Three Novel Methods for Determining Motor Threshold with Transcranial Magnetic Stimulation Outperform Conventional Procedures

Boshuo Wang[1], Angel V. Peterchev[1,2,3,4], Stefan M. Goetz[1,2,4,5*]

[1] Department of Psychiatry and Behavioral Sciences, School of Medicine, Duke University, Durham, NC, USA
[2] Department of Electrical and Computer Engineering, School of Engineering, Duke University, Durham, NC, USA
[3] Department of Biomedical Engineering, School of Engineering, Duke University, Durham, NC, USA
[4] Department of Neurosurgery, School of Medicine, Duke University, Durham, NC, USA
[5] Department of Engineering, School of Technology, University of Cambridge, Cambridge, UK
[*] Corresponding author: stefan.goetz@duke.edu, smg84@cam.ac.uk

Keywords: transcranial magnetic stimulation, motor threshold, threshold finding methods, parametric estimation, Bayesian estimation, stochastic root-finding

Abstract

Objective. Thresholding of neural responses is central to many applications of transcranial magnetic stimulation (TMS), but the stochastic aspect of neuronal activity and motor evoked potentials (MEPs) challenges thresholding methods. We analyzed existing methods for obtaining TMS motor threshold and their variations, introduced new methods from other fields, and compared their accuracy and speed.

Approach. In addition to existing relative-frequency methods, such as the five-out-of-ten method, we examined adaptive methods based on a probabilistic motor threshold model using maximum-likelihood (ML) or maximum *a-posteriori* (MAP) estimation. To improve the performance of these adaptive estimation methods, we explored variations in the estimation procedure and inclusion of population-level prior information. We adapted a Bayesian estimation method which iteratively incorporated information of the TMS responses into the probability density function. A family of non-parametric stochastic root-finding methods with different convergence criteria and stepping rules were explored as well. The performance of the thresholding methods was evaluated with an independent stochastic MEP model.

Main Results. The conventional relative-frequency methods required a large number of stimuli, were inherently biased on the population level, and had wide error distributions for individual subjects. The parametric estimation methods obtained the thresholds much faster and their accuracy depended on the estimation method, with performance significantly improved when population-level prior information was included. Stochastic root-finding methods were comparable to adaptive estimation methods but were much simpler to implement and did not rely on a potentially inaccurate underlying estimation model.

Significance. Two-parameter MAP estimation, Bayesian estimation, and stochastic root-finding methods have better error convergence compared to conventional single-parameter ML estimation, and they all require significantly fewer TMS pulses for accurate estimation than conventional relative-frequency methods. Stochastic root finding appears particularly attractive due to the low computational requirements, simplicity of the algorithmic implementation, and independence from potential model flaws in the parametric estimators.

1. Introduction

1.1. Significance and challenges of thresholding in TMS

Transcranial magnetic stimulation (TMS) is a noninvasive technique to activate neurons in the brain [1], [2]. Certain pulse rhythms can further modulate neural circuits, i.e., change how they process endogenous signals [3]. A key parameter of the stimulation pulses is their strength, which influences not only how many neurons but also which neuron types are stimulated [4]–[6]. As the TMS pulse strength affects the activation of neuron types with different functions, it also influences the efficacy of neuromodulation protocols and can even reverse the direction of the effect [7]–[11]. Finally, the stimulation strength is a key parameter not only for ensuring an effective and reproducible procedure but also for safety, and strength limitations for various modulatory protocols are recommended to avoid seizures [12]–[14].

Typically, the TMS pulse strength is individualized to a specific subject's or patient's anatomy and physiology based on a measured response threshold. A commonly used threshold is the motor threshold based on activation of finger muscles to stimulation of their representation in the primary motor cortex, but other types of thresholds could also be used, such as phosphene threshold of visual artifacts elicited in the visual cortex, threshold of cortical silent period, and transcranial evoked potentials recorded from electroencephalography (TMS-EEG) [15]–[22]. For the motor threshold, the muscle movement following stimulation in the primary motor cortex is observed visually in some cases, typically in clinical practice, whereas in other cases, typically in research, the size of motor evoked potentials (MEPs) is measured with electromyography (EMG) at the target muscle [23]–[25]. The motor threshold is routinely also used as a reference for other brain targets and its detection is recommended before any further neuromodulatory intervention, even for recurring subjects or patients [26]–[28]. Furthermore, the threshold and its changes can serve as a quantitative metric, e.g., for the diagnosis of lesions [29]. The centrality of the motor threshold to TMS procedures underscores the need for a well-defined, rapid, accurate, and unbiased thresholding procedure.

However, the MEPs in response to stimulation in the primary motor cortex are highly stochastic due to the dynamic variability in cortical neurons and the corticospinal tract—a pulse administered twice with identical strength can lead to a maximum MEP response close to saturation once and no response at all in the other case [30]–[32]. Moreover, the variability is far from GAUSSIAN and appears to involve several variability sources [32]–[35]. Consequently, accuracy of the estimation procedure is hard to guarantee and achieving it with a suboptimal estimation algorithm can be time consuming.

1.2. Existing motor threshold estimation methods

The TMS motor threshold, x_{th}, is usually defined as the lowest stimulation pulse amplitude, x, given as a percentage of the maximum stimulator output (MSO), leading to a peripheral muscle response, r, with a 50% probability [27], [28], [36],

$$x_{\mathrm{th}} = \underset{x}{\operatorname{argmin}}(P(r = 1|x) = 50\%). \tag{1}$$

The binary response r is defined as

$$r = \begin{cases} 1 & \text{movement observed, or } V_{\mathrm{pp}} \geq 50\ \mu\mathrm{V} \\ 0 & \text{movement absent, or } V_{\mathrm{pp}} < 50\ \mu\mathrm{V} \end{cases}, \tag{2}$$

where V_{pp} is the peak-to-peak EMG amplitude [37]. Historically, the value of 50 µV was chosen relatively arbitrarily as a low level that is not far from the noise limit of most EMG detection systems [38]–[40]. The most common practice for motor threshold finding, or "threshold hunting", is estimating the relative frequency of positive and negative responses by appropriate stepping of the stimulus amplitude to find the 50% probability level. Therefore, at each test pulse amplitude, a number of stimuli—usually up to ten—have to be administered, hence the name "five-out-of-ten method". Such an approximation of the probability by a relative frequency within just ten stimuli can be rather imprecise, and a larger number of trials, e.g., twenty or more, per stimulation amplitude is required to obtain more reproducible results for clinical and research purposes [28], [41]. Given its simplicity, the relative-frequency method is widely used and often serves as a benchmark for other methods, despite concerns regarding its mathematical validity or usefulness [41]–[43].

Another well-described method based on the relative response frequency was proposed by MILLS and NITHI [39]. Instead of detecting the threshold directly, it uses one point above threshold (*upper threshold*) and one below (*lower threshold*), and the threshold estimate itself is defined as the mean of these two points. Mathematically, this relates to a local linearization of the sigmoid curve of the neuronal response in dependence of the stimulation amplitude in the probability space. Like the relative-frequency method, both upper and lower thresholds are defined statistically—all ten responses suprathreshold and subthreshold, respectively. Both the relative-frequency and MILLS-NITHI methods require a high number of stimuli [40], [44] and consequently a long duration to complete due to the necessary pause between two stimuli for preventing inter-pulse correlation [45]–[47]. Furthermore, a large number of stimuli also bears a risk of unintended neuromodulation and side effects [14].

Instead of determining relative frequencies, an alternative thresholding method performs a maximum-likelihood (ML) test for a predefined model describing the sigmoidal probability distribution of the binary response as a function of stimulation amplitude—most commonly of a cumulative GAUSSIAN function with two parameters, the midpoint/threshold and slope/spread[1]. This maximum likelihood estimation (MLE) method was introduced by Awiszus [48] and was adapted from psychophysics where responses have strong stochastic characteristics as well [49], [50]. The result of such parameter estimation by sequential testing (PEST) provides the most likely parameters, which include the threshold. An essential condition is a good selection of the tested stimulus amplitudes [51]. In

[1] The midpoint or threshold determines the horizontal location of the distribution and the slope at threshold or spread determines its shape, i.e., horizontal stretching. When referring to the distribution's shape parameter, slope and spread are sometimes used interchangeably, although they are inversely related to each other.

contrast to frequency-based methods, each amplitude is usually applied just once. Mishory et al. [44] reported that the number of required stimuli and time for determining threshold was markedly lower for ML-PEST compared to the MILLS–NITHI two-threshold method, whereas both methods have similar accuracy.

1.3. Limitations of existing methods and alternative approaches

Existing motor threshold estimation methods have some limitations that can be addressed to improve the accuracy and speed of the procedure. A key feature that can notably influence performance is the stepping rule [51], which chooses the next stimulation amplitude adaptively by taking into account the information from the previous observations. For the next stimulation amplitude, the ML-PEST [48] chooses the MLE of the threshold, which is obtained from a reduced single-parameter model of the response and uses a fixed nonstochastic relationship for the spread of the probability function. This relationship is assigned based on averages from representative measurements [48], and therefore does not reflect individual variability. A limitation of this approach is that by not allowing any deviation from a fixed spread–threshold relationship, the method may place all stimuli systematically too high or too low with respect to the optimal stepping for an individual subject, which can lead to slow stepping and large threshold misestimations if the initial stimulation amplitude is poorly chosen [52], [53].

The threshold estimation can be improved by a full two-parameter MLE in which both the threshold and spread are free to vary. Incorporating additional statistical information from prior or typical results of the thresholding procedure can further improve the estimation [52], [54], which is a method adapted from the field of psychophysics [51]. Although Qi et al. called it a BAYESIAN method [54], this name is not adopted here since the approach incorporates *a-posteriori* information into a classical FISHER MLE using BAYES' theorem. This has been known in mathematics [55] and applied fields [56] as maximum *a-posteriori* (MAP), a term we adopt in this paper.

While BAYES' theorem has been used for MAP estimation to improve the TMS threshold estimate [52], [54], both ML and MAP methods are point estimators. In contrast to these point estimators, strictly BAYESIAN methods that are consequently and completely premised on BAYESIAN statistics [57]–[59] have not been used in TMS. In this text, the BAYESIAN style particularly refers to the updating rule for the probability function after new observations, which should therefore follow the BAYESIAN learning paradigm. We adapted the BAYESIAN estimation method from KONTSEVICH and TYLER [57] for TMS motor threshold estimation, using the same model as in the point estimators.

Both the relative-frequency methods and the parameter estimation methods adapted from psychophysics reduce the TMS motor responses to binary events (2). Whereas both psychophysics and TMS motor threshold determination by visual observation of finger twitches deal only with binary responses to a stimulus (perceived or not perceived), the use of EMG for measuring muscle activation by TMS provides a continuous response strength and thus richer information. The EMG motor threshold is not a binary transition between no response and response but a point on the dose–response curve close to the lowest stimulation strength that elicits an excitatory response detectable with conventional EMG recording and analysis methods [60]–[62]. Further, the motor threshold is likely

not the lowest stimulation strength to evoke detectable responses, as recent studies demonstrate that the 50 μV threshold is likely closer to the middle of the logarithmic dose–response curve of MEPs when it is separated from the baseline noise [63], [64]. The smoother transition of the expected value of the responses in dependence of the stimulation strength may allow the use of the MEP amplitude value to find the threshold faster. To consider this additional analog information of the response amplitude appropriately, stochastic root-finding methods [65], [66] (also known as stochastic approximation) are adapted and evaluated here.

1.4. Organization of this paper

We first describe existing TMS thresholding methods as well as the aforementioned improvements and introduce the BAYESIAN estimation and stochastics root-find methods, covering a total of three classes and 25 variations of methods. These methods are divided into two categories, based on whether a parametric model describing the motor response underlies the method. To evaluate and compare the thresholding methods, an independent testbed incorporating a detailed stochastic MEP model [61] is used, with 25,000 virtual subjects. The performance of the thresholding methods is analyzed based on their accuracy of the threshold estimate and speed to acquire threshold (relative-frequency methods) or convergence rate of the threshold error (other methods). Both the relative and absolute errors serve for a comparison of the accuracy, and statistics including the median, the upper and lower quartiles and adjacent values, and outliers were calculated for the virtual population. The best performing method of each class are identified and the trade-off between different methods is discussed. The results show that the three alternative approaches—two-parameter MAP estimation, Bayesian estimation, and stochastic root-finding—outperform single-parameter ML estimator, and they all require significantly fewer TMS pulses for accurate motor threshold estimation than the conventional relative-frequency methods.

2. Methods

2.1. Overview of thresholding methods

The estimation of a neuronal property such as the excitation threshold can be performed in two fundamentally different ways. The model-based parameter estimators (ML, MAP, and the BAYESIAN methods) require a probabilistic description of the input–output black box being analyzed in dependence of a certain parameter set. An inaccurate model will very likely deteriorate the quality of the estimate. In contrast, the early approaches, namely the five-out-of-ten or ten-out-of-twenty relative-frequency method based on recommendations of the International Federation of Clinical Neurophysiology (IFCN) [27], [28] or the MILLS–NITHI two-threshold method [39], do not need such a predefined model description as they refer directly to the definition of the threshold. The stochastic root-finding methods can also be counted among this class of non-parametric systems. In the following, the existing and several new or adapted algorithms are shortly described and compared for the application in TMS. Computation and data analysis were performed with MATLAB (versions 2021a and 2023a, The Mathworks, Natick, MA, USA) on the Duke Compute Cluster, with the code of the thresholding methods available at [67].

2.2. Nonparametric methods

2.2.1. IFCN relative-frequency method

The procedure described by the IFCN [27], [28], [68] leaves freedom of interpretation of some details. For the analysis in this paper, the relative-frequency method was defined as follows to realistically represent the implementation in many laboratories:

a) The starting amplitude is known from the hotspot search (see section 2.4.2 Simulation routine for further details on the starting amplitude), in which the corresponding neuron population being responsible for activating the observed muscle group (e.g., the *first dorsal interosseus* or the *abductor pollicis brevis*) is located with test stimuli on the primary motor cortex. The required amplitude adjustments in this step lead to an amplitude not too far above the desired threshold as spatial focality would be compromised with stronger stimuli.

b) At a given stimulation amplitude, a number of stimuli are applied up to a maximum of ten or twenty until either the count of positive responses reaches half the number (five or ten, respectively) or that of negative responses exceed half the number (six or eleven, respectively).

c) Being suprathreshold by definition in the situation of positive responses accounting for at least half the number of stimuli, the stimulation amplitude is stepped downward, typically by 1% or 2% MSO [28], [68], and Step b) is repeated.

d) The latter outcome in Step b) precludes 50% positive responses within the number of stimuli and the amplitude is subthreshold. The threshold itself is defined as the last level before this termination point.

Versions with the number of stimuli of ten and twenty and step size of 1% and 2% MSO were performed to show the effect of the number of stimuli per step and the step size on the accuracy and the number of stimuli required for the methods (Table 1).

Table 1 Summary of relative-frequency-based methods

Method	Variation	Step size (% MSO)
IFCN relative-frequency	Five-out-of-ten	1
		2
	Ten-out-of-twenty	1
		2
MILLS–NITHI	Original	1
		2
	Modified	1
		2

2.2.2. *MILLS–NITHI two-threshold method*

The original MILLS–NITHI method [39] was carried out as follows:

a) The subject is stimulated with pulses starting at an amplitude of 20% MSO and increasing in 10% MSO steps until the first positive response is identified.
b) Applying up to 10 pulses per stimulation amplitude, the stimulation is then decreased in 1% MSO steps when a positive response is observed. The intensity at which all 10 stimuli produce no response is designated as the lower threshold.
c) Starting from the lowest amplitude that has not resulted in a "no response", i.e., the last amplitude that has one stimulus with positive response during Step b), and increasing the stimulation in 1% MSO steps, the upper threshold is designated as the minimum intensity at which 10 stimuli all produce positive responses.
d) The threshold is calculated as the mean of the lower and upper thresholds.

In our study, a modified version of the MILLS–NITHI method was performed using the same starting amplitude as the IFCN relative-frequency method. The first negative response was identified by decreasing the amplitude in 10% steps and the LT and UT were then similarly obtained as in the original methods but in reversed order. Step sizes of 1% and 2% MSO were tested (Table 1).

2.2.3. *Stochastic root-finding methods*

The IFCN relative-frequency and MILLS–NITHI methods described above both require binary information reduction through quantization of the response (above versus below 50 μV). In contrast to such quantization, robust numeric root-finding algorithms that are still relatively unknown in TMS can directly act on analog data, e.g., the voltage amplitude of the EMG response. Preserving valuable continuous information in the response, such root-finding approaches could enable an improvement of convergence. The sequential formulation proposed here was derived from the method of ROBBINS and MONRO [65], [66], with the neuronal variability playing the role of a noise source. The requirements for this approach are rather weak and comprise constant monotonicity on average [65].

The next stimulation amplitude, x_{i+1}, was defined relative to the current one using an iterative analog control sequence (ACS) that was intended to converge towards the threshold:

$$x_{i+1} = x_i - a_i \cdot \Delta y_i = x_i - a_i \cdot (y_i - y_{\mathrm{th}}), \tag{3}$$

where $i \geq 1$ was the step number, a_i was the step size of the control sequence, x_i denoted the stimulation amplitude, and y_i and y_{th} were logarithms of the response ($y_i = \lg(V_i/[1\ \mathrm{V}])$) observed at step i and at the real threshold V_{th} (50 μV peak-to-peak by definition), respectively. The log-transformation accounted for the stochastic characteristics of EMG responses and the first-approximation log-normal distribution of the response amplitude following an inductive stimulus in the motor cortex [46]. The resulting distribution of y_i had reduced skewness and should be closer to GAUSSIAN and therefore statistically well-behaved. Performing the whole procedure in the

logarithmic space also granted the requirement of constant monotonicity. The iteration for the EMG response thus became

$$x_{i+1} = x_i - a_i \lg(V_i/V_{\text{th}}). \tag{4}$$

We implemented several versions of the sequence of step sizes a_i to compare the convergence properties (Table 2). In the basic version ACS1-H, the step size followed a harmonic series with almost sure (a.s.) convergence[2]

$$a_i = a_0/i, \quad \text{for } i \geq 1. \tag{5}$$

Following the style of Delyon and Juditsky [69], adaptive methods were utilized to accelerate convergence by adjusting the control sequence at step i ($i > 1$) only when the response changed from subthreshold to suprathreshold or vice versa from the previous step ($i - 1$), i.e., following a sign change in Δy_i: $\text{sgn}(\Delta y_i) \neq \text{sgn}(\Delta y_{i-1})$. For harmonic convergence, the corresponding adaptive control sequence ACS1-HA was

$$a_i = a_0/(c_i + 1), \quad \text{for } i \geq 1, \tag{6}$$

where c_i was the number of sign changes of Δy_i up to step i,

$$c_i = \begin{cases} c_{i-1} & \text{sgn}(\Delta y_i) = \text{sgn}(\Delta y_{i-1}) \\ c_{i-1} + 1 & \text{sgn}(\Delta y_i) \neq \text{sgn}(\Delta y_{i-1}) \end{cases}, \quad \text{for } i > 1, \tag{7}$$

with $c_1 \triangleq 0$. Unlike classical mathematical problems, the TMS sequence has only a very limited number of steps. Thus, quick convergence seems more attractive here rather than any statement for the case of infinite duration. To test faster geometric convergence of $2^{-i/2}$, ACS1-GA used an adaptive control sequence of

$$a_i = a_0 \cdot 2^{-c_i/2}. \tag{8}$$

Further variations of these methods were designed to examine the convergence rate and performance. To take advantage of both fast convergence and the a.s. convergence properties, a hybrid method ACS1-GHA was designed to switch the control sequence a_i from geometric to harmonic convergence when the threshold estimate was close to the true threshold and the expected change in x_i ($|\Delta x_i| = |a_i \cdot \Delta y_i|$) was small—less than 1.5% MSO. Quantized versions, i.e., digital control sequences (DCS), were implemented for comparison as well. The analog value Δy_i was reduced using a modified sign function to $\delta y_i = \text{sgn}(\Delta y_i) \in \{-1, +1\}$ for the two different cases of below threshold and above (or equal) threshold. DCS allowed the stochastic root-finding method to be used also without EMG recordings, just "response" and "no response" inputs, in a manner compatible with the other estimation methods. Second-order versions of the analog control sequences (ACS2) were also implemented, in which the next stimulation amplitude was defined using information from the previous two steps

$$x_{i+1} = x_i - a_i \cdot (\Delta y_i + w \cdot \Delta y_{i-1}), \quad \text{for } i > 1, \tag{9}$$

with w being the weight for the second-order term.

---

[2] Defined as convergence with probability one: $x_i \xrightarrow{\text{a.s.}} \tilde{x}: \forall \epsilon > 0, \exists\, I, \delta: P(|x_i - \tilde{x}| > \delta) < \epsilon, \forall\, i > I$.

Table 2 Summary of root-finding methods

Method	Control sequence	Order	Convergence of a_i
ACS1-H	Analog	1	Harmonic (i^{-1}), non-adaptive
DCS1-H	Digital	1	
ACS2-H	Analog	2	
ACS1-HA	Analog	1	Harmonic (i^{-1}), adaptive
DCS1-HA	Digital	1	
ACS2-HA	Analog	2	
ACS1-GA	Analog	1	Geometric ($2^{-i/2}$), adaptive,
DCS1-GA	Digital	1	
ACS2-GA	Analog	2	
ACS1-GHA	Analog	1	Geometric ($2^{-i/2}$) switching [a] to harmonic (i^{-1}), adaptive

[a] Transition when the maximum change of x_i, i.e.,$|\Delta x_i| = |a_i \cdot \Delta y_i|$, is expected to be less than 1.5% MSO.

The definition of appropriate initial step size a_0 was assisted by the iteration formulation (3), which resembled a slope [70] after rearrangement

$$a_i^{-1} = \frac{y_{\text{th}} - y_i}{x_{i+1} - x_i}. \tag{10}$$

For reaching the threshold quickly, a_0 should accordingly be not too far from the cotangent of the line connecting the present stimulation–response (x_i, y_i) towards the threshold $(x_{\text{th}}, y_{\text{th}})$. Larger values increase the speed of approach of x_i towards x_{th}, but may cause the stepping to overshoot and oscillate around the target, whereas smaller numbers promote lengthy one-sided convergence. Statistical information about typical and maximum slopes can inform the choice of a_0. For the simulations in this paper, the default initial step size was chosen such that it approximated typical cotangents around the inflection point of the recruitment curve (see section 2.4.1 Response model)

$$a_0 = \frac{0.2}{\lg(10^{-2}) - \lg(10^{-5})} \approx 0.067. \tag{11}$$

Here $\lg(10^{-5})$ and $\lg(10^{-2})$ (i.e., 10 μV and 10 mV, respectively) were approximate response amplitudes near the EMG noise floor and saturation level at the lower and higher ends of the input–output curve, respectively, and 0.2 (20% MSO) was a sufficient change in stimulus strength to cover the transition of the typical sigmoidal response curve. Initial step size from one tenth to ten times the default values evenly sampled on a logarithmic scale were also tested. In second-order sequences, the weight of the second-order term also affects convergence. Larger/positive w speeds up the approach to threshold, but could generate oscillations depending on the value of a_0; smaller/negative w reduces the speed of convergence, but could help to mitigate the oscillation if a_0 is chosen too large. A default value of −0.1 was used and values between −1 and 1 were tested for ACS2-HA.

2.3. Parametric methods

2.3.1. Basic maximum-likelihood estimation method

The implementation of the basic MLE method (ML-PEST) followed substantially the suggestions and definitions by Awiszus [48]. The probability space was described by a cumulative GAUSSIAN model with two degrees of freedom—the threshold $t = x_{\mathrm{th}}$ and the spread s:

$$p(r_i = 1|x_i, t, s) = \frac{1}{s\sqrt{2\pi}} \int_{-\infty}^{x_i} \exp\left(-\frac{(x-t)^2}{2s^2}\right) \mathrm{d}x. \tag{12}$$

Here, r_i denoted the binary response (2) of a single stimulus x_i, and the slope at the threshold, $\left(s\sqrt{2\pi}\right)^{-1}$, was inversely related to the spread.

