Analysis and Comparison of Methods for Determining Motor Threshold with Transcranial Magnetic Stimulation

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, Duke University, 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. Transcranial magnetic stimulation (TMS) is a noninvasive neuromodulation method with both clinical and research applications. However, the stochastic aspect of neuronal activity and motor evoked potentials (MEPs) challenges TMS thresholding methods. We analyzed existing methods for obtaining TMS motor thresholds and their variations, introduced new methods from other fields, and compared their accuracy and speed.

Approach. In addition to existing relative-frequency-based methods, such as the five-out-of-ten method and related approximation of the probability definition of TMS threshold, we examined adaptive methods based on a probabilistic motor threshold model and maximum-likelihood (ML) or maximum *a-posteriori* (MAP) estimation. To improve the performance of the 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 and, despite good accuracy on the population level, 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 is included. Stochastic root-finding methods were comparable to adaptive estimation methods but were much simpler to implement and did not rely on an underlying estimation model.

Significance. Two-parameter MAP estimation, Bayesian estimation, and stochastic root-finding methods require fewer TMS pulses for accurate estimation than conventional relative-frequency and single-parameter ML estimation. 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. Similar to 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 probability distribution of the binary response for a given stimulation—most commonly of a cumulative GAUSSIAN sigmoid function with two parameters, the midpoint/threshold and slope. 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 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 slope of the probability function at the threshold. 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 slope–threshold regression line, 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 slope 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] to TMS motor threshold estimation using the same model as in the point estimators above and compare its performance with that of the latter.

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] are adapted and evaluated here.

1.4. Organization of this paper

We first describe existing TMS thresholding methods as well as 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; 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—require fewer TMS pulses for accurate motor threshold estimation than the conventional relative-frequency and single-parameter MLE methods.

2. Methods

2.1. Classification 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. The early approaches, in contrast, namely the aforementioned five-out-of-ten or ten-out-of-twenty relative-frequency method based on recommendations of the International Federation of Clinical Neurophysiology (IFCN) [27] or the MILLS–NITHI two-threshold method [39], do not need such a predefined model description. 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.

2.2. Nonparametric methods

2.2.1. IFCN relative-frequency method

The procedure described by the IFCN [27] 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, stimuli are applied up to a maximum of ten or twenty until either positive responses of half the number (five or ten, respectively) or negative responses of more than half the number (six or eleven, respectively) are observed.

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, 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).

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 (LT).

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 (UT) is designated as the minimum intensity at which 10 stimuli all produce positive responses.

d) The threshold is calculated as the mean of LT and UT.

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).

Table 1 Summary of relative-frequency 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

2.2.3. 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:

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

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 ($y_i = \lg(V_i/[1\ \text{V}])$) 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 aforementioned 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 in order to compare the convergence properties (Table 2). In the basic version ACS1-H, the control sequence followed a harmonic series with almost sure (a.s.) convergence[1]

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

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

Following the style of Delyon and Juditsky [67], 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: $\mathrm{sgn}(\Delta y_i) \neq \mathrm{sgn}(\Delta y_{i-1})$. 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,

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

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

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

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.

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 to a modified sign function $\Delta y_i = \mathrm{sgn}(y_{\mathrm{th}} - y_i) \in \{-1, +1\}$ for the two different cases of below threshold and above (or equal) threshold. DCS allowed the stochastic root-finding method to be used without EMG recordings in a compatible manner as 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.

The definition of appropriate initial step size a_0 was assisted by the iteration formulation (3), which resembled a slope [68] after rearrangement

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

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 $(x_{\mathrm{th}}, y_{\mathrm{th}})$. 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 $\lg(10^{-5})$ (10 μV) and $\lg(10^{-2})$ (10 mV) 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 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 GAUSSIAN model with two degrees of freedom—the threshold $t = x_{\mathrm{th}}$ and the slope s at threshold:

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

Here, r_i denoted the binary response (2) of a single stimulus x_i.

The likelihood function was defined for independent events

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

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.

Previously, the estimation (13) was simplified by using only the threshold parameter t, whereas the slope parameter was fixed to the value $s = 0.07 \cdot t$ [48,69]. We replicated this method (MLE-1) and also performed a full MLE with both parameters (MLE-2) (Table 3), implementing and comparing two approaches 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 were chosen for $s \in [0.005, 0.5]$. 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 and only the estimation model was modified to use 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]

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

in which $\mathbf{E}$ 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 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) and (18) 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 [48,70] 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), with parameters mimicking real measurements [71]. The analog amplitude response, i.e., the peak-to-peak MEP amplitude V_{PP} in units of volt, was defined as

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

where x was the stimulation amplitude in percent MSO, $p_{1,j}$ to $p_{9,j}$ were the individual response parameters for subject j obtained from an experimental population [34]; 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 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 [34,71,72]. Therefore, stimulation amplitudes higher than the nominal 100% MSO could be achieved by increasing the capacitor voltage. 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. The ground truth threshold for each subject was determined from the corresponding *noiseless* model (see Appendix) by applying root-finding methods to the response curve to identify the stimulation that generated a 50 μV response (Figure 1B). By applying MLE to each subject using 10,000 simulation–response pairs, the ground truth parameters t and s were obtained and analyzed (Figure S1). The probability density distribution of thresholds and slopes $p(t, s)$ 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.

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 [52,73,74]. To avoid ambiguity and circular definition between threshold and hotspot [75], the latter was chosen as the stimulation level that generated a 500 μV response in the noiseless model (Figure S2). The starting amplitude was on average 10% MSO higher than the ground truth threshold (1.14 times threshold), but a portion of the virtual subjects had very steep input–output curve that resulted in a starting amplitude very close to the true threshold. On the other hand, a few subjects with very flat input–output curve had very high starting amplitudes, and for three such outliers the starting amplitude was reduced to 130% MSO, the maximum stimulation amplitude in this study. 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 ROBBINS–MONRO-like algorithms 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 stochastic root-finding methods and parametric estimation algorithms 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.

Computation and data analysis was performed with MATLAB (version2021a, The Mathworks, Natick, MA, USA) on the Duke Compute Cluster. The code is available online (DOI: 10.5281/zenodo.6483601).

3. Results

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

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

For the IFCN relative-frequency methods (Figures 2 and S3, panels A_i, A_{ii}, B_i, and B_{ii}), the number of stimuli until the method terminated and reported a threshold value for a given subject had a wide distribution, ranging from around five to twenty steps (in which a single step below the starting amplitude was already considered subthreshold by the method, terminating the procedure practically immediately after the start) to a few hundred steps. The errors overall had a relatively small bias (e.g., median 0.86%, mean and SD 0.94%±1.72% for δx_{th} of the ten-out-of-

twenty method with 1% MSO stepping size), skewing slightly towards the positive end due to more positive outliers (Figure S4, panels B_i and B_{ii}). Estimates obtained with fewer stimuli (e.g., within 40 stimuli for ten-out-of-twenty method with 1% MSO stepping size) especially tended to overestimate the threshold due to early false-positive fulfillment of the termination rule. As the number of stimuli increased, the error distribution became more symmetric, becoming relatively stable (e.g., between 100 and 250 stimuli), and only narrowed for thresholds obtained with a very large number of steps (e.g., over 250 steps). Increasing the stepping size from 1% to 2% MSO and/or using a coarser five-out-of-ten instead of ten-out-of-twenty for the relative frequency approximation decreased the number of stimuli drastically without changing the error distribution qualitatively. The threshold accuracy was also decreased (i.e., increased median, mean, and standard deviation of errors). While both changes were statistically significant, the difference in errors had little practical meaning for any individual subject given that they had similar ranges for both ten-out-of-twenty or five-out-of-ten methods and either step sizes.

The MILLS–NITHI methods (Figure 2 and S3, panels C_i, C_{ii}, D_i, and D_{ii}) had narrow distributions of the number of stimuli, requiring at least 25 stimuli but rarely exceeding 100. The numbers of stimuli for all four variations were on average smaller than the IFCN five-out-of-ten method with 1% MSO step size but larger than that with 2% MSO step size. For any individual subject, the IFCN relative-frequency method could be much faster if the starting point for the latter was already close to the threshold (i.e., steep slope near the threshold), whereas the original MILLS–NITHI methods did not use the information obtained from the hotspot search. The modified MILLS–NITHI methods exploited this additional information but only reduced the number of stimuli by 4 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. The errors of the original MILLS–NITHI methods were also centered closer to zero (e.g., median 0.08%, mean and SD 0.01%±2.46% for 1% MSO step size) compared to the IFCN five-out-of-ten method, but had a much wider error distribution and more outliers. In contrast, the modified methods biased the errors slightly towards negative values (e.g., median −0.19%, mean and SD −0.21%±2.04% for 1% MSO step size) but had fewer outliers. Increasing the step size to 2% MSO for either original or modified method reduced the number of steps on average by 14 steps and narrowed its distribution but had negligible effect on the accuracy.

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 S5A). The median $|\delta x_{\mathrm{th}}|$ and the interquartile range (IQR) of δx_{th} (Figure 3B, open and filled markers, respectively) 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 S6). Therefore, the median $|\delta x_{\mathrm{th}}|$ captured the distribution of the majority of errors, with the remaining outliers summarized as a percentage of the total number of subjects (Figure 3C). The IFCN five-out-of-ten method with 2% MSO step size, which obtained a threshold estimate with a median 40 steps and median $|\delta x_{\mathrm{th}}|$ of 1.88%, was used as reference.

The basic MLE method [48], which used a single-parameter MLE for threshold estimation and stepping (MLE-1, Figure 3, dark blue lines), reached the same error level as the reference IFCN five-out-of-ten method with only 15 steps (Table 4). At 20 steps, which is considered to produce a sufficiently accurate estimate for all subjects using existing ML-PEST (MLE-1) [48], the median $|\delta x_{\mathrm{th}}|$ of MLE-1 was only smaller than that of the reference method by one fifth, but further decreased to about half that of the reference method at 40 steps. Evaluating the likelihood function explicitly had little effect on the threshold estimation performance (Table 4, results in parentheses, and Figure S7) and computational runtime (Figure S8) compared to implicit maximization. Introducing prior information of the parameters of the population (MAP-1, Figure 3, light blue lines) 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 slightly worse than MLE-1 as the parameters of some subjects in our model were far from such a 0.07 relationship (Figure S1, panels D and F).

Table 4 Error convergence of parametric estimation methods

Method	Steps reaching median $\lvert\delta x_{\mathrm{th}}\rvert$ of 1.88% [a]	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 20 [b]	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 40 [c]
MLE-1 [d]	15 (16)	1.57% (1.62%)	1.06% (1.07%)
MAP-1	16	1.61%	1.09%
MLE-2 [d]	30 (29)	2.34% (2.32%)	1.59% (1.57%)
MAP-2	14	1.50%	1.036%
BAYESIAN	13	1.49%	1.033%
BAYESIAN-NP	29	2.36%	1.51%
BAYESIAN-TS	14	1.51%	1.05%

[a] Median $\lvert\delta x_{\mathrm{th}}\rvert$ of the reference IFCN five-out-of-ten method with 2%MSO step size.
[b] Number of steps considered to produce a sufficiently accurate estimate for all subjects using ML-PEST [48].
[c] Median number of steps of the reference IFCN five-out-of-ten method with 2%MSO step size.
[d] Results obtained with direct maximization of the likelihood followed by those using explicit evaluation in parentheses.

The two-parameter MLE method (MLE-2, Figure 3, dark red lines) had a slower convergence compared to MLE-1 and reached the error level of the IFCN reference slower (30 steps). This 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 S7). However, the explicit evaluation with two parameters resulted in a much higher computational runtime per step that was about ten times slower (median 1 s vs. 0.1 s, Figure S8). Using a two-parameter MAP estimator for the threshold (MAP-2, Figure 3, orange lines) significantly improved the performance of MLE-2, reaching the error level of the IFCN five-out-of-ten method with only 14 steps, and had lower median $\lvert\delta x_{\mathrm{th}}\rvert$ compared to either MLE-1 or MAP-1. As the slope 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 similar to their implicit MLE counterparts (median 0.07 s and 0.1 s, respectively, Figure S8).

