Stochastic conditionin g of matrix functions

Serge Gratton, David Titley-Peloquin

Stochastic Conditioning of Matrix Functions

Serge Gratton[†] and David Titley-Peloquin[‡]

Abstract. We investigate the sensitivity of matrix functions to random noise in their input. We propose the notion of a *stochastic condition number*, which determines, to first order, the sensitivity of a matrix function to random noise. We derive an upper bound on the stochastic condition number that can be estimated efficiently by using "small-sample" estimation techniques. The bound can be used to estimate the median, or any other quantile, of the error in a function's output when its input is subjected to random noise. We give numerical experiments illustrating the effectiveness of our stochastic error estimate.

Key words. sensitivity analysis, perturbation analysis, conditioning, uncertainty propagation, matrix functions

DOI. 10.1137/140973827

1. Introduction.

How sensitive are matrix functions to perturbations in their input? This is a fundamental question in numerical linear algebra. By "matrix function" we mean a general mapping $F : \Omega \to \mathbb{R}^{p\times q}$, where Ω is an open subset of $\mathbb{R}^{m\times n}$. As for the perturbations, these might represent uncertainties in the data or rounding errors arising from computations in finite precision arithmetic. The goal is to quantify the effect that such uncertainties might have on the computed function value.

Sensitivity analyses in numerical linear algebra are usually deterministic in nature. Worst-case error bounds are obtained that hold asymptotically in the small-perturbation limit. (For an excellent overview, we recommend [10, 11].) To the best of our knowledge, few attempts have been made in the literature to rigorously quantify the sensitivity of matrix functions to *random* noise. We believe that it is important to address this issue for the following reasons.

First, although rounding errors are known not to be random, some random models of rounding have been proposed, and, in fact, some automatic error analysis software does use random perturbations. (See, for example, the discussion in [10, sections 2.8, 26.5].) Therefore, it seems worthwhile to attempt to gain at least some insight into the effect of rounding errors on computed function values by using a stochastic analysis. Furthermore, in many applications, the uncertainty due to rounding errors is dominated by measurement noise in the data, which is explicitly modeled as random. In this case, it is a natural idea to take into account the random nature of the noise in the sensitivity analysis.

This research was supported by the "Fondation Sciences et Technologies pour l'Aéronautique et l'Espace (FCS STAE)," within the "Réseau Thématique de Recherche Avancée (RTRA), Toulouse, France.

[†]INPT-IRIT-ENSEEIHT, 31071 Toulouse, France (serge.gratton@enseeiht.fr).
[‡]CERFACS, 31057 Toulouse, France (titleypelo@cerfacs.fr).

For example, there has been some recent interest in the data assimilation community in the sensitivity of Krylov method iterates to random noise [22, 16]. In these applications, the Krylov iteration forms the inner loop of a nested outer-inner iterative scheme, and the uncertainty must be propagated through the inner loop. Specifically, let $x^{(j)}(b+h)$ denote the jth iterate of the method of conjugate gradients (CG) applied to solving $Ax = b + h$, where $A \in \mathbb{R}^{n\times n}$ is deterministic, symmetric, and positive definite, $b \in \mathbb{R}^n$ is deterministic, and h is modeled as a normally distributed random vector, $h \sim \mathcal{N}(0, \Sigma)$. Ideally, one would solve the linear system exactly. Of course, since the exact solution is a linear function of h, the distribution of $x(b + h) = A^{-1}(b + h)$ is easily derived. However, in large-scale applications, one often has no choice but to terminate CG after very few iterations, long before convergence has occurred. In this setting, statistical information about $x^{(j)}(b + h)$, for a small fixed j, as opposed to $x(b + h)$, is required. This is complicated by the fact that, in general, $x^{(j)}(b)$ is known to be very nonlinear in b. This example motivated us to study the problem in a more general context.

First, let us review the classical deterministic analysis. Let Ω represent an open subset of $\mathbb{R}^{m\times n}$, $F : \Omega \to \mathbb{R}^{p\times q}$ be a matrix function defined everywhere on Ω, and $A \in \Omega$ be given data. We seek to estimate some measure of the difference $F(A + H) - F(A)$, for instance, its norm or the magnitude of some of its elements, given some limited information about the perturbation H. For example, suppose we know that in some norm $\|H\|$ is bounded by δ, and from this information we wish to bound $\|F(A + H) - F(A)\|$. In this case the quantity of interest would be

$$c_\delta = \sup_{\|H\|\le\delta} \|F(A + H) - F(A)\|. \tag{1.1}$$

Unfortunately, it is usually unfeasible to compute c_δ. Instead, the normwise absolute condition number of F at A, defined as

$$\kappa = \lim_{\delta\to 0} \frac{c_\delta}{\delta} = \lim_{\delta\to 0} \sup_{\|H\|\le\delta} \frac{\|F(A + H) - F(A)\|}{\delta}, \tag{1.2}$$

leads to the first-order estimate (FOE)

$$\|F(A + H) - F(A)\| \lesssim \kappa \|H\|. \tag{1.3}$$

Rice [19] has shown that if F is Fréchet differentiable at A, then κ is the operator norm of the Fréchet derivative of F at A. There is a large body of literature dedicated to computing or bounding normwise and componentwise condition numbers of matrix functions.

Despite its widespread use, sensitivity analysis using condition numbers has some drawbacks. First, taking the supremum over $\|H\| \le \delta$ as in (1.1) and (1.2) may lead to an unnecessarily pessimistic estimate of the typical sensitivity of F because of pathological values of H that are highly unlikely to occur in practice. Furthermore, the FOE (1.3) is only valid asymptotically for "sufficiently small" $\|H\|$. In practice it is usually hard to determine how small is sufficiently small, and noise present in real data is often *not* sufficiently small.

This paper addresses the first above-mentioned drawback. Specifically, we model H as a random matrix and attempt to quantify some of the statistical properties of the random

variable $\|F(A+H)-F(A)\|_F$, where $\|\cdot\|_F$ denotes the matrix Frobenius norm. One idea that has appeared in the literature [21, 6, 20] is estimating to first order the root-mean-squared (RMS) error

$$\sqrt{\mathbb{E}\big\{\|F(A+H)-F(A)\|_F^2\big\}}.$$

(Throughout, $\mathbb{E}\{x\}$ denotes the expected value of x and $\mathbb{V}\{x\} = \mathbb{E}\big\{(x-\mathbb{E}\{x\})(x-\mathbb{E}\{x\})^T\big\}$.) Unfortunately, the above RMS error is infinite for many common matrix functions and common distributions. We argue that much more insight into the sensitivity of F can be obtained by considering the quantiles of the random variable $\|F(A+H)-F(A)\|_F$ rather than its RMS value. In particular, a useful measure of central tendency is the 0.5-quantile, i.e., the median.

As a first step, in analogy with (1.2), we only consider small perturbation asymptotics. We define a *stochastic condition number* $\tilde{\kappa}_\Sigma$ in terms of the median, as follows. Let the elements of H be random variables following some distribution such that $\mathbb{E}\{H\} = 0$ and $\mathbb{V}\{\mathrm{vec}(H)\} = \sigma^2\Sigma$ for a given positive semidefinite matrix $\Sigma \in \mathbb{R}^{mn\times mn}$. Then

$$\tilde{\kappa}_\Sigma \equiv \limsup_{\sigma\to 0} \frac{\mathrm{Med}\big\{\,\|F(A+H)-F(A)\|_F\big\}}{\sigma}.$$

We shall obtain an upper bound on the stochastic condition number that holds regardless of the distribution of the elements of H. With minor modifications the bound applies to other quantiles as well. The results are obtained by bounding the probability

$$\mathrm{Prob}\Big\{\|F(A+H)-F(A)\|_F \geq \tau\Big\}$$

for any given $\tau > 0$, to first order in σ. Our hope is that this investigation will be the first step toward obtaining *global* bounds that hold for arbitrary $\sigma > 0$, at least for some classes of matrix functions. We shall make further comments about this point in the conclusion.

Our upper bound on the stochastic condition number can be estimated efficiently, for example, by using the small-sample estimates of Kenney et. al. [14, 9]. The resulting error estimate is much more representative of the sensitivity of F to random perturbations than an FOE based on the deterministic condition number.

The rest of the paper is organized as follows. In section 2 we review some basic notions that are useful in later derivations. In section 3 we present some measures of sensitivity to random noise and show how these compare to the deterministic condition number (1.2) and FOE (1.3). Section 4 covers the efficient computation of our stochastic error estimate. Numerical experiments illustrating the theory are given in section 5, and we conclude with a short discussion in section 6.

2. Preliminaries.

First we establish our notation and review some basic notions that are used in later sections.

Notation. We use uppercase letters to denote matrices and lowercase Roman letters for vectors and indices. Subscripts are used to denote elements of a vector or matrix, while superscripts denote terms in a sequence. For example, A_{ij} is the (i,j)th element of the matrix A, and h_j is the jth element of the vector h, while $u^{(i)} \in \mathbb{R}^n$ is the ith term in a sequence of vectors. Otherwise, scalars are designated by lowercase Greek letters.