The likelihood function was defined for independent events

$$p(\mathbf{r}|\mathbf{x}, t, s) = \prod_{\{i|r_i=1\}} p(r_i = 1|x_i, t, s) \prod_{\{i|r_i=0\}} \left(1 - p(r_i = 0|x_i, t, s)\right), \tag{13}$$

where $\mathbf{x}$ and $\mathbf{r}$ were the lists of stimuli and corresponding responses, respectively. According to the classical FISHER approach, the maximum of this expression for the arguments t and s provided the most likely parameter set ($\mathrm{argmax}_{(t,s)}\, p(\mathbf{r}|\mathbf{x}, t, s)$). Due to the strict monotonicity of the logarithm, the product can be split into an equivalent sum, which formed the log-likelihood and was also numerically more robust. The MLE of t provided both the threshold estimate at the current step and stimulation amplitude at the next step.

Previous implementations of MLE (13) (e.g., MTAT and ATH-tool, [48], [71], [72]) were simplified by using only the threshold parameter t, whereas the spread parameter was fixed to the value $s = 0.07 \cdot t$. We replicated this method (MLE-1) and also performed a full MLE with both parameters (MLE-2) (Table 3). Two approaches were implemented and compared to find the threshold t that maximize the likelihood function. The first approach directly maximized the likelihood function over t (using MATLAB function `fmincon`, applied on the negative of the log-likelihood). This approach, however, could be unstable during the initial steps when there were few amplitude and response samples $(\mathbf{r}, \mathbf{x})$ and the likelihood function had a plateau, i.e., a range of t or (t, s) combinations resulted in the same maximum. The second approach explicitly evaluated the likelihood function and calculated the mean of the argument t over the range of the plateau [48]. For the explicit evaluation, the parameter space was discretized with $\Delta t = 0.002$ for $t \in (0, 1.3]$, and 100 logarithmical-distributed samples s were chosen from $[0.005, 0.5]$ for MLE-2. The upper bound of the threshold parameter exceeded one, which is explained in Section 2.4.1 Response model.

2.3.2. Maximum-a-posterior estimation method

Modified estimation methods (MAP-1 and MAP-2, Table 3) were designed based on a MAP-like approach which inversed the likelihood ratio in (13) [52], [54] and directed the likelihood function backwards from the

perspective of the causality of the model. For incorporating statistical information, BAYES' theorem was used to modify the likelihood function (13) to improve the performance for common cases per

$$p(\boldsymbol{\theta}|\mathbf{x},\mathbf{r}) = \frac{p(\boldsymbol{\theta}) \cdot p(\mathbf{r}|\mathbf{x},\boldsymbol{\theta})}{p(\mathbf{r}|\mathbf{x})}, \tag{14}$$

in which the vector $\boldsymbol{\theta} = (t, s)$ denoted the parameter set. The additional information about the distribution of typical parameters $p(\boldsymbol{\theta})$ had to be provided in advance (see section 2.4.1 Response model). Despite that, this was not *a-priori* information from an epistemological point of view, since the algorithm can be designed to collect the information itself. The last factor $p(\mathbf{r}|\mathbf{x})$ can be evaluated using

$$p(\mathbf{r}|\mathbf{x}) = \int_{\mathbb{R}_+^n} p(\boldsymbol{\theta}) p(\mathbf{r}|\mathbf{x},\boldsymbol{\theta}) \mathrm{d}^{\mathrm{n}}\theta, \tag{15}$$

with n=2 being the number of parameters. For maximizing the logarithm of (14) with respect to $\boldsymbol{\theta}$, however, the denominator $p(\mathbf{r}|\mathbf{x})$ was a constant and therefore was omitted.

2.3.3. Kontsevich–Tyler Bayesian estimation method

The BAYESIAN methods (Table 3) here were adapted from KONTSEVICH and TYLER [57]. All central elements were taken as described, using the GAUSSIAN black box (12) as in the other ML and MAP estimators for consistency. Based on this, new information was embedded into the probability density function $p(r|x,\boldsymbol{\theta})$ at every step i using BAYES' theorem following

$$p_i(\boldsymbol{\theta}|x,r) = \frac{p_i(\boldsymbol{\theta}) p(r|x,\boldsymbol{\theta})}{\int_{\mathbb{R}_+^n} p_i(\boldsymbol{\theta}) p(r|x,\boldsymbol{\theta}) \mathrm{d}^n\theta}. \tag{16}$$

The next stimulation amplitude x_{i+1} was determined using an information-theoretical approach which minimized the expected entropy at the next step [57]

$$x_{i+1} = \underset{x}{\operatorname{argmin}}\, \mathbf{E}(H_i(x,r)\,). \tag{17}$$

Here, $\mathbf{E}$ denotes expectation and the entropy $H_i(x,r)$ was defined as

$$H_i(x,r) = -\int_{\mathbb{R}_+^n} p_i(\boldsymbol{\theta}|x,r) \log\big(p_i(\boldsymbol{\theta}|x,r)\big)\, \mathrm{d}^n\theta, \tag{18}$$

and the expectation was a sum of the entropy $H_i(x,r)$ weighted by the probability $p_i(r|x)$ of sub- and supra-threshold responses

$$p_i(r|x) = \int_{\mathbb{R}_+^n} p_i(r|x,\boldsymbol{\theta}) p_i(\boldsymbol{\theta}) \mathrm{d}^n\theta \;, \quad r \in \{0,1\}. \tag{19}$$

Each stimulus–response pair narrowed down the posterior probability density distribution to

$$p_{i+1}(\boldsymbol{\theta}) = p_i(\boldsymbol{\theta}|x_{i+1}, r_{i+1}). \tag{20}$$

The expected value provided the estimates for the parameters at each step per

$$\theta_i^{(l)} = \int_{\mathbb{R}_+^n} \theta^{(l)} p_i(\boldsymbol{\theta}) \mathrm{d}^n\theta, \tag{21}$$

where l was the parameter index (i.e., $l = 1,2$ for $\boldsymbol{\theta} = (t, s)$). In contrast to the point estimator methods, the probably density functions (12) and (16) were explicitly computed and stored for the whole parameter space: $r \in \{0,1\}$, $x \in (0,1.3]$, $t \in (0,1.3]$, and $s \in [0.005, 0.5]$. For computational purposes, the parameter space was discretized with $\Delta x = 0.01$ and $\Delta t = 0.002$, and 100 logarithmical-distributed samples were chosen for s, corresponding to 8.5 million points for $p(r = 1|x, \boldsymbol{\theta})$. The initial distribution $p_0(\boldsymbol{\theta})$ was chosen as the population distribution $p(\boldsymbol{\theta})$ (see section 2.4.1 Response model).

Two variations of the BAYESIAN method were implemented. The first one had no prior (NP) population information—using a uniform distribution for $p_0(\boldsymbol{\theta})$—to evaluate the contribution of the prior information to convergence. To compare the entropy minimization stepping rule against the threshold stepping (TS) of the MLE and MAP methods, the second variation skipped the entropy calculation ((17)–(19)) and used the latter's choice of the threshold estimate at step i (21) for the stimulation amplitude for the next step: $x_{i+1} = t_i = \theta_i^{(1)}$.

Table 3 Summary of parametric estimation methods

Method	Threshold estimation	Stepping
MLE-1[a]	Single-parameter MLE	
MAP-1	Single-parameter MAP	
MLE-2[a]	Two-parameter MLE	
MAP-2	Two-parameter MAP	
BAYESIAN, BAYESIAN-NP[b]	Expectation of threshold	Minimum expected entropy
BAYESIAN-TS	Expectation of threshold	

[a] Implemented using both direct maximization and explicit evaluation of the log-likelihood function.
[b] Implemented with no prior population information.

2.4. Testbed

2.4.1. Response model

We implemented a realistic MEP model [61] which included variability patterns such as amplifier noise, saturation, and neural variability behavior [30], [34], [35]. The model's increased complexity compared to other numerical studies concerning threshold-estimation methods [48], [73] was intended not only to reproduce more faithfully MEP behavior but also to avoid using the very same models underlying the parameter estimation methods for the test procedure, which would risk numerical favoritism for those particular methods. The response model provided analog EMG amplitude values based on a stochastic recruitment curve (Figure 1A$_i$), with parameters

mimicking real measurements [74]. The analog amplitude response, i.e., the peak-to-peak MEP amplitude V_{PP} in units of volt, was defined as

$$V_{\mathrm{PP}}(x) = \epsilon_{\mathrm{yadd},ij}\left(p_{1,j}, p_{8,j}, p_{9,j}\right) + 10^{\epsilon_{\mathrm{ymult},ij}(p_{6,j})} \cdot 10^{p_{2,j}/\left(1+\frac{p_{3,j}}{\left(x-p_{5,j}+\epsilon_{\mathrm{xadd},ij}(p_{7,j})\right)^{p_{4,j}}}\right)}, \tag{22}$$

where $p_{1,j}$ to $p_{9,j}$ were the individual response parameters for subject j obtained from an experimental population [34], and intra-individual variability sources $\epsilon_{\mathrm{xadd},ij}$ and $\epsilon_{\mathrm{ymult},ij}$ and the additive noise $\epsilon_{\mathrm{yadd},ij}$ varied between subjects j and for pulses i along distributions with subject-dependent parameters. The distributions of recruitment parameters, intra-individual variability, and noise are given in the Appendix. While nominally the stimulation amplitude x should be in the range $[0,1]$, the response model was fitted to experimental data obtained with a controllable pulse parameter TMS (cTMS) device using 60 µs pulse width with the maximum capacitor voltage chosen to match stimulation strength to conventional TMS devices for a specific muscle target [34], [74], [75]. Therefore, stimulation amplitudes higher than the nominal 100% MSO could be achieved by increasing the capacitor voltage or with a different TMS device. To focus on the performance of the thresholding algorithms themselves and eliminate influence from the device and choice of stimulation target, we extended the range of possible stimulation amplitudes up to 130% MSO in the model, which was 30% above the maximum threshold within the virtual population.

Model parameters were generated for a total of 25,000 virtual subjects. Each subject's input–output curve was characterized using 100,000 stimulation–response pairs, with a higher number of pulses applied near the amplitudes that generated a 50 µV response in the corresponding *noiseless* model (see Appendix). The ground truth threshold for each subject was determined by linear regression of the relative frequency of suprathreshold responses to find the amplitude with 50% response probability (Figure 1, panels A_{ii}, B_i, and B_{ii}). The regression also provided the slope at the threshold, which was converted to the spread for the corresponding GAUSSIAN model (Figure 1, panels B_i and B_{iii}). The ratio between the spread and threshold (Figure $1B_{iv}$, median, mean, and SD of 0.076, 0.077, and 0.012, respectively) was close to but on average ~ 10% larger than the fixed nominal value of 0.07 from a previous study [48] and also had a wide distribution. The probability density distribution of threshold and spread $p(t,s)$ can be fitted with normal and lognormal distributions, respectively, with details given in the Appendix. This distribution $p(t,s)$ served as prior information $p(\boldsymbol{\theta})$ or $p_0(\boldsymbol{\theta})$ for the MAP and BAYESIAN methods, respectively. By applying MLE to each subject, the parameters t and s for the Gaussian model were obtained and analyzed (Figure S1A). For approximately 90% of subjects, the response probability distributions estimated with GAUSSIAN MLE matched their ground truth well (Figure S1B). For the remaining 10% of subjects, however, MLE resulted in large errors (especially for the spread parameter), showing that the GAUSSIAN model was not a universal fit.

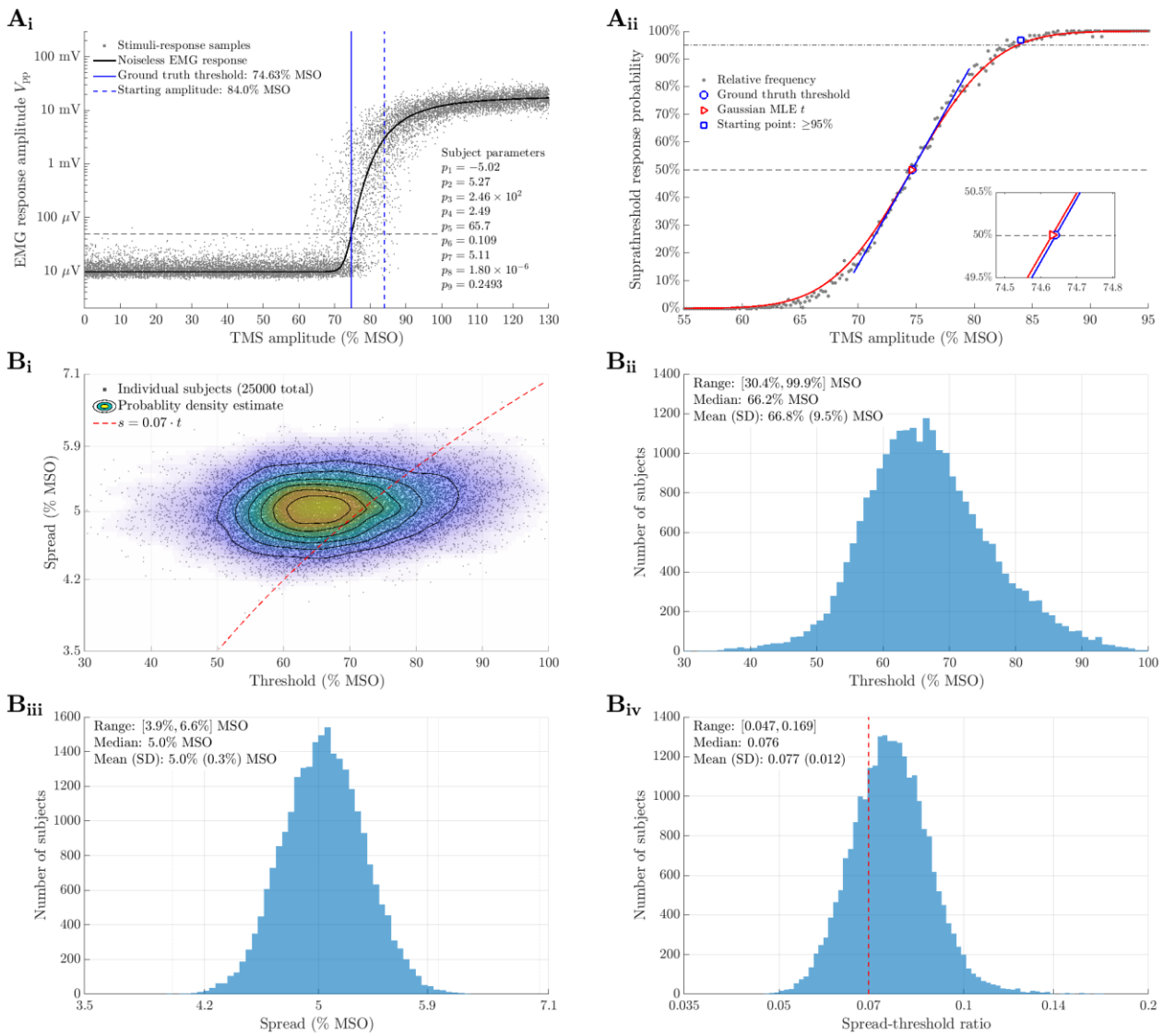

Figure 1: TMS response model. A. An example of the TMS sigmoidal recruiting behavior that is mimicked by the model. **i.** In the EMG space, this virtual stimulator provides analog values, which some of the algorithms can utilize. The response model provides uncorrelated EMG response, which are independent from the tested black-box models in the parameter estimators to avoid model-determined advantages. Subsequent responses are not correlated. **ii.** In the probability space, the threshold is determined from the relative frequency of responses by linear regression to find the 50% probability amplitude. The starting amplitude was determined as the minimum amplitude at and above which the response probably was no less than 95%. A Gaussian response model can also be obtained by applying MLE to the stimulation-response pairs in A_i, which fitted well to the relative frequencies for the majority of subject. **B**. The ground truth thresholds and the slope/spread at the threshold are obtained by linear regression for the 25,000 virtual subjects. **i.** Joint distribution of threshold and spread. **ii.-iv.** Distributions of thresholds, spread, and spread-to-threshold ratio. The distributions are normal or log-normal. The nominal 0.07-ratio is shown by red dashed lines.

2.4.2. Simulation routine

The testbed comprised all threshold detection algorithms defined previously and the response model. In order to meet the real conditions of a threshold search, each algorithm was given information from the hotspot detection procedure, i.e., a starting amplitude which was known to be above the threshold [52], [76], [77]. To avoid ambiguity and circular definition between threshold and hotspot [78], the latter was chosen as the amplitude that generated a response for at least 95% of the stimuli (Figure S2). The starting amplitude was on average 8.7% MSO higher than the ground truth threshold (1.13 times threshold). For the algorithms that use specific integer amplitudes (i.e., IFCN relative frequency and MILLS–NITHI methods), the starting amplitude was rounded to 1% MSO.

The responses for subsequent stimuli were provided to the algorithms by the full MEP model. Whereas the analog versions of the stochastic root-finding methods directly processed this continuous information, all other methods quantized the response into the binary representation of above and below threshold first. To examine both their short- and long-term behaviors, the parametric estimation algorithms and stochastic root-finding methods were tested with a maximum of 200 steps each. The absolute and relative errors of each threshold estimate $x_{\mathrm{th,est}}$ versus the ground truth $x_{\mathrm{th,GT}}$ were analyzed with

$$\begin{aligned} \Delta x_{\mathrm{th}} &= x_{\mathrm{th,est}} - x_{\mathrm{th,GT}}, \\ \delta x_{\mathrm{th}} &= \Delta x_{\mathrm{th}} / x_{\mathrm{th,GT}}. \end{aligned} \tag{23}$$

For comparison with previous studies in the literature, the absolute values of the errors $|\Delta x_{\mathrm{th}}|$ and $|\delta x_{\mathrm{th}}|$ were also analyzed, of which $|\delta x_{\mathrm{th}}|$ was previously referred to as the "relative error" [48], [54]. The results mostly present the relative errors δx_{th}, as they provide normalized distance of the threshold estimate from the true value expressed as a percentage of the latter. The absolute error Δx_{th} overall had similar distributions and are shown in the supplement.

3. Results

3.1. IFCN relative-frequency method and MILLS–NITHI two-threshold method

As both methods have been previously established in the field, only the high-level performance is described here, with more details and full statistics provided in the supplement (Tables S1–S7).

For the IFCN relative-frequency methods (Figures 2 and S3–S5), the number of stimuli until the method terminated and reported a threshold value for a given subject had a wide distribution. On the lower end, the methods stopped after reducing the stimulation amplitude by two or three steps below the starting amplitude, which could be as few as 15 to 20 stimuli for the five-out-of-ten methods or between 40 and 60 for the ten-out-of-twenty methods. On the other hand, some subjects required a larger number of stimuli, reaching over 100 or even 200 depending on the method. Estimates obtained with fewer stimuli overestimated the threshold due to false early fulfillment of the termination rule, whereas those with large numbers of stimuli leaned towards underestimation due to false continuation at an amplitude with response probability less than 50%. The errors were linearly dependent on the

number of stimuli with coefficients of determination (r^2) more than 0.7, and over overall had a positive bias (Figure S5, panels B_i and B_{ii}; e.g., median, mean, and SD of 0.91%, 1.00%, and 1.75%, respectively, for δx_{th} of the ten-out-of-twenty method with 1% MSO stepping size). Such linear relationship and bias were inherently determined by the binomial probability distribution to either terminate or continue (Figure S4; see method for calculating the probability in Supplement), consistent with a previous argument against relative-frequency methods [43]. The number of stimuli drastically decreased by approximately 40% and 50% when the stepping size was increased from 1% to 2% MSO and the coarser five-out-of-ten rule was used instead of the ten-out-of-twenty rule for the relative frequency approximation, respectively, although the shape of the distribution was not affected qualitatively. These changes overall decreased the threshold accuracy (i.e., increased median, mean, and standard deviation of errors) as intuitively expected. However, the errors had similar ranges for both ten-out-of-twenty or five-out-of-ten methods with either step sizes. Therefore, it appears ineffective to use smaller step size and/or a larger number of stimuli at a given amplitude to improve the accuracy, as the threshold error of individual subjects does not necessarily decrease but the required number of pulses increases significantly.

The MILLS–NITHI methods (Figures 2, S3 (panels C_i, C_{ii}, D_i, and D_{ii}), and S5) had narrow distributions of the number of stimuli, requiring at least 20 stimuli but rarely exceeding 100. The numbers of stimuli for all four variations were on average similar to or smaller than the IFCN five-out-of-ten method with 1% MSO step size but larger than that with 2% MSO step size. In comparison, the IFCN relative-frequency method could be much faster for any individual subject due to its false early termination and/or if the starting point was already close to the threshold (i.e., steep slope near the threshold). The modified MILLS–NITHI methods exploited the additional information obtained from the hotspot search but only reduced the number of stimuli by three or four on average compared to the original method using the same step size. Nevertheless, none of the MILLS–NITHI methods resulted in extremely large numbers of stimuli. Due to the averaging of the upper and lower thresholds, the errors of the MILLS–NITHI methods were independent of the number of stimuli ($r^2 < 0.01$) and, compared to the IFCN relative-frequency methods, had tighter distributions and were centered closer to zero. Compared to the modified methods, the original MILLS–NITHI methods had wider error distributions and many outliers with large errors. Increasing the step size to 2% MSO for either the original or modified method reduced the number of stimuli on average by 14 or 15 steps and narrowed its distribution. Increasing the step size also widened the error distribution slightly.