Among all parametric estimation methods, the original KONTSEVICH–TYLER method (BAYESIAN, Figure 3, yellow lines) reached the IFCN five-out-of-ten method's median $|\delta x_{\mathrm{th}}|$ the fastest at step 13 and had the lowest error at any given step. The BAYESIAN method achieved more accurate estimation 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. Removing the prior population information (BAYESIAN-NP, Figure 3, purple lines) resulted in slower convergence and performance similar to that of MLE-2 with 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) 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) and (14). The BAYESIAN methods were the most computationally costly and slowest (Figure S8), with runtime on the order of one second per step and more than ten times 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.

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, 5, and S14A, and absolute error Δx_{th} in Figures S9, S12, and S14B). All root-finding methods were provided with the same above-threshold amplitude from the hotspot search and thus had the same distribution at the zeroth step (gray elements in panels A, A_i, or B_i of Figures 4, 5, S9, S12, and S14). The ratio between the IQR of δx_{th} and the median $|\delta x_{\mathrm{th}}|$ (Figures S10 and S13) 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) except for ACS1-H became negatively biased within ten 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 S9A). 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 2% MSO step size with only 14 steps (Figures 4B and Table 5), whereas ACS1-H needed 35 steps and was only slightly faster.

The non-adaptive ACS1-H (Figure 4, orange 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, yellow lines) outperformed its non-adaptive counterpart by reaching a lower median $|\delta x_{\mathrm{th}}|$ faster, with distributions of δx_{th} being more negatively biased and having very few positive 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 1.40%, 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 a.s. harmonic convergence and performed very similarly compared to ACS1-HA in terms of median

$|\delta x_{\mathrm{th}}|$. With increasing number of steps, the percentage of outliers increased for ACS1-GHA, though they were still less than those for either ACS1-H or ACS1-HA.

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 S11, 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 S11, 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 and/or 40 for which the default was not optimal and the performance indeed improved (Figures 5 and S12 and Table 5). 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 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 $4A_{ii}$). For the short-range behavior within 20 steps, however, the adaptive analog versions performed better. Thus, the adaptive analog versions overall had faster initial convergence of $|\delta x_{\mathrm{th}}|$, utilizing continuous information. But 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 only positive bias for δx_{th}, with the bias for the two adaptive methods quickly decaying close to zero after ten steps. In contrast, the relatively large negative bias of their analog counterparts plateaued and did not decay. With optimal a_0, the

difference in median $|\delta x_{\mathrm{th}}|$ between the analog and digital versions (Figure 5B) became smaller for short term behavior. The convergence of $|\delta x_{\mathrm{th}}|$ improved for the digital versions, with a slight positive bias of the error distribution for the adaptive methods (Figure 5A).

Table 5 Root-finding methods and their error convergence

Method	Step reaching median $\lvert\delta x_{\mathrm{th}}\rvert$ of 1.88% [a]	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 20 [b]	Median $\lvert\delta x_{\mathrm{th}}\rvert$ at step 40 [c]	Optimal a_0 (ratio to default value)
ACS1-H[d]	35 (15)	2.28% (1.61%)	1.80% (1.22%)	$\times 10^{0.4} \approx \times 2.51$
ACS2-H	43	2.43%	1.91%	–
DCS1-H[d]	16 (14)	1.62% (1.53%)	1.08% (1.05%)	$\times 10^{0.1} \approx \times 1.26$
ACS1-HA	14	1.565%	1.18%	1
ACS2-HA	14	1.562%	1.195%	–
DCS1-HA[d]	17 (15)	1.69% (1.562%)	1.201% (1.06%)	$\times 10^{-0.3} \approx \times 0.50$
ACS1-GA	14	1.60%	1.41%	1
ACS2-GA	14	1.63%	1.45%	–
DCS1-GA[e]	17 (15)	1.67% (1.64%)	1.36% (1.38%)	$\times 10^{-0.3} \approx \times 0.50$
ACS1-GHA[d]	14 (15)	1.59% (1.59%)	1.22% (1.19%)	$\times 10^{-0.1} \approx \times 0.79$

[a] Median $|\delta x_{\mathrm{th}}|$ of the reference IFCN five-out-of-ten method with 2%MSO step size.
[b] Number of steps considered to produce a sufficiently accurate estimate for all subjects using ML-PEST [48]
[c] Median number of steps of the reference IFCN five-out-of-ten method with 2%MSO step size.
[d] Results obtained with default a_0 followed by those with optimal a_0 for lowest median $|\delta x_{\mathrm{th}}|$ at step 40 in parentheses.
[e] Results obtained with default a_0 followed by those with optimal a_0 for lowest median $|\delta x_{\mathrm{th}}|$ at step 20 in parentheses.

The methods with second-order control sequences (Figure S14, lighter lines with square markers) again could perform either better (ACS2-HA) or worse (ACS2-H and ACS2-GA) 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 S15 and S16). 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.

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 involving the probability density function itself, and ten stochastic root-finding methods utilizing 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 6 and S17). All relative-frequency methods had similar error performance, and therefore the IFCN five-out-of-ten methods with 2% MSO step size with the minimum number of steps on average would be preferred among those. 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. For either the default or optimal initial step size a_0, the best stochastic root-finding methods using analog control sequences was ACS1-HA, followed closely by ACS1-GHA. For digital control sequences, DCS1-H and DCS1-HA both had similar strong performance: DCS1-H can be considered to be slightly superior as it reached the smallest median $|\delta x_{\mathrm{th}}|$ with the fastest speed, whereas DCS1-HA had a tighter error distribution with fewer outliers. 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 40 steps and 1.88% median $|\delta x_{\mathrm{th}}|$), achieving significant better accuracy (1.03% to 1.18%) at the same number of steps (Figure 6A, right) or requiring much fewer steps (within 15 steps, with the novel methods 1 or 2 steps faster than the single-parameter MLE) to reach the same accuracy (Figure 6B, top). Compared to MLE-1, the best-performing novel methods also demonstrated slightly better accuracy at 20 steps (Figures 6A and S17), which is a typical number used to terminate the single-parameter MLE methods [48] for example in the software tool MTAT 2.0 [76].

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 [77]. 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 S4) or more advanced estimation methods (Figures 3A, 4A, and 5A). 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, as well as detailed visualization and descriptive statistics of the entire distribution. In contrast, many previous studies only 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. These were all important characteristics of the thresholding methods. For example, reducing the stepping size or increasing the number of stimuli at a given amplitude for the relative-frequency method appear to improve the accuracy of the threshold estimate on the population level but does not justify the significant increase in required stimuli given the similar wide error distributions. The original MILLS–NITHI methods had a large number of outliers with negative error and the modified methods had negative biases on the population level, contradicting assertions about its reliability [42].

Our implementation of the various methods all started from a suprathreshold simulation amplitude with a fixed 500 μV response to avoid ambiguity and circular definition of hot-spot and threshold. This, however, is often not the case in practice. Starting at an amplitude that evokes a smaller EMG response closer to 50 μV could accelerate

the threshold search, especially for subject with shallow input–output curves. Sometimes, a default stimulator output is used as starting point for all subjects. However, both these situations risk convergence errors in some algorithms when a positive response is observed for a stimulator output well below the true threshold [52,53].

Existing estimation methods previously were tested using virtual subjects that had the same response model underlying the parameter estimation methods. Such tests of estimators against their own estimation model can be misleading due to statistical bias, similar to the previously criticized circular definitions of hotspot and threshold. 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 [32,78–80]. 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- and two- parameter MLE methods. The improvement in accuracy for the BAYESIAN method 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 slope 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 non-parametric stochastic root-finding methods overall had only slightly higher errors compared to the best parametric estimation methods. Nevertheless, they 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 did not necessarily rely on toolboxes (e.g., [48,69]). The best performing root-finding methods, e.g., DCS1-H and ACS1–HA, even with the default initial step size, were only slightly outperformed in terms of speed and accuracy by the best parametric estimation methods (two-parameter MAP estimation and KONTSEVICH–TYLER BAYESIAN methods, both with prior information). The root-finding methods are a promising addition to the available threshold methods that motivate further studies and improvements.

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 were far from optimal due to their large number of required stimuli and wide distribution of errors. Their accuracy was outperformed by parametric estimators and stochastic root-finding methods with a fraction of the stimuli. As our MEP response model differed from the one underlying parametric estimation methods, some methods did not perform well for a subset of subjects in the population. Including all parameters in the estimation and incorporating prior information further boosted performance. Stochastic root-finding methods were comparable to parametric estimation methods, but were much simpler to implement and did not rely on an underlying model.

Code availability statement

The method used for the analysis and instructions are available at 10.5281/zenodo.6483601 to the community for free use and further development.

Acknowledgments

This work was supported by grants No. R01NS117405 and No. RF1MH124943 from the National Institutes of Health of the United States of America. 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.

Conflicts of interests

S. M. Goetz and A. V. Peterchev are inventors on patents and patent applications on TMS technology. S. M. Goetz has received research funding from Magstim as well as royalties from Rogue Research. A. V. Peterchev has equity options in 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, Soterix Medical, Magnetic Tides, and Ampa. 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:

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

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 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 MLE threshold and slope

The MLE of threshold t and slope s (Figure S1D) for the GAUSSIAN model (12) can be fitted by a joint distribution

$$\begin{pmatrix} t_{\mathrm{MLE}} \\ \lg(s_{\mathrm{MLE}}) \end{pmatrix} \sim \mathcal{N}\left(\mu = \begin{bmatrix} 0.65647 \\ -1.3028 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_t^2 & \rho\sigma_t\sigma_{\lg(s)} \\ \rho\sigma_t\sigma_{\lg(s)} & \sigma_{\lg(s)}^2 \end{bmatrix}\right),$$

where $\sigma_t = 9.8215 \cdot 10^{-2}$, $\sigma_{\lg(s)} = 3.9674 \cdot 10^{-2}$, and the correlation coefficient $\rho = 0.0921$. Given the very small ρ, the distribution was simplified and approximated by independent normal and log-normal distributions, which were fitted as

$$t_{\mathrm{MLE}} \sim \mathcal{N}(\mu = 0.65626, \sigma = 9.7944 \cdot 10^{-2}),$$
$$\lg(s_{\mathrm{MLE}}) \sim \mathcal{N}(\mu = -1.3027, \sigma = 3.9662 \cdot 10^{-2}).$$

This distribution served as $p(\boldsymbol{\theta})$ for the MAP methods and $p_0(\boldsymbol{\theta})$ for the BAYESIAN methods.

References

[1] Goetz S M and Deng Z-D 2017 The development and modelling of devices and paradigms for transcranial magnetic stimulation *Int. Rev. Psychiatry* **29** 115–45

[2] Koponen L M and Peterchev A V 2020 Transcranial Magnetic Stimulation: Principles and Applications *Neural Engineering* ed B He (Cham: Springer International Publishing) pp 245–70

[3] Zhong G, Yang Z and Jiang T 2021 Precise Modulation Strategies for Transcranial Magnetic Stimulation: Advances and Future Directions *Neurosci. Bull.* **37** 1718–1734

[4] Aberra A S, Wang B, Grill W M and Peterchev A V 2020 Simulation of transcranial magnetic stimulation in head model with morphologically-realistic cortical neurons *Brain Stimulat.* **13** 175–89

[5] Di Lazzaro V, Profice P, Ranieri F, Capone F, Dileone M, Oliviero A and Pilato F 2012 I-wave origin and modulation *Brain Stimulat.* **5** 512–25

[6] Kallioniemi E, Säisänen L, Könönen M, Awiszus F and Julkunen P 2014 On the estimation of silent period thresholds in transcranial magnetic stimulation *Clin. Neurophysiol.* **125** 2247–52

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

[8] Lang N, Harms J, Weyh T, Lemon R N, Paulus W, Rothwell J C and Siebner H R 2006 Stimulus intensity and coil characteristics influence the efficacy of rTMS to suppress cortical excitability *Clin. Neurophysiol.* **117** 2292–301

[9] Sasaki T, Shirota Y, Kodama S, Togashi N, Sugiyama Y, Tokushige S, Inomata-Terada S, Terao Y, Ugawa Y, Toda T and Hamada M 2018 Modulation of motor learning by a paired associative stimulation protocol inducing LTD-like effects *Brain Stimulat.* **11** 1314–21