Fréchet derivative. Throughout we assume that $F : \Omega \subseteq \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is Fréchet differentiable at $A \in \Omega$. The Fréchet derivative of F at A is the unique bounded linear operator $F'(A)$ defined by the relation

$$F(A+H) = F(A) + F'(A)(H) + R(H), \quad \lim_{H\to 0} \frac{\|R(H)\|}{\|H\|} = 0. \tag{2.1}$$

The matrix representation of $F'(A)$ in the standard basis is the Jacobian matrix $J_A \in \mathbb{R}^{pq\times mn}$. In other words,

$$\operatorname{vec}(F'(A)(H)) = J_A \operatorname{vec}(H). \tag{2.2}$$

Multi-index notation. A multi-index is simply a vector of natural numbers,

$$v = [v_1, v_2, \dots, v_n]^T \in \mathbb{N}_0^n,$$

where $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. The magnitude and factorial of an n-dimensional multi-index are defined as

$$|v| = v_1 + \cdots + v_n, \qquad v! = v_1! \cdots v_n!,$$

while, for $v \in \mathbb{N}_0^n$ and $h = [h_1, \dots, h_n]^T \in \mathbb{R}^n$,

$$h^v = h_1^{v_1} h_2^{v_2} \cdots h_n^{v_n}.$$

Higher-order partial derivatives can be written compactly as

$$\partial_x^v = \frac{\partial^{v_1}}{\partial x_1^{v_1}} \cdots \frac{\partial^{v_n}}{\partial x_n^{v_n}}.$$

In this notation, the Taylor series of an analytic function $g : \mathbb{R}^n \to \mathbb{R}$ can be expressed as

$$g(x+h) = \sum_{v\in\mathbb{N}_0^n} \frac{\partial_x^v g(x)}{v!} h^v = g(x) + \nabla g(x)^T h + \sum_{\substack{v\in\mathbb{N}_0^n \\ |v|\geq 2}} \frac{\partial_x^v g(x)}{v!} h^v. \tag{2.3}$$

Probability notation. We work with the standard probability space consisting of the sample space $\mathbb{R}$, its σ-algebra of Borel sets, and Lebesgue measure, with straightforward extension to random vectors and matrices. We use $\operatorname{Prob}\{X\}$ to denote the probability of an event X. $\mathbb{E}\{x\}$ and $\mathbb{V}\{x\}$ stand for the expected value and covariance matrix of a random vector x. As shorthand for $\mathbb{E}\{x\} = u$ and $\mathbb{V}\{x\} = \Sigma$ we write $x \sim (u, \Sigma)$, and if additionally x follows a multivariate normal distribution, $x \sim \mathcal{N}(u, \Sigma)$. A median of a random variable α is any scalar τ satisfying both

$$\operatorname{Prob}\{\alpha \leq \tau\} \geq \frac{1}{2} \quad \text{and} \quad \operatorname{Prob}\{\alpha \geq \tau\} \geq \frac{1}{2}. \tag{2.4}$$

If α is a continuous random variable, then $\operatorname{Prob}\{\alpha = \tau\} = 0$. In this case, (2.4) is equivalent to

$$\operatorname{Prob}\{\alpha \leq \tau\} = \operatorname{Prob}\{\alpha \geq \tau\} = \frac{1}{2}.$$

If additionally α has a strictly positive probability density function, then its median is unique. In any case, when we write *the* median of α, or $\mathrm{Med}\{\alpha\}$, we are referring to

$$\mathrm{Med}\{\alpha\} = \sup\big\{\tau \,:\, (2.4) \text{ holds}\big\}.$$

More generally, for $q \in (0,1)$, the qth quantile of α is any scalar τ satisfying

$$\mathrm{Prob}\{\alpha \le \tau\} \ge q, \quad \mathrm{Prob}\{\alpha \ge \tau\} \ge 1-q. \tag{2.5}$$

When we write *the* qth quantile of α, we mean

$$Q_q\{\alpha\} = \sup\big\{\tau \,:\, (2.5) \text{ holds}\big\}.$$

(The median is the 0.5-quantile.) In contrast to the expected value, the quantiles of a random variable are always finite.

Basic probability inequalities. If α and β are random variables such that $\alpha \le \beta$, then for any $\tau \in \mathbb{R}$,

$$\mathrm{Prob}\{\alpha \ge \tau\} \le \mathrm{Prob}\{\beta \ge \tau\}\,, \quad \mathrm{Prob}\{\beta \le \tau\} \le \mathrm{Prob}\{\alpha \le \tau\}\,. \tag{2.6}$$

For random variables α and β, for any $\tau, \epsilon \in \mathbb{R}$,

$$\mathrm{Prob}\{\alpha + \beta \ge \tau\} \le \mathrm{Prob}\{\alpha \ge \tau(1-\epsilon)\} + \mathrm{Prob}\{\beta \ge \tau\epsilon\}\,. \tag{2.7}$$

Markov's inequality. Given a nonnegative random variable α, then for any $\tau > 0$,

$$\mathrm{Prob}\{\alpha \ge \tau\} \le \frac{\mathbb{E}\{\alpha\}}{\tau}. \tag{2.8}$$

Quadratic forms. If $M \in \mathbb{R}^{m\times n}$ and $x \sim (0, \Sigma)$,

$$\mathbb{E}\{\|Mx\|_2\} \le \sqrt{\mathbb{E}\big\{\|Mx\|_2^2\big\}} = \|M\Sigma^{1/2}\|_F. \tag{2.9}$$

If $H \in \mathbb{R}^{m\times n}$ and $\mathrm{vec}(H) \sim (0, \Sigma)$,

$$\mathbb{E}\{\|H\|_F\} \le \sqrt{\mathbb{E}\big\{\|H\|_F^2\big\}} = \|\Sigma^{1/2}\|_F. \tag{2.10}$$

3. Notions of stochastic conditioning.

3.1. "Expected conditioning," revisited. As mentioned in the introduction, there is not a large body of literature dealing with the sensitivity of matrix functions to random noise. We could trace the idea back to Turing [21], who considers the function $F(A) = A^{-1}b$ for nonsingular $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. It is straightforward to verify that

$$(A+H)^{-1}b - A^{-1}b = -A^{-1}HA^{-1}b + o(\|H\|).$$

Turing ignores the $o(\|H\|)$ term and gives an expression for the RMS value of $\|A^{-1}HA^{-1}b\|_2$. Fletcher considers the same problem in [6]. (See also [10, page 136].) In Fletcher's approach,

the elements H_{ij} of H are independent random variables with mean 0 and variance $\sigma^2 A_{ij}^2$. Under these assumptions,

$$\sqrt{\mathbb{E}\{\|A^{-1}HA^{-1}b\|_2^2\}} = \sigma \big\| [A^{-1}][A][A^{-1}b] \big\|_1^{1/2},$$

where $[A]$ denotes the matrix whose entries are A_{ij}^2. Fletcher calls the above the *expected condition number* of $F(A) = A^{-1}b$. He also generalizes the above to other matrix functions. Once again, the idea is to take a first-order expansion,

$$F(A+H) - F(A) = F'(A)(H) + o(\|H\|),$$

ignore the $o(\|H\|)$ term, and find an expression for

$$\sqrt{\mathbb{E}\{\|F'(A)(H)\|_F^2\}}. \tag{3.1}$$

Stewart [20] independently derives a similar result. In Stewart's case, the random perturbation has the form $H = S_c G S_r^T$, where G is a random matrix whose elements are uncorrelated with mean 0 and variance 1, and S_c and S_r are deterministic scalings.

The quantity in (3.1) is easy to analyze because it involves a quadratic form in the elements of H. Suppose that

$$\mathrm{vec}(H) \sim (0, \sigma^2 \Sigma),$$

where $\sigma \in \mathbb{R}$ and $\Sigma \in \mathbb{R}^{mn \times mn}$ is symmetric nonnegative definite. Then from (2.2) and (2.9),

$$\mathbb{E}\{\|F'(A)(H)\|_F^2\} = \mathbb{E}\{\|\mathrm{vec}(F'(A)(H))\|_2^2\} = \mathbb{E}\{\|J_A \mathrm{vec}(H)\|_2^2\} = \sigma^2 \|J_A \Sigma^{1/2}\|_F^2.$$

Thus, assuming that one can indeed ignore the $o(\|H\|)$ term in the expansion (2.1), one obtains

$$\sqrt{\mathbb{E}\{\|F(A+H) - F(A)\|_F^2\}} \approx \sqrt{\mathbb{E}\{\|F'(A)(H)\|_F^2\}} = \sigma \|J_A \Sigma^{1/2}\|_F. \tag{3.2}$$

This reduces to the previously mentioned results of Fletcher and Stewart when the covariance matrix Σ is chosen according to their respective approaches.