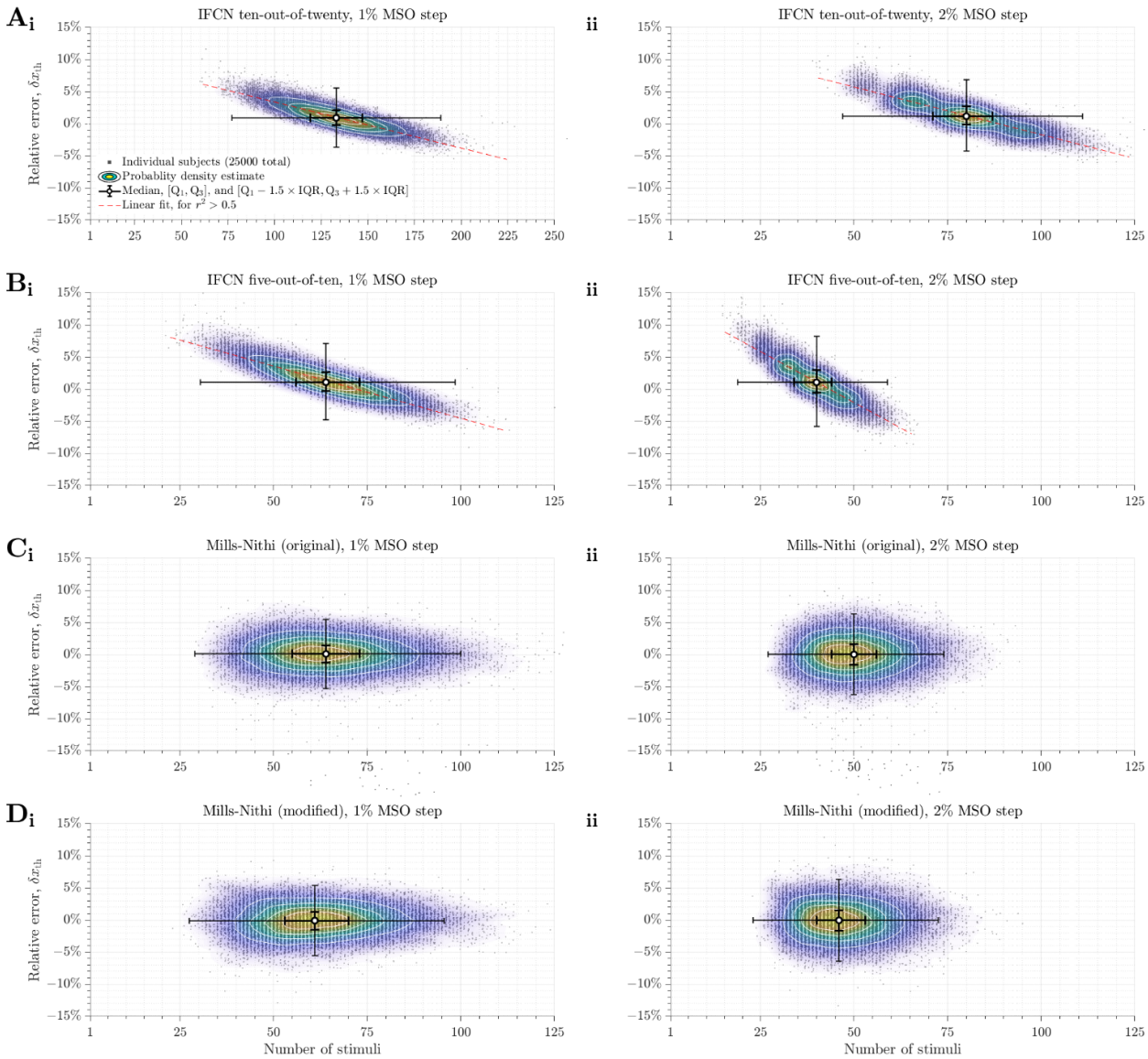

Figure 2: Relative threshold errors of the IFNC relative frequency and MILLS–NITHI methods. The relative errors δx_{th} are plotted against the number of stimuli required for all the virtual subjects as dots. Several statistics are shown, with median as white circle, interquartile range between first and third quartiles (Q_1 and Q_3) as thick error bars, and ranges extending to the lower and upper adjacent values—which are 1.5 times the interquartile range below Q_1 and above Q_3 and cover approximately ± 2.7 standard deviations for normally-distributed data—as thin error bars. A probability density estimate based on a normal kernel function is also shown with the contour lines. Linear regressions are shown for the IFCN relative-frequency methods as red lines. **A.** IFNC ten-out-of-twenty method. **i.** 1% MSO step size. **ii.** 2% MSO step size. **B.** IFNC five-out-of-ten method. **i.** 1% MSO step size. **ii.** 2% MSO step size. **C.** Original MILLS–NITHI methods. **i.** 1% MSO step size. **ii.** 2% MSO step size. **D.** Modified MILLS–NITHI methods. **i.** 1% MSO step size. **ii.** 2% MSO step size. Note different horizontal scale in panel A_i versus the others.

3.2. Parametric estimation

Unlike the relative-frequency methods, the parametric estimation methods generated a threshold estimate and thus an error at every step (relative error δx_{th} in Figure 3A and absolute error Δx_{th} in Figure S6A). The median $|\delta x_{\mathrm{th}}|$ (Figure 3B) and the interquartile range (IQR) of δx_{th} characterized how close the errors approach zero and how wide the errors were distributed. The IQR of δx_{th}, and therefore the range between the lower and upper adjacent values (which were 1.5×IQR below and above the first and third quartile) as well, correlated well with the median $|\delta x_{\mathrm{th}}|$: the ratio between these two statistics after a few steps was approximately 2 and was very stable (Figure S7). Therefore, the median $|\delta x_{\mathrm{th}}|$ captured the distribution of most errors, with the remaining outliers summarized as a percentage of the total number of subjects (Figure 3C). Besides the step-by-step distributions of errors, a few metrics were used to benchmark and compare the performance (Table 4). Estimation speed was quantified by the number of steps for a method to reach the same median $|\delta x_{\mathrm{th}}|$ as the IFCN five-out-of-ten method with 1% MSO step size (median of 64 steps and median $|\delta x_{\mathrm{th}}|$ of 1.6%). Whereas recent IFCN guidelines [28] recognized that the relative-frequency methods require ten-out-of-twenty trials to generate reliable results, the five-out-of-ten method is still widely used [79]–[82] and therefore chosen as the reference. For accuracy, the median $|\delta x_{\mathrm{th}}|$ was compared at 20 and 30 steps, which are respectively considered to produce a sufficiently accurate estimate for all subjects using existing ML-PEST (MLE-1) [41], [48] and commonly used as the upper limit to terminate adaptive thresholding procedures [53], [71], [72].

The basic MLE method [48], which used a single-parameter MLE for threshold estimation and stepping (MLE-1, Figure 3, dark blue lines with circle markers), reached the same error level as the reference IFCN five-out-of-ten method with only 20 steps (Table 4), and the error further decreased by about 20% at 30 steps. Evaluating the likelihood function explicitly had little effect on the threshold estimation performance (Table 4, results in parentheses, and Figure S8) and computational runtime (Figure S9) compared to implicit maximization. Introducing prior information of the parameters of the population (MAP-1, Figure 3, light blue lines with square markers) did not improve the performance for the single-parameter estimation. On the contrary, with the empirical relationship $s = 0.07 \cdot t$ from another population imposed on the parameter distribution of the virtual population in this study, MAP-1 could not fully utilize the prior distribution. Its performance was considerably worse than MLE-1 as the parameters of some subjects in our model were far from such a 0.07 relationship (Figure $1B_{iv}$).

Compared to MLE-1, the two-parameter MLE method (MLE-2, Figure 3, dark red lines with circle markers) had a slower convergence and reached the median error level of the IFCN reference with more (40) pulses. The reduced speed of convergence reflected the exploration of the full parameter space: the likelihood function initially had a large plateau region, which reduced in size as more stimulation–response pairs were recorded. Only after a sufficient number of steps did the likelihood plateau narrow to generate a clear optimal estimate. Similar to MLE-1, the method for maximizing the likelihood function had little effect on performance for MLE-2 (Table 4, results in parentheses, and Figure S8). However, the explicit evaluation with two parameters resulted in a much higher

computational runtime per step that was two order-of-magnitudes slower (Figure S9). Using a two-parameter MAP estimator for the threshold (MAP-2, Figure 3, orange lines with square markers) significantly improved the performance of MLE-2, reaching the error level of the IFCN five-out-of-ten method with only 18 steps, and had lower median $|\delta x_{\mathrm{th}}|$ compared to either MLE-1 or MAP-1. As the spread parameter was no longer fixed to the threshold, MAP-2 could match the wide range of behaviors of the subjects in our model and improved the threshold estimate for all subjects. The prior information significantly influenced the likelihood landscape, rendering one parameter pair (t, s) in the otherwise flat maximum plateau unique. The computational runtime of MAP-1 and MAP-2, which used implicit evaluation, were slightly faster their implicit MLE counterparts (Figure S9).

The original KONTSEVICH–TYLER method (BAYESIAN, Figure 3, yellow lines with diamond markers) had the least bias in error distribution (Figure 3A), reached the IFCN five-out-of-ten method's median $|\delta x_{\mathrm{th}}|$ quickly at step 18 and had faster convergence for small number (< 20) of stimuli (Figure 3B). The BAYESIAN method achieved more accurate estimation early on by considering the entire function space of the probability density functions. Whereas ML and MAP estimators evaluated the underlying likelihood function in every step anew based on the same key equation, the BAYESIAN approach added information into the explicitly evaluated and stored distributions from the last iteration. However, such advantages disappeared with more pulses as the point estimator methods gathered sufficient information. Removing the prior population information (BAYESIAN-NP, Figure 3, purple lines with diamond markers) resulted in slower convergence and performance comparable to that of MLE-2 with a similar slow down comparing MAP-2 to MLE-2, showing that the prior population information narrowed the distribution of parameters in the probability space alike for the point and distribution estimators. Utilizing the threshold estimate for stepping (BAYESIAN-TS, Figure 3, green lines with diamond markers) had similar performance compared to the original method of minimizing the expected entropy, demonstrating that the threshold was at or very close to the stimulation amplitude that provided the maximum information gain. This provided further validation for choosing the threshold estimate as the next stimulation amplitude for the MLE and MAP methods, despite the difference of how the estimate was obtained—for the BAYESIAN methods, the threshold was estimated as the expected value based on its probability distribution (21), whereas for MLE and MAP methods the estimation came from maximizing the likelihood functions (13)–(14). The BAYESIAN methods were the most computationally costly and slowest (Figure S9), with runtime around four seconds per step and more than two order-of-magnitudes slower than the implicitly evaluated MLE/MAP methods. This was mostly due to updating of the entire probability distribution (16), as the runtime without the entropy calculation in BAYESIAN-TS was only slightly reduced and still on the same order-of-magnitude. The BAYESIAN methods were also highly sensitive to rounding-error propagation and sampling of the parameter space, which was a disadvantage compared to MLE and MAP methods.

Table 4 Error convergence of parametric estimation and root-finding methods

Method	Step reaching median $\lvert\delta x_{\mathrm{th}}\rvert$ of 1.6%	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 20	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 30	Optimal a_0 (ratio to default value)
MLE-1 [a]	20 (21)	1.57% (1.62%)	1.25% (1.26%)	–
MAP-1	27	1.96%	1.48%	
MLE-2 [a]	40 (39)	2.26% (2.27%)	1.85% (1.83%)	
MAP-2	18	1.506%	1.21%	
BAYESIAN	18	1.48%	1.198%	–
BAYESIAN-NP	37	2.37%	1.81%	
BAYESIAN-TS	18	1.50%	1.196%	
ACS1-H [b]	24 (18)	1.69% (1.509%)	1.46% (1.31%)	×1.6
ACS2-H	28	1.77%	1.56%	–
DCS1-H [c]	17 (16)	1.450% (1.446%)	1.17% (1.18%)	×0.79
ACS1-HA [c]	18 (18)	1.517% (1.512%)	1.280% (1.30%)	×0.79
ACS2-HA	18	1.521%	1.27%	–
DCS1-HA [b]	22 (19)	1.67% (1.54%)	1.37% (1.24%)	×0.63
ACS1-GA [c]	20 (20)	1.59% (1.584%)	1.45% (1.47%)	×0.79
ACS2-GA	19	1.581%	1.439%	–
DCS1-GA [c]	22 (21)	1.68% (1.62%)	1.43% (1.444%)	×0.63
ACS1-GHA [b]	19 (18)	1.53% (1.514%)	1.283% (1.29%)	×0.63

[a] Results obtained with direct maximization of the likelihood followed by those using explicit evaluation in parentheses.

[b] Results obtained with default a_0 followed by those with optimal a_0 for lowest median $\lvert\delta x_{\mathrm{th}}\rvert$ at both step 20 and 30 in parentheses.

[c] Results obtained with default a_0 followed by those with optimal a_0 for lowest median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 20 in parentheses, with default a_0 being optimal for lowest median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 30.

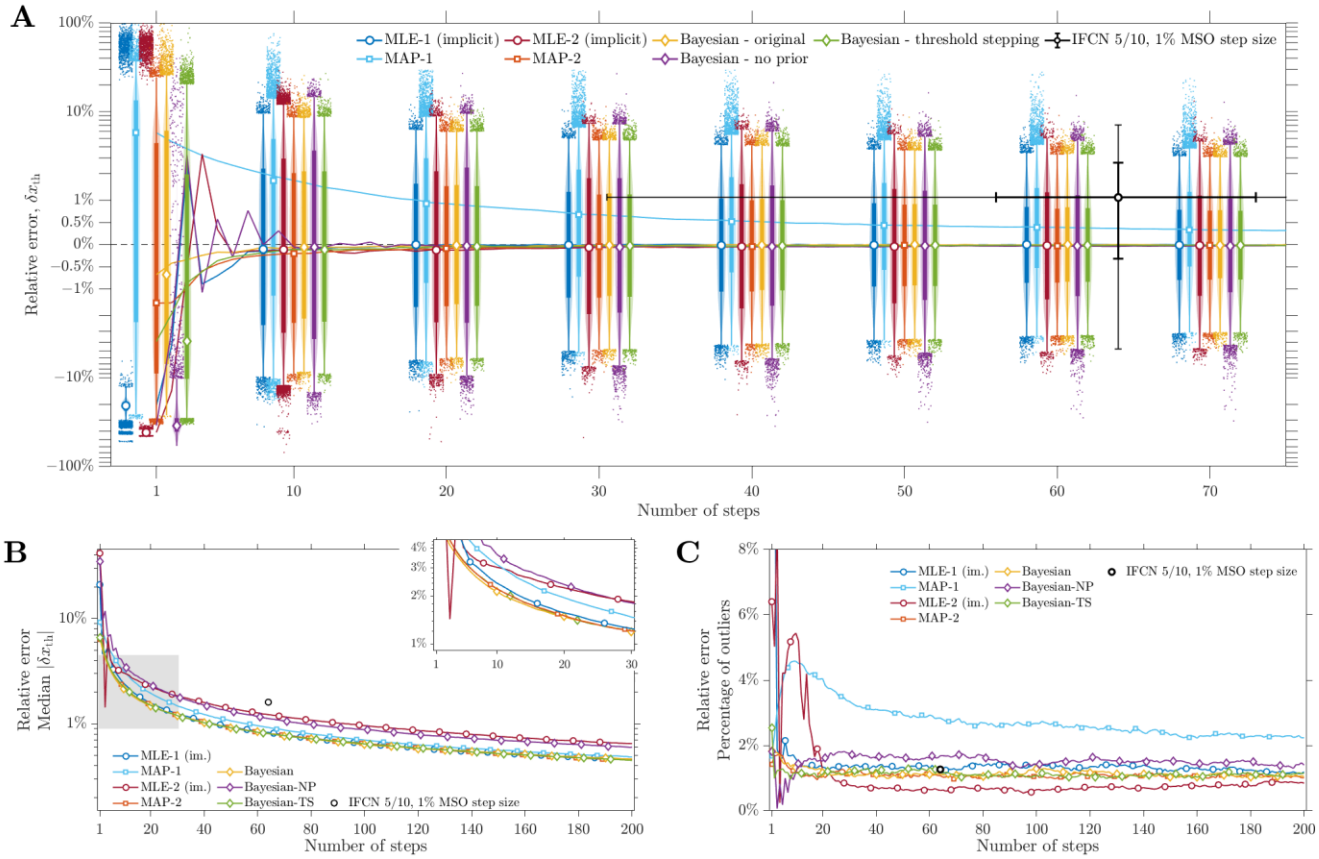

Figure 3: Relative threshold errors of the parametric estimation methods. For MLE methods, results from implicit (im.) maximization of the likelihood function are shown. **A.** Statistics are shown up to the seventieth steps as boxes-and-whiskers at the first step and every ten steps starting from the tenth step, with median as marker, interquartile range as thick vertical line, upper and lower adjacent values as whiskers (thin vertical lines), and outliers as dots. The lines show the median at every step. The boxes-and-whiskers for different methods are slightly offset for clarity. The log-lin-log mixed scale is used to better visualize the errors around zero, with linear scaling in the range of [−1%, 1%] and logarithmic scaling of amplitudes outside this range. The median, interquartile range, and range of adjacent values of the IFCN five-out-of-ten method with 2%MSO step size are shown as a circle and black error bars as in Figure 2B$_{\mathrm{ii}}$ for reference. **B.** The median $|\delta x_{\mathrm{th}}|$ on logarithmic scale, with the inset showing the details in the first 30 steps (the gray area in the main panel). **C.** The percentage of outliers (dots in panel A) outside the upper and lower adjacent values (whiskers in panel A).

3.3. Stochastic root-finding methods

The parameter-free stochastic root-finding methods were based on the threshold definition only and therefore had the important advantage over the parametric estimation approaches, which depended on a potentially biased or inaccurate model. Their simple stepping rule was also much faster to compute, with runtime per step on the

millisecond scale. Like the parametric methods, the root-finding methods also generated a threshold estimate and an error at every step (relative error δx_{th} in Figures 4, S13A, and S15A, and absolute error Δx_{th} in Figures S10, S13B, and S15B). The ratio between the IQR of δx_{th} and the median $|\delta x_{\mathrm{th}}|$ (Figures S11 and S14) was also centered on 2 after a few steps, but showed larger variations compared to those of parametric estimation methods.

The error distributions of all first-order analog methods (Figure 4, lines with circle markers) became negatively biased after a certain number of steps, i.e., the median was less than zero, the absolute values of the first quartile and lower adjacent value were larger than the third quartile and upper adjacent value, respectively, and there were more negative than positive outliers (Figures 4A and S10A). ACS1-H took about 15 and 50 steps for the bias to become negative and plateau, respectively, whereas the process for the other methods was much quicker (within ten steps). The bias originated from the asymmetry of the log-transformed MEP distribution near the threshold [61, Fig. 2]. The 50 μV threshold (which is the median of the distribution according to the probability definition of stimulation threshold) was smaller than the average of the right-skewed distribution and near the EMG noise floor (around 10 μV). Therefore, subthreshold responses generated a smaller Δy to increase a subthreshold stimulation amplitude towards the threshold. In contrast, suprathreshold MEPs can be orders-of-magnitudes larger than the threshold, resulting in large Δy that decreased the stimulation amplitude much lower than the threshold, especially during the early steps. Three of the analog root-finding methods (ACS1-HA, ACS1-GA, and ACS1-GHA) with default parameters reached the accuracy level of the IFCN five-out-of-ten method with 1% MSO step size with only 20 steps or less (Figures 4B and Table 4), whereas ACS1-H needed 24 steps.

The non-adaptive ACS1-H (Figure 4, green lines) had a monotonically declining median $|\delta x_{\mathrm{th}}|$ for the entire range of steps. The percentage of outliers, however, increased with the steps, showing that the error did not improve for some subjects, which became a larger proportion of the total population as the errors of the other subjects in the population became more tightly distributed. The adaptive ACS1-HA (Figure 4, orange lines) outperformed its non-adaptive counterpart by reaching a lower median $|\delta x_{\mathrm{th}}|$ faster, but with distributions of δx_{th} being more negatively biased and having larger proportion of outliers. The ACS1-GA method (Figure 4, blue lines) with higher rate of convergence had a rapid exponential decay of the control sequence and did not further head for the threshold as soon as a_i was too small for a notable influence. It indeed plateaued earlier, at around 40 steps and at a higher median $|\delta x_{\mathrm{th}}|$ of 1.4%, albeit with a smaller portion of outliers as its distribution did not continue to tighten. In comparison, ACS1-GHA (Figure 4, purple lines) switched from fast geometric convergence that provided the rough estimate to $\mathrm{a.s.}$ harmonic convergence and performed very similarly compared to ACS1-HA in terms of median $|\delta x_{\mathrm{th}}|$ and percentage of outliers.

Besides the convergence of the control sequence a_i, the root-finding methods' performance was also determined by the initial value a_0 of the sequence. The default value of a_0 (11) was chosen to approximate the typical slope of the recruitment curve but its definition was inevitably empirical and potentially suboptimal. Examining the performance as a function of the initial step size indeed revealed that this was the case for some of

the root-finding methods. The change from overshoot and oscillatory behavior for large a_0 towards lengthy one-sided convergence for smaller numbers was clearly visible (Figure S12, panels A_i–G_i). Some methods did not have one single optimal a_0, however, as the minimum of the median $|\delta x_{\mathrm{th}}|$ at different steps (Figure S12, panels A_{ii}–G_{ii}, black dashed line and dots) may not be obtained with the same a_0. Instead of evaluating asymptotic behavior, we chose the optimal a_0 that minimized the error at step 20 for which the default was not optimal and the performance indeed improved (Table 4, results in parentheses, and Figure S13). The optimal a_0 for most methods was close to the default value, showing that using the slope of typical input–output curves (11) indeed provided an appropriate a_0. For ACS1-H, the optimal a_0 also resulted in a negative bias of error distribution in just a few steps, similar to the other analog versions.

When the default a_0 was used, the digital versions of the root-finding methods (Figure 4, darker lines with triangle markers) could asymptotically perform either better than (DCS1-H and DCS1-HA) or very similar to (DCS1-GA) their analog counterpart in terms of median $|\delta x_{\mathrm{th}}|$ (Figure 4B). As the root-finding approached the target, the difference between the actual and target responses became very small ($|a_i \Delta y_i| \ll 1$). The step towards the target $|\Delta x_i|$ was therefore smaller for the analog version ($|a_i \Delta y_i|$) compared to the digital version ($|a_i \delta y_i| = a_i$) and thus approached the target slower than the latter. It should be further noticed that all digital versions had mostly positive bias for δx_{th} that decayed to zero, with the decay being especially quick for the two adaptive methods. The lack of asymptotic bias in contrast to the analog versions was due to the sign function of the digital versions equalizing the stepping towards threshold in both directions. With optimal a_0, the digital versions had a slower decay of the positive bias of the error distribution compared to those for default a_0 (Figure S13A_i) but had improved convergence of $|\delta x_{\mathrm{th}}|$ (Figure S13A_{ii}).

The methods with second-order control sequences (Figure S15, lighter lines with square markers) could perform either slightly better (ACS2-HA and ACS2-GA) or worse (ACS2-H) than their first-order counterpart when the default parameters were used. Their performance depended on both the initial step size a_0 and the second-order weight w and the two parameters interacted as demonstrated with ACS2-HA (Figures S16 and S17). Large/positive w sped up the approach towards the threshold, but generated oscillations similar to large a_0. Small/negative w, like small a_0, however, reduced the speed of convergence. The second-order weight could compensate a suboptimal a_0 to generate an optimal approach, and vice versa. Compared to the slope-like a_0 (10), an optimal w was difficult to determine without empirical information or performing extensive parameter sweeps. Therefore, the addition of a second-order term to the root-finding would be less preferable than optimizing the initial step size for the first-order methods.

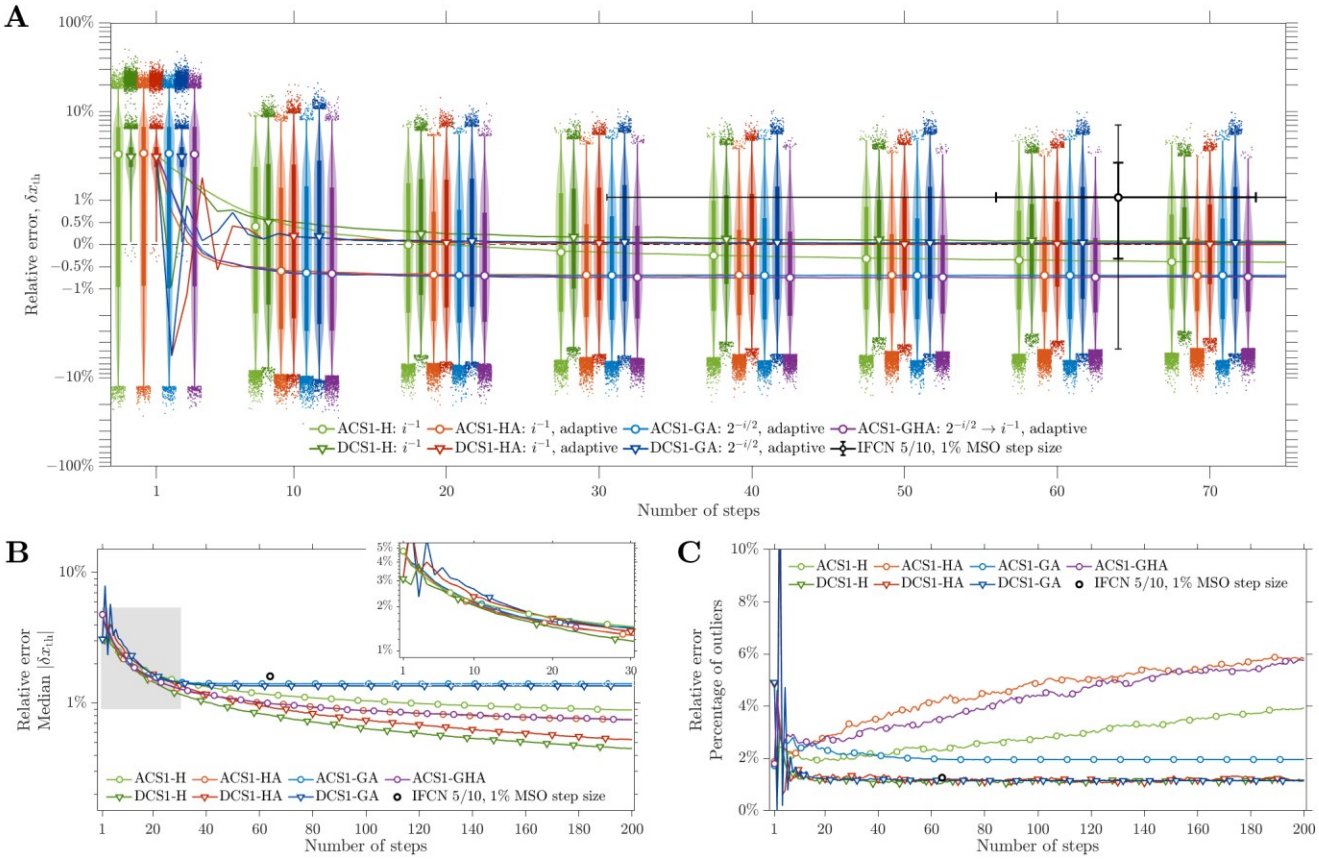

Figure 4: Relative threshold errors of root-finding methods using first-order control sequences with default initial step size a_0. Similar to Figure 3.