[10] Sommer M, Alfaro A, Rummel M, Speck S, Lang N, Tings T and Paulus W 2006 Half sine, monophasic and biphasic transcranial magnetic stimulation of the human motor cortex *Clin. Neurophysiol.* **117** 838–44

[11] Turi Z, Lenz M, Paulus W, Mittner M and Vlachos A 2021 Selecting stimulation intensity in repetitive transcranial magnetic stimulation studies: A systematic review between 1991 and 2020 *Eur. J. Neurosci.* **53** 3404–15

[12] Rossi S, Hallett M, Rossini P M and Pascual-Leone A 2009 Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research *Clin. Neurophysiol.* **120** 2008–39

[13] Rossi S, Antal A, Bestmann S, Bikson M, Brewer C, Brockmöller J, Carpenter L L, Cincotta M, Chen R, Daskalakis J D, Di Lazzaro V, Fox M D, George M S, Gilbert D, Kimiskidis V K, Koch G, Ilmoniemi R J, Pascal Lefaucheur J, Leocani L, Lisanby S H, Miniussi C, Padberg F, Pascual-Leone A, Paulus W, Peterchev A V, Quartarone A, Rotenberg A, Rothwell J, Rossini P M, Santarnecchi E, Shafi M M, Siebner H R, Ugawa Y, Wassermann E M, Zangen A, Ziemann U and Hallett M 2020 Safety and recommendations for TMS use in healthy subjects and patient populations, with updates on training, ethical and regulatory issues: Expert Guidelines *Clin. Neurophysiol.* **132** 269–306

[14] Wassermann E M 1998 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.* **108** 1–16

[15] Antal A, Nitsche M A, Kincses T Z, Lampe C and Paulus W 2004 No correlation between oving phosphene and motor thresholds: a transcranial magnetic stimulation study *NeuroReport* **15** 297–302

[16] Deblieck C, Thompson B, Iacoboni M and Wu A D 2008 Correlation between motor and phosphene thresholds: A transcranial magnetic stimulation study *Hum. Brain Mapp.* **29** 662–70

[17] Goetz S M and Kammer T 2022 Neuronavigation *The Oxford Handbook of TMS* ed E M Wassermann, V Walsh, A V Peterchev, U Ziemann, S Lisanby and H Siebner (Oxford: Oxford University Press)

[18] Noda Y, Barr M S, Zomorrodi R, Cash R F H, Lioumis P, Chen R, Daskalakis Z J and Blumberger D M 2021 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.* **11** 54

[19] Poorganji M, Zomorrodi R, Hawco C, Hill A T, Hadas I, Rajji T K, Chen R, Voineskos D, Daskalakis A A, Blumberger D M and Daskalakis Z J 2021 Differentiating transcranial magnetic stimulation cortical and auditory responses via single pulse and paired pulse protocols: A TMS-EEG study *Clin. Neurophysiol.* **132** 1850–8

[20] Stewart L M, Walsh V and Rothwell J C 2001 Motor and phosphene thresholds: a transcranial magnetic stimulation correlation study *Neuropsychologia* **39** 415–9

[21] Tremblay S, Rogasch N C, Premoli I, Blumberger D M, Casarotto S, Chen R, Di Lazzaro V, Farzan F, Ferrarelli F, Fitzgerald P B, Hui J, Ilmoniemi R J, Kimiskidis V K, Kugiumtzis D, Lioumis P, Pascual-Leone A, Pellicciari M C, Rajji T, Thut G, Zomorrodi R, Ziemann U and Daskalakis Z J 2019 Clinical utility and prospective of TMS–EEG *Clin. Neurophysiol.* **130** 802–44

[22] Wassermann E M 2002 Variation in the response to transcranial magnetic brain stimulation in the general population *Clin. Neurophysiol.* **113** 1165–71

[23] Balslev D, Braet W, McAllister C and Miall R C 2007 Inter-individual variability in optimal current direction for transcranial magnetic stimulation of the motor cortex *J. Neurosci. Methods* **162** 309–13

[24] Pridmore S, Filho J A F, Nahas Z, Liberatos C and George M S 1998 Motor threshold in transcranial magnetic stimulation: A comparison of a neurophysiological method and a visualization of movement method *J. ECT* **14** 25–7

[25] Westin G G, Bassi B D, Lisanby S H and Luber B 2014 Determination of motor threshold using visual observation overestimates transcranial magnetic stimulation dosage: Safety implications *Clin. Neurophysiol.* **125** 142–7

[26] Cotovio G, Oliveira-Maia A J, Paul C, Faro Viana F, Rodrigues da Silva D, Seybert C, Stern A P, Pascual-Leone A and Press D Z 2021 Day-to-day variability in motor threshold during rTMS treatment for depression: Clinical implications *Brain Stimulat.* **14** 1118–25

[27] Rossini P M, Barker A T, Berardelli A, Caramia M D, Caruso G, Cracco R Q, Dimitrijević M R, Hallett M, Katayama Y, Lücking C H, Maertens de Noordhout A L, Marsden C D, Murray N M F, Rothwell J C, Swash

M and Tomberg C 1994 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.* **91** 79–92

[28] Rossini P M, Burke D, Chen R, Cohen L G, Daskalakis Z, Di Iorio R, Di Lazzaro V, Ferreri F, Fitzgerald P B, George M S, Hallett M, Lefaucheur J P, Langguth B, Matsumoto H, Miniussi C, Nitsche M A, Pascual-Leone A, Paulus W, Rossi S, Rothwell J C, Siebner H R, Ugawa Y, Walsh V and Ziemann U 2015 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.* **126** 1071–107

[29] Kobayashi M and Pascual-Leone A 2003 Transcranial magnetic stimulation in neurology *Lancet Neurol.* **2** 145–56

[30] Goetz S M and Peterchev A V 2012 A model of variability in brain stimulation evoked responses *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) pp 6434–7

[31] Jung N H, Delvendahl I, Kuhnke N G, Hauschke D, Stolle S and Mall V 2010 Navigated transcranial magnetic stimulation does not decrease the variability of motor-evoked potentials *Brain Stimulat.* **3** 87–94

[32] Kiers L, Cros D, Chiappa K H and Fang J 1993 Variability of motor potentials evoked by transcranial magnetic stimulation *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.* **89** 415–23

[33] Capaday C 2021 On the variability of motor-evoked potentials: experimental results and mathematical model *Exp. Brain Res.* **239** 2979–95

[34] Goetz S M, Luber B, Lisanby S H and Peterchev A V 2014 A Novel Model Incorporating Two Variability Sources for Describing Motor Evoked Potentials *Brain Stimulat.* **7** 541–52

[35] Goetz S M, Howell B, Wang B, Li Z, Sommer M A, Peterchev A V and Grill W M 2022 Isolating two sources of variability of subcortical stimulation to quantify fluctuations of corticospinal tract excitability *Clin. Neurophysiol.* **138** 134–42

[36] Groppa S, Oliviero A, Eisen A, Quartarone A, Cohen L G, Mall V, Kaelin-Lang A, Mima T, Rossi S, Thickbroom G W, Rossini P M, Ziemann U, Valls-Solé J and Siebner H R 2012 A practical guide to diagnostic transcranial magnetic stimulation: Report of an IFCN committee *Clin. Neurophysiol.* **123** 858–82

[37] Badran B W, Ly M, DeVries W H, Glusman C E, Willis A, Pridmore S and George M S 2019 Are EMG and visual observation comparable in determining resting motor threshold? A reexamination after twenty years *Brain Stimulat.* **12** 364–6

[38] Anderson B S and George M S 2009 A review of studies comparing methods for determining transcranial magnetic stimulation motor threshold: observation of movement or electromyography assisted *J. Am. Psychiatr. Nurses Assoc.* **15** 304–313

[39] Mills K R and Nithi K A 1997 Corticomotor threshold to magnetic stimulation: Normal values and repeatability *Muscle Nerve* **20** 570–6

[40] Tranulis C, Guéguen B, Pham-Scottez A, Vacheron M N, Cabelguen G, Costantini A, Valero G and Galinovski A 2006 Motor threshold in transcranial magnetic stimulation: comparison of three estimation methods *Neurophysiol. Clin. Neurophysiol.* **36** 1–7

[41] Awiszus F 2012 On relative frequency estimation of transcranial magnetic stimulation motor threshold *Clin. Neurophysiol.* **123** 2319–20

[42] Awiszus F 2011 Fast estimation of transcranial magnetic stimulation motor threshold: is it safe? *Brain Stimulat.* **4** 58–9

[43] Awiszus F 2014 Using relative frequency estimation of transcranial magnetic stimulation motor threshold does not allow to draw any conclusions about true threshold *Clin. Neurophysiol.* **125** 1285–6

[44] Mishory A, Molnar C, Koola J, Li X, Kozel F A, Myrick H, Stroud Z, Nahas Z and George M S 2004 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* **20** 160–165

[45] Möller C, Arai N, Lücke J and Ziemann U 2009 Hysteresis effects on the input–output curve of motor evoked potentials *Clin. Neurophysiol.* **120** 1003–8

[46] Nielsen J 1996 Logarithmic Distribution of Amplitudes of Compound Muscle Action Potentials Evoked by Transcranial Magnetic Stimulation *J. Clin. Neurophysiol.* **13** 423–34

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

[48] Awiszus F 2003 Chapter 2 TMS and threshold hunting *Supplements to Clinical Neurophysiology* Transcranial Magnetic Stimulation and Transcranial Direct Current Stimulation vol 56, ed W Paulus, F Tergau, M A Nitsche, J G Rothwell, U Ziemann and M Hallett (Elsevier) pp 13–23

[49] Pentland A 1980 Maximum likelihood estimation: the best PEST. *Percept. Psychophys.*

[50] Treutwein B 1995 Adaptive psychophysical procedures *Vision Res.* **35** 2503–2522

[51] Strasburger H and Treutwein B 1999 Fitting the psychometric function *Percept. Psychophys.* **61** 87–106

[52] Goetz S, Whiting P and Peterchev A 2011 Threshold estimation with transcranial magnetic stimulation: algorithm comparison *Clin. Neurophysiol.* **122** S197

[53] Koponen L M and Peterchev A V 2022 Preventing misestimation of transcranial magnetic stimulation motor threshold with MTAT 2.0 *Brain Stimul. Basic Transl. Clin. Res. Neuromodulation* **15** 1073–6

[54] Qi F, Wu A D and Schweighofer N 2011 Fast estimation of transcranial magnetic stimulation motor threshold *Brain Stimulat.* **4** 50–7

[55] Fisher R A 1922 On the mathematical foundations of theoretical statistics *Philos. Trans. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character* **222** 309–368

[56] Proakis J G and Salehi M 2008 *Digital communications* (McGraw-Hill.)