One might question whether there is a rigorous theoretical justification for dropping the $o(\|H\|)$ term in the above. For instance, because it was obtained from a first-order expansion in H, we might conclude that (3.2) is tight for perturbations with sufficiently small covariance matrix, i.e., for sufficiently small σ. In fact, as noted by Stewart [20], this is not the case, and (3.2) is often meaningless. For example, with $F(A) = A^{-1}b$ and $\mathrm{vec}(H) \sim \mathcal{N}(0, \sigma^2 I_{mn})$, the left-hand side of (3.2) is infinite for any $\sigma > 0$, while the right-hand side tends to 0 as $\sigma \to 0$. Stewart [20, Theorem 2.8] gives some arguments as to why the lower-order terms can be ignored. Nevertheless, it is not immediately clear under which conditions the approximation (3.2) does indeed make sense.

In order to guarantee that (3.2) is tight, at least in the small σ limit, we found it necessary to make the following assumptions on F and the nature of the random noise H. (We abuse notation and interchangeably use $F(A)$ and $F(a)$, where $a = \mathrm{vec}(A)$.)

(3.3)
- $F : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is entire, i.e., each component $F_{ij}(a)$ of F has a Taylor series (2.3) that is absolutely convergent for all $a \in \mathbb{R}^{mn}$.
- $\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$, and the elements H_{ij} of H have their kth moment bounded by $c\sigma^k$ for some constant $c > 0$.

In Theorem 3.2 we show that, under the above assumptions, (3.2) is indeed valid to first order in σ. Specifically,

$$\lim_{\sigma\to 0} \frac{\sqrt{\mathbb{E}\{\|F(A+H) - F(A)\|_F^2\}}}{\sigma} = \|J_A\Sigma^{1/2}\|_F.$$

Thus, if the assumptions (3.3) hold, (3.2) is indeed a valid FOE of the RMS error:

$$\sqrt{\mathbb{E}\{\|F(A+H) - F(A)\|_F^2\}} \approx \sigma\|J_A\Sigma^{1/2}\|_F. \tag{3.4}$$

This is a stochastic analogue of the deterministic FOE in (1.3).

In the derivation of this result we will need the following lemma, whose proof is given in the appendix.

Lemma 3.1. *Let $\xi(x)$ represent a power series in multi-index notation which is absolutely convergent for all $x \in \mathbb{R}^n$:*

$$\xi(x) = \sum_{\substack{v\in\mathbb{N}_0^n \\ |v|\geq 2}} \alpha_v x^v \in \mathbb{R}, \qquad \alpha_v \in \mathbb{R}, \quad x \in \mathbb{R}^n.$$

If h is a random vector whose elements h_i have their kth moment bounded by $c\sigma^k$, then for any $u \in \mathbb{R}^n$,

$$\lim_{\sigma\to 0} \frac{\left|\mathbb{E}\{(u^T h)\xi(h)\}\right|}{\sigma^2} = 0 \qquad \text{and} \qquad \lim_{\sigma\to 0} \frac{\mathbb{E}\{\xi(h)^2\}}{\sigma^2} = 0.$$

Theorem 3.2. *Suppose that $F : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is entire, and let J_A denote the Jacobian (2.2) at $A \in \mathbb{R}^{m\times n}$. Additionally, let the elements of H be random variables whose kth moments are bounded by $c\sigma^k$, and*

$$h = \mathrm{vec}(H) \sim (0, \sigma^2\Sigma).$$

Then

$$\lim_{\sigma\to 0} \frac{\sqrt{\mathbb{E}\{\|F(A+H) - F(A)\|_F^2\}}}{\sigma} = \|J_A\Sigma^{1/2}\|_F. \tag{3.5}$$

Proof. Denote $a = \mathrm{vec}(A) \in \mathbb{R}^{mn}$. For each element $F_{ij}(a)$ of F, define

$$\xi_{ij}(h) = \sum_{\substack{v\in\mathbb{N}_0^{mn} \\ |v|\geq 2}} \frac{\partial_a^v F_{ij}(a)}{v!} h^v \in \mathbb{R}.$$

Then, from (2.3),

$$\begin{aligned}\big(F_{ij}(a+h)-F_{ij}(a)\big)^2 &= \big(\nabla F_{ij}(a)^T h+\xi_{ij}(h)\big)^2\\ &= (\nabla F_{ij}(a)^T h)^2+2(\nabla F_{ij}(a)^T h)\xi_{ij}(h)+\xi_{ij}(h)^2.\end{aligned}$$

In the above,

$$\mathbb{E}\big\{(\nabla F_{ij}(a)^T h)^2\big\}=\mathbb{V}\big\{\nabla F_{ij}(a)^T h\big\}=\sigma^2\nabla F_{ij}(a)^T\Sigma\nabla F_{ij}(a)=\sigma^2\|\Sigma^{1/2}\nabla F_{ij}(a)\|_2^2,$$

while, after dividing by σ^2 and taking the limit as $\sigma\to 0$, from Lemma 3.1, the other two terms tend to 0. Therefore, the limit in (3.5) is

$$\begin{aligned}&\lim_{\sigma\to 0}\frac{\sqrt{\sum_{i=1}^p\sum_{j=1}^q\mathbb{E}\Big\{\big(F_{ij}(a+h)-F_{ij}(a)\big)^2\Big\}}}{\sigma}\\ &\quad=\lim_{\sigma\to 0}\frac{\sqrt{\sum_{i=1}^p\sum_{j=1}^q\sigma^2\|\Sigma^{1/2}\nabla F_{ij}(a)\|_2^2}}{\sigma}=\left\|\begin{bmatrix}\nabla F_{11}(a)^T\\ \vdots\\ \nabla F_{pq}(a)^T\end{bmatrix}\Sigma^{1/2}\right\|_F=\|J_A\Sigma^{1/2}\|_F.\end{aligned}$$

■

Theorem 3.2 is applicable to many matrix functions, such as matrix polynomials and the matrix exponential. However, in order to generalize the result so that it is applicable to more general classes of functions and random perturbations, a different framework is required. This is the topic of the following section.

3.2. A more general stochastic condition number. In this section we provide a generalization of Theorem 3.2. Our motivation is the following. First, recall that Theorem 3.2 was derived under the two assumptions (3.3), one of which being that F is entire. This assumption is rather restrictive: many interesting functions in numerical linear algebra are Fréchet differentiable but *not* entire. Furthermore, even if the assumptions (3.3) required by Theorem 3.2 do hold, more insight about the sensitivity of F can be gained from not only the RMS value of $\|F(A+H)-F(A)\|_F$ but also its quantiles. For example, we might require that the error remain below a certain threshold with probability 95%, in which case we would work with the 0.95-quantile.

Suppose that $F:\Omega\subseteq\mathbb{R}^{m\times n}\to\mathbb{R}^{p\times q}$ is Fréchet differentiable at $A\in\Omega$ and that

$$\operatorname{vec}(H)\sim(0,\sigma^2\Sigma),$$

without any other assumption on the distribution or the moments of the elements of H. In this more general setting, it is no longer possible to obtain a meaningful first-order RMS estimate as in Theorem 3.2. Consider the following quantity:

$$\tilde{\kappa}_\Sigma\equiv\limsup_{\sigma\to 0}\frac{\operatorname{Med}\big\{\,\|F(A+H)-F(A)\|_F\big\}}{\sigma}. \tag{3.6}$$

This is analogous to the limit in (3.5), but with the RMS replaced by the median. In analogy to (1.2), we call $\tilde{\kappa}_\Sigma$ the Frobenius-norm stochastic condition number of F at A with respect to $(0,\Sigma)$ perturbations, or simply the stochastic condition number of F. We show below that

$$\tilde{\kappa}_\Sigma\le 2\|J_A\Sigma^{1/2}\|_F.$$

With minor modifications the result is applicable not only to the median but to other quantiles as well; see Remark 3.2 below. The inequality is a consequence of the following theorem.

