plots from these experiments as well as an additional numerical example in the context of a Darcy flow model can be found in the supplementary material.

Finally, in Section 5 we provide some final conclusions and new potential areas of research.

2. Standing assumptions and error analysis. In this section, we provide error bounds for IRGNM (1.9) in the general data model and introduce the assumptions needed. Note that the same bounds also apply for (1.10) since, as discussed earlier, the only difference between (1.9) and (1.10) lies in the data they used. To treat both cases at the same time, let us denote by \( W \in \{ Z_N, Z_n, Y_n \} \) the available data, define

\[
\mathcal{J}[u, \hat{u}_n, \hat{u}_0, \alpha_n, W] := S(F(\hat{u}_n) + F'(\hat{u}_n)(u - \hat{u}_n); W) + \alpha_n \| u - \hat{u}_0 \|_X^2,
\]

and consider

\[
\hat{u}_{n+1} := \arg\min_{u \in X} \mathcal{J}[u, \hat{u}_n, \hat{u}_0, \alpha_n, W].
\]

If \( W = Z_N \), this equals (1.9), and if \( W = Z_n \), then this equals (1.10). Furthermore, the following analysis will also illustrate why the naive choice \( W = Y_n \) does not allow for an assimilation of the sequential data (1.6) and theoretically will not provide a convergent algorithm unless the noise vanishes.

Our analysis here closely follows the general approach to error bounds for variational regularization methods described in [27]. For brevity we additionally introduce

\[
\mathcal{T}(g; y^\dagger) = \frac{1}{2} \| g - y^\dagger \|_Y^2.
\]

In what follows, we make use of different constants. The letter \( C \) is used for a generic constant \( C \in (0, \infty) \), which might differ from line to line. To keep better track of other (fixed) constants we provide a summary in the following Table 1.

2.1. Preliminary error decomposition. Let us assume that the \( n \)-th iterate \( \hat{u}_n \in D(F) \) is well defined. As a first step, we aim to provide an error bound for \( \hat{u}_{n+1} \)
defined by (2.2). The minimality condition of (2.2) implies

\[ \alpha_n \left[ \| \tilde{u}_{n+1} - \tilde{u}_0 \|_X^2 - \| u^t - \tilde{u}_0 \|_X^2 \right] \leq S ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) ; W ) - S ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) ; W ) \]

Introducing the effective noise level

\[ \text{err} ( g ) := T ( g ; y^t ) - ( S ( g ; W ) - S ( y^t ; W ) ) , \quad g \in \mathcal{Y}, \]

we rewrite the right-hand side of (2.3) by

\[ S ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) ; W ) - S ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) ; W ) \]

\[ = ( S ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) ; W ) - S ( y^t ; W ) ) \]

\[ - ( S ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) ; W ) - S ( y^t ; W ) ) \]

\[ = \frac{1}{2} \| F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) - y^t \|_Y^2 - \text{err} ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) ) \]

\[ - \frac{1}{2} \| F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) - y^t \|_Y^2 + \text{err} ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) ) . \]

Note that in each of the settings \( W \in \{ Z_N, Z_n, Y_n \} \) we can derive an explicit formulation for \( \text{err}(g) \), namely

\[ \text{err} ( g ) = \begin{cases} 
\frac{\sigma}{N} \sum_{i=1}^N \langle \xi_i, g - y^t \rangle & \text{if } W = Z_N, \\
\frac{\sigma}{n} \sum_{i=1}^n \langle \xi_i, g - y^t \rangle & \text{if } W = Z_n, \\
\sigma \langle \xi_n, g - y^t \rangle & \text{if } W = Y_n, 
\end{cases} \]

for \( g \in \mathcal{Y} \).

Now we continue by denoting

\[ \text{err}_n := \text{err} ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) ) - \text{err} ( F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) ) , \]

and obtain by (2.3) that

\[ \alpha_n \left[ \| \tilde{u}_{n+1} - \tilde{u}_0 \|_X^2 - \| u^t - \tilde{u}_0 \|_X^2 \right] + \frac{1}{2} \| F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( \tilde{u}_{n+1} - \tilde{u}_n ) - y^t \|_Y^2 \]

\[ \leq \text{err}_n + \frac{1}{2} \| F ( \tilde{u}_n ) + F' [ \tilde{u}_n ] ( u^t - \tilde{u}_n ) - y^t \|_Y^2 . \]

To proceed further, we need the following variational source condition, which has been first formulated in [24] and has become a standard assumption in the analysis of variational regularization methods. In many cases variational source conditions are necessary and sufficient for convergence rates [25]. Note that, as for most general source conditions, the smoothness of \( u^t \) is measured relative to the smoothing properties of \( F \).
Assumption 1 (Variational source condition). There exists a concave index function $\varphi$ (i.e. $\varphi(0) = 0$ and $\varphi \not\equiv 0$) such that for all $u \in D(F)$ it holds

\begin{equation}
\|u - u^\dagger\|_X^2 \leq \|u - \hat{u}_0\|_X^2 - \|u^\dagger - \hat{u}_0\|_X^2 + \varphi \left( \frac{1}{2} \| F(u) - F(u^\dagger) \|_Y^2 \right).
\end{equation}

Plugging Assumption 1 into (2.6) with $u = \hat{u}_{n+1}$ yields

\begin{equation}
\alpha_n \|\hat{u}_{n+1} - u^\dagger\|_X^2 + \frac{1}{2} \| F(\hat{u}_n) + F'(\hat{u}_n) (\hat{u}_{n+1} - \hat{u}_n) - y^\dagger\|_Y^2
\end{equation}

\begin{equation}
\leq \| \varphi \| F(\hat{u}_n) + F'(\hat{u}_n) (u^\dagger - \hat{u}_n) - y^\dagger\|_Y^2.
\end{equation}

In order to further treat the nonlinearity, we employ the following assumption.

Assumption 2 (Tangential cone condition). There exist constants $C_{tc} \geq 1$ and $\eta > 0$ sufficiently small such that

\begin{equation}
\frac{1}{C_{tc}} \left| \| F(v) - y^\dagger\|_Y^2 - \eta \| F(u) - y^\dagger\|_Y^2 \right| \leq \| F(u) + F'(u) (v - u) - y^\dagger\|_Y^2
\end{equation}

\begin{equation}
\leq C_{tc} \| F(v) - y^\dagger\|_Y^2 + \eta \| F(u) - y^\dagger\|_Y^2.
\end{equation}

Remark 2.1. This tangential cone condition follows from the standard tangential cone condition with some $C_{tc}$, see [26, Lemma 5.2]. If $\varphi \not\equiv \sqrt{t}$ as $t \to 0$, it can - using the techniques from [51] - be replaced by a Lipschitz-type assumption.

The tangential cone condition gives for the second term on the left-hand side of (2.8) that

\begin{equation}
\frac{1}{C_{tc}} \| F(\hat{u}_{n+1}) - y^\dagger\|_Y^2 - \eta \| F(\hat{u}_n) - y^\dagger\|_Y^2 \leq \| F(\hat{u}_n) + F'(\hat{u}_n) (\hat{u}_{n+1} - \hat{u}_n) - y^\dagger\|_Y^2,
\end{equation}

and for the third term on the right-hand side, with (1.1), that

\begin{equation}
\| F(\hat{u}_n) + F'(\hat{u}_n) (u^\dagger - \hat{u}_n) - y^\dagger\|_Y^2 \leq C_{tc} \| F(u^\dagger) - y^\dagger\|_Y^2 + \eta \| F(\hat{u}_n) - y^\dagger\|_Y^2
\end{equation}

\begin{equation}
= \eta \| F(\hat{u}_n) - y^\dagger\|_Y^2.
\end{equation}

Inserting above two inequalities into (2.8) we obtain the recursive error estimate

\begin{equation}
\alpha_n \|\hat{u}_{n+1} - u^\dagger\|_X^2 + \frac{1}{2C_{tc}} \| F(\hat{u}_{n+1}) - y^\dagger\|_Y^2
\end{equation}

\begin{equation}
\leq \| \varphi \| F(\hat{u}_n) + F'(\hat{u}_n) (u^\dagger - \hat{u}_n) - y^\dagger\|_Y^2.
\end{equation}

We then abbreviate

\begin{equation}
d_n := \| \hat{u}_n - u^\dagger\|_X^2 \quad \text{and} \quad t_n := \frac{1}{2} \| F(\hat{u}_n) - y^\dagger\|_Y^2,
\end{equation}

and estimate

\begin{equation}
\varphi(t) - \frac{t}{\alpha} \leq \sup_{\tau \geq 0} \left[ \varphi(\tau) - \frac{\tau}{\alpha} \right] = (-\varphi)^* \left( -\frac{1}{\alpha} \right) =: \Psi(\alpha),
\end{equation}

for $t = \frac{1}{2} t_{n+1}$ with the Fenchel conjugate $(-\varphi)^*$ of the convex function $-\varphi$. We have proven the following:
Lemma 2.1 (Preliminary error estimate). Let Assumptions 1 and 2 hold and assume that \( \hat{u}_n \in D(F) \) is well defined. Then the next iterate \( \hat{u}_{n+1} \) defined by (2.2) satisfies a.s. a preliminary error decomposition

\[
\alpha_n d_{n+1} + \frac{1}{2C_{tc}} \tau_{n+1} \leq \text{err}_n + \alpha_n \Psi (2C_{tc}\alpha_n) + 2\eta t_n,
\]

with

\[
\text{err}_n := \langle W, F' [\hat{u}_n] (\hat{u}_{n+1} - u^l) \rangle.
\]

Note that (2.10) is a recursive error estimate similar to those obtained for the standard IRGNM, see e.g. [26]. However, since the right-hand side still depends implicitly on \( \hat{u}_n \) and \( \hat{u}_{n+1} \) via \( \text{err}_n \), it is not immediatly clear at this point whether we can derive absolute error bounds for \( d_{n+1} \). Hence, why we refer to these estimates as “preliminary”, while we prepare to further employ these estimates in our further analysis.