[57] Kontsevich L L and Tyler C W 1999 Bayesian adaptive estimation of psychometric slope and threshold *Vision Res.* **39** 2729–37

[58] Paninski L 2005 Asymptotic Theory of Information-Theoretic Experimental Design *Neural Comput.* **17** 1480–507

[59] Wichmann F A and Hill N J 2001 The psychometric function: I. Fitting, sampling, and goodness of fit *Percept. Psychophys.* **63** 1293–313

[60] Bawa P and Lemon R N 1993 Recruitment of motor units in response to transcranial magnetic stimulation in man. *J. Physiol.* **471** 445–64

[61] Goetz S M, Alavi S M M, Deng Z and Peterchev A V 2019 Statistical Model of Motor-Evoked Potentials *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 1539–45

[62] Mathias J P, Barsi G I, van de Ruit M and Grey M J 2014 Rapid Acquisition of the Transcranial Magnetic Stimulation Stimulus Response Curve *Brain Stimulat.* **7** 59–65

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

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

[65] Lai T L 2003 Stochastic approximation *Ann. Stat.* **31** 391–406

[66] Robbins H and Monro S 1951 A Stochastic Approximation Method *Ann. Math. Stat.* **22** 400–7

[67] Delyon B and Juditsky A 1993 Accelerated stochastic approximation *SIAM J. Optim.* **3** 868–881

[68] Nemirovski A, Juditsky A, Lan G and Shapiro A 2009 Robust stochastic approximation approach to stochastic programming *SIAM J. Optim.* **19** 1574–1609

[69] Julkunen P 2019 Mobile Application for Adaptive Threshold Hunting in Transcranial Magnetic Stimulation *IEEE Trans. Neural Syst. Rehabil. Eng.* **27** 1504–10

[70] Borckardt J J, Nahas Z, Koola J and George M S 2006 Estimating Resting Motor Thresholds in Transcranial Magnetic Stimulation Research and Practice: A computer Simulation Evaluation of Best Methods *J. ECT* **22** 169–175

[71] Peterchev A V, Goetz S M, Westin G G, Luber B and Lisanby S H 2013 Pulse width dependence of motor threshold and input–output curve characterized with controllable pulse parameter transcranial magnetic stimulation *Clin. Neurophysiol.* **124** 1364–72

[72] Peterchev A V, Jalinous R and Lisanby S H 2008 A Transcranial Magnetic Stimulator Inducing Near-Rectangular Pulses With Controllable Pulse Width (cTMS) *IEEE Trans. Biomed. Eng.* **55** 257–66

[73] Goetz S M 2012 *Magnetic Neurostimulation from a Physical Perspective* Dissertation (Munich, Germany: Technical University of Munich)

[74] Goetz S M, Luber B, Lisanby S H, Murphy D L K, Kozyrkov I C, Grill W M and Peterchev A V 2016 Enhancement of Neuromodulation with Novel Pulse Shapes Generated by Controllable Pulse Parameter Transcranial Magnetic Stimulation *Brain Stimulat.* **9** 39–47

[75] Awiszus F 2014 Of thresholds and "hot spots". Quo vadis transcranial magnetic stimulation? *Clin. Neurophysiol.* **125** 2451–3

[76] Borckardt J J and Awiszus F 2011 *TMS motor threshold assessment tool (MTAT)* (https://www.clinicalresearcher.org/software.htm)

[77] Zacharias L R, Peres A S C, Souza V H, Conforto A B and Baffa O 2019 Method to assess the mismatch between the measured and nominal parameters of transcranial magnetic stimulation devices *J. Neurosci. Methods* **322** 83–7

[78] Darling W G, Wolf S L and Butler A J 2006 Variability of motor potentials evoked by transcranial magnetic stimulation depends on muscle activation *Exp. Brain Res.* **174** 376–85

[79] Orth M and Rothwell J C 2004 The cortical silent period: intrinsic variability and relation to the waveform of the transcranial magnetic stimulation pulse *Clin. Neurophysiol.* **115** 1076–82

[80] Roy Choudhury K, Boyle L, Burke M, Lombard W, Ryan S and McNamara B 2011 Intra subject variation and correlation of motor potentials evoked by transcranial magnetic stimulation *Ir. J. Med. Sci.* **180** 873–80

Figures

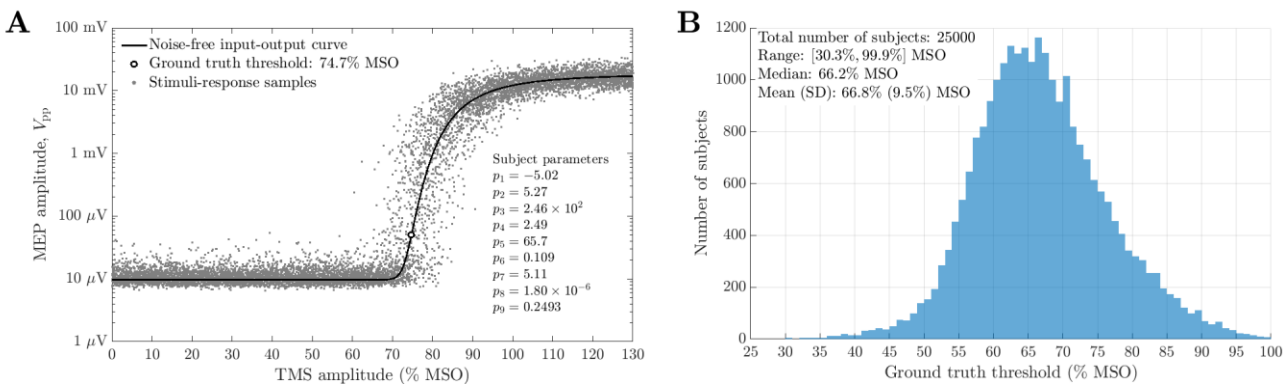

Figure 1: TMS response model.

A. An example of the TMS recruiting behavior that is mimicked by the model. This virtual stimulator provides analog values, which some of the algorithms use for a faster convergence. The used equations are independent from the tested black-box models in the parameter estimators in order to avoid model-determined advantages. Subsequent responses are not correlated. **B**. The ground truth thresholds obtained from the noiseless model of the 25,000 virtual subjects shown in a histogram with bin width of 1% MSO.

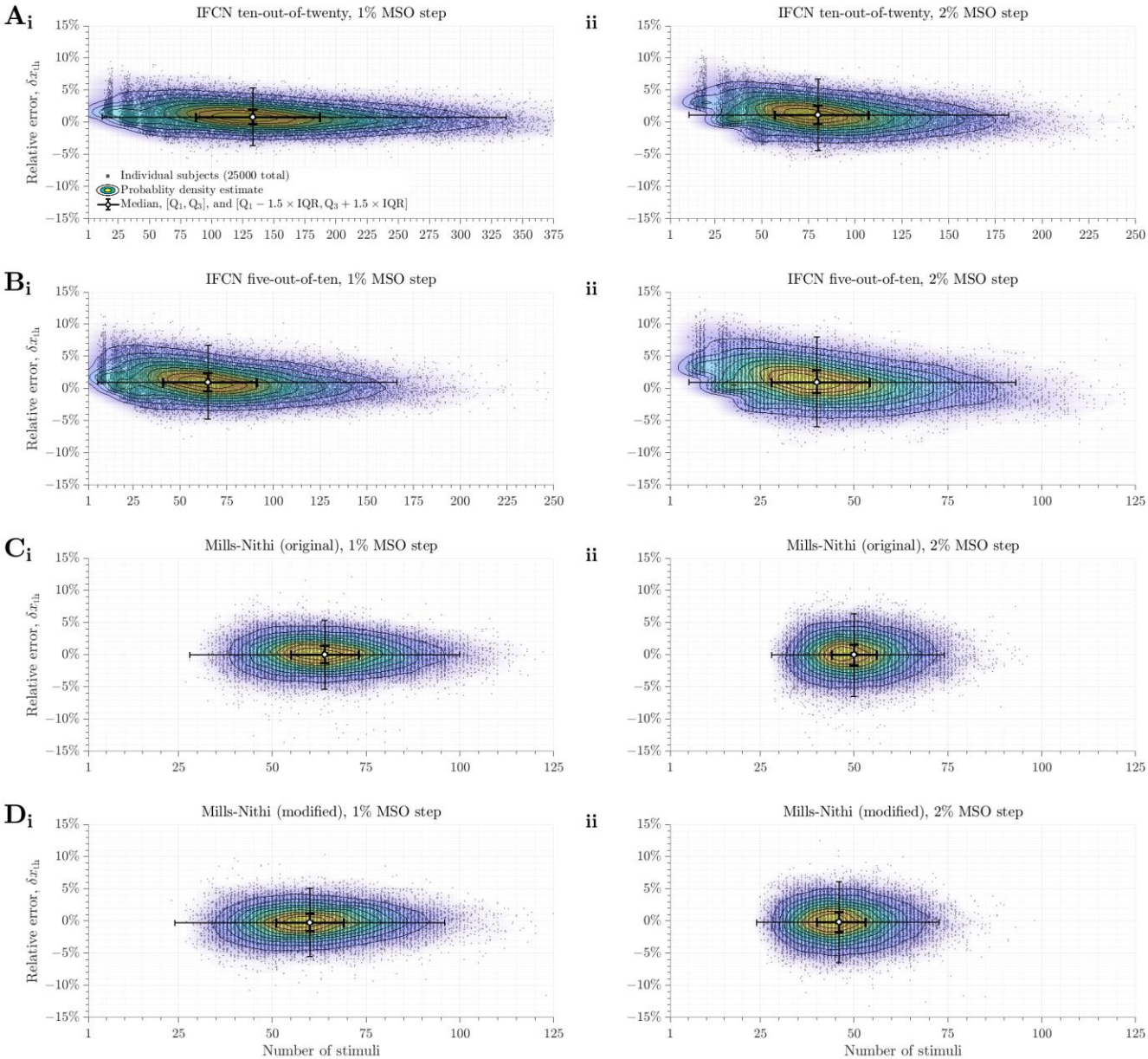

Figure 2: Relative threshold errors of the IFNC relative frequency and MILLS–NITHI methods

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. **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 panels A_i, A_{ii}, and B_i versus the others.

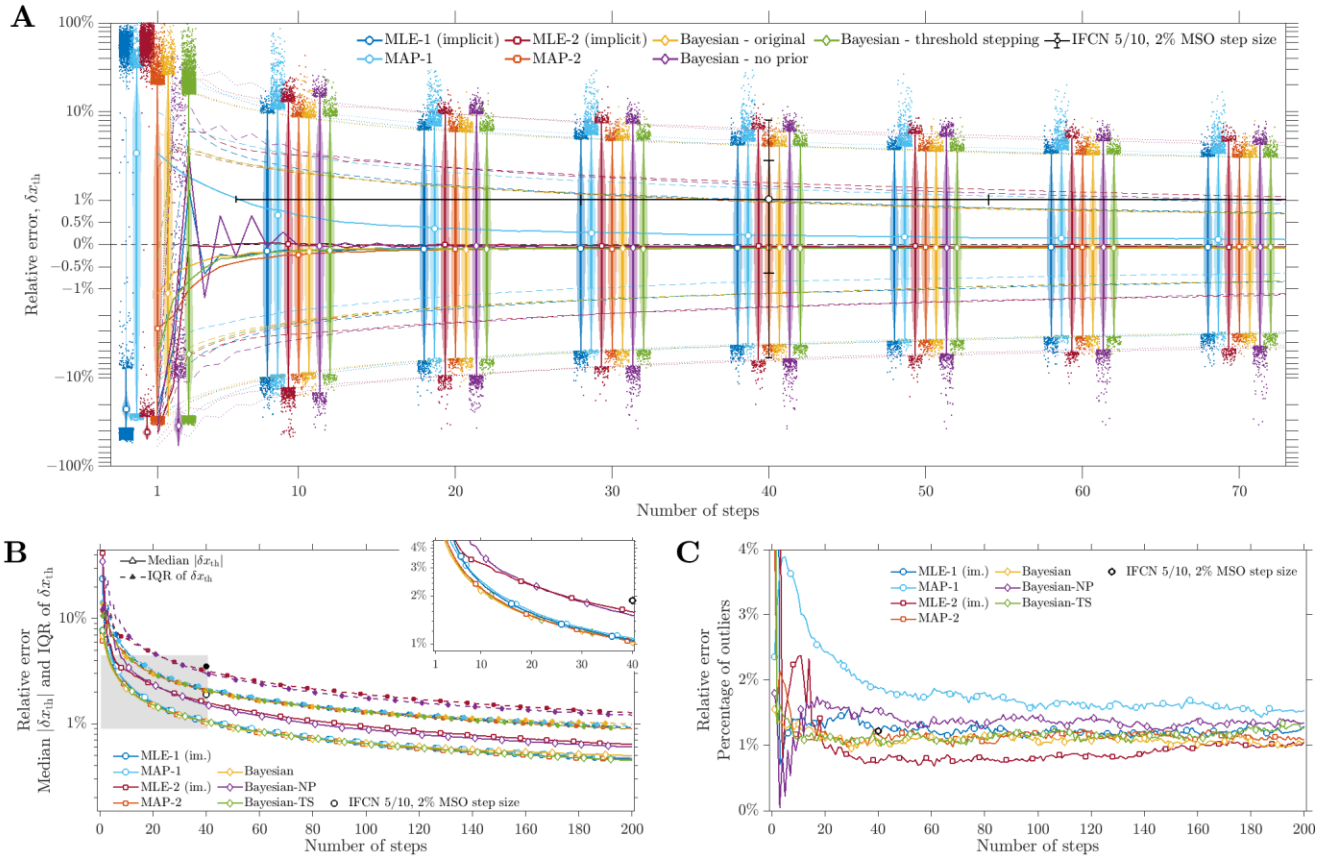

Figure 3: Relative threshold errors of the parametric estimation methods

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, and as lines for every step with median as thick lines, first and third quartiles as dashed lines, and whiskers as dotted line. 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 and interquartile range 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$_{ii}$ for reference. **B.** The median $|\delta x_{\text{th}}|$ and the interquartile range (IQR) of δx_{th} (length of whiskers in panel A) on logarithmic scale. The inset shows the details of the median $|\delta x_{\text{th}}|$ in the first 20 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).