Theorem 3.3. *Suppose that $F : \Omega \subseteq \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$ is Fréchet differentiable at $A \in \Omega$, and let J_A denote the Jacobian (2.2) at A. If $\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$, then for any $\tau > 0$,*

$$\limsup_{\sigma\to 0} \mathrm{Prob}\left\{\frac{\|F(A+H)-F(A)\|_F}{\sigma} \geq \tau\right\} \leq \frac{\|J_A\Sigma^{1/2}\|_F}{\tau}.$$

Proof. Denote

$$\mathcal{P}_{\tau\sigma} = \mathrm{Prob}\left\{\frac{\|F(A+H)-F(A)\|_F}{\sigma} \geq \tau\right\}.$$

From (2.1) along with (2.6) and (2.7), for any $\epsilon \in (0,1)$ we can decompose $\mathcal{P}_{\tau\sigma}$ as follows:

$$\begin{aligned}\mathcal{P}_{\tau\sigma} &= \mathrm{Prob}\left\{\frac{\|F'(A)(H)+R(H)\|_F}{\sigma} \geq \tau\right\} \\ &\leq \mathrm{Prob}\left\{\frac{\|F'(A)(H)\|_F}{\sigma} \geq \tau(1-\epsilon)\right\} + \mathrm{Prob}\left\{\frac{\|R(H)\|_F}{\sigma} \geq \tau\epsilon\right\}.\end{aligned} \tag{3.7}$$

Under the assumption that $\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$, from (2.2), (2.8), and (2.9) we obtain

$$\begin{aligned}\mathrm{Prob}\left\{\frac{\|F'(A)(H)\|_F}{\sigma} \geq \tau(1-\epsilon)\right\} &\leq \frac{\mathbb{E}\{\|F'(A)(H)\|_F\}}{\sigma\tau(1-\epsilon)} \\ &= \frac{\mathbb{E}\{\|J_A\mathrm{vec}(H)\|_2\}}{\sigma\tau(1-\epsilon)} \leq \frac{\|J_A\Sigma^{1/2}\|_F}{\tau(1-\epsilon)}.\end{aligned} \tag{3.8}$$

Next we need to show that for any $\tau > 0$ and any $\epsilon \in (0,1)$, the residual term in (3.7) is bounded above by ϵ if σ is sufficiently small. Actually, this was shown by Stewart in [20, Theorem 2.8]. For completeness, we include a short proof here. For any $\beta > 0$,

$$\begin{aligned}\mathcal{R}_{\tau\sigma\epsilon} &= \mathrm{Prob}\left\{\frac{\|R(H)\|_F}{\sigma} \geq \tau\epsilon\right\} \\ &= \mathrm{Prob}\left\{\frac{\|R(H)\|_F}{\sigma} \geq \tau\epsilon \cap \|H\|_F < \beta\right\} \\ &\quad + \mathrm{Prob}\left\{\frac{\|R(H)\|_F}{\sigma} \geq \tau\epsilon \cap \|H\|_F \geq \beta\right\} \\ &\leq \mathrm{Prob}\left\{\frac{\|R(H)\|_F}{\sigma} \geq \tau\epsilon \cap \|H\|_F < \beta\right\} + \mathrm{Prob}\{\|H\|_F \geq \beta\}.\end{aligned}$$

From (2.1), for any $\alpha > 0$, there exists β such that

$$\frac{\|R(H)\|_F}{\|H\|_F} \leq \alpha \quad \text{when} \quad \|H\|_F \leq \beta. \tag{3.9}$$

Therefore, for any $\alpha > 0$, there is a corresponding β such that

$$\begin{aligned}\mathcal{R}_{\tau\sigma\epsilon} &\le \text{Prob}\left\{\frac{\alpha\|H\|_F}{\sigma} \ge \tau\epsilon \cap \|H\|_F < \beta\right\} + \text{Prob}\{\|H\|_F \ge \beta\} \\ &\le \text{Prob}\left\{\frac{\alpha\|H\|_F}{\sigma} \ge \tau\epsilon\right\} + \text{Prob}\{\|H\|_F \ge \beta\} \\ &\le \frac{\alpha\mathbb{E}\{\|H\|_F\}}{\sigma\tau\epsilon} + \frac{\mathbb{E}\{\|H\|_F\}}{\beta} \\ &\le \frac{\alpha\|\Sigma^{1/2}\|_F}{\tau\epsilon} + \frac{\sigma\|\Sigma^{1/2}\|_F}{\beta}.\end{aligned}$$

(The last two inequalities follow from (2.8) and (2.10).) Set $\alpha = \tau\epsilon^2/(2\|\Sigma^{1/2}\|_F)$ and obtain the corresponding β such that (3.9) holds. Note that β depends on α but is independent of σ. Then, for all $\sigma \le \sigma^*(\epsilon) = \epsilon\beta/(2\|\Sigma^{1/2}\|_F)$,

$$\mathcal{R}_{\tau\sigma\epsilon} \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Substituting the above and (3.8) into (3.7) we see that for any $\tau > 0$ and any $\epsilon \in (0, 1)$,

$$\mathcal{P}_{\tau\sigma} \le \frac{\|J_A\Sigma^{1/2}\|_F}{\tau(1-\epsilon)} + \epsilon = \gamma(\epsilon) \qquad \text{when } \sigma \le \sigma^*(\epsilon). \tag{3.10}$$

It follows that

$$\limsup_{\sigma\to 0} \mathcal{P}_{\tau\sigma} \le \gamma(0) = \frac{\|J_A\Sigma^{1/2}\|_F}{\tau}. \qquad \blacksquare$$

We can use Theorem 3.3 to bound the condition number $\tilde{\kappa}_\Sigma$ as follows. Setting τ to be the median in (3.10), so that $\mathcal{P}_{\tau\sigma} \ge 1/2$, we obtain

$$\frac{1}{2} \le \frac{\|J_A\Sigma^{1/2}\|_F}{\text{Med}\left\{\frac{\|F(A+H)-F(A)\|_F}{\sigma}\right\}(1-\epsilon)} + \epsilon. \tag{3.11}$$

Rearranging the above shows that for any $\epsilon \in (0, 1/2)$, when σ is sufficiently small,

$$\text{Med}\left\{\frac{\|F(A+H)-F(A)\|_F}{\sigma}\right\} \le \frac{2\|J_A\Sigma^{1/2}\|_F}{(1-2\epsilon)(1-\epsilon)}.$$

Therefore, as in the proof of Theorem 3.3, we have

$$\tilde{\kappa}_\Sigma = \limsup_{\sigma\to 0} \frac{\text{Med}\{\|F(A+H)-F(A)\|_F\}}{\sigma} \le 2\|J_A\Sigma^{1/2}\|_F. \tag{3.12}$$

This leads to the the FOE

$$\text{Med}\big\{\|F(A+H)-F(A)\|_F\big\} \lesssim 2\sigma\|J_A\Sigma^{1/2}\|_F. \tag{3.13}$$

We conclude this section with two additional generalizations of the stochastic condition number, given in the following remarks.

Remark 3.1. *It might also be of interest to measure the sensitivity of a linear function of F, namely,*

$$\left\|W_1^T\big(F(A+H)-F(A)\big)W_2\right\|_F,$$

for given $W_1 \in \mathbb{R}^{p\times r_1}$ and $W_2 \in \mathbb{R}^{q\times r_2}$. For instance, we might be interested in the error in just one element of F, in which case W_1 and W_2 would be standard basis vectors. It can be verified that the results of Theorem 3.3 carry through in this case:

$$\limsup_{\sigma\to 0}\frac{\operatorname{Med}\left\{\left\|W_1^T\big(F(A+H)-F(A)\big)W_2\right\|_F\right\}}{\sigma} \le 2\|(W_2^T\otimes W_1^T)J_A\Sigma^{1/2}\|_F,$$

where $W_2^T\otimes W_1^T \in \mathbb{R}^{r_1r_2\times pq}$ denotes the Kronecker product.

Remark 3.2. *We have chosen to define $\tilde{\kappa}_\Sigma$ in terms of the median for two reasons. The first is so that it may be easily comparable with the RMS approach described in section 3.1. The second is that we think it is legitimate to be interested in a measure of "central tendency" or "typical error." Besides the median, we could bound in the same way any other quantile of the random variable $\|F(A+H)-F(A)\|_F$. Replacing $1/2$ in the left-hand side of (3.11) by an arbitrary $1-q\in(0,1)$, from (2.5) we obtain*

$$\limsup_{\sigma\to 0}\frac{Q_q\{\|F(A+H)-F(A)\|_F\}}{\sigma} \le \frac{1}{1-q}\|J_A\Sigma^{1/2}\|_F.$$

For example, we might require that the error remain below a certain threshold with probability 95%, in which case we would set $q=0.95$. The resulting error estimate is only a factor 10 larger than $\tilde{\kappa}_\Sigma$.

3.3. Sharpness of the bounds. Compare the bound on the stochastic condition number (3.12) and the resulting FOE (3.13) with their RMS counterparts (3.5) and (3.4). Recall that (3.12) and (3.13) apply much more generally: the only required assumptions are that F be Fréchet differentiable at A and that $\operatorname{vec}(H)\sim(0,\sigma^2\Sigma)$, without any restriction on the distribution of the elements of H. Furthermore, (3.12) and (3.13) can easily be modified to apply to any other quantile—not just the median.

Being more generally applicable, (3.12) and (3.13) are less sharp than (3.5) and (3.4). The additional factor 2 in (3.12) and (3.13) is not significant. Typically one is not interested in the precise value of the error but only in its order of magnitude. Furthermore, our numerical tests indicate that the factor 2 can typically be ignored.

On the other hand, in (3.5) there is equality, whereas (3.12) is merely an upper bound. It would be useful to obtain a nontrivial lower bound for $\tilde{\kappa}_\Sigma$. In fact, this is impossible to achieve without additional assumptions on the distribution of the elements of H. The following simple example shows that $\tilde{\kappa}_\Sigma$ can be arbitrarily smaller than $\|J_A\Sigma^{1/2}\|_F$.