Before we continue, let us briefly provide more insight on this error estimate. For the case of noise-free observations \( Y_n = F(u^l) \), we have \( \xi_n = 0 \) (corresponding to \( W = 0 \) here) and hence \( \text{err}_n = 0 \). This shows that the first term in (2.10) is in fact a (preliminary) propagated data noise error. Correspondingly, if \( F \) is linear, we have \( \eta = 0 \) and hence the third term in (2.10) is an upper bound for the nonlinearity error. The remaining second term \( \alpha_n \Psi (2C_{tc}\alpha_n) \) in (2.10) is a bound for the approximation error, which can clearly be made arbitrarily small by letting \( \alpha_n \searrow 0 \). In the following we will discuss the individual error terms.

2.1.1. The approximation error. To simplify the approximation error bound \( \alpha_n \Psi (2C_{tc}\alpha_n) \), we need an additional restriction on the source function \( \varphi \):

Assumption 3 (Saturation of source functions). For the function \( \varphi \) from the variational source condition (2.7) there exists an \( \epsilon > 0 \) such that \( \varphi^{1+\epsilon} \) is concave.

Assumption 3 implies that

\[
\varphi (C\tau)^{1+\epsilon} = \varphi (C\tau + (1 - C)0)^{1+\epsilon} \leq C\varphi (\tau)^{1+\epsilon} + (1 - C) \varphi (0)^{1+\epsilon} = C\varphi (\tau)^{1+\epsilon},
\]

for all \( \tau \geq 0 \) and \( C \geq 1 \), and thus the monotonicity of \( \varphi \) yields

\[
\varphi (C\tau) \leq \max \left\{ 1, C\frac{\tau}{1+\epsilon} \right\} \varphi (\tau) \quad \text{for all} \quad C, \tau \geq 0.
\]

For the approximation error function \( \Psi \), this implies (2.11)

\[
\Psi (C\alpha) = \sup_{\tau \geq 0} \left[ \varphi (\tau) - \frac{\tau}{C\alpha} \right] = \sup_{s \geq 0} \left[ \varphi \left( C^{1+\epsilon} s \right) - \frac{C^{1+\epsilon}}{\alpha} s \right] \leq \max \left\{ 1, C\frac{\tau}{1+\epsilon} \right\} \Psi (\alpha).
\]

Consequently, under Assumption 3 we can simplify (2.10) to

\[
\alpha_n d_{n+1} + \frac{1}{2C_{tc}} \tau_{n+1} \leq \text{err}_n + \left( 2C_{tc} \right)^{\frac{1}{2}} \alpha_n \Psi (\alpha_n) + 2\eta t_n.
\]

2.1.2. The propagated data noise error. The propagated data noise error in (2.10) is somewhat more difficult to handle, as \( \text{err}_n \) depends (implicitly) on \( \hat{u}_{n+1} \), which is why (2.10) should only be considered as a preliminary error estimate. To derive a more helpful bound without implicit dependencies, we have to bound \( \text{err}_n \) and factorize over \( \hat{u}_{n+1} \). Therefore note that, in view of (2.4), \( \text{err}(g) \) is always of the
form $\text{err}(g) = \lambda_n \langle \Xi_n, g - y^l \rangle$ with a scalar $\lambda_n$ and some Hilbert space process $\Xi_n$. Unfortunately, we cannot estimate the term $\langle \Xi_n, g - y^l \rangle$ by means of the Cauchy-Schwarz inequality, as $\Xi_n \notin \mathcal{Y}$. This can be avoided by considering a smaller space $\mathcal{V} \subset \mathcal{Y}$ such that the embedding $\iota : \mathcal{V} \hookrightarrow \mathcal{Y}$ is a Hilbert-Schmidt operator. This yields a Gelfand triple $(\mathcal{V}, \mathcal{Y}, \mathcal{V}')$ with the topological dual space $\mathcal{V}'$ of $\mathcal{V}$ and for the adjoint $\iota^* : \mathcal{Y} \rightarrow \mathcal{V}'$ we obtain

$$E\|\Xi_n\|^2_{\mathcal{V}'} = E\|\iota^*\Xi_n\|^2_{\mathcal{Y}} = \text{trace}(\iota^* \text{Cov} [\Xi_n]\iota) \leq \text{trace}(\iota^* \iota) < \infty,$$

i.e. $\|\Xi_n\|_{\mathcal{Y}} < \infty$ a.s.. Ideally, one should choose $\mathcal{V}$ as the largest (or vice versa $\mathcal{V}'$ as the smallest) possible space with this property, but for our subsequent analysis this is not required. We only employ the resulting inequality

$$\text{err}(g) \leq \lambda_n \|\Xi_n\|_{\mathcal{V}'} \|g - y^l\|_{\mathcal{V}}, \quad \forall g \in \mathcal{Y}.$$  

In order to continue with our analysis, we use (2.4) and (2.5) to obtain

$$\text{err}_n = \text{err}(F(\hat{u}_n) + F'(\hat{u}_n)(\hat{u}_{n+1} - \hat{u}_n)) \leq \lambda_n \langle \Xi_n, F(\hat{u}_n) + F'(\hat{u}_n)(\hat{u}_{n+1} - \hat{u}_n) \rangle$$

$$\leq \lambda_n \langle \Xi_n, F(\hat{u}_n) + F'(\hat{u}_n)(\hat{u}_{n+1} - \hat{u}_n) - F(\hat{u}_n) + F'(\hat{u}_n)(u^l - \hat{u}_n) \rangle$$

$$= \lambda_n \langle \Xi_n, F'(\hat{u}_n)(\hat{u}_{n+1} - u^l) \rangle$$

$$\leq \lambda_n \|\Xi_n\|_{\mathcal{V}'} \|F'(\hat{u}_n)(\hat{u}_{n+1} - u^l)\|_{\mathcal{V}},$$

with

$$\lambda_n = \begin{cases} \frac{\varepsilon}{n} & \text{if } W = Z_N, \\ \frac{\varepsilon}{n} & \text{if } W = Z_n, \\ \sigma & \text{if } W = Y_n, \end{cases} \quad \Xi_n = \begin{cases} \sum_{i=1}^{N} \xi_i \sqrt{N} & \text{if } W = Z_N, \\ \sum_{i=1}^{n} \xi_i \sqrt{n} & \text{if } W = Z_n, \\ \xi_n & \text{if } W = Y_n, \end{cases}$$

since we assume that the noises $\xi_i, 1 \leq i \leq N$ are independently identically distributed.

To bound the second term in (2.12), we employ the following assumption.

**Assumption 4 (V-smoothing).** There exist a parameter $\theta \in (0,1)$ and a constant $C_\theta \geq 1$ such that

$$\|F'(\hat{u}_n)(u - u^l)\|_{\mathcal{V}} \leq C_\theta \|F'(\hat{u}_n)(u - u^l)\|_{\mathcal{Y}}^{\theta} \|u - u^l\|_{X}^{1-\theta},$$

for all $u \in X$.

**Remark 2.2.** This assumption is e.g. satisfied if $F'(\hat{u}_n)$ maps Lipschitz continuously into a smoother space than $\mathcal{V}$ which obeys a classical interpolation inequality, see e.g. [27, Rem. 2.6]. Thus, Assumption 4 characterizes in fact the smoothing properties of the forward operator $F$ in terms of its derivative $F'$.

Together with Young’s inequality, i.e.

$$ab \leq \varepsilon a^r + \frac{1}{r} \left( \frac{1}{r\varepsilon} \right)^\frac{r}{r'} b^{r'},$$

for all $a, b \in [0,\infty)$. 