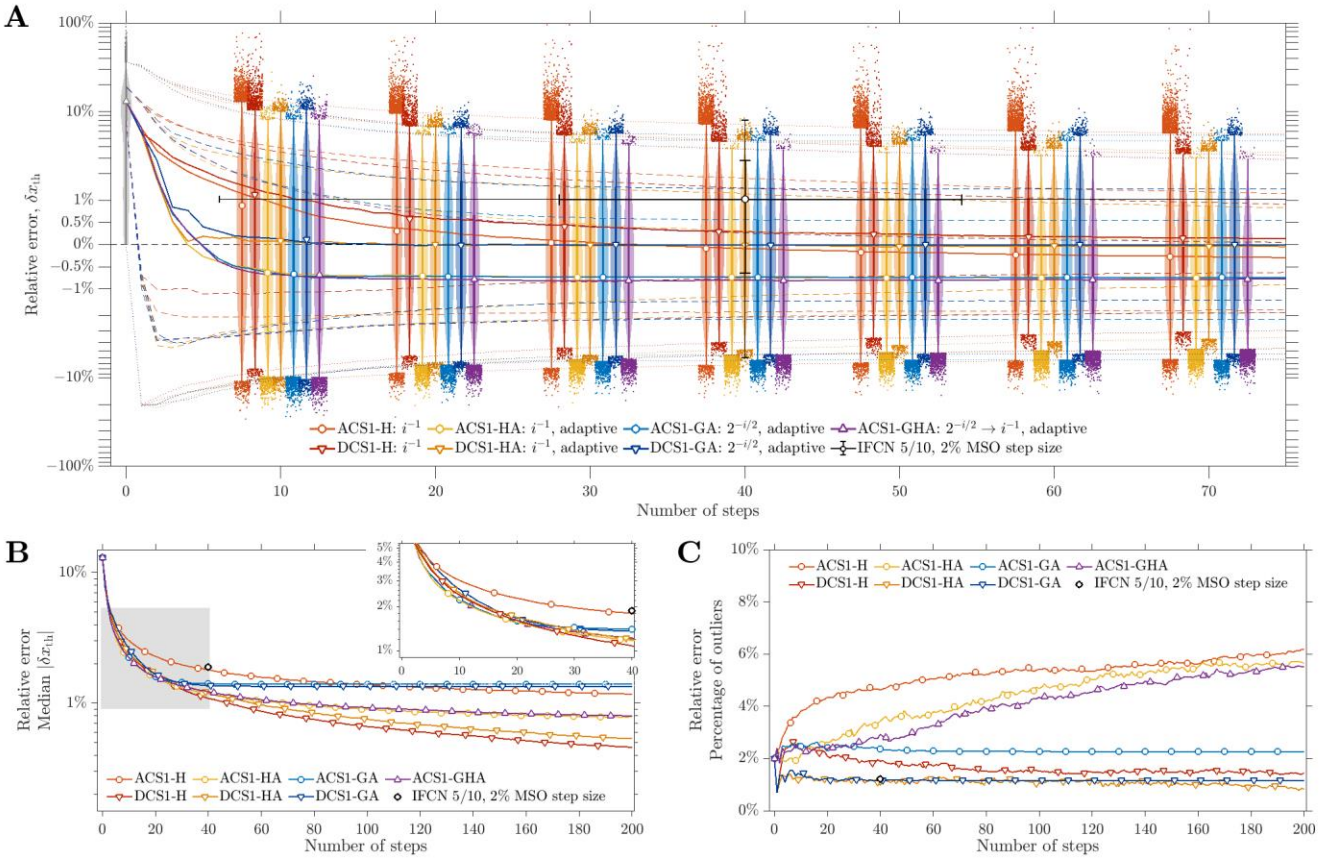

Figure 4: Relative threshold errors of root-finding methods using first-order control sequences with default initial step size a_0

Similar to Figure 3, without the interquartile range of δx_{th} in panel B.

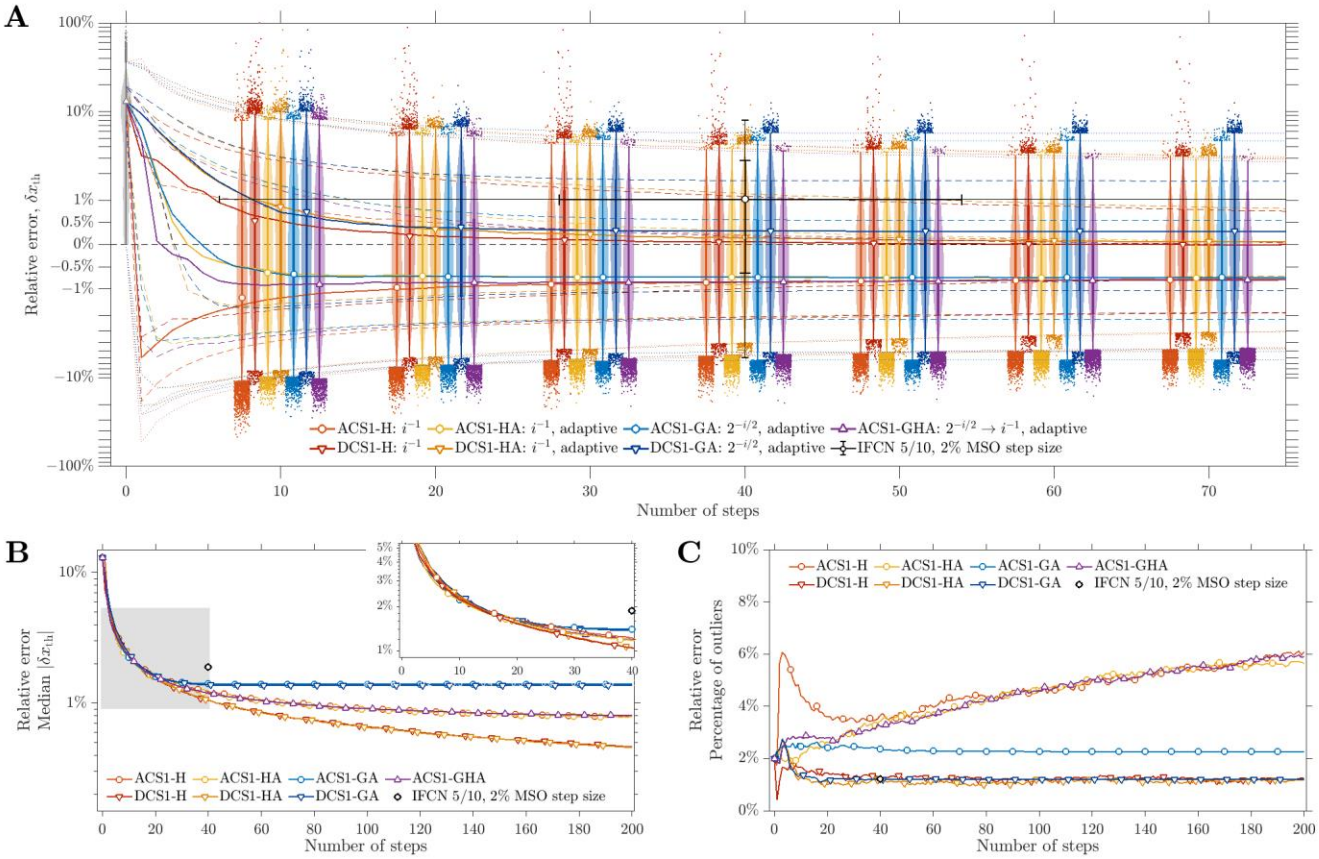

Figure 5: Relative threshold errors of root-finding methods using first-order control sequences with optimal initial step size a_0

Similar to Figure 4.

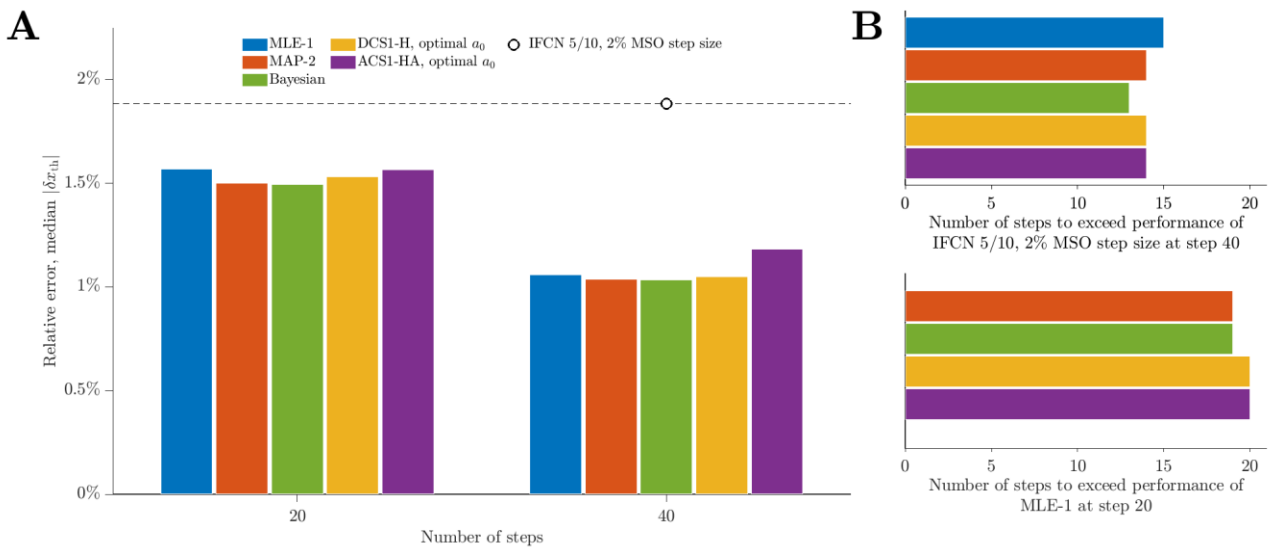

Figure 6: Best performing methods

Selected statistics of the best performing methods, with full statistics shown in Figure S17. The fastest relative-frequency method is the IFCN five-out-of-ten method with 2% MSO step size, whereas all relative-frequency methods have similar errors. 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 are ACS1-HA and DCS1-H for analog and digital control sequence, respectively, with results for the optimal initial a_0 given. **A.** Median $|\delta x_{\mathrm{th}}|$ at steps 20 and 40. The IFCN five-out-of-ten method with 2% MSO step size and the single-parameter MLE (MLE-1) are shown for reference. **B.** The number of steps needed for a given method to outperform the reference IFCN five-out-of-ten method at 40 steps (top) and the single-parameter MLE at 20 steps (bottom).