4. Discussion

We examined a variety of existing and new TMS thresholding methods and compared their performance in a virtual population of 25,000 subjects. The methods included eight based on relative frequencies using either a single threshold (IFCN relative-frequency methods) or two thresholds (MILLS–NITHI methods) and either 1% MSO or 2% MSO step size, seven parametric estimation methods, of which four used different combinations of single- or two-parameter ML or MAP estimation and three were based on BAYESIAN estimation that updated the probability density function itself, and ten stochastic root-finding methods utilizing control sequences with different convergence and stepping rules.

Although there were some trade-offs between different methods, several threshold estimation methods can be considered to have overall better performance within their class or even across all methods (Figures 5 and S18). All relative-frequency methods had similar error performance and should be avoided due to the large number of pulses required and the presence of inherent error bias or large error outliers (see Figure S5). The two-parameter MAP estimation demonstrated the best performance among the point estimators considering the speed of convergence,

median error, and computational complexity. The KONTSEVICH–TYLER BAYESIAN estimation performed even better regarding the first two aspects, but with increased computational cost, which might mostly matter if implemented on embedded microcontrollers. The best stochastic root-finding methods using analog control sequences was ACS1-HA, whereas DCS1-H had the strongest performance among both analog and digital versions. Unsurprisingly, the best-performing novel methods as well as the previous single-parameter MLE all outperformed the reference IFCN five-out-of-ten method (median 64 steps and 1.6% median $|\delta x_{\mathrm{th}}|$), requiring much fewer steps (within 20 steps, with the novel methods a few steps faster than the single-parameter MLE) to reach the same accuracy (Figure 5B, top). Compared to MLE-1, the best-performing novel methods also demonstrated better accuracy at 20 steps (Figures 5A and S18), which is a typical number used to terminate the single-parameter MLE methods [41], [48], for example in ATH-tool [72].

Despite their wide range of behavior and difference in performance, all methods performed remarkably accurate in terms of threshold error on the population level. Either δx_{th} and Δx_{th} (or $|\delta x_{\mathrm{th}}|$ and $|\Delta x_{\mathrm{th}}|$) had mean or median on the level of 1% and 1% MSO, respectively. In comparison, the pulse amplitude variability of TMS machines themselves have exactness (deviation from nominal value) and precision (variation between pulses) in comparable ranges (1% to 3%) depending on the device [83]. The errors of individual subjects, however, had a large spread, and their relative error may not fall within the ±5% criteria for safety [12], [42] for either relative-frequency methods (Figure S5) or more advanced estimation methods (Figures 3A, 4A, and S13A$_{\mathrm{i}}$). For some methods, the estimation may not improve with increasing number of TMS pulses for certain subjects (e.g., ACS1-H) or the entire population (e.g., ACS1-GA) after a certain number of steps.

We analyzed both the errors and their absolute values, and provided detailed visualization and descriptive statistics of the entire distribution. In contrast, many previous studies quantified the absolute values of the relative error ($|\delta x_{\mathrm{th}}|$, simply referring to it as the relative error) and/or presented mean and standard deviation only [48], [54]. The use of absolute values for errors provided the convenience to compare a positive metric that decreased with stepping and allowed the use of logarithmic scale for visualization. Whereas the median of the absolute values can capture the spread of the errors very well—our results showed that the ratio between the IQR of the errors and the median of the errors' absolute value was quite consistently around two for the methods examined—the absolute value of errors, however, could not capture the bias of the estimated thresholds or its symmetry around zero. Similarly, the mean and standard deviation did not capture outliers and skewness of the distributions, which were all important characteristics of the thresholding methods. For example, the relative-frequency methods had inherent bias regardless of stepping size or the number of stimuli at a given amplitude, and the original MILLS–NITHI methods had a large number of outliers with negative error, contradicting assertions about its reliability [42].

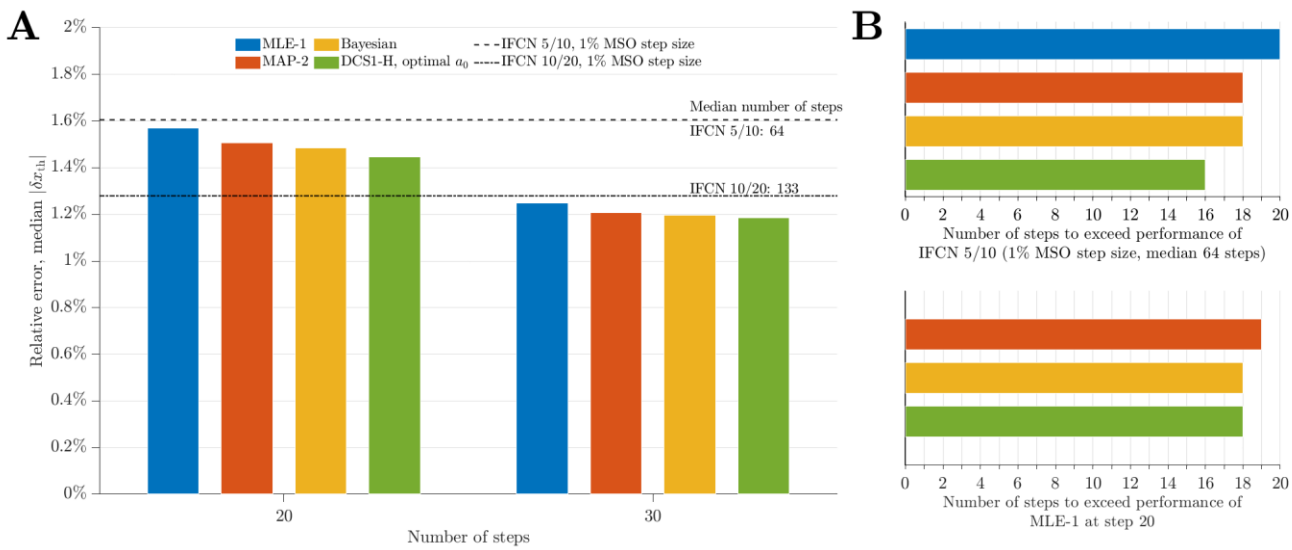

Figure 5: Summary of best performing methods. Selected statistics of the best performing methods, with full statistics shown in Figure S18. The best parametric methods are two-parameter MAP estimation (MAP-2) and BAYESIAN estimation, with the single-parameter MLE method (MLE-1) given for comparison. The best stochastic root-finding methods is DCS1-H, with optimal a_0 shown. The MAP-2 and BAYESIAN methods utilize prior information about the threshold distribution. **A.** Median $|\delta x_{\text{th}}|$ at steps 20 and 30. The dashed and dotted lines shows for reference the IFCN five-out-of-ten and ten-out-of-twenty methods with 1% MSO step size, respectively. **B.** The number of steps needed for a given method to outperform the reference IFCN five-out-of-ten method (top) and the single-parameter MLE at 20 steps (bottom).

Our implementation of the various methods all started from a suprathreshold simulation amplitude with a minimum 95% response probability to avoid ambiguity and circular definition of hot-spot and threshold. This, however, is often not the case in practice. A lower starting amplitude could be used to accelerate the threshold search for some methods, especially for subjects with shallow input–output curves, whereas a default stimulator output is sometimes used as starting point for all subjects. The stimulation amplitudes of stochastic root-finding methods are constrained by previous pulses, and different starting points change the error convergence and require a different optimal initial step size a_0 (results not shown). In contrast, the starting point does not affect the performance of parametric estimation methods on the population level, e.g., using a fixed 50% MSO amplitude (results not shown), as their stimulation amplitudes determined by the likelihood maximization are not constrained and can freely "jump" arbitrary distances from pulse to pulse. However, both of these situations risk convergence errors when a positive response is observed for a stimulator output well below the true threshold, especially early during the procedure [52], [53].

Existing estimation methods previously were tested using virtual subjects that had the same response model underlying the parameter estimation methods [48]. Such tests of estimators against their own estimation model can be misleading due to statistical bias. We therefore used an independently derived MEP model from the literature, which was previously trained to the statistics of experimental data. The response model used here, although more sophisticated than most other approaches, may still have limitations in representing all features and statistical properties of stimulation-evoked EMG responses due to the limited data on which it was based [32], [84]–[86]. In addition, complex temporal effects, such as inter-pulse correlation caused by stimulation-induced short-term plasticity and also stimulation-independent excitability fluctuations [45] may become important, particularly for short inter-pulse intervals of only a few seconds used in some labs. Accurate knowledge of the physiologic system, however, is essential for the thresholding problem. This does not only influence the design of a realistic testbed, but also directly determines, for example, how well the GAUSSIAN model used in all parameter estimators suits the requirements or if another model, such as logistic distribution [40], [50], is needed in order to avoid bias.

The performance of some estimation methods demonstrated substantial improvement over existing methods. For example, the two-parameter MAP and KONTSEVICH–TYLER BAYESIAN estimation methods all outperformed the single-parameter MLE methods. The improvement in accuracy for the BAYESIAN method, however, comes at a cost of increased runtime per step and larger memory requirement. While these costs may still be challenging for many microcontrollers depending on the specific application and potential implementation as an embedded version on a TMS device, they are readily manageable on personal computers which many recent TMS machines anyway include. The BAYESIAN method could become more appealing with improved implementation efficiency, for example by reducing the range and/or resolution of parameter space. All parametric estimators, however, had the potential weakness that the underlying model could be incorrect. Some of the parametric methods further relied on accurate prior information of the model parameters (threshold and spread in the GAUSSIAN model) to achieve the presented accuracy and improvement over other methods. Such prior information on the population level could require significant effort to collect or may be completely unavailable.

The newly introduced family of non-parametric stochastic root-finding methods overall had similar or even lower errors compared to the parametric estimation methods. Comparing the best performing methods between the two classes, DCS1-H was superior in terms of both speed and accuracy compared to the two-parameter MAP estimation and KONTSEVICH–TYLER BAYESIAN methods, which both required prior information. Furthermore, stochastic root-finding methods did not rely on potentially incorrect models and thus could work for subjects whose MEP response curves did not follow the GAUSSIAN model. The rule for choosing and updating the step size a_i was also simple enough to be performed by hand or a calculator and would not necessarily rely on *installed* software (e.g., [48], [72]). Hence, the root-finding methods are a promising addition to the available threshold methods that motivate further studies and improvements. To enable the TMS community to use this method, we implemented

DCS1-H as an easy-to-use web-based tool SAMT (Stochastic Approximator of Motor Threshold [87]) using HTML/JavaScript, which is accessible via web browsers on desktop computers and mobile devices.

5. Conclusion

This work analyzed three classes and a total of 25 variations of TMS thresholding methods and compared their performance, including their accuracy and speed. As pointed out in previous critiques, the relative-frequency methods, such as five-out-of-ten, were far from optimal due to their large number of required stimuli, wide distribution of errors, and inherent error bias. Their accuracy was outperformed by parametric estimators and stochastic root-finding methods with a fraction of the stimuli. As the response model underlying parametric estimation methods did not always fit well with the virtual subject data, some parametric methods performed poorly for a small subset of subjects in the population. Including all model parameters in the estimation and incorporating prior information boosted performance. Stochastic root-finding methods were comparable or better in performance than parametric estimation methods, were much simpler to implement, and did not rely on an underlying model.

Code availability statement

The MATLAB code of the thresholding methods is available at [67] (DOI: 10.5281/zenodo.6483601). SAMT is available to the community online at [87]: https://tms-samt.github.io.

Acknowledgments

Research reported in this publication was supported by the U.S.A. National Institutes of Health under Award Numbers R01NS117405 and RF1MH124943. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency. The authors thank Dr. Philip Whiting and Dr. Lari M. Koponen for helpful discussions. Computational support was provided by the Duke Compute Cluster.

S. M. G. conceived, supervised, and secured funding for the study, and developed the theoretical framework and original simulation code. B.W. reconstructed, improved, and optimized the computational framework, performed the simulations, data analysis, and visualization. A.V.P. contributed concepts, supervised, and secured funding and computational resources for the study. B. W. and S. M. G. wrote the manuscript, and all authors revised, commented on, and approved the final version of the manuscript. Preliminary results of this study were presented at the 5th International Brain Stimulation Meeting [88] (February, 2023, Lisbon, Portugal).

Conflicts of interests

S. M. Goetz and A. V. Peterchev are inventors on patents and patent applications on TMS technology. S. M. Goetz has previously received research funding from Magstim as well as royalties from Rogue Research. A. V. Peterchev has equity options in and serves on the scientific advisory board of Ampa and has received research funding, travel support, patent royalties, consulting fees, equipment loans, hardware donations, and/or patent

application support from Rogue Research, Magstim, MagVenture, Neuronetics, BTL Industries, Magnetic Tides, Ampa, and Soterix Medical. B. Wang declares no relevant conflict of interest.

Appendix

Response model parameters

The individual recruitment parameters $p_{1,j}$ to $p_{5,j}$ followed normal and lognormal distributions [61]:

$$p_{1,j} \sim \mathcal{N}(\mu = -0.5818, \sigma^2 = 3.5705 \cdot 10^{-3}),$$
$$p_{2,j} \sim \mathcal{N}(\mu = \ \ 4.5323, \sigma^2 = 0.12619),$$
$$\lg \begin{pmatrix} p_{3,j} \\ p_{4,j} \\ p_{5,j} \end{pmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} 3.6466 \\ 0.42639 \\ 1.6665 \end{bmatrix}, \Sigma = \begin{bmatrix} 7.2027 & 0.56974 & -0.19746 \\ 0.56974 & 4.7426 \cdot 10^{-2} & -1.5113 \cdot 10^{-2} \\ -0.19746 & -1.5113 \cdot 10^{-2} & 7.1771 \cdot 10^{-2} \end{bmatrix}\right).$$

The intra-individual variability sources $\epsilon_{\mathrm{xadd},ij}$ and $\epsilon_{\mathrm{ymult},ij}$ followed normal distributions with subject-specific standard deviation at each step i:

$$\epsilon_{\mathrm{xadd},ij} \sim \mathcal{N}\left(\mu = 0, \sigma^2 = p_{6,j}^2\right),$$
$$\epsilon_{\mathrm{ymult},ij} \sim \mathcal{N}\left(\mu = -7, \sigma^2 = p_{7,j}^2\right).$$

in which the standard deviations for subjects j followed lognormal distributions:

$$\lg\left(p_{6,j}\right) \sim \mathcal{N}(\mu = -0.96453, \sigma^2 = 2.2759 \cdot 10^{-2}),$$
$$\lg\left(p_{7,j}\right) \sim \mathcal{N}(\mu = \ \ 0.68827, \sigma^2 = 2.3671 \cdot 10^{-2}).$$

The additive noise $\epsilon_{\mathrm{yadd},ij}$ followed a generalized extreme-value (GEV) distribution, with subject-specific parameters

$$\epsilon_{\mathrm{yadd},ij} \sim \mathrm{GEV}(\mu = 10^{\,p_{1,j}}, \sigma = \ 1.55 \cdot 10^{-6} \cdot 1.5^{p_{8,j}}, k = \ 0.14 \cdot 2^{p_{9,j}}),$$

where $p_{8,j}$ and $p_{9,j}$ were independent normal distributions $\mathcal{N}(0,1)$, and the GEV distribution had a probability density function of

$$f_{\mathrm{GEV}} = \frac{1}{\sigma}\left(1 + \frac{k(x-\mu)}{\sigma}\right)^{-\left(1+\frac{1}{k}\right)} \cdot \exp\left(-\left(1 + \frac{k(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}\right).$$

The noiseless model had noise parameters $\epsilon_{\mathrm{xadd},ij}$ and $\epsilon_{\mathrm{ymult},ij}$ set to zero and $\epsilon_{\mathrm{yadd},ij} = 10^{\,p_{1,j}}$, which approached the mean of the GEV for the parameters $p_{1,j}$, $p_{8,j}$, and $p_{9,j}$.

Distribution of threshold and spread

The threshold t and spread s (Figure 1B) was fitted by a joint distribution

$$\begin{pmatrix} t \\ \lg(s) \end{pmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} 0.65847 \\ -1.2989 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_t^2 & \rho\sigma_t\sigma_{\lg(s)} \\ \rho\sigma_t\sigma_{\lg(s)} & \sigma_{\lg(s)}^2 \end{bmatrix}\right),$$

with $\sigma_t = 8.9703 \cdot 10^{-2}$, $\sigma_{\lg(s)} = 2.7576 \cdot 10^{-2}$, and the correlation coefficient $\rho = 0.130$. This distribution served as $p(\boldsymbol{\theta})$ for the MAP methods and $p_0(\boldsymbol{\theta})$ for the BAYESIAN methods.

References

[1] S. M. Goetz and Z.-D. Deng, "The development and modelling of devices and paradigms for transcranial magnetic stimulation," *Int. Rev. Psychiatry*, vol. 29, no. 2, pp. 115–145, Mar. 2017, doi: 10.1080/09540261.2017.1305949.

[2] L. M. Koponen and A. V. Peterchev, "Transcranial Magnetic Stimulation: Principles and Applications," in *Neural Engineering*, B. He, Ed., Cham: Springer International Publishing, 2020, pp. 245–270. doi: 10.1007/978-3-030-43395-6_7.

[3] G. Zhong, Z. Yang, and T. Jiang, "Precise Modulation Strategies for Transcranial Magnetic Stimulation: Advances and Future Directions," *Neurosci. Bull.*, vol. 37, no. 12, pp. 1718–1734, Dec. 2021, doi: 10.1007/s12264-021-00781-x.

[4] A. S. Aberra, B. Wang, W. M. Grill, and A. V. Peterchev, "Simulation of transcranial magnetic stimulation in head model with morphologically-realistic cortical neurons," *Brain Stimulat.*, vol. 13, no. 1, pp. 175–189, Jan. 2020, doi: 10.1016/j.brs.2019.10.002.

[5] V. Di Lazzaro *et al.*, "I-wave origin and modulation," *Brain Stimulat.*, vol. 5, no. 4, pp. 512–525, Oct. 2012, doi: 10.1016/j.brs.2011.07.008.

[6] E. Kallioniemi, L. Säisänen, M. Könönen, F. Awiszus, and P. Julkunen, "On the estimation of silent period thresholds in transcranial magnetic stimulation," *Clin. Neurophysiol.*, vol. 125, no. 11, pp. 2247–2252, Nov. 2014, doi: 10.1016/j.clinph.2014.03.012.

[7] I. Halawa, Y. Shirota, A. Neef, M. Sommer, and W. Paulus, "Neuronal tuning: Selective targeting of neuronal populations via manipulation of pulse width and directionality," *Brain Stimulat.*, Apr. 2019, doi: 10.1016/j.brs.2019.04.012.

[8] N. Lang *et al.*, "Stimulus intensity and coil characteristics influence the efficacy of rTMS to suppress cortical excitability," *Clin. Neurophysiol.*, vol. 117, no. 10, pp. 2292–2301, Oct. 2006, doi: 10.1016/j.clinph.2006.05.030.

[9] T. Sasaki *et al.*, "Modulation of motor learning by a paired associative stimulation protocol inducing LTD-like effects," *Brain Stimulat.*, vol. 11, no. 6, pp. 1314–1321, Nov. 2018, doi: 10.1016/j.brs.2018.07.054.

[10] M. Sommer *et al.*, "Half sine, monophasic and biphasic transcranial magnetic stimulation of the human motor cortex," *Clin. Neurophysiol.*, vol. 117, no. 4, pp. 838–844, Apr. 2006, doi: 10.1016/j.clinph.2005.10.029.

[11] Z. Turi, M. Lenz, W. Paulus, M. Mittner, and A. Vlachos, "Selecting stimulation intensity in repetitive transcranial magnetic stimulation studies: A systematic review between 1991 and 2020," *Eur. J. Neurosci.*, vol. 53, no. 10, pp. 3404–3415, 2021, doi: 10.1111/ejn.15195.

[12] S. Rossi, M. Hallett, P. M. Rossini, and A. Pascual-Leone, "Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research," *Clin. Neurophysiol.*, vol. 120, no. 12, pp. 2008–2039, Dec. 2009, doi: 10.1016/j.clinph.2009.08.016.

[13] S. Rossi *et al.*, "Safety and recommendations for TMS use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert Guidelines," *Clin. Neurophysiol.*, vol. 132, pp. 269–306, Oct. 2020, doi: 10.1016/j.clinph.2020.10.003.

[14] E. M. Wassermann, "Risk and safety of repetitive transcranial magnetic stimulation: report and suggested guidelines from the International Workshop on the Safety of Repetitive Transcranial Magnetic Stimulation, June 5–7, 1996," *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.*, vol. 108, no. 1, pp. 1–16, Jan. 1998, doi: 10.1016/S0168-5597(97)00096-8.

[15] A. Antal, M. A. Nitsche, T. Z. Kincses, C. Lampe, and W. Paulus, "No correlation between moving phosphene and motor thresholds: a transcranial magnetic stimulation study," *Neuroreport*, vol. 15, no. 2, pp. 297–302, Feb. 2004, doi: 10.1097/00001756-200402090-00017.

[16] C. Deblieck, B. Thompson, M. Iacoboni, and A. D. Wu, "Correlation between motor and phosphene thresholds: A transcranial magnetic stimulation study," *Hum. Brain Mapp.*, vol. 29, no. 6, pp. 662–670, Jun. 2008, doi: 10.1002/hbm.20427.

[17] S. M. Goetz and T. Kammer, "Neuronavigation," in *The Oxford Handbook of TMS*, E. M. Wassermann, V. Walsh, A. V. Peterchev, U. Ziemann, S. Lisanby, and H. Siebner, Eds., 2nd ed.Oxford: Oxford University Press, 2022. [Online]. Available: 10.1093/oxfordhb/9780198832256.013.7

[18] Y. Noda *et al.*, "Single-Pulse Transcranial Magnetic Stimulation-Evoked Potential Amplitudes and Latencies in the Motor and Dorsolateral Prefrontal Cortex among Young, Older Healthy Participants, and Schizophrenia Patients," *J. Pers. Med.*, vol. 11, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/jpm11010054.

[19] M. Poorganji *et al.*, "Differentiating transcranial magnetic stimulation cortical and auditory responses via single pulse and paired pulse protocols: A TMS-EEG study," *Clin. Neurophysiol.*, vol. 132, no. 8, pp. 1850–1858, Aug. 2021, doi: 10.1016/j.clinph.2021.05.009.

[20] L. M. Stewart, V. Walsh, and J. C. Rothwell, "Motor and phosphene thresholds: a transcranial magnetic stimulation correlation study," *Neuropsychologia*, vol. 39, no. 4, pp. 415–419, Jan. 2001, doi: 10.1016/S0028-3932(00)00130-5.

[21] S. Tremblay *et al.*, "Clinical utility and prospective of TMS–EEG," *Clin. Neurophysiol.*, vol. 130, no. 5, pp. 802–844, May 2019, doi: 10.1016/j.clinph.2019.01.001.