Let H be a random scalar such that

$$H=\begin{cases}\pm\epsilon, & \text{each with probability } 1/3,\\ \pm\sqrt{3}, & \text{each with probability } 1/6,\end{cases}$$

for $\epsilon > 0$. As required in Theorem 3.3, H has mean 0 and finite variance $\Sigma = 1 + 2\epsilon^2/3$. Consider the scalar linear function $F(A) = A$. Then, for all $A \in \mathbb{R}$, the stochastic condition number $\tilde{\kappa}_\Sigma$ in (3.6) simplifies to

$$\tilde{\kappa}_\Sigma = \text{Med}\{|H|\} = \epsilon,$$

while the upper bound in (3.12) is

$$2\|J_A\Sigma^{1/2}\|_F = 2\sqrt{\Sigma} \geq 2.$$

Therefore,

$$\frac{\tilde{\kappa}_\Sigma}{\|J_A\Sigma^{1/2}\|_F} \to 0$$

as $\epsilon \to 0$, showing that no nontrivial lower bound in (3.12) is possible in general. The above example can easily be extended to the case in which the distribution of H is continuous.

The limiting behavior of $\mathcal{P}_{\tau\sigma}$ in Theorem 3.3 is given by

$$\text{Prob}\left\{\frac{\|F'(A)(H)\|_F}{\sigma} \geq \tau\right\} = \text{Prob}\left\{\frac{\|J_A\text{vec}(H)\|_2^2}{\sigma^2} \geq \tau^2\right\}.$$

Even if one makes very strong assumptions on the nature of H, it seems difficult to derive a useful relationship between $\|J_A\text{vec}(H)\|_2^2/\sigma^2$ and its expected value $\|J_A\Sigma^{1/2}\|_F^2$. For example, if $\text{vec}(H) \sim \mathcal{N}(0, \sigma^2\Sigma)$, then it is straightforward to verify that $\|J_A\text{vec}(H)\|_F^2$ is a linear combination of independent χ^2 random variables with one degree of freedom:

$$\frac{\|J_A\text{vec}(H)\|_2^2}{\sigma^2} = \sum_{i=1}^{r} \sigma_i^2 \chi_i^2,$$

where r and σ_i are, respectively, the rank and nonzero singular values of $J_A\Sigma^{1/2}$. Tail bounds for linear combination of independent χ^2 random variables do exist (e.g., [15, Lemma 1] or [4, Theorem 5.2]). For example, in our notation, [15, Lemma 1] shows that for any $\tau > 0$,

$$\text{Prob}\left\{\frac{\|J_A\text{vec}(H)\|_2^2}{\|J_A\Sigma^{1/2}\|_F^2} \leq 1 - 2\frac{\sqrt{\tau\sum_{i=1}^r \sigma_i^4}}{\sum_{i=1}^r \sigma_i^2}\right\} \leq \exp(-\tau).$$

Therefore, if

$$\sqrt{\sum_{i=1}^{r} \sigma_i^4} \ll \sum_{i=1}^{r} \sigma_i^2,$$

then $\|J_A\text{vec}(H)\|_2$ is very unlikely to be much smaller than $\|J_A\Sigma^{1/2}\|_F$, so (3.12) and (3.13) are likely to be sharp. Unfortunately, such bounds depend on the coefficients in the linear combination (here the unknown singular values σ_i of $J_A\Sigma^{1/2}$) and/or are generally quite pessimistic. The matter is even more complicated if $\text{vec}(H)$ is not normally distributed.

To summarize, it is possible to create distributions such that our bound (3.12) and FOE (3.13) are arbitrarily larger than the actual stochastic condition number. Furthermore,

even in simple specific cases, it appears to be quite difficult to fully quantify the sharpness of (3.12) and (3.13). Nevertheless, our numerical experiments indicate that these give reasonable order-of-magnitude estimates of error quantiles, and we believe that they are suitable for practical use.

3.4. Comparison with a deterministic error estimate. Using (2.1) and (2.2), we can write the Frobenius-norm deterministic condition number (1.2) as follows:

$$\kappa = \lim_{\delta\to 0} \sup_{\|H\|_F\le\delta} \frac{\|F(A+H)-F(A)\|_F}{\delta} = \sup_{\|H\|_F\le 1} \|F'(A)(H)\|_F = \|J_A\|_2. \tag{3.14}$$

Therefore, in the Frobenius norm, the deterministic FOE (1.3) is

$$\|F(A+H)-F(A)\|_F \lesssim \|J_A\|_2\|H\|_F. \tag{3.15}$$

In general, it is difficult to compare the above to the stochastic FOE from (3.13):

$$\mathrm{Med}\{\|F(A+H)-F(A)\|_F\} \lesssim 2\sigma\|J_A\Sigma^{1/2}\|_F.$$

In fact, (3.14) and (3.15) are not particularly suitable for measuring the sensitivity of F to random noise since they are based on the condition $\|H\|_F \le \delta$. In general, from the distribution of the elements of H, not much is known about the distribution of $\|H\|_F$.

Nevertheless, some insight can be gained from some specific cases. For example, suppose that $\mathrm{vec}(H) \sim \mathcal{N}(0, \sigma^2 I_{mn})$. In this case, it is highly unlikely that $\|H\|_F$ lies far from its RMS value, and (3.15) leads to

$$\|F(A+H)-F(A)\|_F \lesssim \sigma\sqrt{mn}\|J_A\|_2 \equiv \delta_{\mathrm{det}}.$$

Compare the above to the stochastic FOE (3.13) of the median with $\Sigma = I_{mn}$:

$$\mathrm{Med}\{\|F(A+H)-F(A)\|_F\} \lesssim 2\sigma\|J_A\|_F \equiv \delta_{\mathrm{med}}.$$

Because $J_A \in \mathbb{R}^{pq\times mn}$, $\|J_A\|_F \le \|J_A\|_2 \min\{\sqrt{mn}, \sqrt{pq}\}$, and it follows that

$$\frac{\delta_{\mathrm{med}}}{\delta_{\mathrm{det}}} = \frac{2\|J_A\|_F}{\sqrt{mn}\|J_A\|_2} \le \frac{2\min\{\sqrt{mn}, \sqrt{pq}\}}{\sqrt{mn}}. \tag{3.16}$$

In particular, if

- $\|J_A\|_F \ll \|J_A\|_2 \min\{\sqrt{mn}, \sqrt{pq}\}$, i.e., the Jacobian matrix has a few singular values that are very large relative to its remaining singular values; and/or
- $pq \ll mn$, i.e., the domain of F is in a much larger space than its range,

then $\delta_{\mathrm{med}}/\delta_{\mathrm{det}} \ll 1$, and there can be a very large difference between the worst-case sensitivity and the typical sensitivity of F to random noise, at least asymptotically in the small σ limit. Numerical examples comparing stochastic and deterministic FOEs are given in section 5.

4. Estimating the stochastic condition number. Recall that the deterministic Frobenius-norm condition number κ in (3.14) involves the operator norm of $F'(A)$ or, equivalently, the spectral norm of the Jacobian matrix:

$$\kappa = \sup_{\|H\|_F \le 1} \|F'(A)(H)\|_F = \|J_A\|_2.$$

The above can be computed using the power method; see, e.g., [11, Algorithm 3.20]. This requires the evaluation of the Fréchet derivative $F'(A)(H)$ and its adjoint at a few different values of H or, equivalently, the computation of a few matrix-vector products with the Jacobian matrix and with its transpose. These can be computed by standard automatic differentiation techniques. (See, for example, [18, section 7.2] or [8] for an introduction to automatic differentiation.) A number of specialized algorithms have also recently been developed for this purpose; see, e.g., [12, section 7], [2], and [1, 3, 13] for methods specific to some important matrix functions.

On the other hand, the upper bound $2\|J_A\Sigma^{1/2}\|_F$ on the stochastic condition number in (3.12) involves the matrix Frobenius norm of the scaled Jacobian matrix $J_A\Sigma^{1/2} \in \mathbb{R}^{pq\times mn}$. Several randomized estimators can be used for this purpose. For a recent survey, we recommend [4]. An attractive feature of these methods is that they are adjoint-free: only matrix-vector products with J_A (and not its transpose) are required.

In our numerical experiments we use the "small-sample" estimator of Gudmundsson, Kenney, and Laub [9, 14], defined as

$$\eta_k(J_A\Sigma^{1/2}) = \sqrt{\frac{mn}{k}}\|J_A\Sigma^{1/2}Q\|_F = \sqrt{\frac{mn}{k}\sum_{i=1}^{k}\|J_A\Sigma^{1/2}q^{(i)}\|_2^2}, \tag{4.1}$$

where $Q = [q^{(1)}, \ldots, q^{(k)}] \in \mathbb{R}^{mn\times k}$ is the "thin" Q factor in the QR decomposition of an $mn \times k$ matrix whose elements are mutually independent standard normal variables. Small values of k are required to obtain good order of magnitude estimates with a high probability. In our tests we use $k = 5$.