Supplement

Analysis and Comparison of Methods for Determining Motor Threshold with Transcranial Magnetic Stimulation

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

[1] Department of Psychiatry and Behavior 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, Duke University, 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:

- Table S1
- Figures S1–S17

Supplementary tables

Table S1 Statistics of number of steps, absolute error, and relative error of IFCN and MILLS–NITHI methods (related to Figures 2, S3, and S4)

Group number	Method	Step size (% MSO)	Number of steps Mean±S.D.	Number of steps Median	Relative error δx_{th} Mean±S.D.	Relative error δx_{th} Median	Absolute error Δx_{th}, (% MSO) Mean±S.D.	Absolute error Δx_{th}, (% MSO) Median
1	IFCN, ten-out-of-twenty	1	141.7±75.4	133	0.94%±1.72%	0.86%	0.61±1.12	0.57
2	IFCN, ten-out-of-twenty	2	84.3±38.1	80	1.23%±2.11%	1.14%	0.81±1.36	0.76
3	IFCN, five-out-of-ten	1	69.0±37.7	65	1.10%±2.21%	0.99%	0.73±1.44	0.66
4	IFCN, five-out-of-ten	2	41.9±19.4	40	1.13%±2.66%	1.03%	0.74±1.73	0.69
5	MILLS–NITHI, original	1	64.7±13.3	64	0.01%±2.46%	0.08%	0.00±1.66	0.06
6	MILLS–NITHI, original	2	50.6± 9.4	50	−0.07%±2.80%	0.00%	−0.05±1.87	0.00
7	MILLS–NITHI, modified	1	61.0±13.3	60	−0.21%±2.04%	−0.19%	−0.14±1.33	−0.12
8	MILLS–NITHI, modified	2	47.0± 9.4	46	−0.15%±2.38%	−0.13%	−0.10±1.55	−0.08

ANOVA, number of steps

Source	SS	df	MS	F	Prob>F
Groups	1.795×10^{8}	7	2.550×10^{7}	2.158×10^{4}	0
Error	2.364×10^{8}	199992	1.182×10^{3}		
Total	4.149×10^{8}	199999			

ANOVA, relative error

Source	SS	df	MS	F	Prob>F
Groups	7.652×10^{4}	7	1.093×10^{4}	2.101×10^{3}	0
Error	1.040×10^{6}	199992	5.202		
Total	1.117×10^{6}	199999	0		

ANOVA, absolute error

Source	SS	df	MS	F	Prob>F
Groups	3.219×10^{4}	7	4.599×10^{3}	1.978×10^{3}	0
Error	4.650×10^{5}	199992	2.325		
Total	4.972×10^{5}	199999	0		

Post-hoc multiple comparison test for number of steps

Group A	Group B	Mean (Group A – Group B)	95% CI (Group A – Group B)		p-value
1	2	57.36	56.43	58.29	0
1	3	72.58	71.65	73.51	0
1	4	99.71	98.78	100.64	0
1	5	76.95	76.01	77.88	0
1	6	91.07	90.13	92.00	0
1	7	80.63	79.70	81.56	0
1	8	94.65	93.72	95.58	0
2	3	15.22	14.29	16.16	0
2	4	42.35	41.42	43.28	0
2	5	19.59	18.65	20.52	0
2	6	33.71	32.78	34.64	0
2	7	23.27	22.34	24.20	0
2	8	37.29	36.36	38.22	0
3	4	27.13	26.19	28.06	0
3	5	4.36	3.43	5.30	0
3	6	18.48	17.55	19.42	0
3	7	8.05	7.12	8.98	0
3	8	22.07	21.13	23.00	0
4	5	−22.76	−23.70	−21.83	0
4	6	−8.64	−9.57	−7.71	0
4	7	−19.08	−20.01	−18.15	0
4	8	−5.06	−5.99	-4.13	0
5	6	14.12	13.19	15.05	0
5	7	3.69	2.75	4.62	0
5	8	17.70	16.77	18.64	0
6	7	−10.44	−11.37	−9.50	0
6	8	3.58	2.65	4.51	0
7	8	14.02	13.09	14.95	0

Post-hoc multiple comparison test for relative error

Group A	Group B	Mean (Group A – Group B)	95% CI (Group A – Group B)		p-value
1	2	−0.30	−0.36	−0.23	0
1	3	−0.16	−0.23	−0.10	0
1	4	−0.20	−0.26	−0.13	0
1	5	0.93	0.87	0.99	0
1	6	1.00	0.94	1.06	0
1	7	1.15	1.09	1.21	0
1	8	1.15	1.09	1.21	0
2	3	0.13	0.07	0.19	2×10^{-9}
2	4	0.10	0.04	0.16	2×10^{-5}
2	5	1.23	1.16	1.29	0
2	6	1.30	1.24	1.36	0
2	7	1.44	1.38	1.51	0
2	8	1.44	1.38	1.51	0
3	4	−0.03	−0.09	0.03	**0.78**
3	5	1.09	1.03	1.16	0
3	6	1.17	1.10	1.23	0
3	7	1.31	1.25	1.37	0
3	8	1.31	1.25	1.37	0
4	5	1.13	1.06	1.19	0
4	6	1.20	1.14	1.26	0
4	7	1.34	1.28	1.40	0
4	8	1.34	1.28	1.40	0
5	6	0.07	0.01	0.13	8×10^{-3}
5	7	0.22	0.16	0.28	0
5	8	0.22	0.16	0.28	0
6	7	0.14	0.08	0.21	3×10^{-11}
6	8	0.14	0.08	0.21	3×10^{-11}
7	8	0.00	−0.06	0.06	**1**

Post-hoc multiple comparison test for absolute error

Group A	Group B	Mean (Group A – Group B)	95% CI (Group A – Group B)		p-value
1	2	−0.19	−0.23	−0.15	0
1	3	−0.11	−0.15	−0.07	0
1	4	−0.13	−0.17	−0.09	0
1	5	0.61	0.57	0.65	0
1	6	0.66	0.62	0.70	0
1	7	0.75	0.71	0.79	0
1	8	0.71	0.67	0.75	0
2	3	0.08	0.04	0.12	10^{-7}
2	4	0.06	0.02	0.10	2×10^{-4}
2	5	0.81	0.76	0.85	0
2	6	0.85	0.81	0.90	0
2	7	0.94	0.90	0.99	0
2	8	0.90	0.86	0.94	0
3	4	−0.02	−0.06	0.02	**0.88**
3	5	0.73	0.68	0.77	0
3	6	0.77	0.73	0.82	0
3	7	0.86	0.82	0.91	0
3	8	0.82	0.78	0.86	0
4	5	0.74	0.70	0.79	0
4	6	0.79	0.75	0.83	0
4	7	0.88	0.84	0.92	0
4	8	0.84	0.80	0.88	0
5	6	0.05	0.01	0.09	8×10^{-3}
5	7	0.14	0.10	0.18	0
5	8	0.10	0.05	0.14	4×10^{-11}
6	7	0.09	0.05	0.13	10^{-9}
6	8	0.05	0.01	0.09	0.01
7	8	−0.04	−0.08	−0.00	**0.04**

Supplementary figures

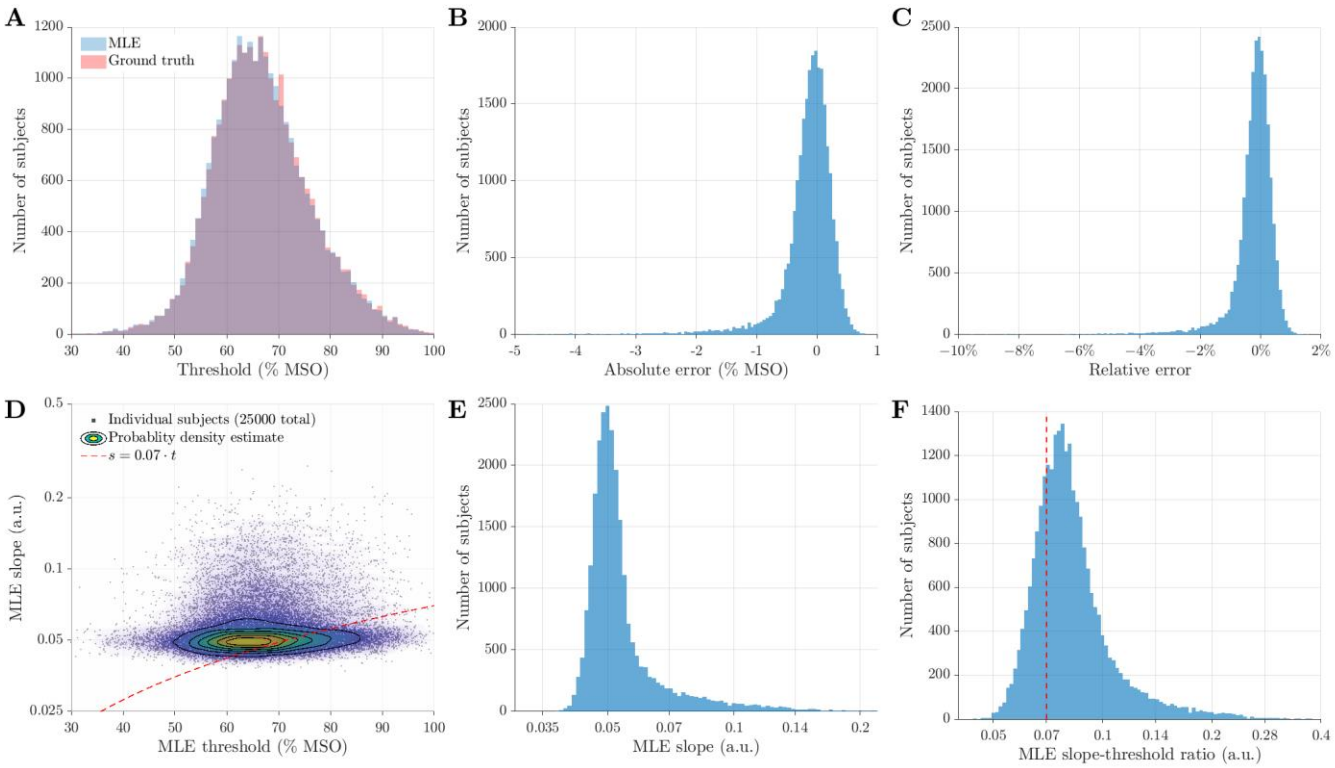

Figure S1: Maximum likelihood estimation (MLE) thresholds and slopes of model population

A. Distribution of MLE thresholds t (blue), with ground truth thresholds in red (as in Figure 1B) for comparison. **B, C.** Distributions of absolute and relative errors of MLE thresholds compared to ground truth thresholds, respectively. **D.** Joint distribution of MLE thresholds and slopes. **E.** Distribution of MLE slopes s. **F.** Distribution of ratios between MLE slopes and thresholds. The slopes and slope–threshold ratios are shown on logarithmic scale in panels D–F.

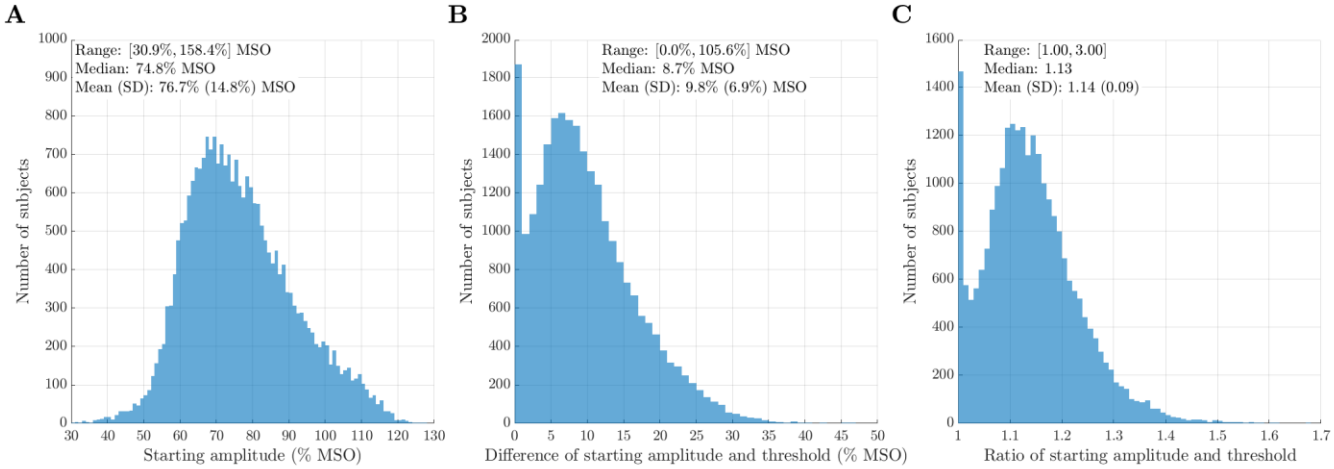

Figure S2: Starting amplitudes

A. Distribution of the starting amplitudes. Less than 0.15% of the subjects had starting amplitudes larger than 120% MSO, with three outliers above 130% MSO. **B.** Difference of starting amplitudes and ground truth threshold. Less than 0.1% of the subjects had starting amplitudes more than 40% MSO higher their thresholds. **C.** Ratio of starting amplitudes and ground truth threshold. Less than 0.1% of the subjects had starting amplitudes more than 1.6 times the thresholds.