[22] E. M. Wassermann, "Variation in the response to transcranial magnetic brain stimulation in the general population," *Clin. Neurophysiol.*, vol. 113, no. 7, pp. 1165–1171, Jul. 2002, doi: 10.1016/S1388-2457(02)00144-X.

[23] D. Balslev, W. Braet, C. McAllister, and R. C. Miall, "Inter-individual variability in optimal current direction for transcranial magnetic stimulation of the motor cortex," *J. Neurosci. Methods*, vol. 162, no. 1, pp. 309–313, May 2007, doi: 10.1016/j.jneumeth.2007.01.021.

[24] S. Pridmore, J. A. F. Filho, Z. Nahas, C. Liberatos, and M. S. George, "Motor threshold in transcranial magnetic stimulation: A comparison of a neurophysiological method and a visualization of movement method," *J. ECT*, vol. 14, no. 1, pp. 25–27, 1998, doi: 10.1097/00124509-199803000-00004.

[25] G. G. Westin, B. D. Bassi, S. H. Lisanby, and B. Luber, "Determination of motor threshold using visual observation overestimates transcranial magnetic stimulation dosage: Safety implications," *Clin. Neurophysiol.*, vol. 125, no. 1, pp. 142–147, Jan. 2014, doi: 10.1016/j.clinph.2013.06.187.

[26] G. Cotovio *et al.*, "Day-to-day variability in motor threshold during rTMS treatment for depression: Clinical implications," *Brain Stimulat.*, vol. 14, no. 5, pp. 1118–1125, Sep. 2021, doi: 10.1016/j.brs.2021.07.013.

[27] P. M. Rossini *et al.*, "Non-invasive electrical and magnetic stimulation of the brain, spinal cord and roots: basic principles and procedures for routine clinical application. Report of an IFCN committee," *Electroencephalogr. Clin. Neurophysiol.*, vol. 91, no. 2, pp. 79–92, Aug. 1994, doi: 10.1016/0013-4694(94)90029-9.

[28] P. M. Rossini *et al.*, "Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.C.N. Committee," *Clin. Neurophysiol.*, vol. 126, no. 6, pp. 1071–1107, Jun. 2015, doi: 10.1016/j.clinph.2015.02.001.

[29] M. Kobayashi and A. Pascual-Leone, "Transcranial magnetic stimulation in neurology," *Lancet Neurol.*, vol. 2, no. 3, pp. 145–156, Mar. 2003, doi: 10.1016/S1474-4422(03)00321-1.

[30] S. M. Goetz and A. V. Peterchev, "A model of variability in brain stimulation evoked responses," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2012, pp. 6434–6437. doi: 10.1109/EMBC.2012.6347467.

[31] N. H. Jung, I. Delvendahl, N. G. Kuhnke, D. Hauschke, S. Stolle, and V. Mall, "Navigated transcranial magnetic stimulation does not decrease the variability of motor-evoked potentials," *Brain Stimulat.*, vol. 3, no. 2, pp. 87–94, Apr. 2010, doi: 10.1016/j.brs.2009.10.003.

[32] L. Kiers, D. Cros, K. H. Chiappa, and J. Fang, "Variability of motor potentials evoked by transcranial magnetic stimulation," *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.*, vol. 89, no. 6, pp. 415–423, Dec. 1993, doi: 10.1016/0168-5597(93)90115-6.

[33] C. Capaday, "On the variability of motor-evoked potentials: experimental results and mathematical model," *Exp. Brain Res.*, vol. 239, no. 10, pp. 2979–2995, Oct. 2021, doi: 10.1007/s00221-021-06169-7.

[34] S. M. Goetz, B. Luber, S. H. Lisanby, and A. V. Peterchev, "A Novel Model Incorporating Two Variability Sources for Describing Motor Evoked Potentials," *Brain Stimulat.*, vol. 7, no. 4, pp. 541–552, Jul. 2014, doi: 10.1016/j.brs.2014.03.002.

[35] S. M. Goetz *et al.*, "Isolating two sources of variability of subcortical stimulation to quantify fluctuations of corticospinal tract excitability," *Clin. Neurophysiol.*, vol. 138, pp. 134–142, Jun. 2022, doi: 10.1016/j.clinph.2022.02.009.

[36] S. Groppa *et al.*, "A practical guide to diagnostic transcranial magnetic stimulation: Report of an IFCN committee," *Clin. Neurophysiol.*, vol. 123, no. 5, pp. 858–882, May 2012, doi: 10.1016/j.clinph.2012.01.010.

[37] B. W. Badran *et al.*, "Are EMG and visual observation comparable in determining resting motor threshold? A reexamination after twenty years," *Brain Stimulat.*, vol. 12, no. 2, pp. 364–366, Mar. 2019, doi: 10.1016/j.brs.2018.11.003.

[38] B. S. Anderson and M. S. George, "A Review of Studies Comparing Methods for Determining Transcranial Magnetic Stimulation Motor Threshold: Observation of Movement or Electromyography Assisted," *J. Am. Psychiatr. Nurses Assoc.*, vol. 15, no. 5, pp. 304–313, Oct. 2009, doi: 10.1177/1078390309347372.

[39] K. R. Mills and K. A. Nithi, "Corticomotor threshold to magnetic stimulation: Normal values and repeatability," *Muscle Nerve*, vol. 20, no. 5, pp. 570–576, 1997, doi: 10.1002/(SICI)1097-4598(199705)20:5<570::AID-MUS5>3.0.CO;2-6.

[40] C. Tranulis *et al.*, "Motor threshold in transcranial magnetic stimulation: comparison of three estimation methods," *Neurophysiol. Clin. Neurophysiol.*, vol. 36, no. 1, pp. 1–7, Jan. 2006, doi: 10.1016/j.neucli.2006.01.005.

[41] F. Awiszus, "On relative frequency estimation of transcranial magnetic stimulation motor threshold," *Clin. Neurophysiol.*, vol. 123, no. 11, pp. 2319–2320, Nov. 2012, doi: 10.1016/j.clinph.2012.04.014.

[42] F. Awiszus, "Fast estimation of transcranial magnetic stimulation motor threshold: is it safe?," *Brain Stimulat.*, vol. 4, no. 1, pp. 58–59, Jan. 2011, doi: 10.1016/j.brs.2010.09.004.

[43] F. Awiszus, "Using relative frequency estimation of transcranial magnetic stimulation motor threshold does not allow to draw any conclusions about true threshold," *Clin. Neurophysiol.*, vol. 125, no. 6, pp. 1285–1286, Jun. 2014, doi: 10.1016/j.clinph.2013.09.046.

[44] A. Mishory *et al.*, "The Maximum-likelihood Strategy for Determining Transcranial Magnetic Stimulation Motor Threshold, Using Parameter Estimation by Sequential Testing Is Faster Than Conventional Methods With Similar Precision," *J. ECT*, vol. 20, no. 3, pp. 160–165, Sep. 2004, doi: 10.1097/00124509-200409000-00007.

[45] C. Möller, N. Arai, J. Lücke, and U. Ziemann, "Hysteresis effects on the input–output curve of motor evoked potentials," *Clin. Neurophysiol.*, vol. 120, no. 5, pp. 1003–1008, May 2009, doi: 10.1016/j.clinph.2009.03.001.

[46] J. Nielsen, "Logarithmic Distribution of Amplitudes of Compound Muscle Action Potentials Evoked by Transcranial Magnetic Stimulation," *J. Clin. Neurophysiol.*, vol. 13, no. 5, pp. 423–434, Sep. 1996, doi: 10.1097/00004691-199609000-00005.

[47] E. Kallioniemi, F. Awiszus, M. Pitkänen, and P. Julkunen, "Fast acquisition of resting motor threshold with a stimulus-response curve – possibility or hazard for transcranial magnetic stimulation applications?," *Clin. Neurophysiol. Pract.*, Dec. 2021, doi: 10.1016/j.cnp.2021.10.005.

[48] F. Awiszus, "Chapter 2 TMS and threshold hunting," in *Supplements to Clinical Neurophysiology*, W. Paulus, F. Tergau, M. A. Nitsche, J. G. Rothwell, U. Ziemann, and M. Hallett, Eds., in Transcranial Magnetic Stimulation and Transcranial Direct Current Stimulation, vol. 56. Elsevier, 2003, pp. 13–23. doi: 10.1016/S1567-424X(09)70205-3.

[49] A. Pentland, "Maximum likelihood estimation: the best PEST," *Percept. Psychophys.*, vol. 28, pp. 377–9, 1980, doi: 10.3758/BF03204398.

[50] B. Treutwein, "Adaptive psychophysical procedures," *Vision Res.*, vol. 35, no. 17, pp. 2503–2522, 1995, doi: 10.1016/0042-6989(95)00016-X.

[51] H. Strasburger and B. Treutwein, "Fitting the psychometric function," *Percept. Psychophys.*, vol. 61, no. 1, pp. 87–106, 1999, doi: 10.3758/BF03211951.

[52] S. Goetz, P. Whiting, and A. Peterchev, "Threshold estimation with transcranial magnetic stimulation: algorithm comparison," *Clin. Neurophysiol.*, vol. 122, p. S197, Jun. 2011, doi: 10.1016/S1388-2457(11)60712-8.

[53] L. M. Koponen and A. V. Peterchev, "Preventing misestimation of transcranial magnetic stimulation motor threshold with MTAT 2.0," *Brain Stimul. Basic Transl. Clin. Res. Neuromodulation*, vol. 15, no. 5, pp. 1073–1076, Sep. 2022, doi: 10.1016/j.brs.2022.07.057.

[54] F. Qi, A. D. Wu, and N. Schweighofer, "Fast estimation of transcranial magnetic stimulation motor threshold," *Brain Stimulat.*, vol. 4, no. 1, pp. 50–57, Jan. 2011, doi: 10.1016/j.brs.2010.06.002.

[55] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character*, vol. 222, no. 594–604, pp. 309–368, 1922, doi: 10.1098/rsta.1922.0009.

[56] J. G. Proakis and M. Salehi, *Digital communications*, 5th ed. New York: McGraw-Hill, 2008.

[57] L. L. Kontsevich and C. W. Tyler, "Bayesian adaptive estimation of psychometric slope and threshold," *Vision Res.*, vol. 39, no. 16, pp. 2729–2737, Aug. 1999, doi: 10.1016/S0042-6989(98)00285-5.

[58] L. Paninski, "Asymptotic Theory of Information-Theoretic Experimental Design," *Neural Comput.*, vol. 17, no. 7, pp. 1480–1507, Jul. 2005, doi: 10.1162/0899766053723032.

[59] F. A. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Percept. Psychophys.*, vol. 63, no. 8, pp. 1293–1313, Nov. 2001, doi: 10.3758/BF03194544.

[60] P. Bawa and R. N. Lemon, "Recruitment of motor units in response to transcranial magnetic stimulation in man.," *J. Physiol.*, vol. 471, no. 1, pp. 445–464, 1993, doi: 10.1113/jphysiol.1993.sp019909.

[61] S. M. Goetz, S. M. M. Alavi, Z. Deng, and A. V. Peterchev, "Statistical Model of Motor-Evoked Potentials," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1539–1545, Aug. 2019, doi: 10.1109/TNSRE.2019.2926543.

[62] J. P. Mathias, G. I. Barsi, M. van de Ruit, and M. J. Grey, "Rapid Acquisition of the Transcranial Magnetic Stimulation Stimulus Response Curve," *Brain Stimulat.*, vol. 7, no. 1, pp. 59–65, Jan. 2014, doi: 10.1016/j.brs.2013.08.003.

[63] S. M. Goetz, Z. Li, and A. V. Peterchev, "Noninvasive Detection of Motor-Evoked Potentials in Response to Brain Stimulation Below the Noise Floor—How Weak Can a Stimulus Be and Still Stimulate," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, USA: IEEE, 2018, p. 4. doi: 10.1109/EMBC.2018.8512765.

[64] Z. Li, A. V. Peterchev, J. C. Rothwell, and S. M. Goetz, "Detection of motor-evoked potentials below the noise floor: rethinking the motor stimulation threshold," *J. Neural Eng.*, 2022, doi: 10.1088/1741-2552/ac7dfc.

[65] T. L. Lai, "Stochastic approximation," *Ann. Stat.*, vol. 31, no. 2, pp. 391–406, 2003, doi: 10.1214/aos/1051027873.

[66] H. Robbins and S. Monro, "A Stochastic Approximation Method," *Ann. Math. Stat.*, vol. 22, no. 3, pp. 400–407, Sep. 1951, doi: 10.1214/aoms/1177729586.

[67] B. Wang, "TMS_Thresholding," Apr. 13, 2023. https://doi.org/10.5281/zenodo.6483601

[68] J. C. Rothwell, M. Hallett, A. Berardelli, A. Eisen, P. Rossini, and W. Paulus, "Magnetic stimulation: motor evoked potentials. The International Federation of Clinical Neurophysiology," *Electroencephalogr. Clin. Neurophysiol. Suppl.*, vol. 52, pp. 97–103, 1999.

[69] B. Delyon and A. Juditsky, "Accelerated stochastic approximation," *SIAM J. Optim.*, vol. 3, no. 4, pp. 868–881, 1993, doi: 10.1137/0803045.

[70] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009, doi: 10.1137/070704277.

[71] F. Awiszus and J. J. Borckardt, "TMS Motor Threshold Assessment Tool (MTAT)," 2011. https://www.clinicalresearcher.org/software.htm

[72] P. Julkunen, "Mobile Application for Adaptive Threshold Hunting in Transcranial Magnetic Stimulation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 8, pp. 1504–1510, Aug. 2019, doi: 10.1109/TNSRE.2019.2925904.

[73] J. J. Borckardt, Z. Nahas, J. Koola, and M. S. George, "Estimating Resting Motor Thresholds in Transcranial Magnetic Stimulation Research and Practice: A computer Simulation Evaluation of Best Methods," *J. ECT*, vol. 22, no. 3, pp. 169–175, Sep. 2006, doi: 10.1097/01.yct.0000235923.52741.72.

[74] A. V. Peterchev, S. M. Goetz, G. G. Westin, B. Luber, and S. H. Lisanby, "Pulse width dependence of motor threshold and input–output curve characterized with controllable pulse parameter transcranial magnetic stimulation," *Clin. Neurophysiol.*, vol. 124, no. 7, pp. 1364–1372, Jul. 2013, doi: 10.1016/j.clinph.2013.01.011.

[75] A. V. Peterchev, R. Jalinous, and S. H. Lisanby, "A Transcranial Magnetic Stimulator Inducing Near-Rectangular Pulses With Controllable Pulse Width (cTMS)," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 257–266, Jan. 2008, doi: 10.1109/TBME.2007.900540.

[76] S. M. Goetz, "Magnetic Neurostimulation from a Physical Perspective," Dissertation, Technical University of Munich, Munich, Germany, 2012. [Online]. Available: http://mediatum.ub.tum.de?id=1079891

[77] S. M. Goetz *et al.*, "Enhancement of Neuromodulation with Novel Pulse Shapes Generated by Controllable Pulse Parameter Transcranial Magnetic Stimulation," *Brain Stimulat.*, vol. 9, no. 1, pp. 39–47, Jan. 2016, doi: 10.1016/j.brs.2015.08.013.

[78] F. Awiszus, "Of thresholds and 'hot spots'. Quo vadis transcranial magnetic stimulation?," *Clin. Neurophysiol.*, vol. 125, no. 12, pp. 2451–2453, Dec. 2014, doi: 10.1016/j.clinph.2014.03.010.

[79] K. Weise, O. Numssen, A. Thielscher, G. Hartwigsen, and T. R. Knösche, "A novel approach to localize cortical TMS effects," *NeuroImage*, vol. 209, p. 116486, Apr. 2020, doi: 10.1016/j.neuroimage.2019.116486.

[80] N. Schaworonkow, J. Triesch, U. Ziemann, and C. Zrenner, "EEG-triggered TMS reveals stronger brain state-dependent modulation of motor evoked potentials at weaker stimulation intensities," *Brain Stimulat.*, vol. 12, no. 1, pp. 110–118, Jan. 2019, doi: 10.1016/j.brs.2018.09.009.

[81] A. Benussi *et al.*, "Classification accuracy of TMS for the diagnosis of mild cognitive impairment," *Brain Stimulat.*, vol. 14, no. 2, pp. 241–249, Mar. 2021, doi: 10.1016/j.brs.2021.01.004.

[82] I. Laakso, T. Murakami, A. Hirata, and Y. Ugawa, "Where and what TMS activates: Experiments and modeling," *Brain Stimulat.*, vol. 11, no. 1, pp. 166–174, Jan. 2018, doi: 10.1016/j.brs.2017.09.011.

[83] L. R. Zacharias, A. S. C. Peres, V. H. Souza, A. B. Conforto, and O. Baffa, "Method to assess the mismatch between the measured and nominal parameters of transcranial magnetic stimulation devices," *J. Neurosci. Methods*, vol. 322, pp. 83–87, Jul. 2019, doi: 10.1016/j.jneumeth.2019.03.021.

[84] W. G. Darling, S. L. Wolf, and A. J. Butler, "Variability of motor potentials evoked by transcranial magnetic stimulation depends on muscle activation," *Exp. Brain Res.*, vol. 174, no. 2, pp. 376–385, Sep. 2006, doi: 10.1007/s00221-006-0468-9.

[85] M. Orth and J. C. Rothwell, "The cortical silent period: intrinsic variability and relation to the waveform of the transcranial magnetic stimulation pulse," *Clin. Neurophysiol.*, vol. 115, no. 5, pp. 1076–1082, May 2004, doi: 10.1016/j.clinph.2003.12.025.

[86] K. Roy Choudhury, L. Boyle, M. Burke, W. Lombard, S. Ryan, and B. McNamara, "Intra subject variation and correlation of motor potentials evoked by transcranial magnetic stimulation," *Ir. J. Med. Sci.*, vol. 180, no. 4, pp. 873–880, Dec. 2011, doi: 10.1007/s11845-011-0722-4.

[87] B. Wang, L. M. Koponen, S. M. Goetz, and A. V. Peterchev, "SAMT: Stochastic Approximator of Motor Threshold," 2023. https://tms-samt.github.io/ (accessed Apr. 14, 2023).

[88] B. Wang, A. Peterchev, and S. Goetz, "Comparison of thresholding methods for transcranial magnetic stimulation of primary motor cortex," *Brain Stimul. Basic Transl. Clin. Res. Neuromodulation*, vol. 16, no. 1, p. 219, Jan. 2023, doi: 10.1016/j.brs.2023.01.308.

Supplement

Three Novel Methods for Determining Motor Threshold with Transcranial Magnetic Stimulation Outperform Conventional Procedures

Boshuo Wang[1], Angel V. Peterchev[1,2,3,4], Stefan M. Goetz[1,2,4,5*]

[1] Department of Psychiatry and Behavioral Sciences, School of Medicine, Duke University, Durham, NC, USA

[2] Department of Electrical and Computer Engineering, School of Engineering, Duke University, Durham, NC, USA

[3] Department of Biomedical Engineering, School of Engineering, Duke University, Durham, NC, USA

[4] Department of Neurosurgery, School of Medicine, Duke University, Durham, NC, USA

[5] Department of Engineering, School of Technology, University of Cambridge, Cambridge, UK

[*] Corresponding author: stefan.goetz@duke.edu, smg84@cam.ac.uk

Contents:

- Numeric calculation of error distribution of IFCN relative-frequency methods
- Table S1-S7
- Figures S1–S18

Numeric calculation of error distribution of IFCN relative-frequency methods

The distribution of the absolute errors of the IFCN relative-frequency methods can be numerically calculated using probability. The number of stimuli for the procedure is N (even, e.g., 10), and at the lth stimulation amplitude $x_{(l)}$, the suprathreshold response probability $p_{(l)}$ follows the Gaussian distribution (12). Here, the stimulation amplitude is relative to the threshold, so the t parameter is set to zero in (12). The probability that the procedure stops or continues after $k_{(l)}$ stimuli are then given based on binary distributions

$$P_{\text{stop}}\left(N, p_{(l)}, k_{(l)}\right) = \begin{cases} 0 & 0 \le k_{(l)} \le \frac{N}{2} \\ C\left(\frac{N}{2}, k_{(l)} - 1\right)\left(1 - p_{(l)}\right)^{\frac{N}{2}+1} p_{(l)}^{k_{(l)} - \frac{N}{2} - 1} & \frac{N}{2} + 1 \le k_{(l)} \le N \end{cases},$$

$$P_{\text{cont}}\left(N, p_{(l)}, k_{(l)}\right) = \begin{cases} 0 & 0 \le k_{(l)} \le \frac{N}{2} - 1 \\ C\left(\frac{N}{2} - 1, k_{(l)} - 1\right)\left(1 - p_{(l)}\right)^{k_{(l)} - \frac{N}{2}} p_{(l)}^{\frac{N}{2}} & \frac{N}{2} \le k_{(l)} \le N \end{cases}.$$

Here, $C(n, k)$ is the number of k-combinations for a set of n elements. If the procedure continues, the stimulation amplitude is reduced to $x_{(l+1)} = x_{(l)} - \delta x$ with δx being either 1% or 2% MSO, and the response probability is updated to $p_{(l+1)}$ by (12) accordingly. This iterative procedure starts with $x_{(1)}$ that corresponds to $p_{(1)} = 0.95$ for the given spread s and a total probability of one. The probability for the procedure to terminate at a specific amplitude $x_{(l)}$ for given number of total stimuli L can then be calculated as

$$P\left(x_{(l)}, L\right) = \sum_{\sum_{i=1}^{l} k_{(i)} = L} \left(\left(\prod_{i=1}^{l-1} P_{\text{cont}}\left(N, p_{(i)}, k_{(i)}\right) \right) P_{\text{stop}}\left(N, p_{(l)}, k_{(l)}\right) \right).$$

The summation is over all combinations of stimuli numbers $k_{(i)}$ with non-zero P_{cont} or P_{stop} that result in a total number L.

A recursive depth-first-search was used for the calculation. Besides the natural termination due to failing the relative-frequency criteria, the procedure also was also terminated at points where $p_{(l)} < 0.05$, $P_{\text{cont}} < 10^{-6}$, or the total probability to continue is less than 10^{-10} to limit the search depth and speed up calculation. To account for subjects with different response curves (12), i.e., different s, $P\left(x_{(l)}, L\right)$ was calculated for a range of spread parameter, and the final distribution was a weighted sum with weights based on the probability distribution of s (see Appendix).