This manuscript is for review purposes only.
for $a, b \geq 0$, arbitrary $\varepsilon > 0$, and $r, r' \in [1, \infty]$ such that $\frac{1}{r} + \frac{1}{r'} = 1$, the $\mathcal{V}$-smoothing yields

$$
\text{err}_n \leq C_0 \lambda_n \| \Xi_n \|_{\mathcal{V}} \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta \| \tilde{u}_{n+1} - u^1 \|_{\mathcal{V}_\lambda}^{1-\theta}
$$

$$(2.14) \leq \frac{1}{2} \alpha_n d_{n+1} + \left( \frac{\theta + 1}{2} \right) \left( 1 - \theta \right) \left( \frac{\theta}{2} \alpha_n \right) \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta \leq \left( \frac{\theta}{2} \right) \alpha_n \| \Xi_n \|_{\mathcal{V}} \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta \leq \left( \frac{\theta}{2} \right) \alpha_n \| \Xi_n \|_{\mathcal{V}} \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta
$$

$$(2.14) \leq \frac{1}{2} \alpha_n d_{n+1} + \varepsilon \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta + 2^{-(1+\theta)} \left( \frac{\theta}{\varepsilon} \right) \left( 1 - \theta \right) \| \Xi_n \|_{\mathcal{V}} \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta$$

$$+ 2^{-(1+\theta)} \left( \frac{\theta}{\varepsilon} \right) \left( 1 - \theta \right) \| \Xi_n \|_{\mathcal{V}} \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta$$

where in the second inequality we used $\varepsilon = 1$, $r = \frac{2}{1+\theta}$ and $r' = \frac{2}{1+\theta}$, and in the
third inequality we used $r = \frac{1+\theta}{\alpha}$, $r' = 1 + \theta$ and an arbitrary $\varepsilon > 0$. Now the term
$$\| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^2$$
can again be handled by the tangential cone condition, since

$$\| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^2 \leq 2 \| F (\tilde{u}_n) + F' [\tilde{u}_n] (\tilde{u}_{n+1} - \tilde{u}_n) - y^1 \|_{\mathcal{V}}^2$$

$$+ 2 \| F (\tilde{u}_n) + F' [\tilde{u}_n] (u^1 - \tilde{u}_n) - y^1 \|_{\mathcal{V}}^2$$

$$\leq 2 C_{tc} \| F (\tilde{u}_{n+1}) - y^1 \|_{\mathcal{V}}^2 + 4 \eta \| F (\tilde{u}_n) - y^1 \|_{\mathcal{V}}^2$$

Plugging this in, we obtain the error estimate

$$\text{err}_n \leq 4 \varepsilon C_{tc} t_{n+1} + 8 \varepsilon \eta t_n + C \left( \theta, \varepsilon \right) \lambda_n^2 \alpha_n^{\theta - 1} \| \Xi_n \|_{\mathcal{V}}^2 + \frac{1}{2} \alpha_n d_{n+1},$$

with an arbitrary constant $\varepsilon > 0$ and the constant

$$C \left( \theta, \varepsilon \right) = 2^{-(1+\theta)} \left( \frac{\theta}{\varepsilon} \right) \left( 1 - \theta \right) \| \Xi_n \|_{\mathcal{V}} \| F' [\tilde{u}_n] (\tilde{u}_{n+1} - u^1) \|_{\mathcal{V}}^\theta$$

Choosing $\varepsilon$ as the solution of $4 \varepsilon C_{tc} = \frac{1}{4 C_{tc}^2}$, i.e. $\varepsilon := (16 C_{tc}^2)^{-1}$ and combining the
above estimate with (2.10), we obtain the following.

**Lemma 2.2** (Total error estimate). Let Assumptions 1, 2 and 4 hold and suppose
that $\tilde{u}_n \in D(F)$ is well-defined. If $\eta$ and $t_0$ are sufficiently small, then the total error estimate

$$\frac{1}{2} \alpha_n d_{n+1} + \frac{1}{4 C_{tc}^2} t_{n+1} \leq \alpha_n \Psi \left( 2 C_{tc} \alpha_n \right) + 2 \left( 1 + \frac{1}{4 C_{tc}^2} \right) \eta t_n$$

$$+ C \left( \theta, \frac{1}{16 C_{tc}^2} \right) \lambda_n^2 \alpha_n^{\theta - 1} \| \Xi_n \|_{\mathcal{V}}^2$$

with $\lambda_n$ and $\Xi_n$ as in (2.13) according to the specific data model, holds true almost
surely. If additionally Assumption 3 hold, then the total estimate obeys

$$\frac{1}{2} \alpha_n d_{n+1} + \frac{1}{4 C_{tc}^2} t_{n+1} \leq \left( 2 C_{tc} \right)^{1/2} \alpha_n \Psi \left( \alpha_n \right) + 2 \left( 1 + \frac{1}{4 C_{tc}^2} \right) \eta t_n$$

$$+ C \left( \theta, \frac{1}{16 C_{tc}^2} \right) \lambda_n^2 \alpha_n^{\theta - 1} \| \Xi_n \|_{\mathcal{V}}^2$$

almost surely.
The nonlinear term, \( \eta_t u_n \), in the right-hand side of the previous inequality can now be bounded by means of standard techniques. While the details are included in the supplementary material, the following assumption on the decay rate of the regularization parameters is required to reach the final result.

**Assumption 5 (Regularization parameters).** The regularization parameters \( \alpha_n \) are chosen such that

\[
\alpha_0 \leq 1, \quad \alpha_n \searrow 0, \quad 1 \leq \frac{\alpha_n}{\alpha_{n+1}} \leq C_{\text{dec}} \quad \text{for all} \quad n \in \mathbb{N}.
\]

In the supplementary material we prove that, under Assumption 5 and for sufficiently small \( t_0 \) and small \( \eta \), it follows that

\[
t_{n+1} \leq \gamma_{nL} \left( 2C_{\text{tc}} \frac{1}{t} \alpha_n \Psi (\alpha_n) + C \left( \frac{1}{16C_{\text{tc}}^2} \right) \lambda_n^2 \alpha_n^{\theta-1} \| \Xi_n \|^2 \right),
\]

for all \( n \) with

\[
(2.18) \quad \gamma_{nL} := \frac{4C_{\text{tc}}}{1 - 8\eta C_{\text{tc}} C_{\text{bnoise}} (1 + \frac{1}{4C_{\text{tc}}^2})}, \quad \text{and} \quad C_{\text{bnoise}} := C_{\text{dec}}^{1+\frac{1}{2}} + 2.
\]

Substituting this into (2.16) and dropping the \( t_{n+1} \) term on the left-hand side, we get after division by \( \alpha_n \) the following result by adjusting the iteration number accordingly.

**Lemma 2.3 (Final recursive error estimate).** Let Assumptions 1-4 hold and assume that \( \hat{u}_n \in D(F) \) is well-defined. If \( t_0 \) and \( \eta > 0 \) are sufficiently small, then we have the error bound

\[
d_n \leq C_1 (2C_{\text{tc}})^{\frac{1}{2}} \Psi (\alpha_{n-1}) + C_2 \lambda_{n-1}^{\frac{1}{2}} \alpha_{n-1}^{\theta-2} \| \Xi_{n-1} \|^2_{\mathcal{V}},
\]

(2.19)

\[
(2.20) \quad \leq C_1 (2C_{\text{tc}} C_{\text{dec}})^{\frac{1}{2}} \Psi (\alpha_n) + 2C_2 \lambda_n^{\frac{1}{2}} \alpha_n^{\theta-2} \| \Xi_n \|^2_{\mathcal{V}},
\]

a.s. with constants

\[
C_1 := 2 \left( 1 + 2 \left( 1 + \frac{1}{4C_{\text{tc}}^2} \right) \eta \gamma_{nL} (C_{\text{dec}})^{1+\frac{1}{2}} \right),
\]

\[
C_2 := 2 \left( 1 + 4 \left( 1 + \frac{1}{4C_{\text{tc}}^2} \right) \eta \gamma_{nL} \right) C \left( \frac{1}{16C_{\text{tc}}^2} \right).
\]

The above error estimate (2.19) plays an important role in the following analysis and discussion.

**2.2. Well-definedness of the method.** As a first application, we are now in position to prove the well-definedness of the dIRGNM in (2.2). Recall therefore that we have \( W = Y_n \) and hence \( \lambda_n = \sigma \) and \( \Xi_n = \xi_n \) in this case.