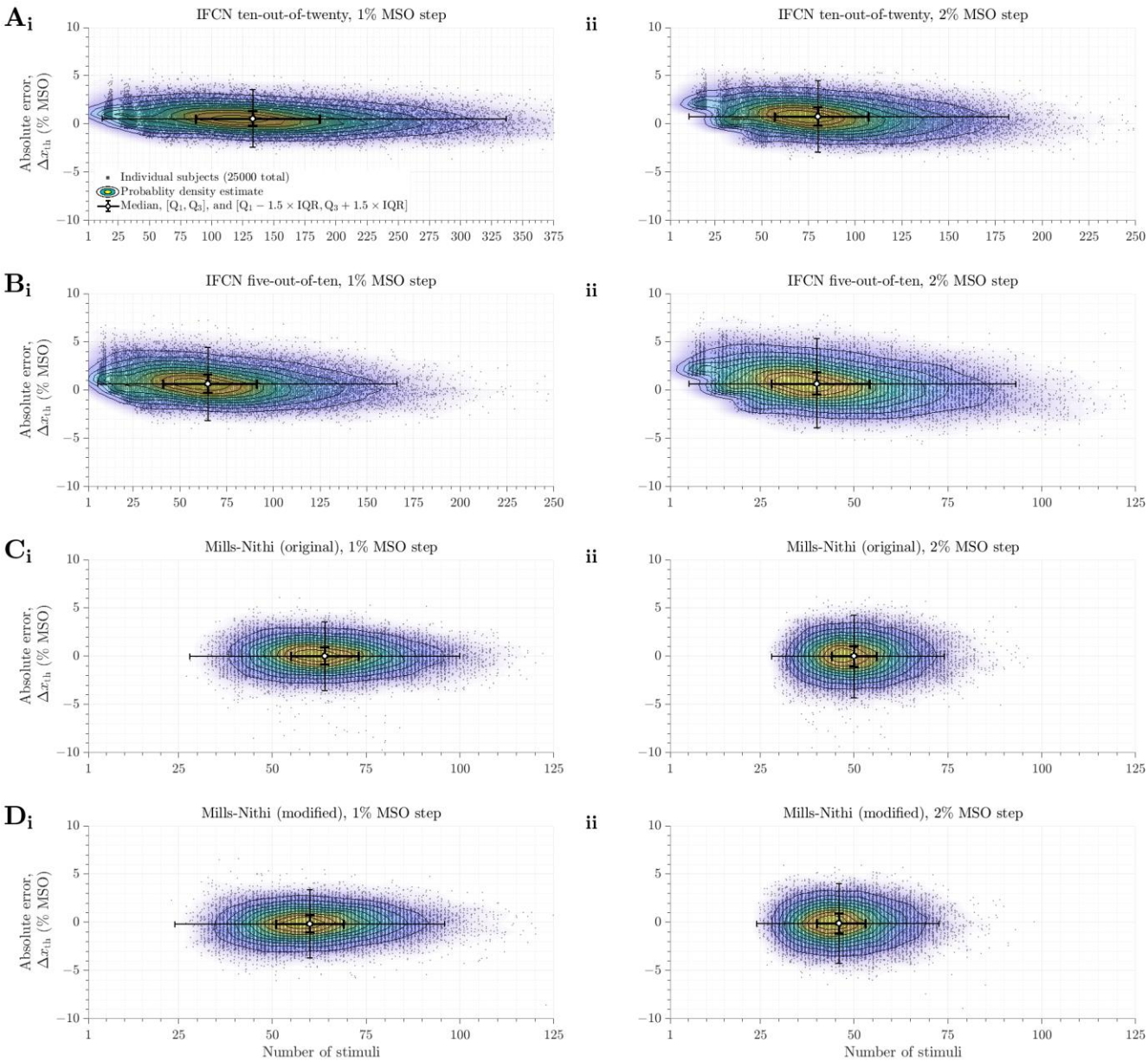

Figure S3: Absolute threshold errors of the IFNC relative frequency and MILLS–NITHI methods
Shown similarly as the relative threshold errors in Figure 2.

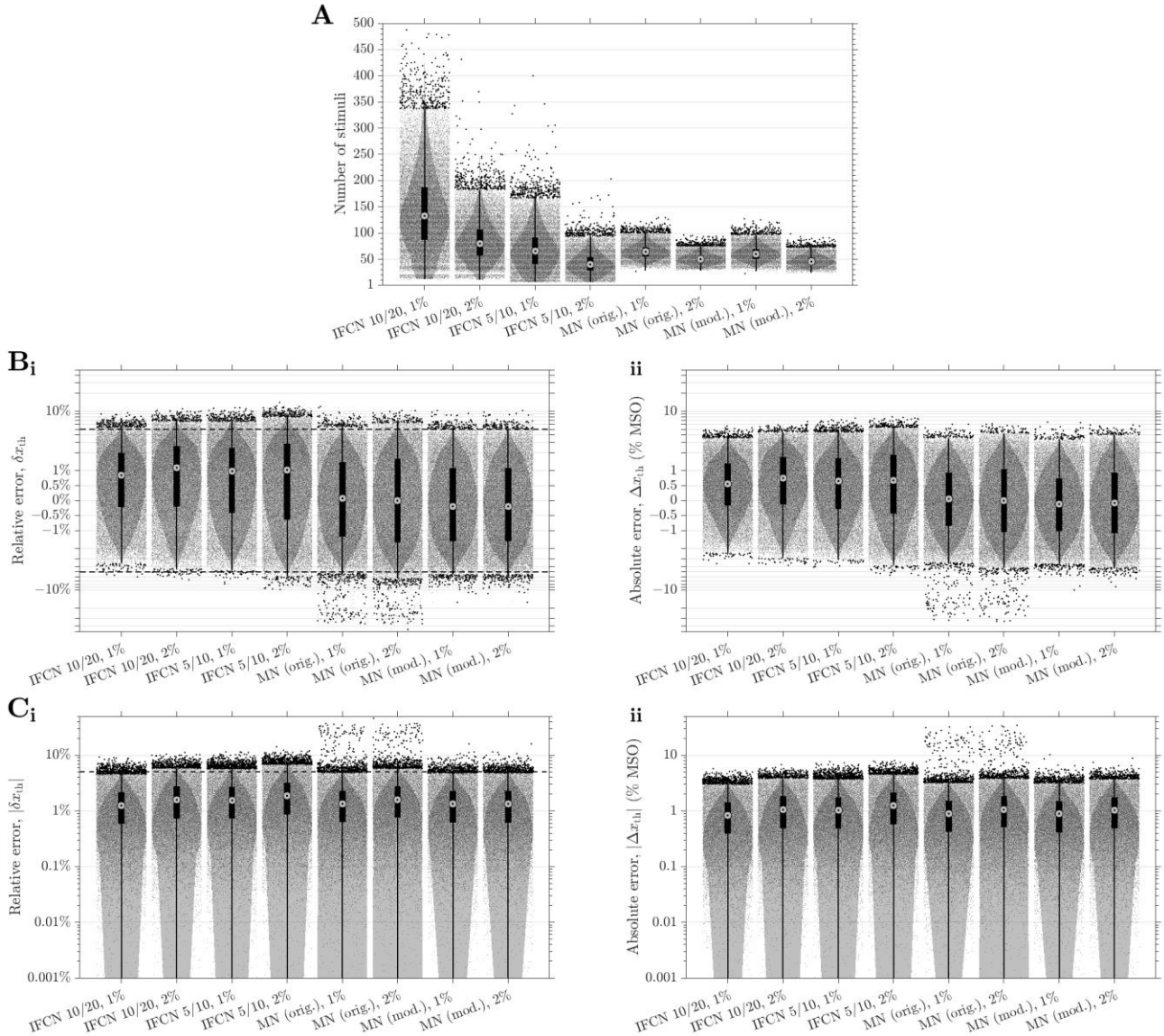

Figure S4: Statistics of IFNC relative frequency and MILLS–NITHI methods

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 dashed line for the relative errors in panels B_i and C_i.

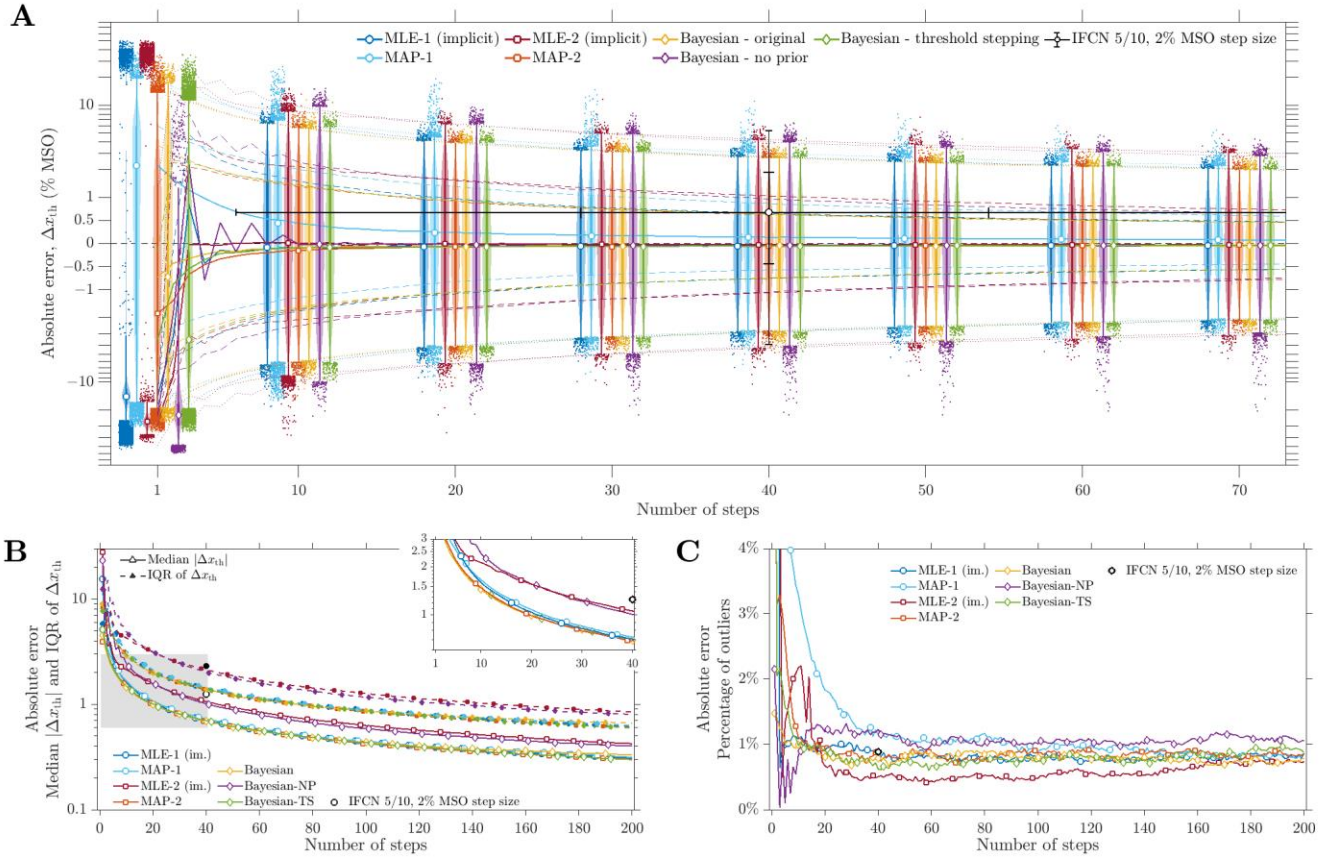

Figure S5: Absolute threshold errors of parametric estimation methods

Shown similarly as the relative threshold error in Figure 3.