Supplementary tables

Table S1 Statistics of number of steps, absolute errors, and relative errors of IFCN and MILLS–NITHI methods (related to Figures 2, S3, and S5)

Group number	Method	Step size (% MSO)	Number of stimuli		Relative error δx_{th}				$\lvert\delta x_{\mathrm{th}}\rvert$	Absolute error Δx_{th} (% MSO)				$\lvert\Delta x_{\mathrm{th}}\rvert$ (% MSO)
			Mean±S.D.	Median	Mean±S.D.	Median	Slope	r^2	Median	Mean±S.D.	Median	Slope	r^2	Median
1	IFCN, ten-out-of-twenty	1	133.0±20.7	133	1.00%±1.75%	0.91%	−0.072%	0.73	1.28%	0.66±1.13	0.60	−0.047	0.74	0.84
2	IFCN, ten-out-of-twenty	2	79.7±12.1	80	1.31%±2.11%	1.17%	−0.15%	0.72	1.58%	0.85±1.37	0.78	−0.097	0.73	1.06
3	IFCN, five-out-of-ten	1	64.5±12.3	64	1.23%±2.28%	1.07%	−0.16%	0.78	1.61%	0.80±1.47	0.72	−0.11	0.79	1.07
4	IFCN, five-out-of-ten	2	39.6± 7.2	40	1.22%±2.66%	1.06%	−0.32%	0.75	1.85%	0.80±1.73	0.71	−0.21	0.76	1.24
5	MILLS–NITHI, original	1	64.6±13.4	64	0.03%±2.45%	0.09%	−0.015%	0.006	1.35%	0.02±1.61	0.06	−0.01	0.007	0.89
6	MILLS–NITHI, original	2	50.5± 9.3	50	−0.04%±2.69%	0.02%	−0.003%	10^{-4}	1.57%	−0.03±1.78	0.01	−0.002	10^{-4}	1.05
7	MILLS–NITHI, modified	1	62.0±13.2	61	−0.11%±2.10%	−0.10%	0.004%	10^{-4}	1.38%	−0.07±1.36	−0.07	0.003	10^{-4}	0.91
8	MILLS–NITHI, modified	2	47.0± 9.2	46	−0.06%±2.41%	−0.05%	−0.012%	0.002	1.59%	−0.04 ±1.56	−0.03	−0.008	0.002	1.05

Table S2 Analysis of variance (ANOVA) of number of stimuli of IFCN and MILLS–NITHI methods (related to Table S1)

Source	SS	df	MS	F	Prob>F
Groups	1.495×10^{8}	7	2.135×10^{7}	1.308×10^{5}	0
Error	3.265×10^{7}	199992	163.2		
Total	1.821×10^{8}	199999			

Table S3 ANOVA of relative errors of IFCN and MILLS–NITHI methods (related to Table S1)

Source	SS	df	MS	F	Prob>F
Groups	7.950×10^{4}	7	1.136×10^{3}	2.171×10^{3}	0
Error	1.046×10^{6}	199992	5.231		
Total	1.126×10^{6}	199999	0		

Table S4 ANOVA of absolute errors of IFCN and MILLS–NITHI methods (related to Table S1)

Source	SS	df	MS	F	Prob>F
Groups	3.338×10^{4}	7	4.768×10^{3}	2.081×10^{3}	0
Error	4.583×10^{5}	199992	2.291		
Total	4.916×10^{5}	199999	0		

Table S5 Post-hoc multiple comparison test of number of stimuli of IFCN and MILLS–NITHI methods (related to Table S1)

Group A	Group B	Mean (Group A – Group B)	95% CI (Group A – Group B)		p-value
1	2	53.30	52.95	53.64	0
1	3	68.54	68.20	68.89	0
1	4	93.44	93.10	93.79	0
1	5	68.40	68.05	68.75	0
1	6	82.53	82.18	82.87	0
1	7	71.00	70.65	71.34	0
1	8	86.01	85.66	86.35	0
2	3	15.25	14.90	15.59	0
2	4	40.15	39.80	40.49	0
2	5	15.10	14.76	15.45	0
2	6	29.23	28.88	29.58	0
2	7	17.70	17.35	18.05	0
2	8	32.71	32.36	33.06	0
3	4	24.90	24.55	25.25	0
3	5	−0.15	−0.49	0.20	**0.91**
3	6	13.98	13.64	14.33	0
3	7	2.45	2.11	2.80	0
3	8	17.46	17.12	17.81	0
4	5	−25.04	−25.39	−24.70	0
4	6	−10.92	−11.26	−10.57	0
4	7	−22.45	−22.79	−22.10	0
4	8	−7.44	−7.78	−7.09	0
5	6	14.13	13.78	14.47	0
5	7	2.60	2.25	2.94	0
5	8	17.61	17.26	17.95	0
6	7	−11.53	−11.88	−11.18	0
6	8	3.48	3.13	3.83	0
7	8	15.01	14.66	15.36	0

Table S6 Post-hoc multiple comparison test of relative errors of IFCN and MILLS–NITHI methods (related to Table S1)

Group A	Group B	Mean (Group A – Group B)	95% CI (Group A – Group B)		p-value
1	2	−0.30	−0.36	−0.24	0
1	3	−0.23	−0.29	−0.16	0
1	4	−0.22	−0.28	−0.16	0
1	5	0.97	0.91	1.04	0
1	6	1.05	0.99	1.11	0
1	7	1.11	1.05	1.17	0
1	8	1.11	1.05	1.17	0
2	3	0.08	0.02	0.14	4×10^{-3}
2	4	0.08	0.02	0.15	10^{-3}
2	5	1.28	1.22	1.34	0
2	6	1.35	1.29	1.41	0
2	7	1.41	1.35	1.48	0
2	8	1.41	1.35	1.48	0
3	4	0.01	−0.06	0.07	**1**
3	5	1.20	1.14	1.26	0
3	6	1.27	1.21	1.33	0
3	7	1.34	1.28	1.40	0
3	8	1.34	1.28	1.40	0
4	5	1.19	1.13	1.26	0
4	6	1.27	1.20	1.33	0
4	7	1.33	1.27	1.39	0
4	8	1.33	1.27	1.39	0
5	6	0.07	0.01	0.13	9×10^{-3}
5	7	0.14	0.08	0.20	4×10^{-10}
5	8	0.14	0.08	0.20	4×10^{-10}
6	7	0.06	0.00	0.13	0.03
6	8	0.06	0.00	0.13	0.03
7	8	0.00	−0.06	0.06	1

Table S7 Post-hoc multiple comparison test of absolute errors of IFCN and MILLS–NITHI methods (related to Table S1)

Group A	Group B	Mean (Group A – Group B)	95% CI (Group A – Group B)		p-value
1	2	−0.20	−0.24	−0.15	0
1	3	−0.14	−0.19	−0.10	0
1	4	−0.14	−0.18	−0.10	0
1	5	0.64	0.60	0.68	0
1	6	0.69	0.65	0.73	0
1	7	0.73	0.69	0.77	0
1	8	0.70	0.66	0.74	0
2	3	0.05	0.01	0.09	4×10^{-3}
2	4	0.05	0.01	0.09	3×10^{-3}
2	5	0.83	0.79	0.87	0
2	6	0.89	0.85	0.93	0
2	7	0.92	0.88	0.96	0
2	8	0.89	0.85	0.93	0
3	4	0.00	−0.04	0.04	**1**
3	5	0.78	0.74	0.82	0
3	6	0.84	0.79	0.88	0
3	7	0.87	0.83	0.91	0
3	8	0.84	0.80	0.88	0
4	5	0.78	0.74	0.82	0
4	6	0.83	0.79	0.88	0
4	7	0.87	0.83	0.91	0
4	8	0.84	0.80	0.88	0
5	6	0.05	0.01	0.09	0.002
5	7	0.09	0.05	0.13	7×10^{-10}
5	8	0.06	0.02	0.10	4×10^{-4}
6	7	0.04	0.00	0.08	**0.12**
6	8	0.01	−0.04	0.05	**1**
7	8	−0.03	−0.07	0.01	**0.29**

Supplementary figures

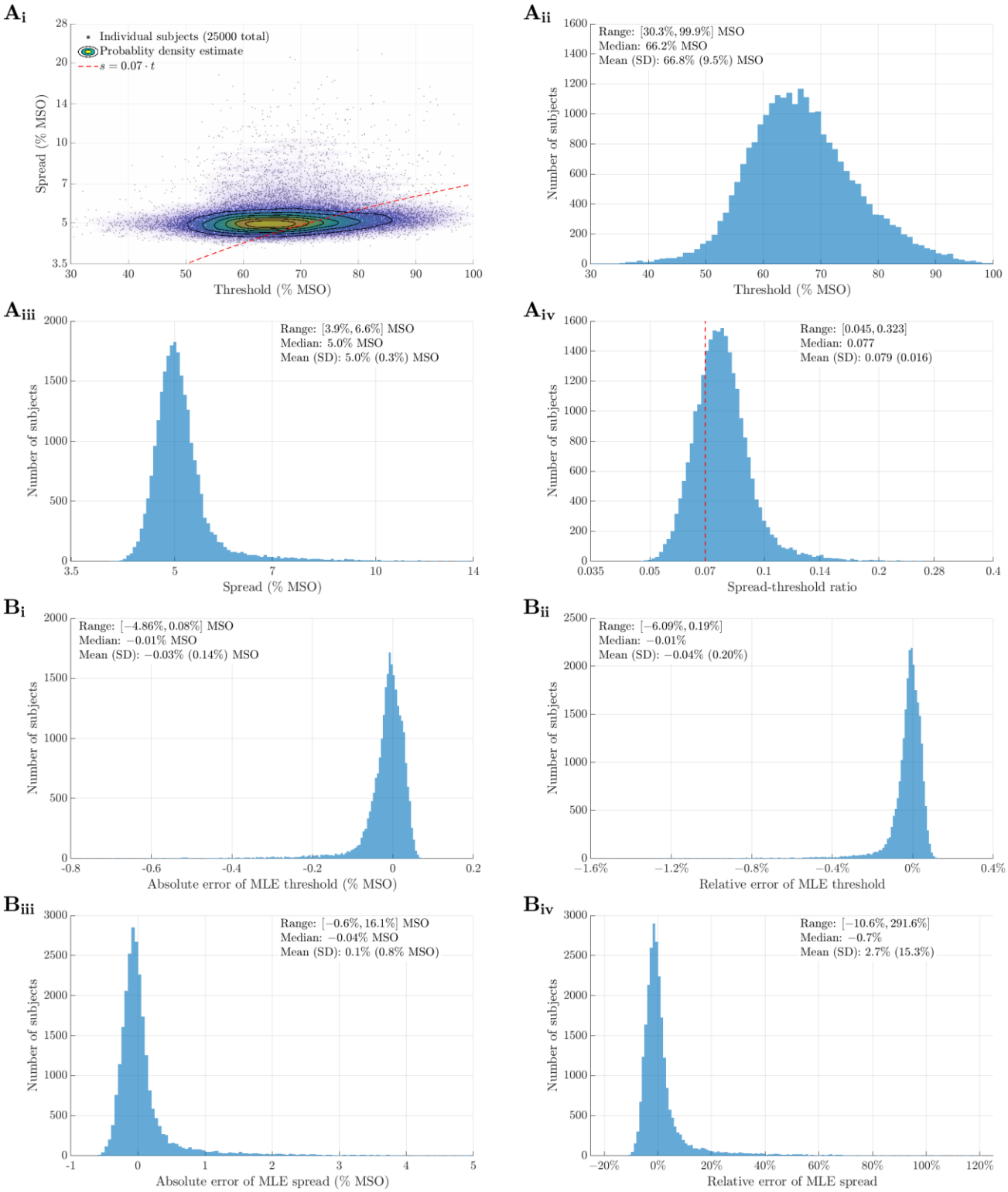

Figure S1: Maximum likelihood estimation (MLE) thresholds and slopes of model population
A. Distribution of MLE thresholds t and spread s, similar to Figure 1B. **B.** Distributions of absolute and relative errors of MLE threshold and spread compared to ground truth obtained from linear regression.

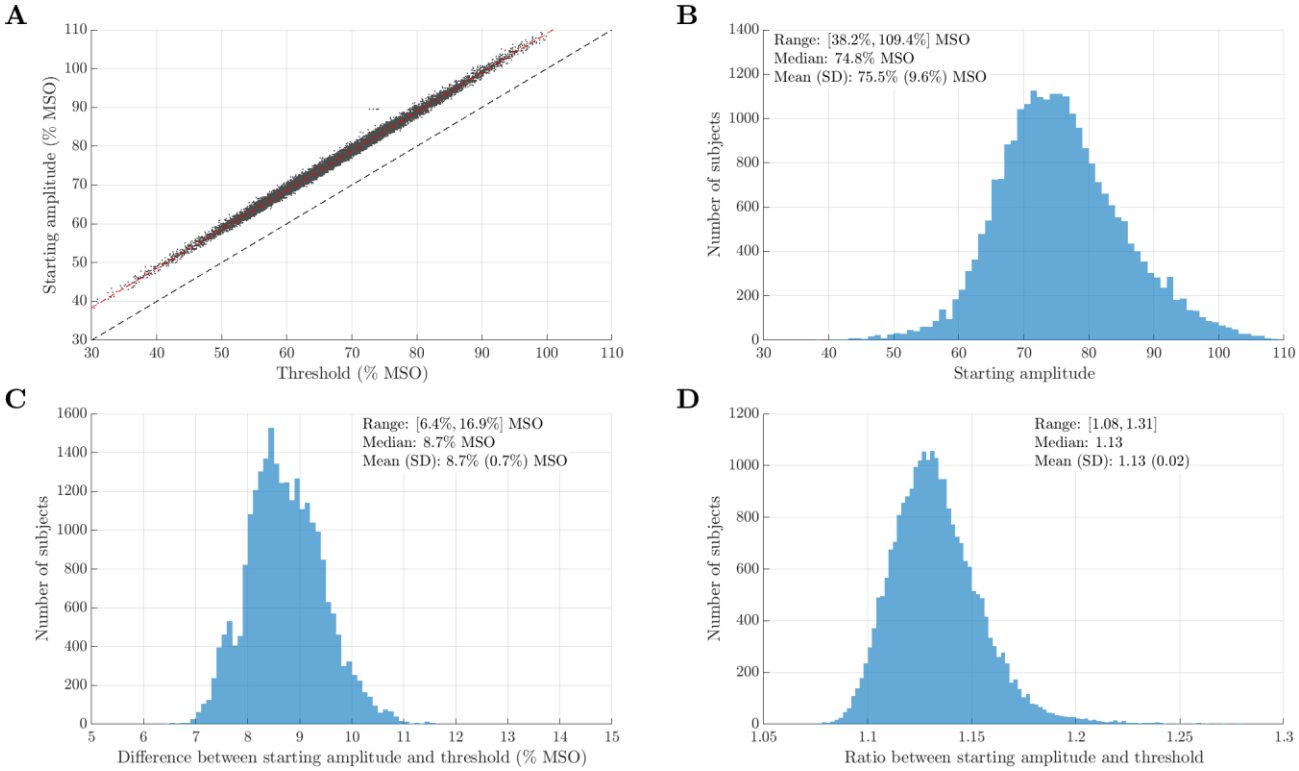

Figure S2: Starting amplitudes

A. Distribution of the starting amplitudes versus ground truth threshold, with a highly linear relationship (coefficient of determination r^2>0.99). **B.** Distribution of the starting amplitudes. **C.** Difference of starting amplitudes and ground truth threshold, centered at 8.7% MSO. **D.** Ratio of starting amplitudes and ground truth threshold.

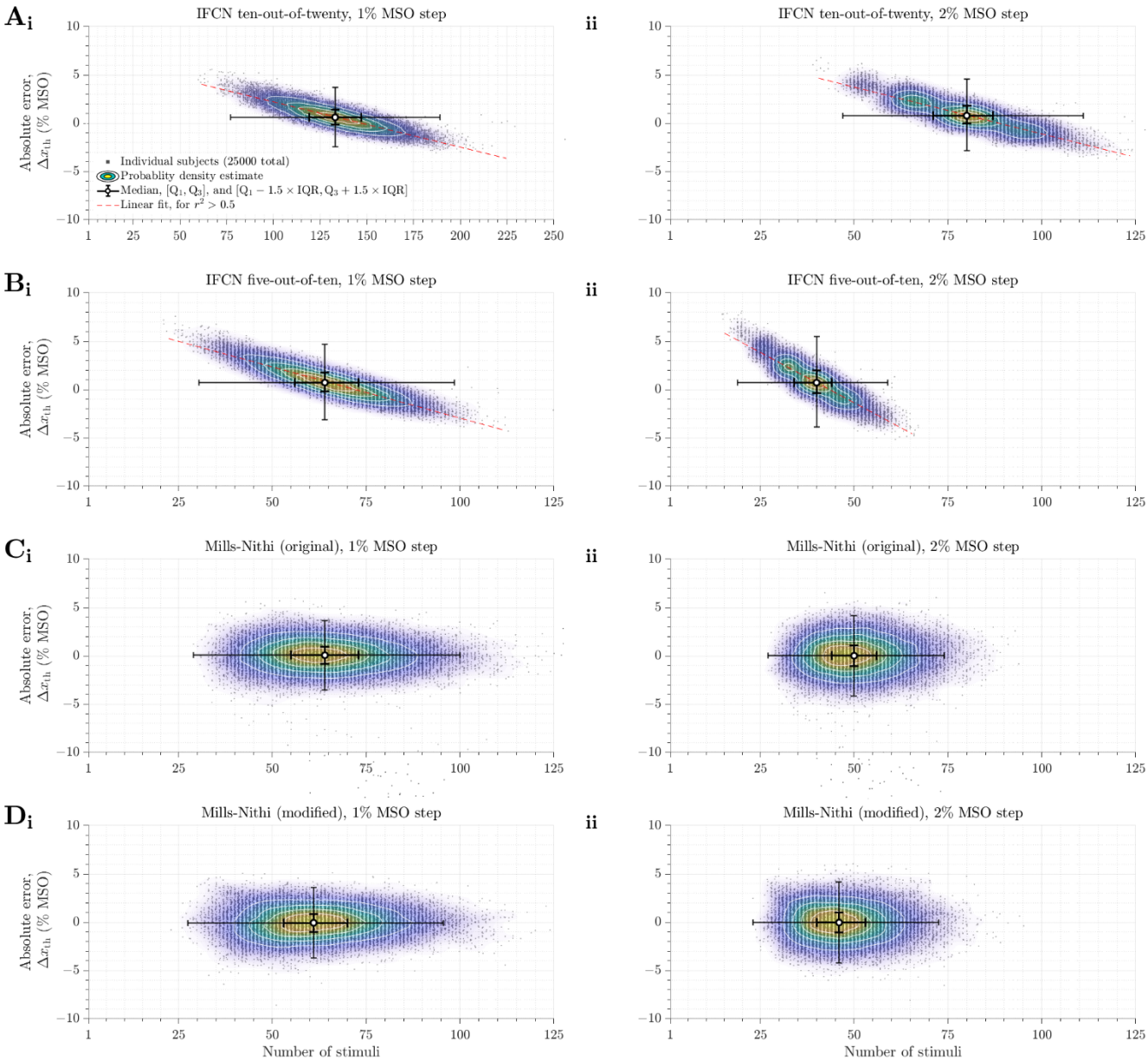

Figure S3: Absolute threshold errors of the IFNC relative frequency and MILLS–NITHI methods
Shown similarly as the relative threshold errors in Figure 2.

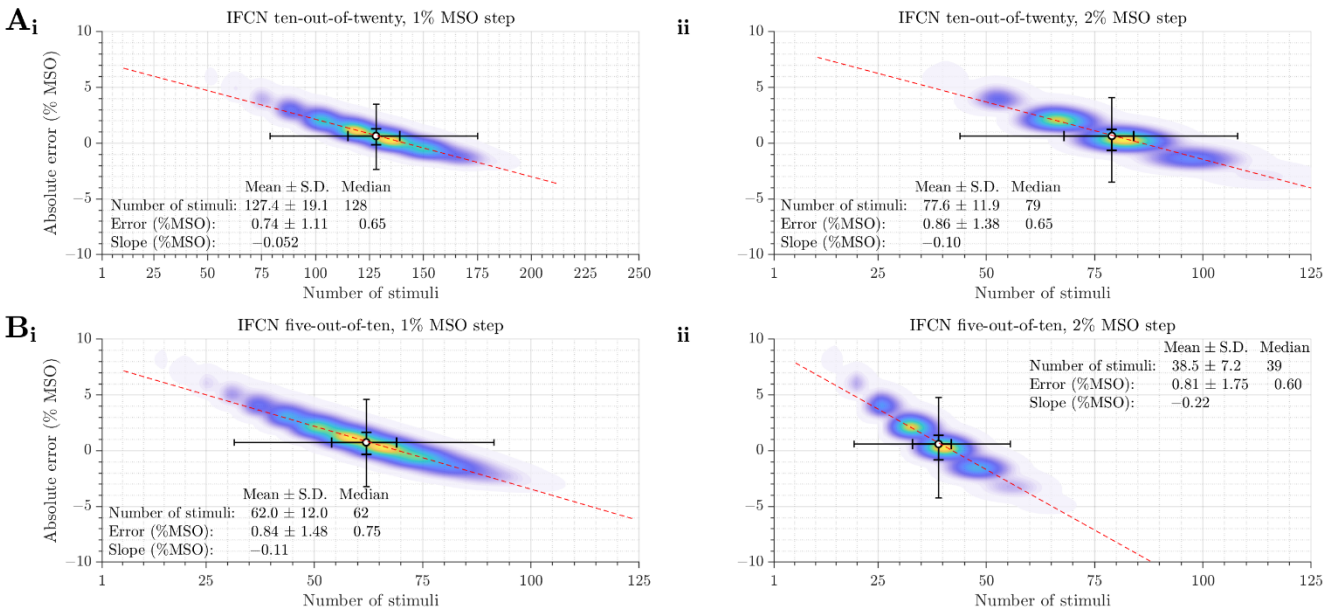

Figure S4: Theoretical distribution of absolute threshold errors of the IFNC relative-frequency methods
Similar to Figure S3.

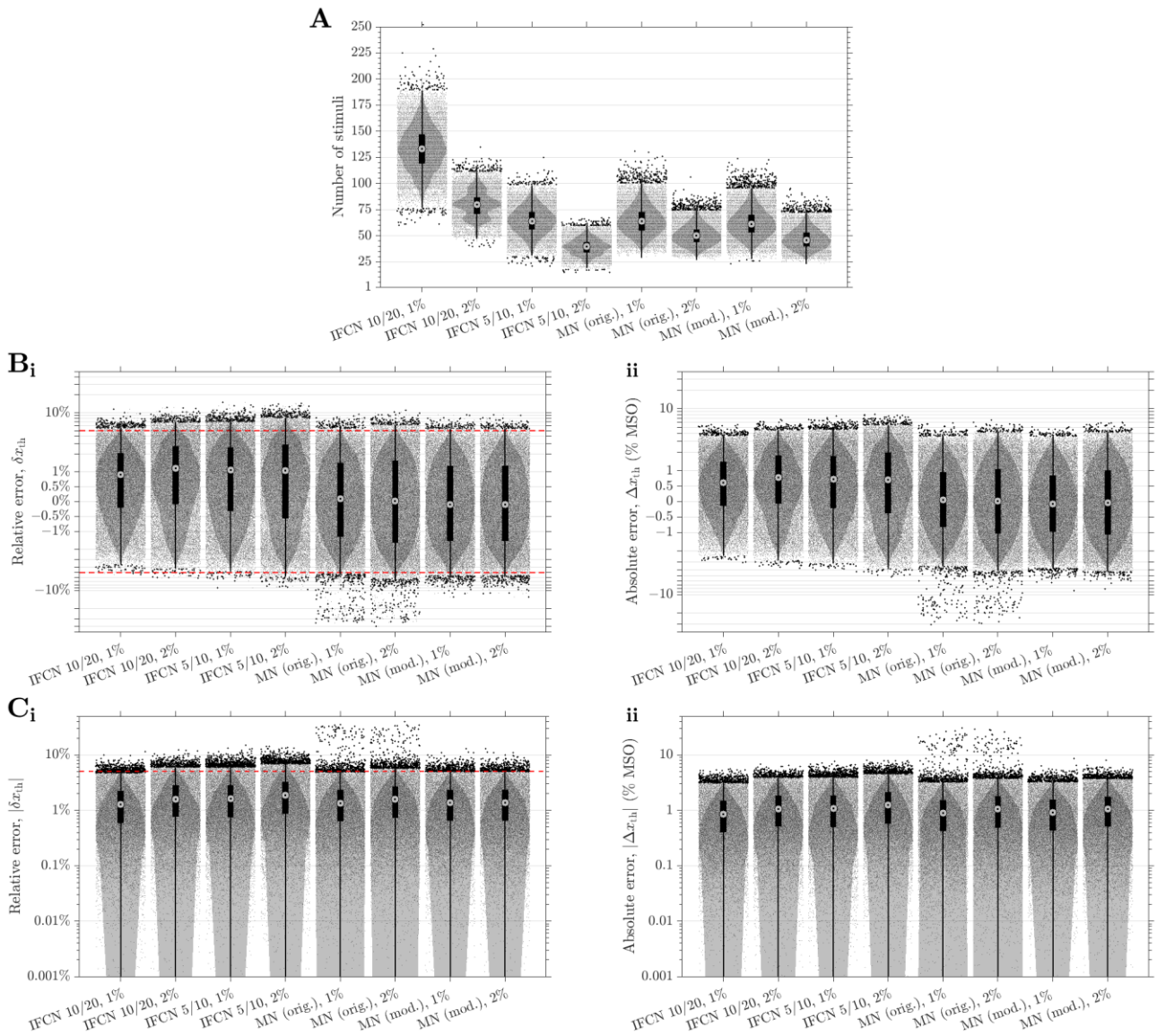

Figure S5: Statistics of IFCN relative frequency and MILLS–NITHI methods

The box-and-whisker plots show the medians (circles), interquartile ranges (thick lines). The ranges extending to the lower and upper adjacent values—1.5 times the interquartile range below and above the first and third quartiles (Q_1 and Q_3), respectively—are shown as whiskers (thin lines). All individual data points are shown, with outliers as darker dots. The gray patches are probability density estimates based on a normal kernel function. **A.** Number of stimuli. **B.** Threshold errors shown on a log-lin-log mixed scale to better visualize the errors around zero, with linear scaling in the range of [−1%, 1%] and [−1, 1] %MSO for the relative and absolute errors, respectively, and logarithmic scaling of amplitudes outside these ranges. **i.** Relative error of the threshold. **ii.** Absolute error of the threshold. **C.** Absolute value of threshold errors shown on logarithmic scale. The whiskers and probability density estimates extend to zero, but are cut off at 0.001% and 0.001% MSO, respectively, for visualization. **i.** Relative error of threshold. **ii.** Absolute error of threshold. The ±5% criteria for safety are shown as red dashed line for the relative errors in panels B_i and C_i.