**Theorem 2.3.** Let Assumptions 1-5 hold and let \( \hat{u}_0 \in D(F) \). Assume that there exists an open ball \( B_r (u^1) \subset D(F) \) around \( u^1 \) in \( D(F) \) and that \( \| \xi_n \|_{\mathcal{V}} \) in (1.6) satisfies the deviation inequality

\[
(2.21) \quad \mathbb{P} \left[ \| \xi_n \|_{\mathcal{V}} - \mathbb{E} \| \xi_n \|_{\mathcal{V}} \right] \geq \delta \right] \leq 2 \exp (-c \delta),
\]

for all \( \delta > 0 \) with some constant \( c > 0 \). Suppose furthermore that \( t_0 \) and \( \eta > 0 \) are sufficiently small. Then, if both \( \alpha_n > 0 \) and \( \sigma^2 \alpha_n^{-2} \) are sufficiently small, then we have also \( \hat{u}_{n+1} \in B_r (u^1) \subset D(F) \) with probability at least

\[
1 - \exp \left( -C \frac{\alpha_n^{\frac{1}{2} - 1}}{\sigma} \left( \frac{r}{C} - \left( \sqrt{\Psi (\alpha_n)} + \frac{\sigma}{\alpha_n^{\frac{1}{2} - 1}} \mathbb{E} \| \xi_n \|_{\mathcal{V}} \right) \right) \right).
\]
Proof. Let \( X := \|\xi_n\|_{\mathcal{Y}} \). By Jensen’s inequality, we have \( |E|X|^2 \leq \|E(X)^2 \| \). As \( E|X|^2 < \infty \) by the Hilbert-Schmidt property of the embedding \( \iota : \mathcal{Y} \to \mathcal{Y} \), this shows that \( E|X| < \infty \). Now suppose that \( \tilde{u}_n \in D(F) \). Then by Lemma 2.3, the error estimate (2.19) holds true. Thus, due to \( \sqrt{a+b} \leq \sqrt{a}+\sqrt{b} \) and adjusting the iteration number appropriately, we have

\[
\|\tilde{u}_{n+1} - u^1\|_X \leq C \left( \sqrt{\Psi(\alpha_n)} + \frac{\sigma}{\alpha_n^{\frac{1}{2}}} \right),
\]

as well, we have

\[
\mathbb{P} \left( \|\tilde{u}_{n+1} - u^1\|_X \leq r \right) \geq \mathbb{P} \left( C - \frac{\sigma}{\alpha_n^{\frac{1}{2}}} \|X - EX\| \leq r - C \left( \sqrt{\Psi(\alpha_n)} + \frac{\sigma}{\alpha_n^{\frac{1}{2}}} \left\|EX\right\| \right) \right).
\]

Remark 2.4. We provide some extended discussion below.

- A deviation inequality of the form (2.21) is e.g. satisfied for Gaussian white noise \( \xi_1 \), see [21, Thm. 2.1.20].
- The assumption that both \( \alpha_n > 0 \) and \( \sigma^2 \alpha_n^{\frac{1}{2}-2} \) are sufficiently small is natural, as both terms tend to 0 anyway during the iteration.
- Note that the stated probability in Theorem 2.3 tends to 0 as \( \alpha_n \to 0 \) and \( \sigma^2 \alpha_n^{\frac{1}{2}-2} \to 0 \). As a corollary, Theorem 2.3 implies that all iterates from a certain \( n \in \mathbb{N} \) onward will be well-defined with overwhelming probability under reasonable assumptions or, in plain words, that all iterates will be well-defined with overwhelming probability if the initial parameters \( \alpha_0, \tilde{u}_0 \) are chosen carefully.
- Theorem 2.3 shows that well-definedness probability of (2.2) with \( W = Z_n \) is larger than that with \( W = Y_n \), which highlights the focus of our current work.

3. Error bounds for different observation models. In this section, we provide discussion on the error bounds or asymptotical behaviour for the different observation models \( W \in \{Z_N, Z_n, Y_n\} \) based on the recursive error estimate in Lemma 2.3.

3.1. Error bound for the non-dynamic IRGNM. Let us start by analyzing the non-dynamic IRGNM, i.e. (2.2) with \( W = Z_N \). According to (2.13) we have

\[
\lambda_n = \sigma/N \quad \text{and} \quad \Xi_n = \sum_{i=1}^{N} \xi_i \overset{D}{=} \sqrt{N} \xi_1, \quad \text{so that (2.17) specializes to}
\]

\[
d_n \leq C_1 (2C_{te}C_{dec})^\frac{1}{2} \Psi(\alpha_n) + 2C_2 \|\xi_1\|_{\mathcal{Y}}^2 \frac{\sigma^2}{N\alpha_n^{\frac{1}{2}}},
\]
where $\|\xi_1\|_Y^2$ can be handled as a (random) constant.

To determine an optimal regularization parameter $\alpha_N$ for (3.1), let us informally search for the infimal value

$$\inf_{\alpha > 0} \left[ \Psi(\alpha) + \frac{\sigma^2}{\alpha^{2-\theta}} \right] = \inf_{\alpha > 0} \left[ (-\varphi)^* \left( -\frac{1}{\alpha} \right) + \frac{\sigma^2}{\alpha^{2-\theta}} \right].$$

If we set

$$\Psi_\theta(t) := (-\varphi)^* \left( -t^{\frac{1}{2-\theta}} \right),$$

then we can compute this infimum by means of Fenchel duality as

$$\inf_{\alpha > 0} \left[ \Psi(\alpha) + \frac{\sigma^2}{\alpha^{2-\theta}} \right] = \inf_{\alpha' > 0} \left[ \Psi_\theta(\alpha') + \alpha' \sigma^2 \right] = - \sup_{\alpha' > 0} \left[ -\Psi_\theta(\alpha') - \alpha' \sigma^2 \right] = (-\Psi_\theta)^* (\sigma^2),$$

and by the equality condition in Young’s inequality, this infimum is attained for

$$\frac{1}{\alpha^{2-\theta}} \in \partial (-\Psi_\theta)^* (\sigma^2).$$

Now we define the stopping criterion as

$$(3.2) \quad n_* = \min \left\{ n \in \mathbb{N} \mid \frac{1}{\alpha_n^{2-\theta}} \geq \partial (-\Psi_\theta)^* \left( \frac{\sigma^2}{N} \right) \right\}.$$ 

and obtain the following

**Theorem 3.1.** Let Assumptions 1-4 hold. Assume that all iterates $\tilde{u}_n$ are well defined with probability larger than $1 - \delta$ and that $\| F(\tilde{u}_0) - y \|_Y$ and $\eta$ are sufficiently small. Then for the single fixed observation model we obtain

$$\| \tilde{u}_{n_*} - u^\dagger \|_X = \mathcal{O} \left( (-\Psi_\theta)^* \left( \frac{\sigma^2}{N} \right) \right),$$

with probability larger than $1 - \delta$ and $n_*$ chosen by (3.2).

The proof is straightforward by the proposed parameter choice rule and we skip it here. In particular, Theorem 3.1 can be viewed as an extended convergence rate result compared with the cIRGNM in the deterministic setting [35].

**3.2. Asymptotical analysis for infinitely many (averaged) observations.**

This part focuses on the most interesting case with infinitely many (averaged) observations, i.e. an infinite sequence of observations $Y_n$ as in (1.6).

Note that in case of $W = Y_n$, i.e. $\lambda_n = \sigma$ and $\Xi_n = \xi_n$, the recursive error bound (2.20) contains the terms $\sigma^2$ and $\|\xi_1\|_Y^2$, which will in general not tend to 0. This shows that no convergence (and hence no assimilation) can be expected theoretically from the corresponding scheme. Meanwhile, if we consider (1.10), where $\lambda_n = \sigma/n$ and $\Xi_n = \sum_{i=1}^n \xi_i \overset{d}{=} \sqrt{n} \xi_1$, the recursive error bound (2.20) becomes

$$(3.3) \quad d_n \leq C_1 \left( 2C_{1c}C_{dec} \right)^{\frac{1}{2}} \Psi(\alpha_n) + 2C_2 \|\xi_1\|_Y^2 \frac{\sigma^2}{n\alpha_n^{2-\theta}},$$

where $\|\xi_1\|_Y^2$ can be handled as a (random) bounded constant.

Thus as a central result, we obtain the following:
Theorem 3.2. Let Assumptions 1-4 hold, suppose that \( \| F(\hat{u}_0) - y^\dagger \|_Y \) and \( \eta \) are sufficiently small, and assume that all iterates \( \hat{u}_n \) are well-defined a.s. If \( \alpha_n \) is chosen such that
\[
\alpha_n \downarrow 0 \quad \text{and} \quad n\alpha_n^{2-\theta} \nearrow \infty,
\]
then for averaged observations we have \( d_n \to 0 \) as \( n \to \infty \) such that dIRGNM converges a.s. for infinitely many averaged observations.

If in addition, Assumption 5 holds and we choose the regularization parameter \( \alpha_n \) such that
\[
\frac{1}{\alpha_n^{2-\theta}} \in \partial (-\Psi_\theta)^* \left( -\frac{\sigma^2}{n} \right),
\]
it then follows that
\[
\| \hat{u}_n - u^\dagger \|_X^2 = O \left( (-\Psi_\theta)^* \left( -\frac{\sigma^2}{n} \right) \right).
\]

Proof. As all iterates are well-defined by assumption, the first result now follows immediately noticing both terms in (3.3), by adopting to the proposed parameter choice rule, vanish when \( n \to \infty \).