From (4.1) we see that computing $\eta_k(J_A\Sigma^{1/2})$ reduces to computing k matrix-vector products with $J_A\Sigma^{1/2}$. This entails the computation of $u^{(i)} = \Sigma^{1/2}q^{(i)}$, which can be very expensive if an appropriate factorization of Σ is not known a priori. Nevertheless, this step can sometimes be performed easily, for example, if Σ is diagonal as in Fletcher's approach discussed in section 3.1. For large-scale problems involving noise following a multivariate normal distribution with general covariance matrices Σ, an iterative strategy for computing $u^{(i)}$ is proposed in [5]. Finally, several methods can be used to compute the matrix-vector products $J_Au^{(i)}$, as mentioned above in our discussion of the power method.

5. Numerical experiments. To illustrate the theory we give numerical examples involving the solution of linear equations, least squares problems, and a Krylov subspace iterative method. Our tests are meant to compare the effectiveness of the FOEs of the error obtained from the deterministic condition number and the stochastic condition number when

$\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$ for a wide range of values of σ.

$$\text{Deterministic FOE (see (3.15))}: \quad \frac{\|F(A+H)-F(A)\|_F}{\|F(A)\|_F} \lesssim \frac{\|J_A\|_2\|H\|_F}{\|F(A)\|_F}. \tag{5.1}$$

$$\text{Stochastic FOE (see (3.13))}: \quad \frac{\|F(A+H)-F(A)\|_F}{\|F(A)\|_F} \lesssim \frac{\sigma\|J_A\Sigma^{1/2}\|_F}{\|F(A)\|_F}. \tag{5.2}$$

Note that, H being random, the deterministic FOE is random, while the stochastic FOE is a deterministic quantity.

5.1. Solution of linear systems of equations. First, we provide examples using the problem discussed in section 3.1 that originally motivated the investigations of Turing, Fletcher, and Stewart: $F(A) = A^{-1}b$ for nonsingular $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$.

In the following examples, $n = 500$ and A is created from its SVD, $A = USV^T$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{n\times n}$ are the Q factors in the QR factorization of random matrices and $S \in \mathbb{R}^{n\times n}$ is diagonal with the n singular values of A, logarithmically equally spaced between $10^{-\ell}$ and 1, on its main diagonal. To highlight the effect of the smallest singular value of A, we give results with $\ell = 2$ and $\ell = 8$. The vector b is formed as follows: $b = Ax$, where $x = \big[\cos(1), \dots, \cos(n)\big]^T$.

We perturb the input A with random noise H satisfying $h = \mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$ for various values of σ. In the first test, the elements H_{ij} are independent normal variables with mean 0 and variance σ^2; in other words, $h \sim \mathcal{N}(0, \sigma^2\Sigma)$ with $\Sigma = I_{n^2}$. In the second test, the elements H_{ij} are mutually independent and equal to $\pm\sigma A_{ij}$ each with probability 1/2; in other words, h follows a scaled Bernoulli distribution with $\Sigma = \mathrm{diag}(\mathrm{vec}([A]))$. This last covariance matrix comes from Fletcher's approach discussed in section 3.1.

For each value of σ we compute 1000 samples of H and of the resulting normwise relative error $\|F(A+H) - F(A)\|_F/\|F(A)\|_F$. We plot the sample median, as well as error bars representing the 5th and 95th sample percentiles, versus σ. Similarly, for the same 1000 random samples of H, we also show the sample median, as well as the 5th and 95th sample percentiles of the deterministic FOE (5.1), versus σ. We also plot the stochastic FOE (5.2), which is linear in σ.

In this simple example, expressions for the Fréchet derivative $F'(A)(H) = -A^{-1}Hx$ and the Jacobian matrix $J_A = -x^T \otimes A^{-1}$ are known. The spectral norm of the Jacobian is $\|J_A\|_2 = \|x\|_2\|A^{-1}\|_2$, which is known from the construction of A and b. As for the scaled norm $\|J_A\Sigma^{1/2}\|_F$, for each given covariance matrix Σ, we use the estimate $\eta_k(J_A\Sigma^{1/2})$ with $k = 5$ samples.

Results are plotted in Figure 1. In all cases, for sufficiently small σ, the stochastic FOE is an excellent estimate of the median relative error. Here "sufficiently small" is roughly $\sigma \approx \sigma_{\min}(A)$, the smallest singular value of A. We have not found a rigorous explanation for this. The deterministic FOE is roughly two orders of magnitude larger than the stochastic FOE. As discussed in section 3.4, the difference between the two may increase with increasing n. In these examples, the deterministic FOE, which involves the random variable $\|H\|_F$, is very concentrated about its mean: the 5th and 95th sample percentiles essentially overlap on the loglog plot. (This is due to the central limit theorem, $\|H\|_F^2$ being a sum of random variables.)

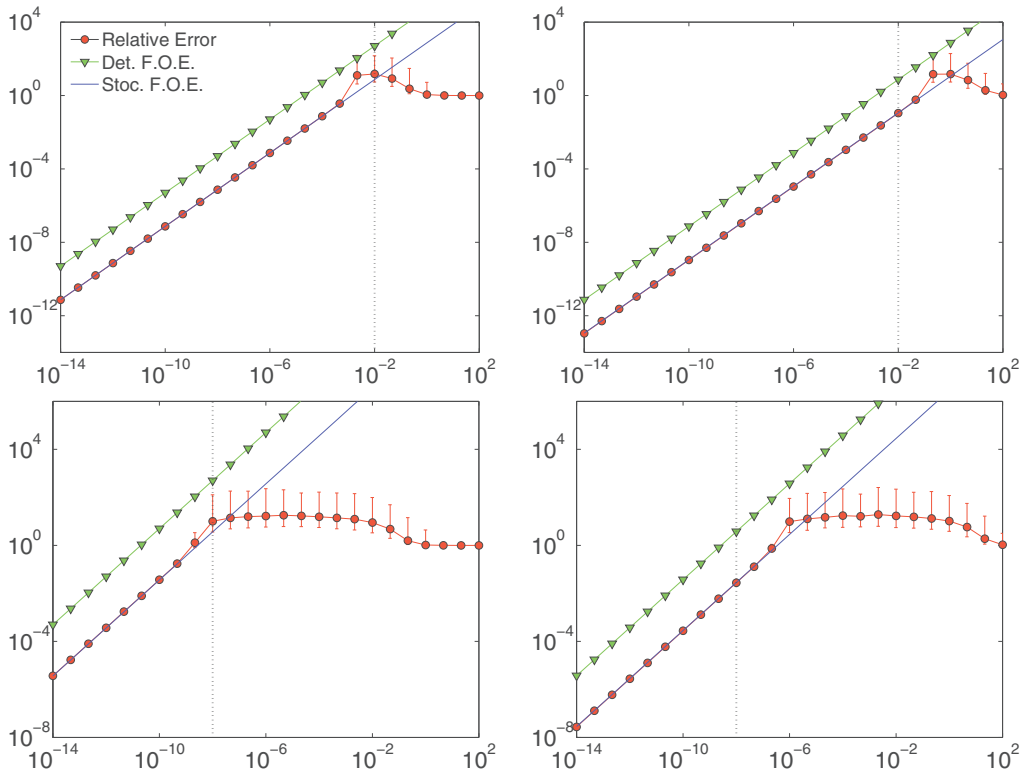

Figure 1. *Relative error, and deterministic and stochastic FOEs plotted versus σ. The function is $F(A) = A^{-1}b$, and the input A is perturbed by H satisfying $\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$. The dotted vertical lines represent $\sigma_{\min}(A)$, the smallest singular value of A. Top: $\sigma_{\min}(A) = 10^{-2}$. Bottom: $\sigma_{\min}(A) = 10^{-8}$. Left: Normal distribution, $\Sigma = I_{n^2}$. Right: Scaled Bernoulli distribution, $\Sigma = \mathrm{diag}(\mathrm{vec}([A]))$.*

For small σ the distribution of the relative error is also very concentrated; see the comments in section 3.3. For larger values of σ, the relative error seems to level off and concentrate at 1. Explaining the *large* σ asymptotics, as well as obtaining nonasymptotic error estimates, is the subject of ongoing research.

5.2. Linear least squares problems. Our next example involves a linear least squares problem,

$$F(b) = \arg\min_x \|b - Ax\|_2 = A^\dagger b,$$

where $A \in \mathbb{R}^{m\times n}$ has full column rank and $A^\dagger = (A^TA)^{-1}A^T$ denotes the Moore–Penrose generalized inverse.

We perturb the input b with random noise $h \sim \mathcal{N}(0, \sigma^2\Sigma)$. However, instead of fixing the problem dimensions and varying σ, in this example we fix $\sigma = 10^{-4}$ and vary m and n. Specifically, we test

- $n = 10^1, \ldots, 10^5$ and $m = 100n$ as well as
- $n = 10$ and $m = 10^1, \ldots, 10^7$.