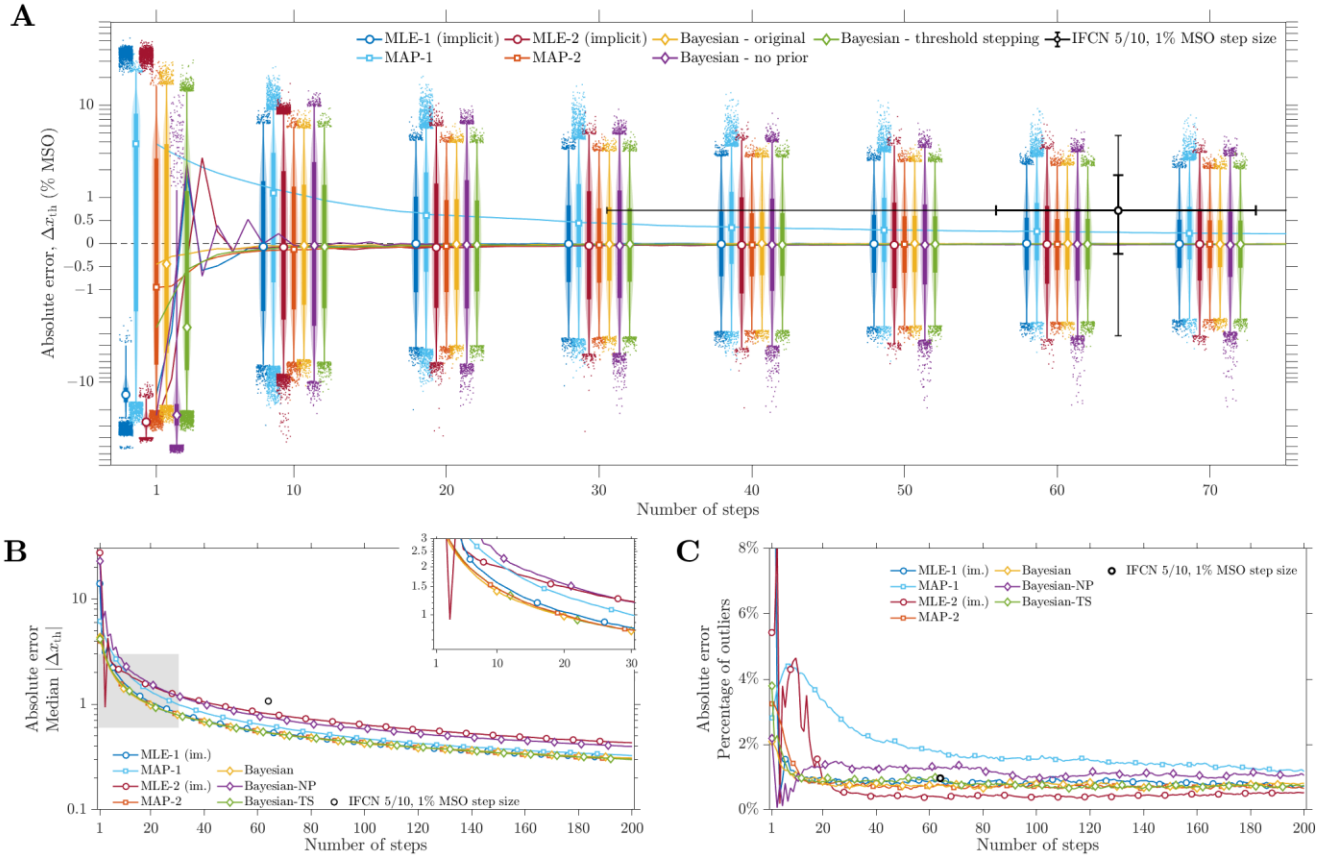

Figure S6: Absolute threshold errors of parametric estimation methods
Shown similarly as the relative threshold error in Figure 3.

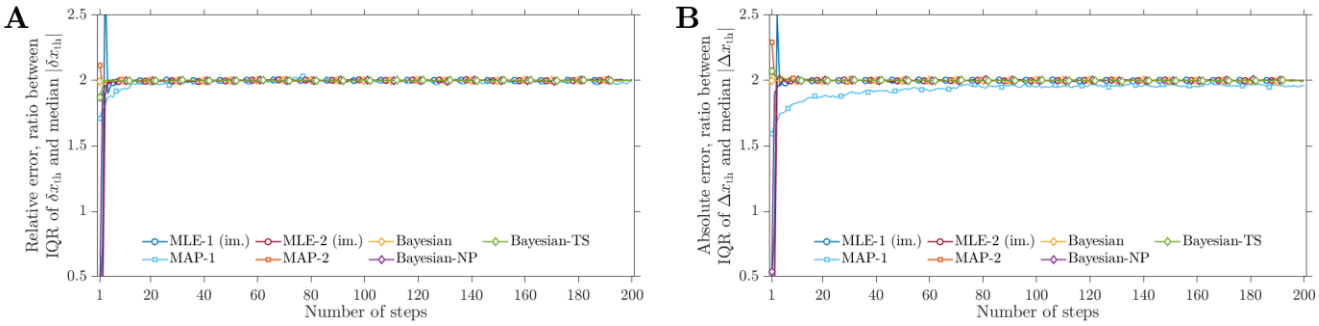

Figure S7: Ratio of median of absolute value of errors and interquartile ranges (IQR) of errors of parametric estimation methods
A. Ratio for relative threshold errors (related to Figures 3). **B.** Ratio for absolute threshold errors (related to Figure S6).

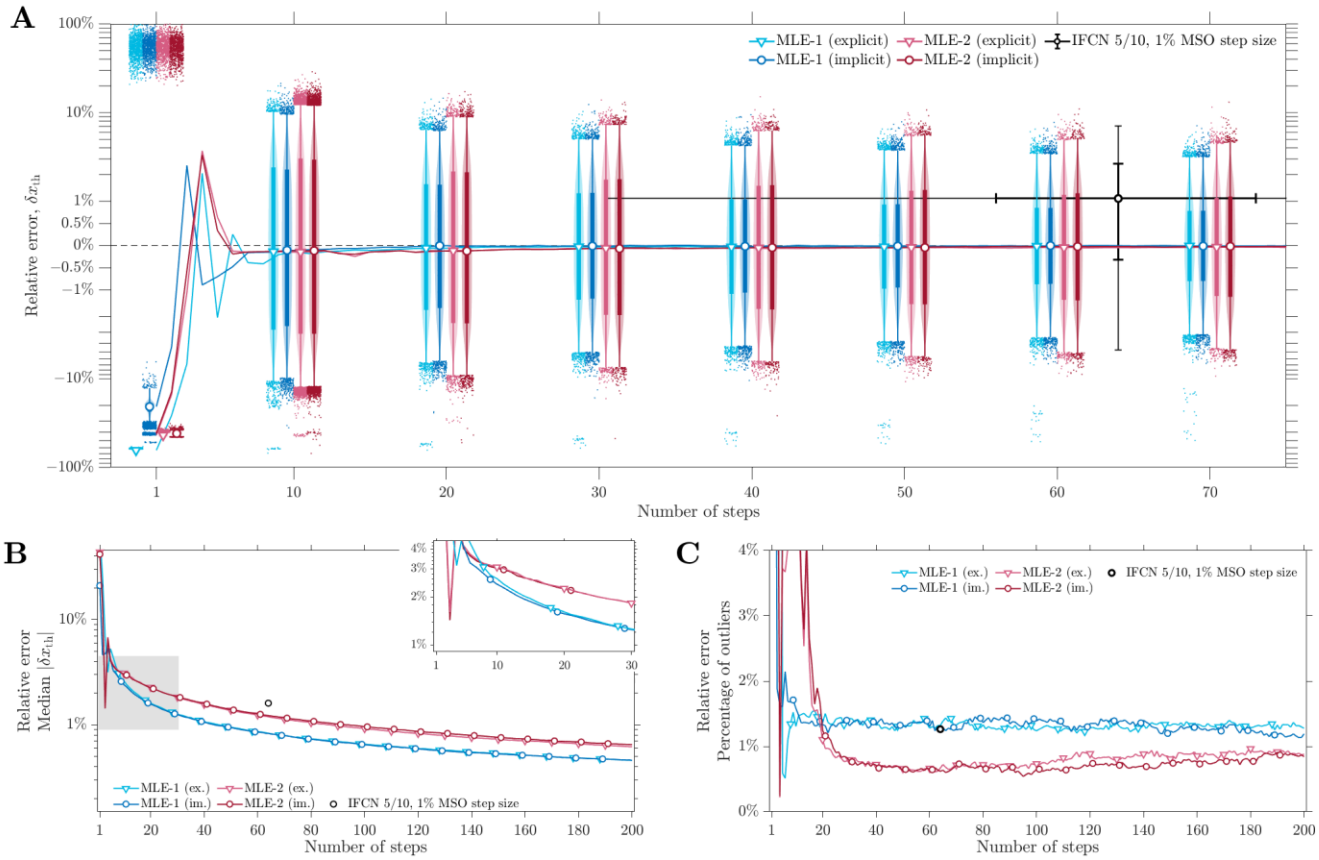

Figure S8: Relative threshold errors of MLE methods using explicit (ex.) and implicit (im.) maximization of the likelihood function

Shown similarly as in Figure 3.

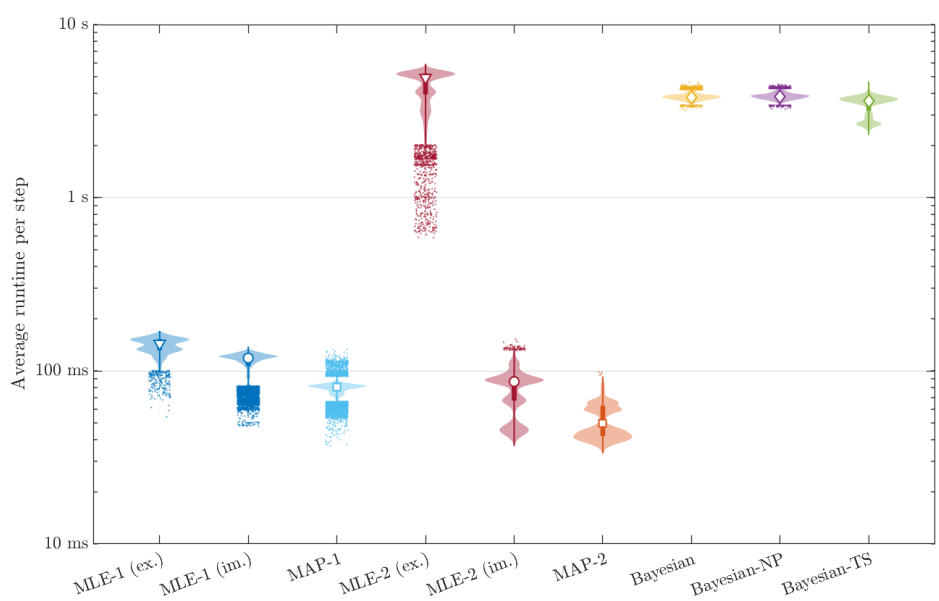

Figure S9: Runtime of parametric estimation methods

The distribution of average runtime per step shown for the 25,000 virtual subjects.

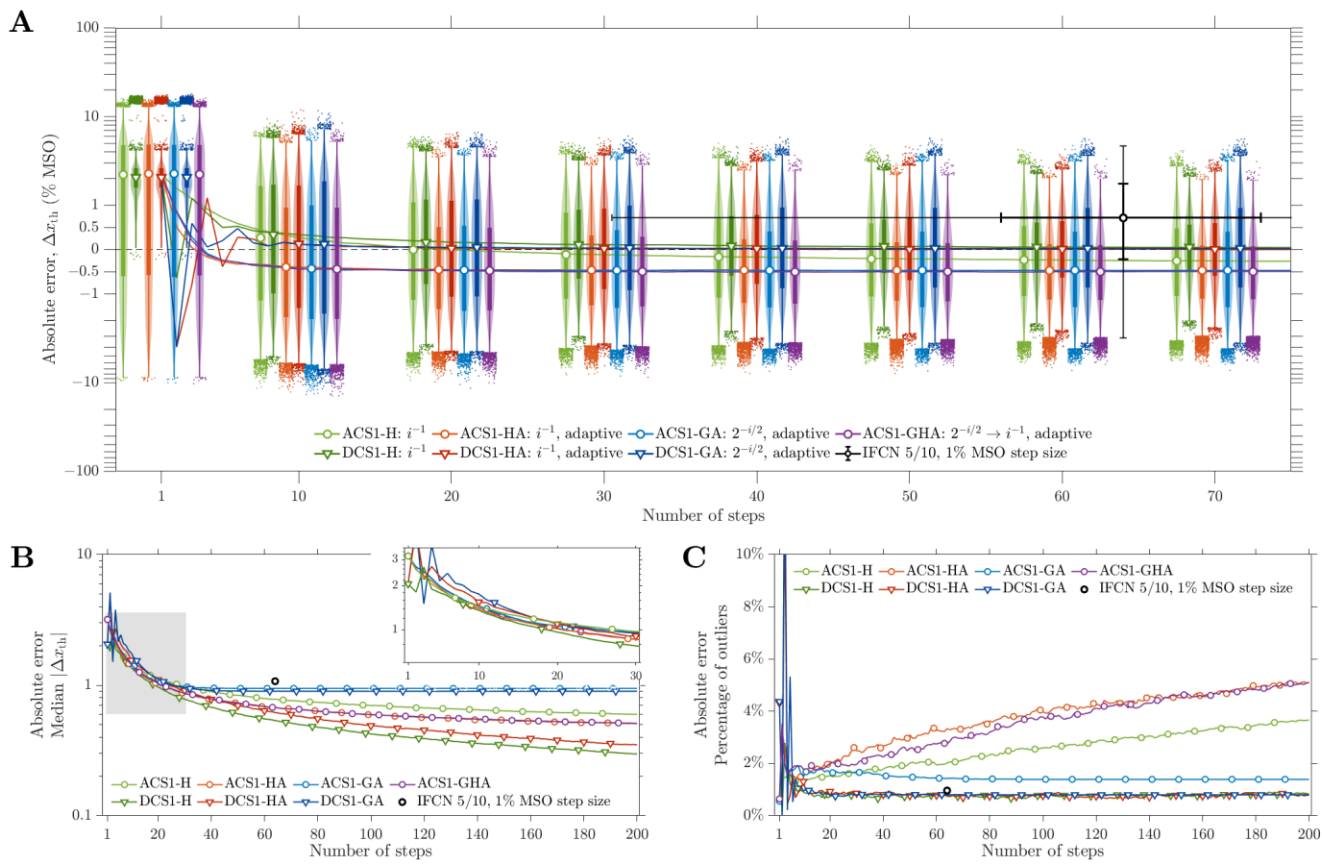

Figure S10: Absolute threshold errors of root finding methods using first-order control sequence with default initial step size a_0

Shown similarly as the relative threshold error in Figure 4.

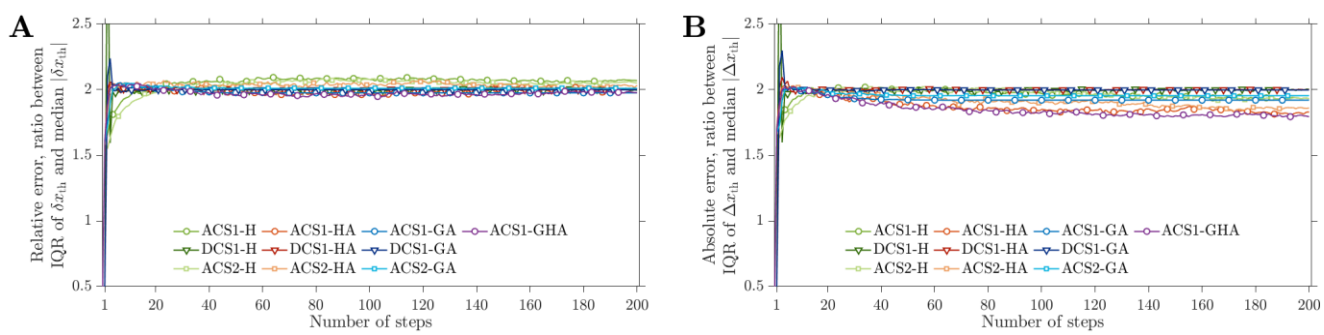

Figure S11: Ratio of median of absolute value of errors and interquartile range (IQR) of errors of root finding methods using default parameters

A. Ratio for relative threshold errors (related to Figures 4 and S15A). **B.** Ratio for absolute threshold errors (related to Figures S10 and S15B).

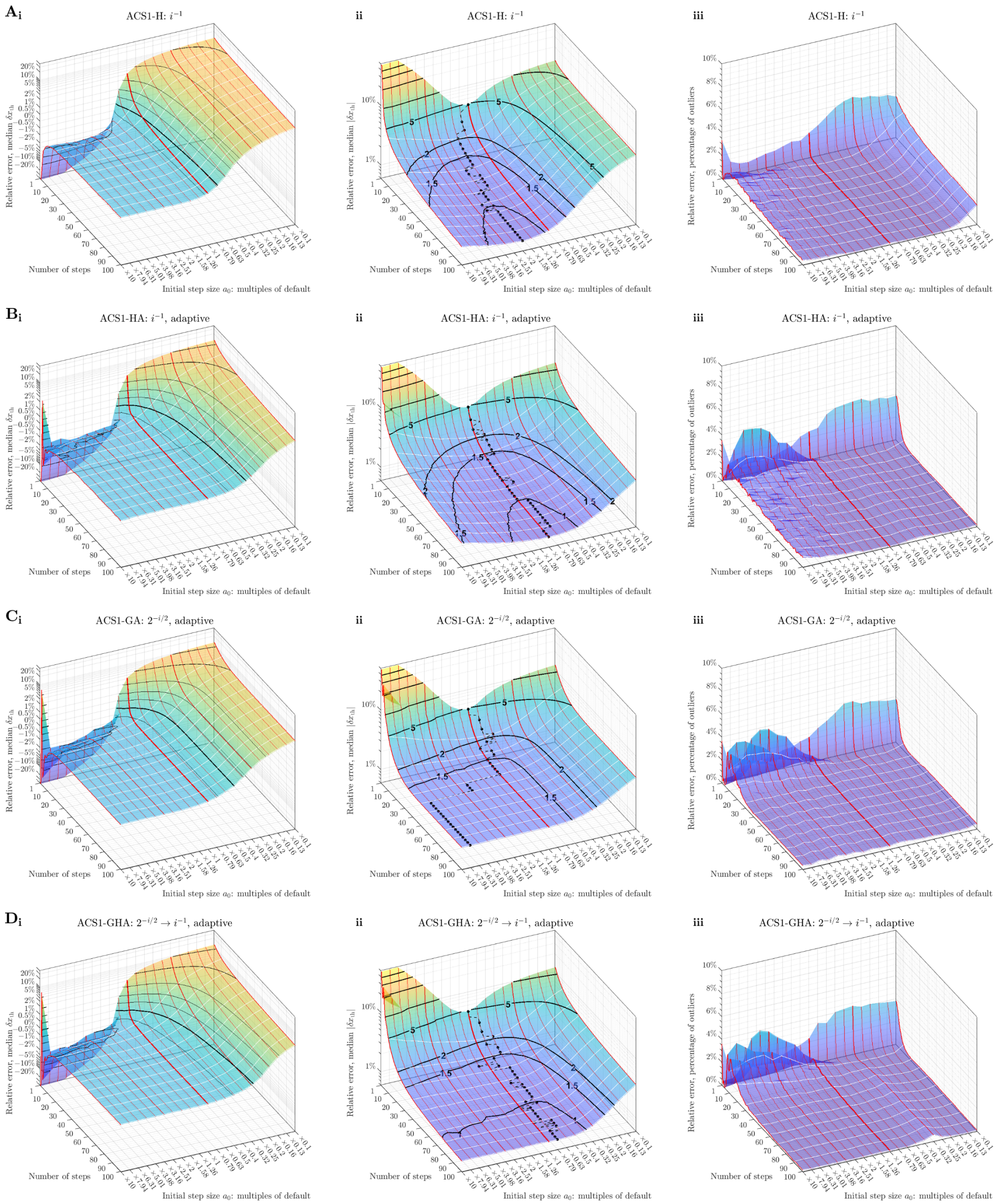

Figure S12: Relative threshold errors of root finding methods using first-order control sequence for different initial step size a_0

The red lines show results for different initial step size a_0, which are provided as multiples of the default value, with the thick red line showing the default value (results in Figures 4). The white lines mark every tenth step. **A.** Results for ACS1-H. **i.** Median of errors shown on log-lin-log scale (as in Figures 4A). Black contour lines are shown at levels corresponding to the numbered ticks on the z-axis, with the thick line indicating zero. **ii.** Median of absolute value of errors shown on logarithmic scale (as in Figures 4B). The minimum at every step is shown with the black dashed line with a black dot at every third step. **iii.** Percentage of outliers (as in Figures 4C). **B.**–**D.** Results for ACS1-HA, ACS1-GA, and ACS1-GHA, respectively.