Concerning the second results, the proposed parameter choice rule then allows us to obtain
\[
d_n \leq C \left( (-\Psi_\theta)^* \left( -\frac{\sigma^2}{n} \right) \right),
\]
which proves the claim.

Remark 3.3. 1. The assumption that all iterates are well-defined a.s. is reasonable in view of Theorem 2.3 and can be interpreted as a conditioning on some event with overwhelming probability. To derive overall rates of convergence in expectation, one would have to specify what is considered as the reconstruction if \( \hat{u}_n \) is no longer well-defined.

2. Theorem 3.2 yields a qualitative result showing that by (2.2) we can obtain a vanishing asymptotical behavior for the dIRGNM (1.10) by choosing the regularization parameter appropriately. Though the index \( \theta \) might be unknown, we can slightly modify the condition on \( \alpha_n \) such that
\[
\alpha_n \downarrow 0 \quad \text{and} \quad n\alpha_n^{2-\theta} \nearrow \infty,
\]
are sufficient to guarantee the same result. A natural choice would be \( \alpha_n \sim n^{-\beta} \), i.e., for \( \beta \in (1/2, 1 - \theta/2) \) and we will examine the numerical performance in Section 4 for different choices of \( \beta \).

3. For Hölder-type source conditions, we have \( \varphi(t) = ct^\nu \) with some \( 0 \leq \nu < 1 \) and \( c > 0 \). Straight-forward computations show
\[
(-\varphi)^* (-s) \sim s^{\frac{\nu}{\nu-1}}, \quad s > 0
\]
and hence
\[
\Psi(\alpha_n) \sim \alpha_n^{\frac{\nu}{\nu-1}}.
\]
Then by choosing \( \alpha_n \sim n^{-\frac{1}{2\nu-\theta(1-\nu)}} \), we obtain an asymptotical decaying rate
\[
\| \hat{u}_n - u^\dagger \|_X^2 = O(n^{-\frac{1}{2\nu-\theta(1-\nu)}}).
\]
3.3. Analysis for finitely many averaged observations. Let us now consider the case that we have access to $N \in \mathbb{N}$ sequential observations. Our aim is to use the dIRGNM iteration. Clearly, the iteration should be stopped after the $N$th iteration, as no further data is available then, and additional iterations are not immediately well-defined. In this case, (3.3) holds true for all $n \leq N$. As a consequence of the above considerations, we obtain the following result:

**Theorem 3.4.** Let Assumptions 1-4 hold. Assume that all iterates $\hat{u}_n$ are well defined with probability larger than $1 - \delta$ and that $\| F(\hat{u}_0) - y \|_Y$ and $\eta$ are sufficiently small. Choose the regularization parameters $\alpha_1, ..., \alpha_{N-1}$ arbitrary and $\alpha_N$ such that

$$\frac{1}{\alpha_N^2} \in \partial(-\Psi_{\theta})^* \left( -\frac{\sigma^2}{N} \right).$$

Then we obtain the final estimate

$$\| \hat{u}_N - u^\dagger \|_X^2 = \mathcal{O} \left( (-\Psi_{\theta})^* \left( -\frac{\sigma^2}{N} \right) \right),$$

with probability larger than $1 - \delta$. The implications of comparing Theorems 3.1 and 3.4 is that the cIRGNM applied to the averaged datum $Z_N$ and the dIRGNM applied to sequential data $Z_1, ..., Z_N$, yield exactly the same rates of convergence. As a consequence, the accuracy provided by both these algorithms is comparable. However, the dIRGNM has several advantages as mentioned in the introduction including that it can be executed alongside the measurement process. Another potential advantage is that the dIRGNM could potentially inform when to stop the experiment based on the (preliminary) reconstructions $\hat{u}_n$ that we obtain on the fly.

The intrinsic advantages of the dIRGNM also allow us to design a blended IRGNM (bIRGNM) that first runs the dIRGNM for the sequential finitely many (averaged) observation $Z_n$ with $n \leq N$. Then we use the preliminary estimate $\hat{U}_{dIRGNM}$ obtained by dIRGNM as the initial estimate that we use to execute the cIRGNM using the final averaged observation $Z_N$.

3.4. Summary of Algorithms. In this subsection we provide the summary of the cIRGNM and its two proposed variants, i.e. dIRGNM and bIRGNM.

---

**Algorithm 3.1** Classical iterated regularized Gauss-Newton method (cIRGNM)

**inputs:** $\hat{u}_0, u^\dagger, \alpha_0, M, W, C_{dec} > 1$

**for** $n = 1, \ldots, M$ **do**

- Compute

$$\hat{u}_n := \arg\min_{\hat{u} \in \mathcal{X}} \left\{ S \left( F(\hat{u}_{n-1}) + F^\prime [\hat{u}_{n-1}] (u - \hat{u}_{n-1}); W \right) + \alpha_n \left\| u - u^\dagger \right\|_X^2 \right\},$$

with $S$ defined in (1.8) and $\alpha_n = \alpha_0 C_{dec}^{-n}$.

**end**

**output:** $\hat{u}_M$.

---

The cIRGNM with generic observations $W$ is displayed in Algorithm 3.1 where we have used the standard choice of regularization parameter $\alpha_n$. Usually, one can
use $\hat{u}_0 = u^*$ to start the iteration but we keep them differently as we need to do so in order to define the bIRGNM later. For the purpose of monitoring performance, we select a maximum number of $M$ iterations. However, we recognise that in practice this algorithm needs to be stopped, for example, via the discrepancy principle.

Algorithm 3.2 Dynamic iterated regularized Gauss-Newton method ($\text{dIRGNM}$)

**inputs**: $\hat{u}_0$, $\alpha_0$, $N$, $\{Y_i\}_{i=1}^N$, $\beta > 0$

**for** $n = 1, \ldots, N$ **do**

1. Collect data $Y_n$.
2. Compute $\hat{u}_n := \arg\min_{\hat{u} \in X} \left[ S(F(\hat{u}_{n-1}) + F'(\hat{u}_{n-1})(u - \hat{u}_{n-1}); Z_n) + \alpha_n \|u - \hat{u}_0\|^2_X \right]$, where $Z_n = n^{-1} \sum_{i=1}^n Y_i$, and $\alpha_n = \alpha_0 n^{-\beta}$.

**end**

**outputs**: $U_N^{\text{dIRGNM}} = \hat{u}_N$ and $\alpha_N$.

Algorithm 3.3 Blended iterated regularized Gauss-Newton method ($\text{bIRGNM}$)

**input**: $\hat{u}_0$, $\alpha_0$, $N$, $M$, $\{Y_i\}_{i=1}^N$, $C_{dec} > 1$, $\beta > 0$.

**First Part**: Compute $(U_N^{\text{dIRGNM}}, \alpha_N) = \text{dIRGNM}(\hat{u}_0, \alpha_0, N, \{Y_i\}_{i=1}^N, \beta)$.

**Second Part**: Set $\hat{w}_0 = U_N^{\text{dIRGNM}}$ and $\tilde{\alpha}_0 = \alpha_N$. Compute $\hat{w}_n = \text{cIRGNM}(\hat{w}_0, \hat{u}_0, \tilde{\alpha}_0, M, Z_N, C_{dec})$ with $Z_N = N^{-1} \sum_{i=1}^N Y_i$.

**output**: $\hat{w}_M$.

The proposed dIRGNM is summarised in Algorithm 3.2. We recall that in contrast to the cIRGNM in which the observations are fixed throughout the entire algorithm, the dIRGNM allows us to use observations $Y_1, Y_2, \ldots, Y_n$ as they become available. More specifically, at each iteration $n$, we use $Z_n$ i.e. the average of the $n$ available observations, in order to produce the estimate $\hat{u}_n$. While the previous section ensures the asymptotic convergence of the dIRGNM, in practical settings we have only access to limited number of experiments. Therefore, we propose the blended version shown in Algorithm 3.3. The first part of this blended IRGNM consists of applying the dIRGNM with $N$ iterations. For the second part we use the cIRGNM using the final estimate of the dIRGNM as starting point, as well as the average of the $N$ measurements collected upon completion of the dIRGNM. Furthermore, for the second part we choose the regularisation parameter $\tilde{\alpha}_n = \alpha_N C_{dec}^{-n}$ where $\alpha_N$ is the final value.
computed with the diRGNM.

As discussed in Section 1, within the classical setting we would have to wait until all observations are acquired, and use the ciRGNM with the average of all these observations (i.e. with $W = Z_N$). However, the numerical experiments from the following section show that the blended version can offer significant computational advantages. Indeed, by the time all measurements have been collected and assimilated with the diRGNM encoded in the blended version, the estimate of the unknown already shows good levels of accuracy. Consequently, convergence of the second part of the bIRGNM is then achieved in much fewer iterations than those required by the ciRGNM. For problems where an iteration of the diRGNM can be computed within the time-scale of measurement acquisition, faster estimates can be obtained using the blended algorithm compared to the classical one.