Although it may seem artificial to create matrices with such a large aspect ratio, we purposefully do so to highlight the fact that when the domain of F is in a much larger space than its range, i.e., when $m \gg n$ in this case, then there can be a very big difference between the deterministic and stochastic FOEs. (See the discussion in section 3.4.)

Once again, A is formed from its SVD: $A = USV^T$. In each case S is diagonal with entries logarithmically equally spaced between 10^{-4} and 1, while U and V are Householder matrices:

$$u = \begin{bmatrix} \cos(1) \\ \vdots \\ \cos(m) \end{bmatrix}, \quad v = \begin{bmatrix} \sin(1) \\ \vdots \\ \sin(n) \end{bmatrix}, \quad U = I - 2uu^\dagger, \quad V = I - 2vv^\dagger.$$

We do not form U, S, V, or A explicitly to avoid running out of computer memory. We let $b = A\left[1, \ldots 1\right]^T + v$, where $v \sim \mathcal{N}(0, \sigma^2\Sigma)$.

The Jacobian matrix is simply $J_b = A^\dagger$, whose spectral norm is known from the construction of A and whose scaled Frobenius norm $\|J_A\Sigma^{1/2}\|_F$ we estimate using $\eta_k(J_A\Sigma^{1/2})$ with $k = 5$ samples.

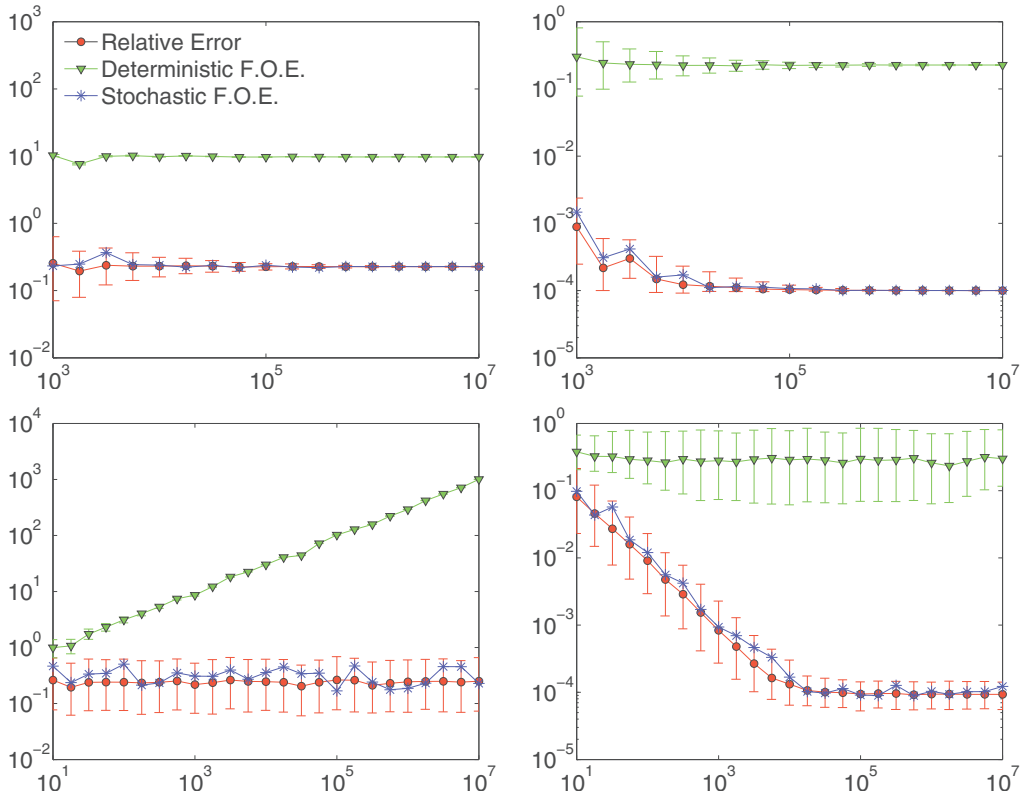

Figure 2. *Relative error, and deterministic and stochastic FOEs plotted versus m. The function is $F(b) = A^\dagger b$, and the input b is perturbed by $h \sim \mathcal{N}(0, \sigma^2\Sigma)$. Top: Constant aspect ratio m/n. Bottom: Increasing aspect ratio m/n. Left: $\Sigma = I_m$. Right: $\Sigma = \mathrm{diag}(\mathrm{vec}([b]))$.*

For each value of m, we compute 1000 samples of $h \sim \mathcal{N}(0, \sigma^2\Sigma)$. In Figure 2 we plot the

sample median as well as the 5th and 95th sample percentiles of the normwise relative error, along with the deterministic and stochastic FOEs, versus m. Once again, in these examples, the stochastic FOE is an excellent example of the median relative error for all values of m tested. On the other hand, the deterministic FOE can be several orders of magnitude larger and sometimes increasingly so with increasing m/n.

5.3. The method of conjugate gradients. As mentioned in the introduction, there has been some recent interest in the data assimilation and optimization communities in the sensitivity of Krylov subspace iterates (e.g., [22, 16, 17]). Let $x^{(j)}$ denote the jth iterate of the method of CG applied for solving $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. In previous work [7], we derived expressions for the Jacobian matrix of $x^{(j)}$ with respect to b. This led to bounds on the (deterministic) condition number of $x^{(j)}$ with respect to perturbations in b. Here we perform a numerical experiment to investigate the sensitivity of CG iterates to random noise in the right-hand side vector b.

Let $x^{(j)} = F(b)$ denote the jth iterate of CG. In this experiment we set $j = 10$. We perturb the input b with random noise h, where $h \sim (0, \sigma^2\Sigma)$ for various values of σ. As in the previous examples, we perform tests with h following a multivariate normal distribution and a scaled Bernoulli distribution, and we test both $\Sigma = I_n$ and $\Sigma = \mathrm{diag}(\mathrm{vec}([b]))$.

In this example, $n = 1000$ and the matrix $A \in \mathbb{R}^{n\times n}$ is created via its spectral decomposition: $A = U\Lambda U^T$. $U \in \mathbb{R}^{n\times n}$ is the Q factor in the QR factorization of a random matrix, and Λ contains the eigenvalues λ_i of A on its main diagonal. The eigenvalues λ_i are chosen logarithmically equally spaced between 0.1 and 1, with one extra eigenvalue 10^{-6} small with respect to the rest of the spectrum. The vector b is taken as $b = \big[\cos(1), \ldots, \cos(n)\big]^T$.

Because of the small problem dimensions, we can explicitly create the $n \times n$ Jacobian matrix J_b using automatic differentiation. We then directly compute its spectral norm and scaled Frobenius norm. As in the previous example, for each value of σ we compute 1000 samples of h and of the resulting normwise relative error $\|F(b+h)-F(b)\|_2/\|F(b)\|_2$. We plot the sample median, as well as the 5th and 95th sample percentiles of the normwise relative error, versus σ, along with the deterministic and stochastic FOEs of the relative error.

Results are plotted in Figure 3. In all four cases, the behavior is very similar. The stochastic FOE is an excellent estimate of the median relative error. This seems to be the case even for large values of σ. In other words, the CG iterates seem to behave linearly with respect to random perturbations. This is very surprising as Krylov methods are known for being highly *nonlinear*. We have not found a theoretical explanation for this behavior. As before, the deterministic FOE is a few orders of magnitude larger than the stochastic FOE.

6. Conclusions. We have defined a notion of stochastic condition number of a matrix function F to random noise H in its input. First we gave sufficient conditions for the first-order RMS approach from [21, 6] to be applicable. Our sensitivity analysis based on the work of Stewart [20] applies much more generally—for all Fréchet differentiable functions and any random noise H such that $\mathrm{vec}(H) \sim (0, \sigma^2\Sigma)$, regardless of the distribution of $\mathrm{vec}(H)$. Consequently, we proposed the stochastic condition number (3.12). Analogously to the deterministic condition number (1.2), this measures the asymptotic sensitivity of F in the small perturbation limit, but in this setting "small" refers to the norm of the covariance matrix of $\mathrm{vec}(H)$. The stochastic condition number can be computed very efficiently by random