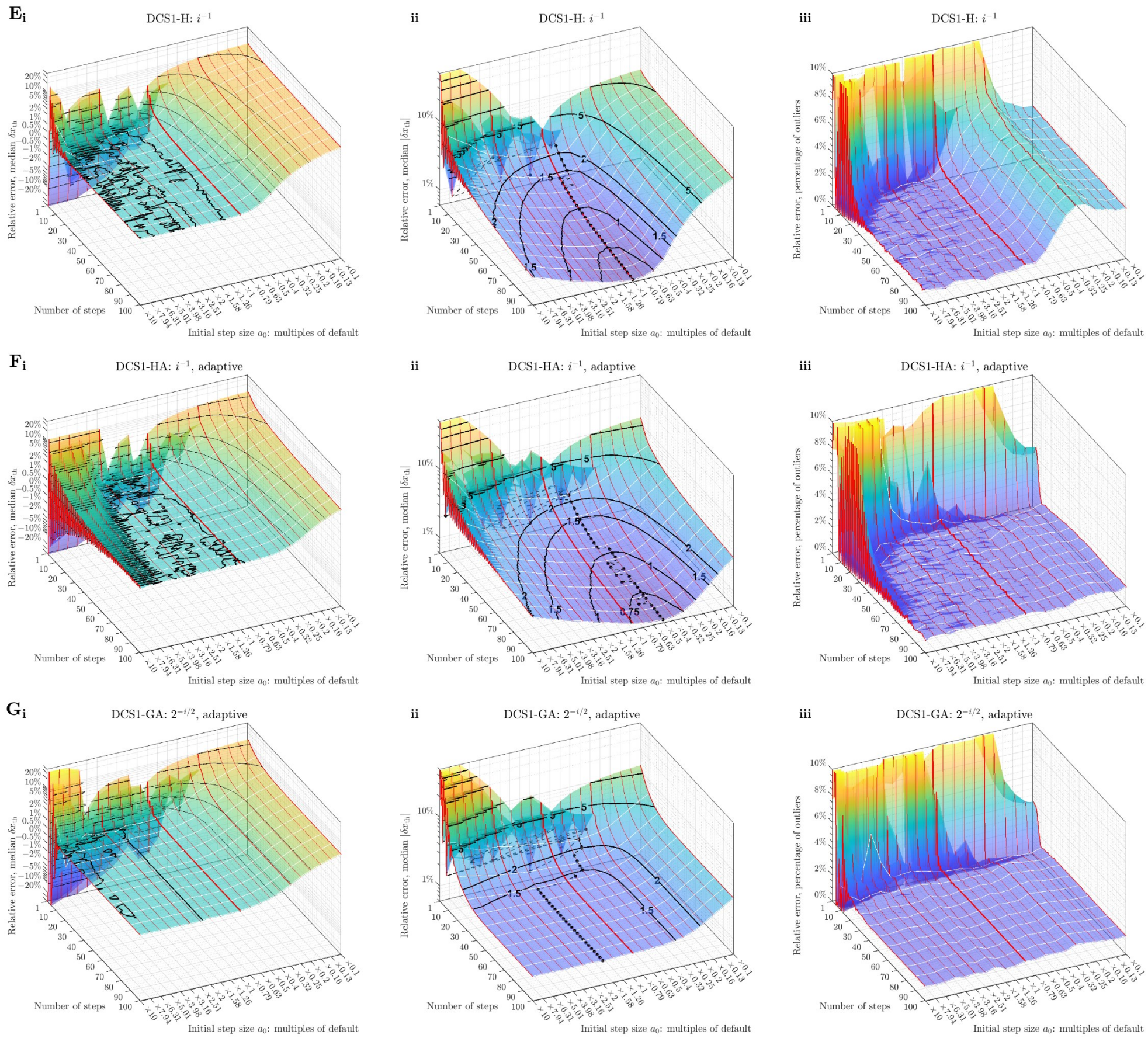

Figure S12: continued.
E.–G. Results for DCS1-H, DACS1-HA, and DCS1-GA, respectively.

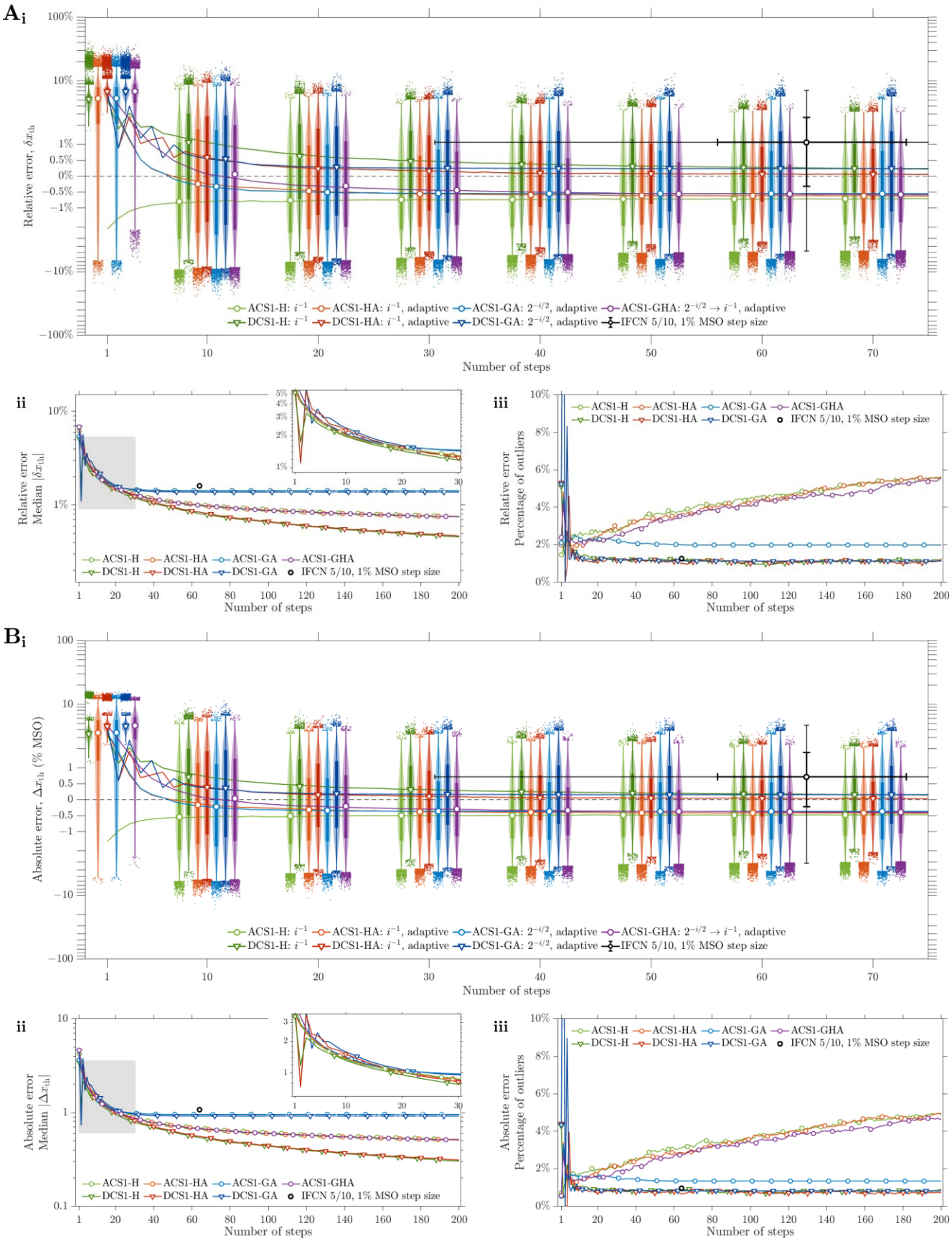

Figure S13: Threshold errors of root finding methods using first-order control sequence with optimal initial step size a_0

Shown similarly as in Figures 4 and S10. **A.** Relative threshold error. **B.** Absolute threshold error.

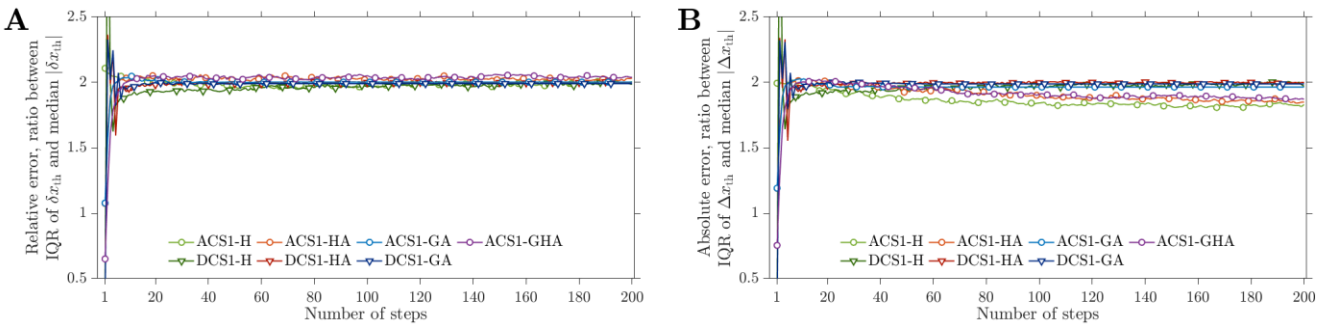

Figure S14: Ratio of median of absolute value of errors and interquartile range (IQR) of errors of first-order root finding methods using optimal initial step size a_0

A. Ratio for relative threshold errors (related to Figures S13A). **B.** Ratio for absolute threshold errors (related to Figures S13B).

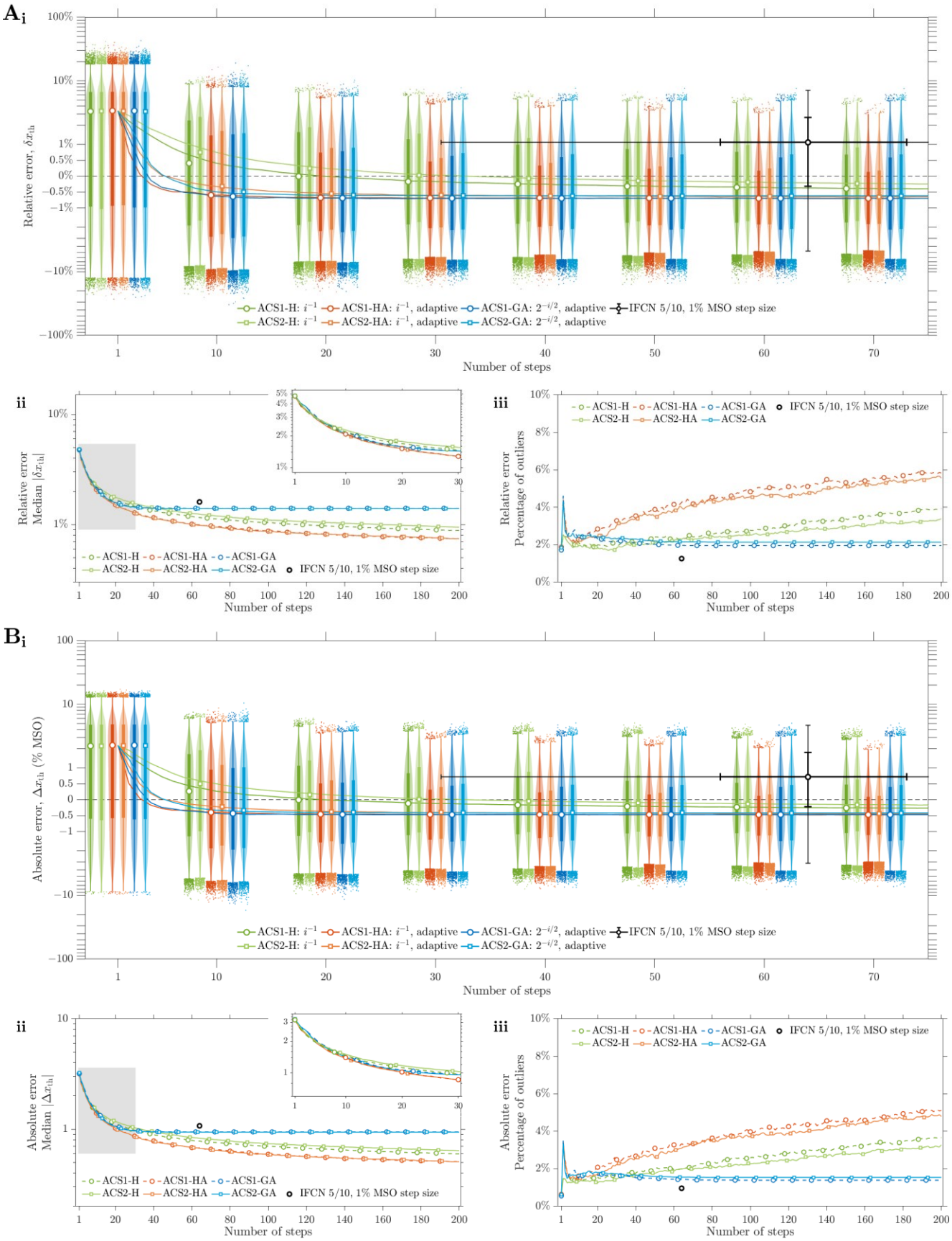

Figure S15: Threshold errors of root finding methods using second-order analog control sequence with default parameters

The corresponding methods using first-order analog control sequence are also shown for comparison (as in Figures 4 and S10). **A.** Relative threshold error. **B.** Absolute threshold error.

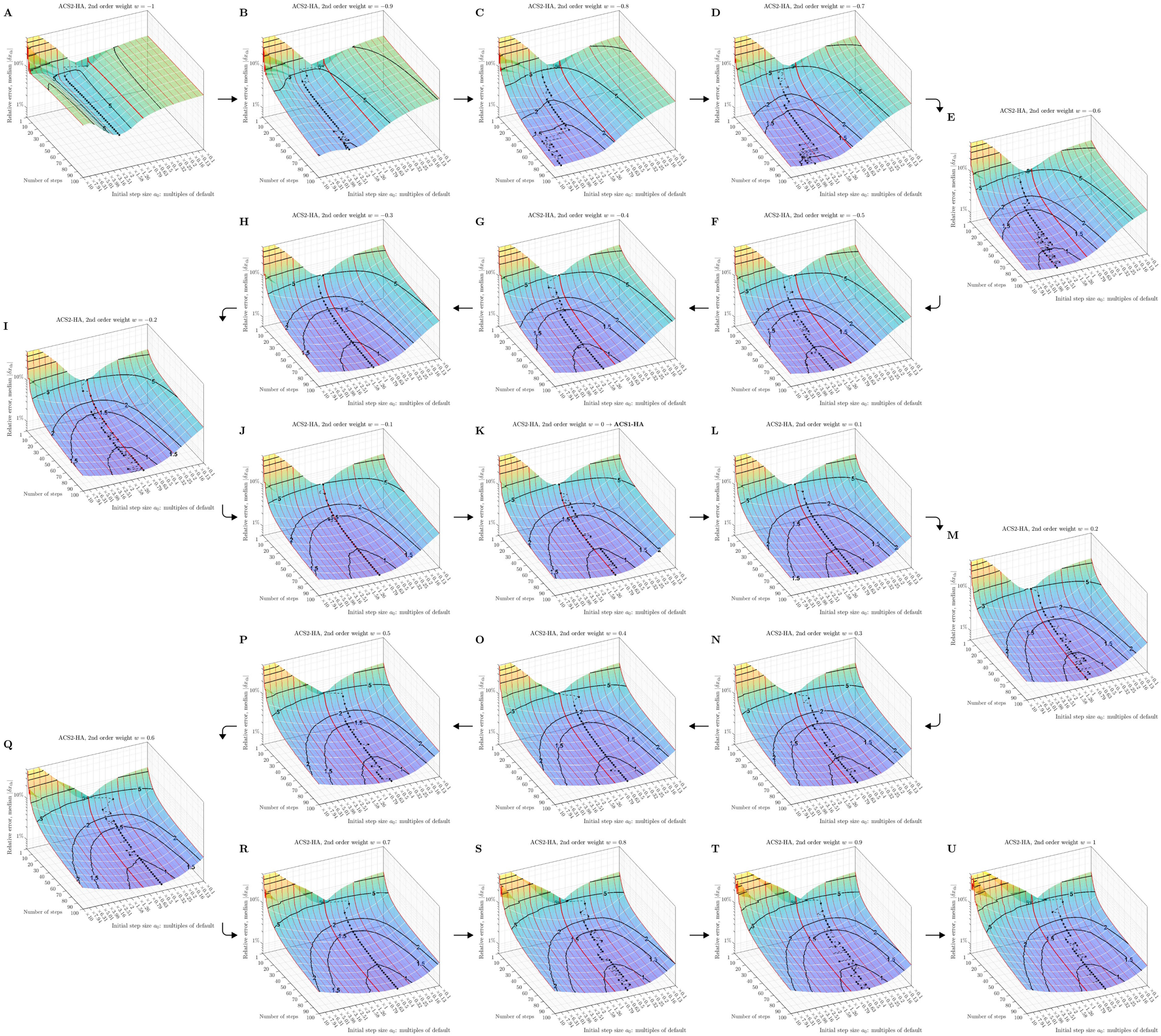

Figure S16: Relative threshold errors of root finding methods using second-order analog control sequence for different initial step size a_0 and given second order weights w
A.–U. Results for second order weights from −1 to 1 in steps of 0.1, shown similarly as in Figure S12A$_{ii}$.

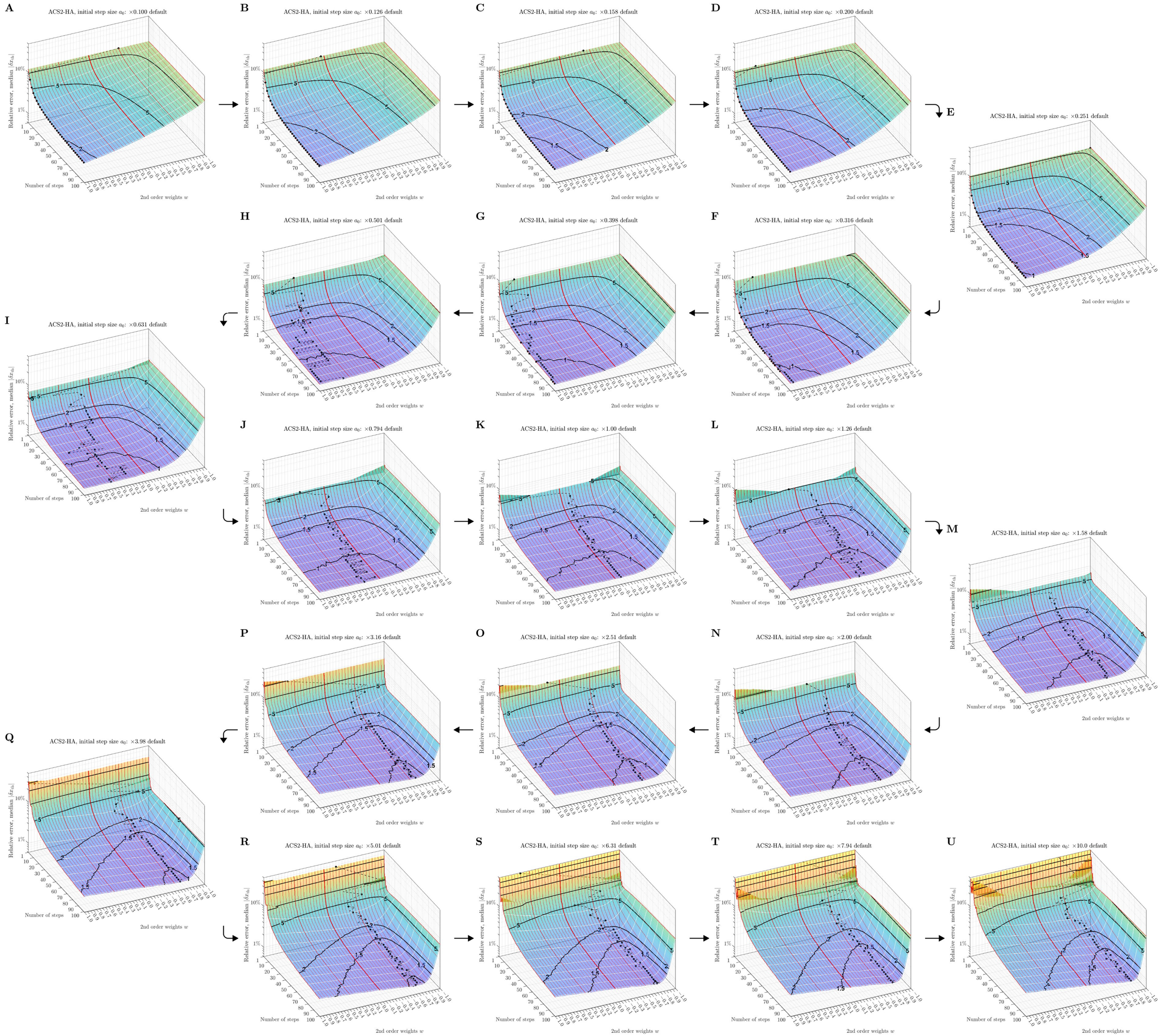

Figure S17: Relative threshold errors of root finding methods using second-order analog control sequence for different second order weights w and given initial step size a_0

A.–U. Results for initial step size from 0.1 times to 10 times the default value, uniformly sampled on logarithmic scale, shown similarly as Figure S12A$_{\mathrm{ii}}$ but with second-order weight w as the parameter. The thick red line indicates w=0, corresponding to the first-order version.

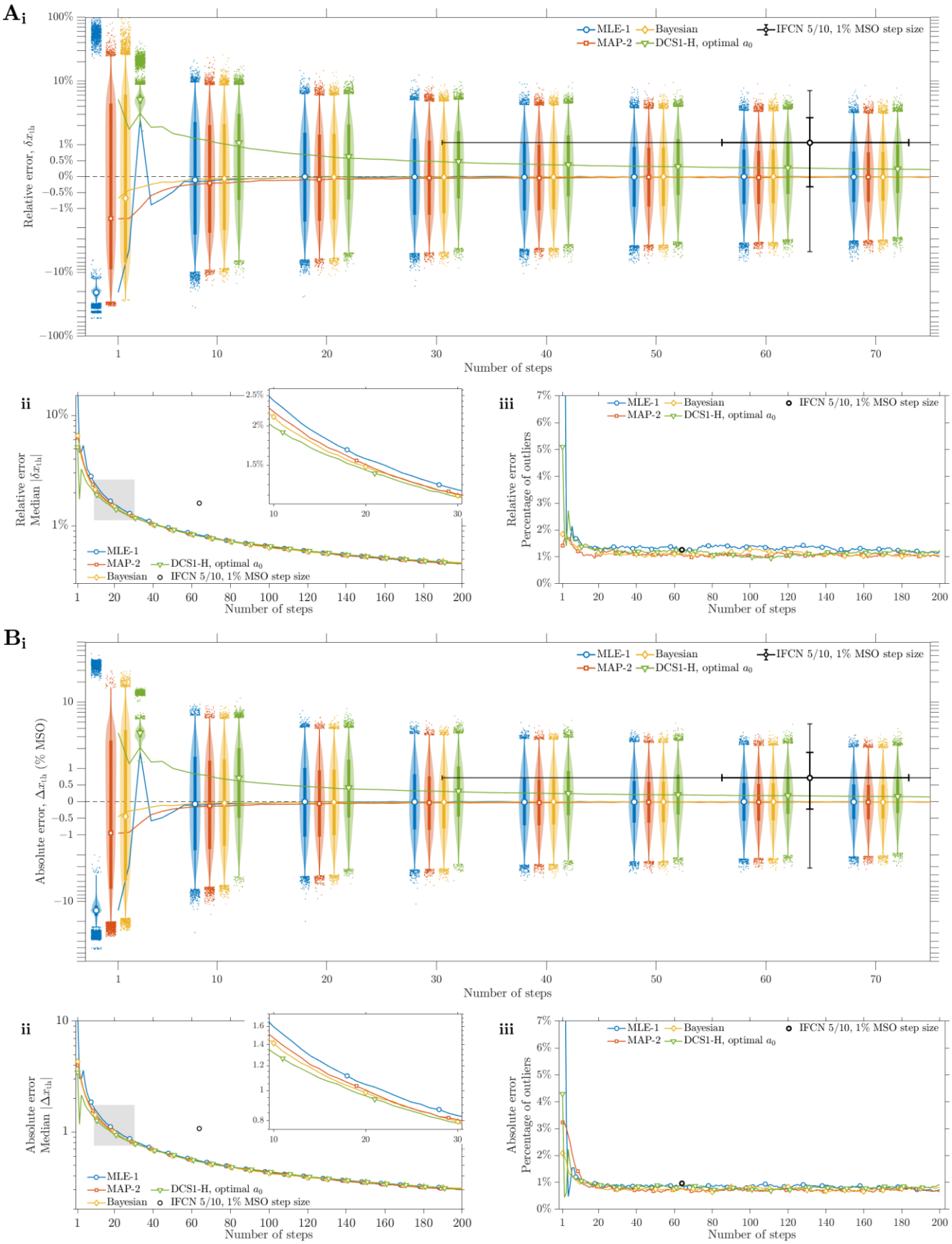

Figure S18: Threshold errors of the best performing methods in each category

Related to Figure 5. Similar format as Figures 3 and S6. **A.** Relative threshold error. **B.** Absolute threshold error.