For all the algorithms we adopt the standard practice of starting the iteration using the same element, $\hat{u}_0$, that appears in the stabilization term of the cost functional (2.1) that we minimize at each iteration of these algorithms. However, it is worth emphasizing, that for the second part of Algorithm 3.3, we initialise the iterations using the estimate from diRGNM while keeping the same initial guess, $\hat{u}_0$, in the stabilisation term.

4. Numerical Experiments. In this section, we provide two numerical examples verifying the theoretical finding of current work. Our focus mainly concentrates on the convergence of diRGNM for infinitely many (averaged) observation, i.e. Theorem 3.2, and the comparison between bIRGNM and ciRGNM when the same finitely many observation is given, i.e. Theorems 3.1 and 3.4.

4.1. Example 1. In the first benchmark example, the unknown is the potential coefficient, $u$, in the following PDE

$$-\Delta p + up = f, \quad \text{in } \Omega,$$

$$p = g, \quad \text{on } \partial \Omega,$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with Lipschitz boundary $\partial \Omega$, $f \in L^2(\Omega)$ and $g \in H^{3/2}(\Omega)$. We define the parameter-to-measurements operator $F : L^2(\Omega) \rightarrow L^2(\Omega)$ via $p = F(u)$, where $p$ is the unique solution of (4.1).

Note that this operator obeys the tangential cone condition as shown in [23, Example 4.2]. Thus Assumption 2 is satisfied. To treat white noise, we choose $\mathcal{V} = H^a(\Omega)$ with $a > 1 = d/2$ to ensure $\|\xi_1\|_{\mathcal{V}} < \infty$ a.s. . Furthermore, the Fréchet derivative $v = F'[u]h$ for $u \in L^2(\Omega), h \in L^2(\Omega)$ can, as also shown in [23, Example 4.2], be expressed as the solution to

$$-\Delta v + uv = -hF(u), \quad \text{in } \Omega,$$

$$v = 0, \quad \text{on } \partial \Omega.$$

Note that the weak form of (4.2) has unique solution $v \in H^2(\Omega) \cap H^1_0(\Omega)$. This representation now allows us to verify Assumption 4 whenever $a < 2$: By means of elliptic regularity theory, the operator $F' : L^2(\Omega) \rightarrow H^2(\Omega)$ is bounded (in fact a homomorphism), and thus it follows from [27, Rem. 2.6] that Assumption 4 is satisfied with $\theta = \frac{a}{2}$ and $C\theta = \|F'[u]\|_{L^2 \rightarrow H^2}$. Finally, we can also verify Assumption 1 similar to [27, Ex. 2.2] by using the tangential cone condition. Precisely, if $u^1 \in H^a(\Omega)$ for some $s > 0$, then (2.7) holds true with $\varphi(\lambda) = C\lambda^{s+1}$ with some constant $C > 0$.

In order to verify our theoretical findings let us first discuss some aspects which are relevant to the implementation of the IRGNM for this example. We start by deriving
the first-order optimality conditions for the minimization procedure in Algorithms 3.1-3.3. We note that the cost functionals in all these algorithms only vary in the measurements that they employ. Hence, here we focus only on the generic form of the minimization given in (2.2) and which, for the example under consideration, can be written as the minimiser of

\[ Q(u, v) := \frac{1}{2} \| F(\hat{u}_n) + v \|_{L^2(\Omega)}^2 - \langle W, F(\hat{u}_n) + v \rangle_{L^2(\Omega)} + \frac{\alpha_n}{2} \| u - \hat{u}_0 \|_{L^2(\Omega)}^2, \]

where \( v \) satisfies the constraint

\[ -\Delta v + \hat{u}_n v = (\hat{u}_n - u) F(u_n), \quad \text{in } \Omega, \]
\[ v = 0, \quad \text{on } \partial \Omega. \]

Let us define the Lagrangian

\[ L(v, \lambda, u) := Q(u, v) + \langle -\Delta v + \hat{u}_n v - (\hat{u}_n - u) F(\hat{u}_n), \lambda \rangle_{L^2(\Omega)}, \]

which we now employ to solve the unconstrained optimisation problem. To this end, we derive expression for the optimality conditions:

\[ D_v L(v, \lambda, u) \tilde{v} = 0, \]
\[ D_\lambda L(v, \lambda, u) \tilde{\lambda} = 0, \]
\[ D_u L(v, \lambda, u) h = 0, \]

for all \((\tilde{v}, \tilde{\lambda}, h) \in V \times V \times L^2(\Omega)\). It follows trivially that the condition (4.5) yields directly the constraint (4.3). Furthermore, note that

\[ D_v L(v, \lambda, u) \tilde{v} = \langle F(\hat{u}_n) + v - W, \tilde{v} \rangle_{L^2(\Omega)} + \langle -\Delta \tilde{v} + \hat{u}_n \tilde{v}, \lambda \rangle_{L^2(\Omega)}, \]

which, after integrating by parts and applying boundary conditions yields

\[ D_v L(v, \lambda, u) \tilde{v} = \langle F(\hat{u}_n) + v - Y_n - \Delta \lambda + \hat{u}_n \lambda, \tilde{v} \rangle_{L^2(\Omega)}. \]

Hence, (4.6) is equivalent to the following adjoint equation for \( \lambda \in V \)

\[ -\Delta \lambda + \hat{u}_n \lambda = W - F(\hat{u}_n) - v, \]

with homogeneous Dirichlet boundary conditions. Finally, it is easy to see that (4.7) is equivalent to

\[ u = \hat{u}_0 - \frac{1}{\alpha_n} \lambda F(\hat{u}_n). \]

We use the previous equation in (4.3) which we then combine with (4.9) to obtain the linear system for \((\lambda, v)\) given by

\[ \begin{pmatrix} -\Delta + \hat{u}_n & I \\ -\alpha_n^{-1}(F(\hat{u}_n))^2 & -\Delta + \hat{u}_n \end{pmatrix} \begin{pmatrix} \lambda \\ v \end{pmatrix} = \begin{pmatrix} W - F(\hat{u}_n) \\ (\hat{u}_n - \hat{u}_0) F(\hat{u}_n) \end{pmatrix}, \]

where \( I \) denotes the identity in \( L^2(\Omega) \). At a given iteration level \( n \), we solve (4.11) and use \( \lambda \) in (4.10) to compute the update \( \hat{u}_{n+1}. \) Replacing with \( W \) with \( Z_N \) and \( Z_n \) gives the corresponding updates for Algorithms 3.1-3.2, respectively.
4.1.1. Numerical results. We consider two experiments using a domain \( \Omega = [0,1]^2 \). For the first set we consider a smooth truth defined by

\[
u^\dagger(x,y) = \exp\left[-100\left((x-0.3)^2 + (y-0.7)^2\right)\right] + \frac{1}{2} \exp\left[-100\left((x-0.35)^2 + (y-0.35)^2\right)\right],
\]

while for the second we use

\[
u^\dagger(x,y) = \begin{cases} 1, & \text{if } (x-0.3)^2 + (y-0.7)^2 < 0.15^2, \\ 0.5, & \text{if } (x,y) \in [0.6,0.8] \times [0.2,0.5], \\ 0, & \text{otherwise}. \end{cases}
\]

In the top row of Figure 1 we show the plots of these functions. For both cases we define \( f(x,y) = (x+y)u^\dagger(x,y) \) and \( g(x,y) = (x+y)|_{\partial \Omega} \), so that the noise-free data is given by \( F(u^\dagger) = p^\dagger(x,y) = (x+y) \).

We implement Algorithms 3.1-3.3 in MATLAB and use pde-toolbox to solve equation (4.1) as well as the linear system (4.11). We use a mesh which consists of 7444 linear elements and 3837 nodes. Using the analytical solution, and thus avoiding inverse crimes, we evaluate the noise free observations on the nodes of the computational mesh, and produce the sequence of synthetic observations \( Y_n \) (see e.g. (1.6)), using a Gaussian random vector \( \xi_n \in \mathbb{R}^{3837} \) with zero mean and standard deviation \( \sigma = 5 \times 10^{-4} \). For all algorithms we use \( \hat{u}_0(x,y) = 0 \) (for all \( (x,y) \in \Omega \)) and \( \alpha_0 = 10^{-3} \). For Algorithm 3.1 and for the second part of Algorithm 3.2 we use \( C_{dec} = 1.5 \) in the definition of \( \alpha_n \).