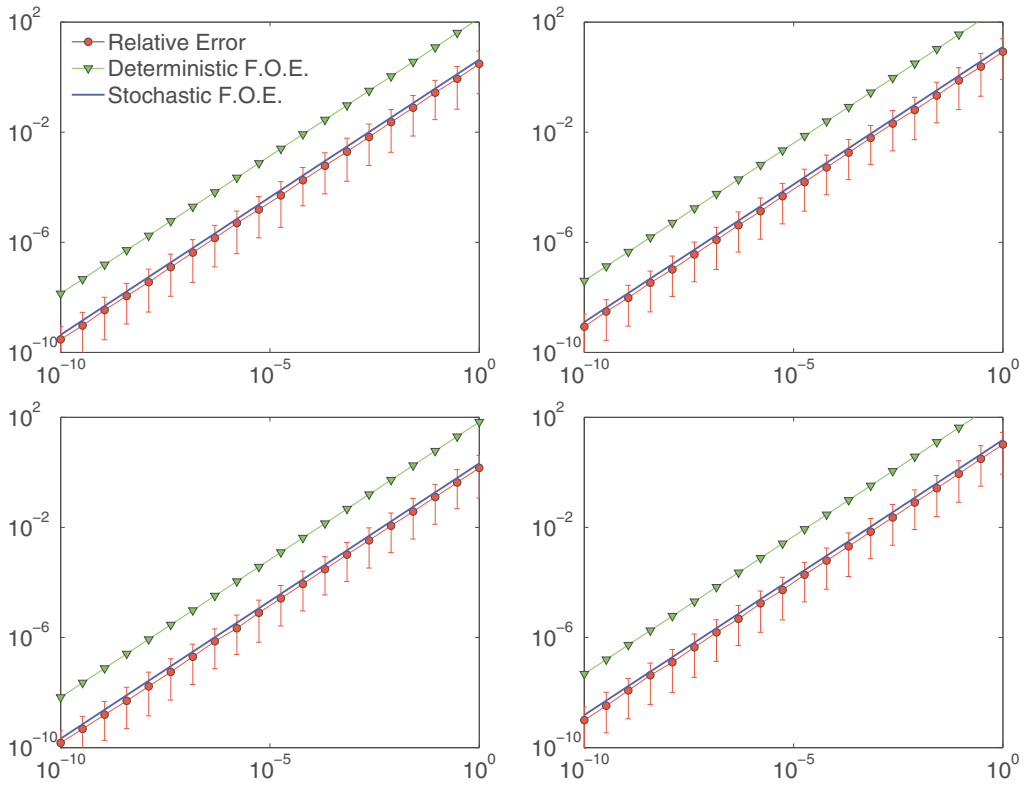

Figure 3. *Relative error, and deterministic and stochastic FOEs plotted versus σ. The function is $x_{10} = F(b)$, and the input b is perturbed by $h \sim (0, \sigma^2\Sigma)$ following two distributions. Top: Normal distribution. Bottom: Scaled Bernoulli distribution. Left: $\Sigma = I_n$. Right: $\Sigma = \mathrm{diag}(\mathrm{vec}([b]))$.*

sampling: few samples are required to obtain excellent order-of-magnitude estimates with probability close to 1. Our numerical experiments indicate that the resulting error estimates are excellent measures of the median error when A is subjected to random noise.

One drawback of the stochastic condition number is that, like the deterministic condition number, it leads only to a first-order error estimate (valid for sufficiently small σ). The behavior of the error for large values of σ is problem dependent. In our examples involving linear systems, the error estimate was only valid for values of σ smaller than roughly the smallest singular value of A. Surprisingly, however, for perturbations to the right-hand side vector in CG, the small σ asymptotics seemed to be descriptive for arbitrary σ. Obtaining *global* bounds on

$$\mathrm{Prob}\Big\{\|F(A+H) - F(A)\|_F \geq \tau\Big\}$$

for these matrix functions is the subject of current research. Finally, in this work we have restricted ourselves to real matrices and real perturbations. An extension to the complex case seems possible, for example, by considering complex random perturbations $H = H_R + iH_I$ (where $i = \sqrt{-1}$ and H_R, H_I are real random matrices) of complex matrices $A \in \mathbb{C}^{m\times n}$. This

topic is also currently under investigation.

Appendix. Proof of Lemma 3.1.

Proof. For the first part, using linearity of expectation, we obtain

$$\big|\,\mathbb{E}\big\{(u^T h)\xi(h)\big\}\big| = \Big|\,\mathbb{E}\Big\{\sum_{j=1}^{n} u_j h_j \sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 2}} \alpha_v h^v\Big\}\Big| = \Big|\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 2}}\sum_{j=1}^{n} \alpha_v u_j \mathbb{E}\{h_j h^v\}\Big|.$$

Repeatedly using the Cauchy–Schwarz inequality $|\,\mathbb{E}\{x_1 x_2\}\,| \leq \sqrt{\mathbb{E}\{x_1^2\}\,\mathbb{E}\{x_2^2\}}$ and the fact that $\mathbb{E}\{h_i^k\} \leq c\sigma^k$, we obtain

$$\mathbb{E}\{h_j h^v\} = \mathbb{E}\{h_j h_1^{v_1}, \ldots, h_n^{v_n}\} \leq c\sigma^{|v|+1}.$$

Therefore, for $\sigma < 1$,

$$\begin{aligned}\frac{\big|\,\mathbb{E}\big\{(u^T h)\xi(h)\big\}\big|}{\sigma^2} &\leq \Big|\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 2}}\sum_{j=1}^{n} \alpha_v u_j c\sigma^{|v|-1}\Big| \\ &\leq c\|u\|_1\Big(\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 2}} |\alpha_v|\Big)\sum_{\omega=2}^{\infty}\sigma^{\omega-1} = c\|u\|_1\Big(\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 2}} |\alpha_v|\Big)\frac{\sigma}{1-\sigma},\end{aligned}$$

which converges to 0 as $\sigma \to 0$. (The last summation is bounded since $\xi(x)$ is absolutely convergent at $x = [1, \ldots, 1]^T$.)

Similarly, for the second part,

$$\begin{aligned}\mathbb{E}\big\{\xi(h)^2\big\} &= \mathbb{E}\Big\{\Big(\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 2}} \alpha_v h^v\Big)\Big(\sum_{\substack{w\in\mathbb{N}_0^n\\ |w|\geq 2}} \alpha_w h^w\Big)\Big\} \\ &= \mathbb{E}\Big\{\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 4}} \beta_v h^v\Big\} = \sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 4}} \beta_v \mathbb{E}\{h^v\} \leq \sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 4}} \beta_v c\sigma^{|v|},\end{aligned}$$

so that when $\sigma < 1$,

$$\frac{\mathbb{E}\big\{\xi(h)^2\big\}}{\sigma^2} \leq c\Big(\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 4}} |\beta_v|\Big)\sum_{\omega=4}^{\infty}\sigma^{\omega-2} = c\Big(\sum_{\substack{v\in\mathbb{N}_0^n\\ |v|\geq 4}} |\beta_v|\Big)\frac{\sigma^2}{1-\sigma},$$

which once again converges to 0 as $\sigma \to 0$. ∎

REFERENCES

[1] A. H. Al-Mohy and N. J. Higham, *Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation*, SIAM J. Matrix Anal. Appl., 30 (2009), pp. 1639–1657.

[2] A. H. Al-Mohy and N. J. Higham, *The complex step approximation to the Fréchet derivative of a matrix function*, Numer. Algorithms, 53 (2010), pp. 133–148.

[3] A. H. Al-Mohy, N. J. Higham, and S. D. Relton, *Computing the Fréchet derivative of the matrix logarithm and estimating the condition number*, SIAM J. Sci. Comput., 35 (2013), pp. C394–C410.

[4] H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix*, J. ACM, 58 (2011), pp. 8:1–8:34.

[5] E. Chow and Y. Saad, *Preconditioned Krylov subspace methods for sampling multivariate Gaussian distributions*, SIAM J. Sci. Comput., 36 (2014), pp. A588–A608.

[6] R. Fletcher, *Expected conditioning*, IMA J. Numer. Anal., 5 (1985), pp. 247–273.

[7] S. Gratton, D. Titley-Peloquin, P. Toint, and J. Tshimanga Ilunga, *Differentiating the method of conjugate gradients*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 110–126.

[8] A. Griewank and G. Corliss, *Automatic Differentiation of Algorithms: Theory, Implementation, and Application*, SIAM, Philadelphia, 1991.

[9] T. Gudmundsson, C. S. Kenney, and A. J. Laub, *Small-sample statistical estimates for matrix norms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 776–792.

[10] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.

[11] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.

[12] N. J. Higham and A. H. Al-Mohy, *Computing matrix functions*, Acta Numer., 19 (2010), pp. 159–208.

[13] N. J. Higham and L. Lin, *An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 1341–1360.

[14] C. S. Kenney and A. J. Laub, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput., 15 (1994), pp. 36–61.

[15] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist., 28 (2000), pp. 1302–1338.

[16] A. M. Moore, H. G. Arango, and G. Broquet, *Estimates of analysis and forecast errors derived from the adjoint of 4D-Var*, Monthly Weather Rev., 140 (2012), pp. 3183–3203.

[17] J. J. Moré and S. M. Wild, *Estimating computational noise*, SIAM J. Sci. Comput., 33 (2011), pp. 1292–1314.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.

[19] J. R. Rice, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[20] G. W. Stewart, *Stochastic perturbation theory*, SIAM Rev., 32 (1990), pp. 579–610.

[21] A. M. Turing, *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math., 1 (1948), pp. 287–308.

[22] Y. Zhu and R. Gelaro, *Observation sensitivity calculations using the adjoint of the gridpoint statistical interpolation (GSI) analysis system*, Monthly Weather Rev., 136 (2008), pp. 335–351.