To assess the convergence of the dIRGNM, we implement Algorithm 3.2 with \( N = 10^4 \) for various selections of \( \beta \) in the definition of \( \alpha_n \). At each iteration we compute the relative error with respect to the truth defined by

\[
E_n = \frac{\|\hat{u}_n - u^\dagger\|_{L^2}}{\|u^\dagger\|_{L^2}}.
\]

As suggested in Item 2 of Remark 3.3, the theory requires to choose \( \beta \in (1/2, 1 - \theta/2) \) to ensure convergence of the dIRGNM. Such a remark is confirmed in Figure 2 where relative error for various choices of \( \beta \) are displayed with the smooth (resp. discontinuous) truth. For validation purposes, in these plots we also display the relative error w.r.t the truth that we obtain from applying the cIRGNM with noise-free observations (i.e. we set \( W = F(u^\dagger) \)). The reconstruction estimates obtained are provided in Figures ??-?? of the supplementary material. Though different choices of \( \beta \) yield decaying relative error in the first hundred iterations, if \( \beta > 1 \), we do
Fig. 2. Example 1. Relative $L^2$-errors obtained using the dIRGNM with various choices of $\beta$ in the definition of $\alpha_n := \alpha_0 n^{-\beta}$. We include both cases with smooth (left) and discontinuous (right) truth.

Fig. 3. Example 1. Left: Relative $L^2$-errors for the case with the smooth (top) and discontinuous (bottom) truth obtained using the cIRGNM with a single data set (blue-circle line), the cIRGNM with averaged data sets (black-diamond line) as well as the noise-free cIRGNM (red-asterisk line). Right: Relative $L^2$-errors for the case with the smooth (top) and discontinuous (bottom) truth obtained from the second part of the bIRGNM (magenta-square line) as well as the noise-free cIRGNM (red-asterisk line).

obtain some amplified relative error when a sufficiently large number of observations are averaged.

We now investigate the case where we have access to only finitely many observations and compare the performance of the (online+offline) bIRGNM and the (offline) cIRGNM. More precisely we consider $N = 500$ synthetic set of observations which we use for (i) the bIRGNM (Algorithm 3.3) with sequential averaged obser-
vations $W = \{Z_l\}_{l=1}^{500}$ and (ii) the cIRGNM (Algorithm 3.1) with $W = Z_{500}$. To further demonstrate the advantage of using multiple observations we also implement the cIRGNM with $W = Y_1$ which corresponds to the standard approach of aiming at reconstructing the unknown with a single data set. The relative error w.r.t the truth obtained with the cIRGNM are shown in the left panels of Figure 3. As we expect when noisy observations are employed, the error starts increasing after several iterations due to the ill-posedness of the inverse problem. Since the noise level of the averaged observation is smaller than that of a single observation, it comes as no surprise that the cIRGNM with $W = Z_{500}$ reaches a lower minimum value (displayed on the plots).

For the dIRGNM encoded in the first part of the bIRGNM we use $\beta = 1.2$ informed by the previous experiments that suggest that we can safely go slightly above the value predicted by the theory in order to achieve faster convergence without the risk of compromising accuracy. The error from the iterations during the first part of the bIRGNM corresponds to the first $N = 500$ iterations shown in Figure 2 (for $\beta = 1.2$). In the right panels of Figure 3 we show the iterations achieved during the second part of the bIRGNM (Algorithm 3.3). The minimum relative error achieved by bIRGNM is shown on the corresponding plots. When we compare left and right panels, we can notice that the minimum relative error value attained by the bIRGNM is very similar to the one obtained using the cIRGNM with $W = Z_{500}$. However, we notice the second part of bIRGNM reaches the minimum after only a few iterations. In fact, for case with a smooth truth, the second part of the bIRGNM takes only one iteration to reach the minimum value.

### 4.2. Example 2

The context of this example is electrical impedance tomography (EIT) [8]. For the forward problem we employ the complete electrode model (CEM) introduced in [45]. We consider a medium with physical domain denoted by $\Omega$ and an electric conductivity $\kappa$. A set of electrodes $\{e_l\}_{l=1}^{m_e}$ are attached on the boundary $\partial\Omega$ with contact impedance $\{z_l\}_{l=1}^{m_e}$. The aim of the CEM is to compute the electric potential $\nu$ inside $\Omega$ as well as the voltages $\{V_l\}_{l=1}^{m_e}$ on the electrodes. The governing equations are

\[
\begin{align*}
\nabla \cdot (\exp(u) \nabla \nu) &= 0, \quad \nu = \log(\kappa), \quad \n \in \Omega, \\
\nu + z_l \exp(u) \nabla \nu \cdot \n & = V_l, \quad \nu = e_l, \quad l = 1, \ldots, m_e, \\
\nabla \nu \cdot \n & = 0, \quad \nu \in \partial\Omega \setminus \bigcup_{l=1}^{m_e} e_l, \\
\int_{e_l} \exp(u) \nabla \nu \cdot \n \, ds & = I_l, \quad \nu = e_l, \quad l = 1, \ldots, m_e,
\end{align*}
\]

where $u = \log(\kappa)$, $\n$ denotes the outward normal vector on the boundary and $I_l$ ($l = 1, \ldots, m_e$) is the current injected through the electrode $e_l$. We require that

$$I = (I_1, \ldots, I_{m_e}) \in \mathbb{R}_{0}^{m_e} \equiv \left\{ V \in \mathbb{R}_{0}^{m_e} \left| \sum_{m=1}^{m_e} V_l = 0 \right. \right\},$$

which implies conservation of charge. For $\kappa = \exp(u) \in C(\overline{\Omega})$, the weak form (4.12) has a unique solution $(\nu, V) \in H^1(\Omega) \times \mathbb{R}_{0}^{m_e}$ [45].

For the inverse problem we employ $n_p$ injection patterns $\mathbf{I}_j = \{I_{j,k}\}_{k=1}^{m_e}$ ($j = 1, \ldots, n_p$), and pose the EIT problem of estimating the unknown (log) conductivity $u$ from measurements of the corresponding voltages $\mathbf{V}_j = \{I_{j,k}\}_{k=1}^{m_e}$ ($j = 1, \ldots, n_p$).
In this example, to treat more a general setting that enables us to incorporate
smoothness of the unknown, we consider the following weighted $L^2$ space

\begin{equation}
\mathcal{X} = \{ u \in L^2(\Omega) \mid ||C^{-1/2} u||_{L^2(\Omega)} \leq \infty \},
\end{equation}

where $C$ is a covariance operator induced by a correlation function as follows

\begin{equation}
C[u](x) = \int_{\Omega} u(x') c(x, y') dx' dx'.
\end{equation}

We choose a Matérn correlation function given by

\begin{equation}
c(x, x') := c_0 \frac{2^{1-\nu}}{\Gamma(\nu)} K_\nu \left( \frac{|x - x'|}{\ell} \right) \left( \frac{|x - x'|}{\ell} \right)^\nu.
\end{equation}

where $c_0 \in \mathbb{R}^+$ is a scaling factor, $\nu \in \mathbb{R}^+$ is a smoothness parameter, $\ell \in \mathbb{R}^+$ denotes the length-scale, $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind. The forward map $F : \mathcal{X} \to \mathbb{R}^{n_e m_e}$ is defined by $F(u) = V \equiv (V_1, \ldots, V_{n_p})$.

The question of whether the tangential cone condition is satisfied in this example has received considerable attention during the recent decade, see e.g. [37, 38] and the references therein. However, whether Assumption 2 holds remains an open problem. Concerning Assumptions 1 and 4, the same comments as in the previous example apply.

For the implementation of the IRGNM and its variants, we use the fact that, for the weighted $L^2$ space, the IRGNM step (1.4) is equivalent to

\begin{equation}
\hat{u}_{n+1} - \hat{u}_n = (F'[\hat{u}_n]^* F'[u_n] + \alpha_n C^{-1})^{-1} \left( F'[\hat{u}_n]^* (W - F(\hat{u}_n)) + \alpha_n C^{-1} (\hat{u}_0 - \hat{u}_n) \right),
\end{equation}

which, in the finite-dimensional case, using the Woodbury Lemma can be written as

\begin{equation}
\hat{u}_{n+1} = u_0 + C F'[u_n]^* \left( C F'[u_n] C F'[u_n]^* + \alpha_n I \right)^{-1} \left( W - F(u_n) - F'[u_n](u_0 - u_n) \right).
\end{equation}

We use a discretised version of the update formula from (4.17). The midpoint rule is applied for the discretisation $C$ in (4.14). The parameters for the Matérn correlation function (4.15) are $c_0 = 25, \nu = 1.0$ and $\ell = 0.1$. For the discretised Fréchet derivative $F'(u_n)$ we use the built-in command in the toolbox EIDORS [1] $\text{calc_jacobian}$ which yields the matrix $D_n V$. Then, via the chain rule we compute $F'[u] = D_n V \exp(u)$.

**4.2.1. Numerical results.** In this example, we implement Algorithms 3.1-3.3 in MATLAB using the toolbox EIDORS to solve (4.12) with the finite element method. Contact impedances $\{z_i\}_{i=1}^{m_z}$ are chosen with value 0.01. We employ $m_e = 16$ electrodes and $n_p = 16$ injection patterns in which current of 0.1 Amps is injected through each pair of adjacent electrodes.

We conduct two set of examples in which we use a smooth and a discontinuous truth shown in Figure 4. Noise free voltages are computed solving (4.12) using the truth and the collection of injection patterns. To avoid inverse crimes we employ a mesh with 9216 elements while a coarser mesh (with 7744 elements) is used for the computations in Algorithms 3.1-3.3. As before, noisy observations $Y_n$ are obtained by adding Gaussian noise to the noise-free measurements as in eq. (1.6). We use
standard deviation of $\sigma = 2.5 \times 10^{-3}$. Furthermore, we use $\hat{u}_0(x,y) = -1$ (for all $(x,y) \in \Omega$), $\alpha_0 = 10^{-3}$, and $C_{\text{dec}} = 1.5$.

Again we focus on the convergence of dIRGNM and the comparison between bIRGNM and IRGNM with the same finitely many observation. The relative error w.r.t. the truth obtained using the dIRGNM for different choices of $\beta$ is shown in Figure 5. Compared with the previous example, dIRGNM seems to be more robust with respect to the choice of $\beta$ where the amplified relative error appears more obvious when $\beta > 2$. We also include the relative error obtained using the cIRGNM with noise-free observations as reference.

To compare the performance of bIRGNM and cIRGNM, we show relative errors obtained by both algorithms in Figure 6. Algorithm 3.1 is realized with the different type of observations (i.e. noise-free, single set and the average observation of $N = 500$). As comparison, we apply the bIRGNM using the same $N = 500$ observations and a value $\beta = 2.0$ which, as mentioned earlier, produced stable results when using the dIRGNM with large $N$. The iterations achieved during the second part of the bIRGNM are shown in the right panels of Figure 6. The value displayed on these plots corresponds to the minimum relative error attained during the second part of bIRGNM. Similar to our previous experiment, we notice that while this value is approximately equal to the value obtained via the cIRGNM with averaged observations, the second part of the bIRGNM reached this minimum value in less number of iterations. In the case with the smooth truth only two iterations sufficed to attained such a minimum value while 14 iterations were required by the cIRGNM. The reconstruction

This manuscript is for review purposes only.
estimates obtained are provided in Figures ??-?? of the supplementary material.

5. Conclusion. The purpose of this work is to investigate a dIRGNM (2.2) solving nonlinear inverse problems with sequential observation. The idea behind our work is highly inspired by the artificial dynamics proposed in [28]. In the context of our setting, those dynamics are of the following augmented form:

\[
\begin{align*}
    u_{n+1} &= u_n, \\
    Y_{n+1} &= F(u_{n+1}) + \sigma \xi_{n+1}, \\
    Z_{n+1} &= \frac{1}{n+1}(nZ_n + Y_{n+1}) = \frac{1}{n+1}\sum_{i=1}^{n+1} Y_i,
\end{align*}
\]

(5.1)

with \( u_0 = u^\dagger \). Such artificial dynamics comprise a steady state equation associated with the unknown variable \( u^\dagger \) and two other observation equations with sequential observations \( \{Y_n\}_{n=1,...} \) and their running averages \( \{Z_n\}_{n=1,...} \). The proposed dIRGNM (2.2) is exactly an online filter algorithm applied to the artificial dynamical system in (5.1). Systematic convergence analysis of this reconstruction algorithm has been provided in Sections 2-3 where the averaged observation \( Z_n \) yields a vanishing asymptotical behavior if the regularization parameter is appropriately chosen. This theoretical outcome verifies that the uncertainty of the nonlinear inverse problem has been dramatically weaken when the averaged observation is taken, i.e. \( Z_n \) in (1.10). Numerical evidence of our findings were presented via solving two parameter identification problems with forward models given by elliptic partial differential equations;
an inverse potential problem and electrical impedance tomography. Not only we verified numerically the rates predicted by the theory, but we also tested the numerical performance of the dIRGNM.

There are various different avenues one can consider for future work. A natural direction would be other nonlinear methodologies such as the Levenburg–Marquardt method (LMM), which is well known and has applications to geophysical sciences [22, 30]. A natural analysis here, as the LMM method commonly relies more on spectral methods, rather than a variational one. Furthermore, it would be interesting to explore effective stopping rules as well as other a-posteriori parameter choice rules for $\alpha_n$. Other common examples aside from Lepskii principle [4, 39], would include the empirical risk minimization. Another direction, related to a-posteriori choices is terminating the dIRGNM procedure once an adequate reconstruction is attained.

This is an advantage of the method, which has not been utilized since future work is required on this, which could have many practical benefits, such as accuracy and computation time. Given the results we have obtained one could aim to characterize the EnKF [18, 19], related of inverse problems [11, 10, 12, 28], in terms of convergence through asymptotic regularization [41]. A final direction would be to combine our setup for sequential inversion with the concept of recycling, which has been used in the context of linear inverse problems, primarily aimed at tomography-based problems [13, 36, 46].

Acknowledgements. The authors thank the reviewers for their helpful suggestions, which have helped to improve the clarity and presentation of the paper.

REFERENCES

[1] A. Adler and W. R. B. Lionheart. Uses and abuses of EIDORS: An extensible software base for EIT. Physiol Meas 27, 25–42, 2006.
[2] A. B. Bakushinskii. The problem of the convergence of the iteratively regularized Gauss–Newton method. Computational Mathematics and Mathematical Physics, 32(9), 1353–1359, 1992.
[3] F. Bauer, T. Hohage and A. Munk. Iteratively regularized Gauss-Newton method for nonlinear inverse problems with random noise. SIAM J. Numer. Anal., 47(3), 1827–1846, 2009.
[4] F. Bauer and T. Hohage. A Lepskii-type stopping rule for regularized Newton methods. Inverse Problems, 21(6):1975, 2005.
[5] B. M. Bell. The iterated Kalman smoother as a Gauss-Newton method. SIAM J. Optim., 4, 626–636, 1994.
[6] B. M. Bell and F. W. Cathey. The iterated Kalman filter update as a Gauss-Newton method. IEEE Transactions on Automatic Control, 43(2):294–297, 1993.
[7] B. Blaschke, A. Neubauer and O. Scherzer. On convergence rates for the iteratively regularized Gauss-Newton method, IMA J. Numer. Anal., 17, 421–436, 1997.
[8] L. Borcea, Electrical impedance tomography. Inverse Problems Series, 18(6), 2002.
[9] F. Cassola and M. Burlando. Wind speed and wind energy forecast through Kalman filtering of numerical weather prediction model output. Applied Energy, 99, 154–166.
[10] N. K. Chada, Y. Chen and D. Sanz-Alonso. Iterative ensemble Kalman methods: a unified perspective with some new variants. Foundations of Data Science, 3(3), 331-369, 2021.
[11] N. K. Chada, M. A. Iglesias, L. Roininen and A. M. Stuart. Parameterizations for ensemble Kalman inversion, Inverse Problems, 34(5), 055009, 2018.
[12] N. K. Chada and X. T. Tong. Convergence acceleration of ensemble Kalman inversion in nonlinear settings. Math. of Comp, 91(335), 1247–1280, 2022.
[13] J. Chung, E. de Sturler and J. Jiang. Hybrid projection methods with recycling for inverse problems. SIAM J Sci Comput., 43(5), 2021.
[14] G. Da Prato. An Introduction to Infinite-Dimensional Analysis. Springer, 2006.
[15] L. Ding, S. Lu and J. Cheng. Weak-norm posterior contraction rate of the 4DVAR method for linear severely ill-posed problems. J. Complexity, 46, 1–18, 2018.
[16] A. C. Duffy. An Introduction to Gradient Computation by the Discrete Adjoint Method Pre-print, 2009.
This manuscript is for review purposes only.
[46] K. M. Soodhalter. Block Krylov subspace recycling for shifted systems with unrelated right-hand sides. *SIAM J Sci Comput.*, 38(1):302–324, 2016.

[47] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, Vol. 19, 451–559, 2010.

[48] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*, Elsevier, 1987.

[49] C. R. Vogel. Sparse matrix computations arising in distributed parameter identification. *SIAM J. Matrix Anal. Appl.*, 20(4), 1027–1037, 1999.

[50] J. A. Wellner. Gaussian white noise models: some results for monotone functions. In: *Crossing boundaries: statistical essays in honor of Jack Hall*, vol. 43 of IMS Lecture Notes Monogr. Ser., Inst. Math. Statist., Beachwood, OH, pp. 87–104, 2003.

[51] F. Werner. On convergence rates for iteratively regularized Newton-type methods under a Lipschitz-type nonlinearity condition. *Journal of Inverse and Ill-Posed Problems* 23 (1), 75–84, 2015.