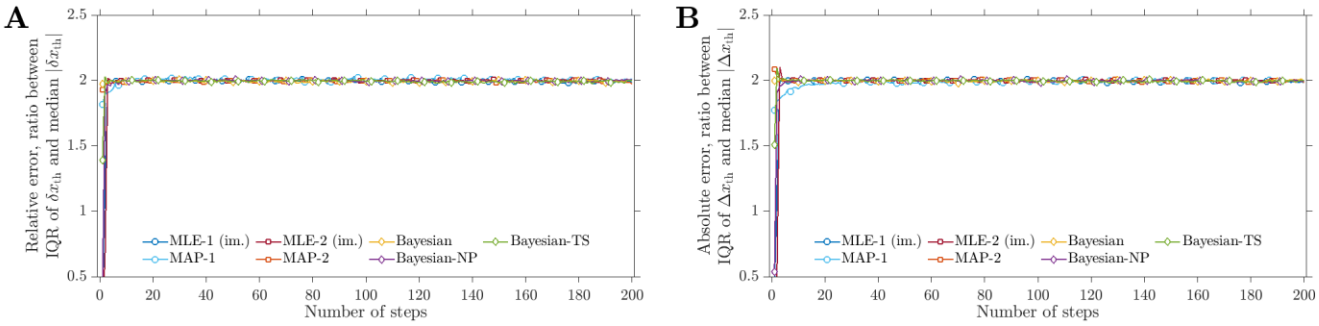

Figure S6: Ratio of median of absolute value of errors and interquartile ranges (IQR) of errors of parametric estimation methods

A. Ratio for relative threshold errors (related to Figures 3). **B.** Ratio for absolute threshold errors (related to Figure S5).

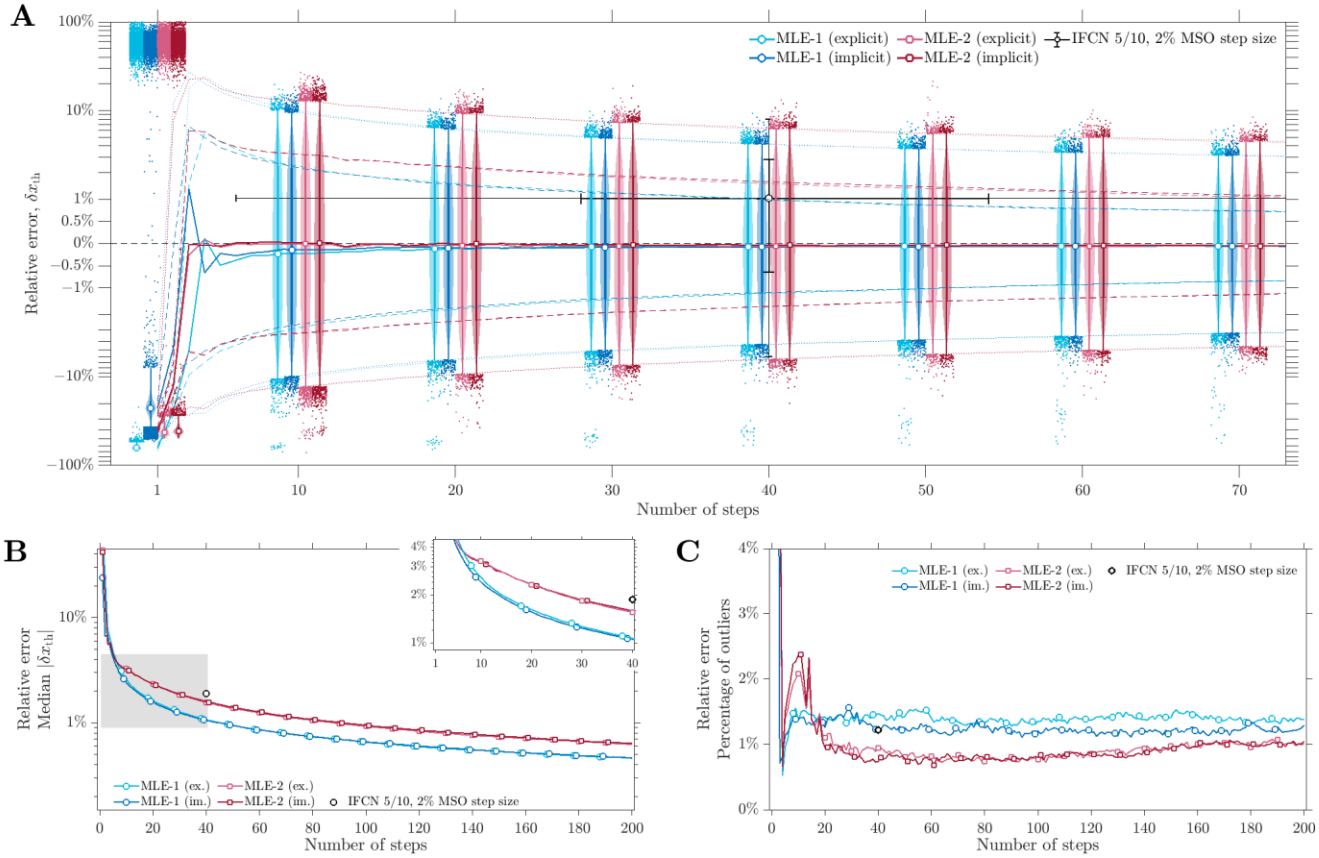

Figure S7: Relative threshold errors of MLE methods using explicit (ex.) and implicit (im.) maximization of the likelihood function

Shown similarly as in Figure 3.

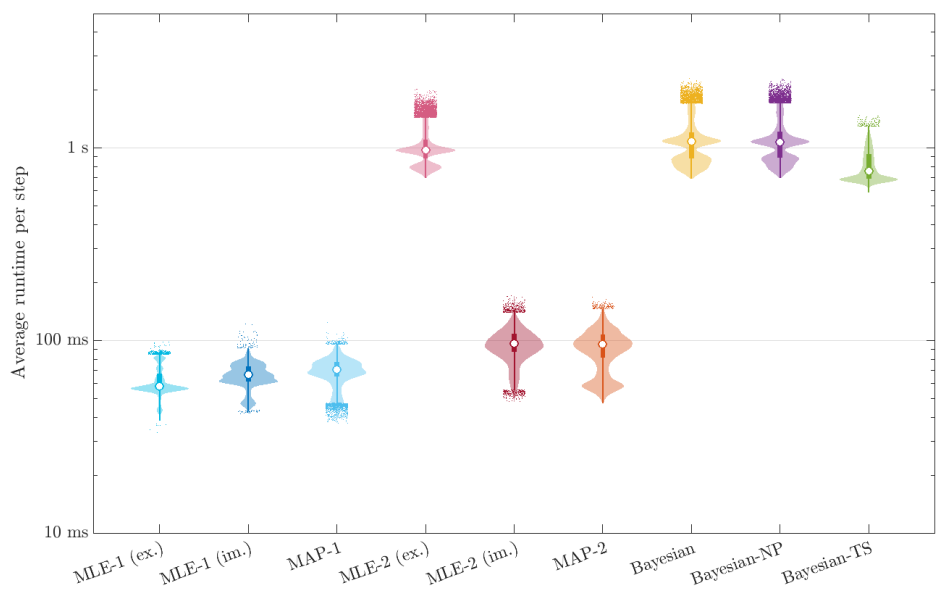

Figure S8: Runtime of parametric estimation methods

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

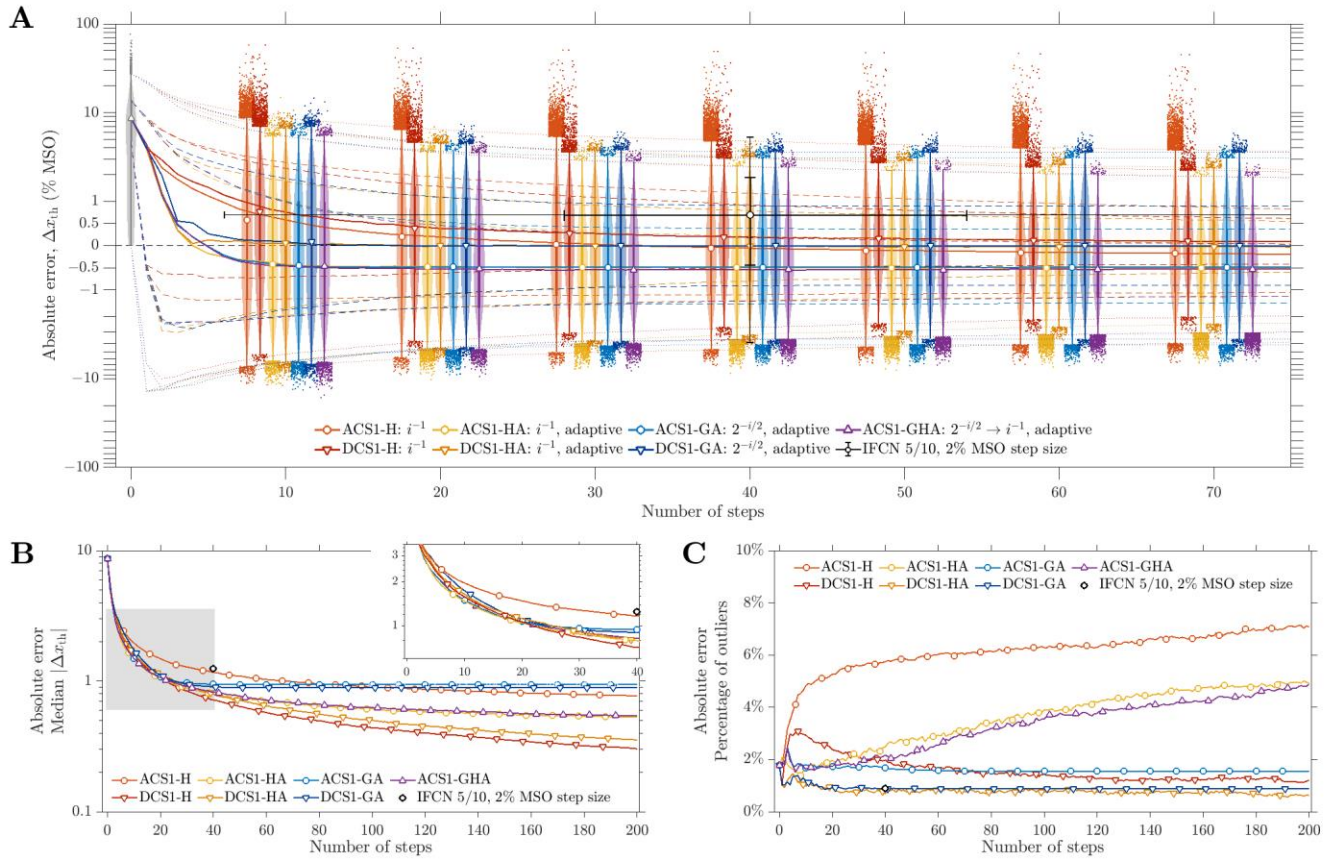

Figure S9: Absolute threshold errors of root finding methods using first-order control sequence with default initial step size a_0

Shown similarly as the relative threshold error in Figure 4.

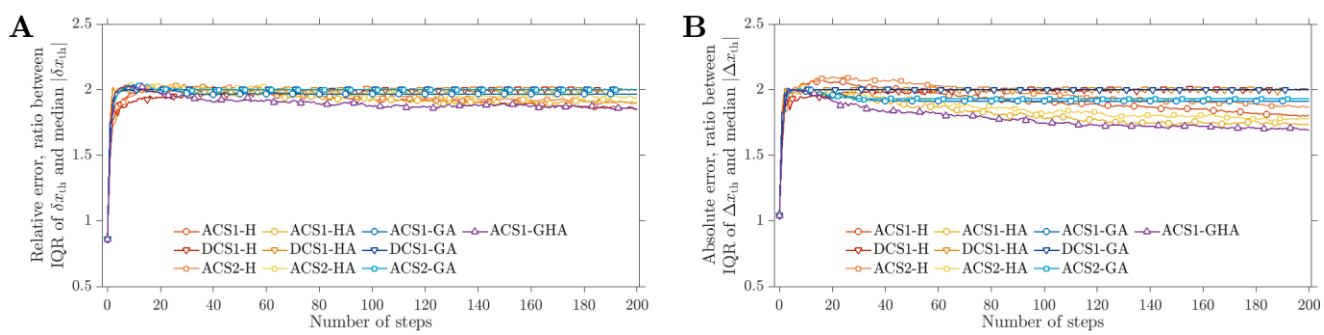

Figure S10: Ratio of median of absolute value of errors and interquartile range (IQR) of errors of root finding methods using default parameters

A. Ratio for relative threshold errors (related to Figures 4 and S14A). **B.** Ratio for absolute threshold errors (related to Figures S9 and S14B).

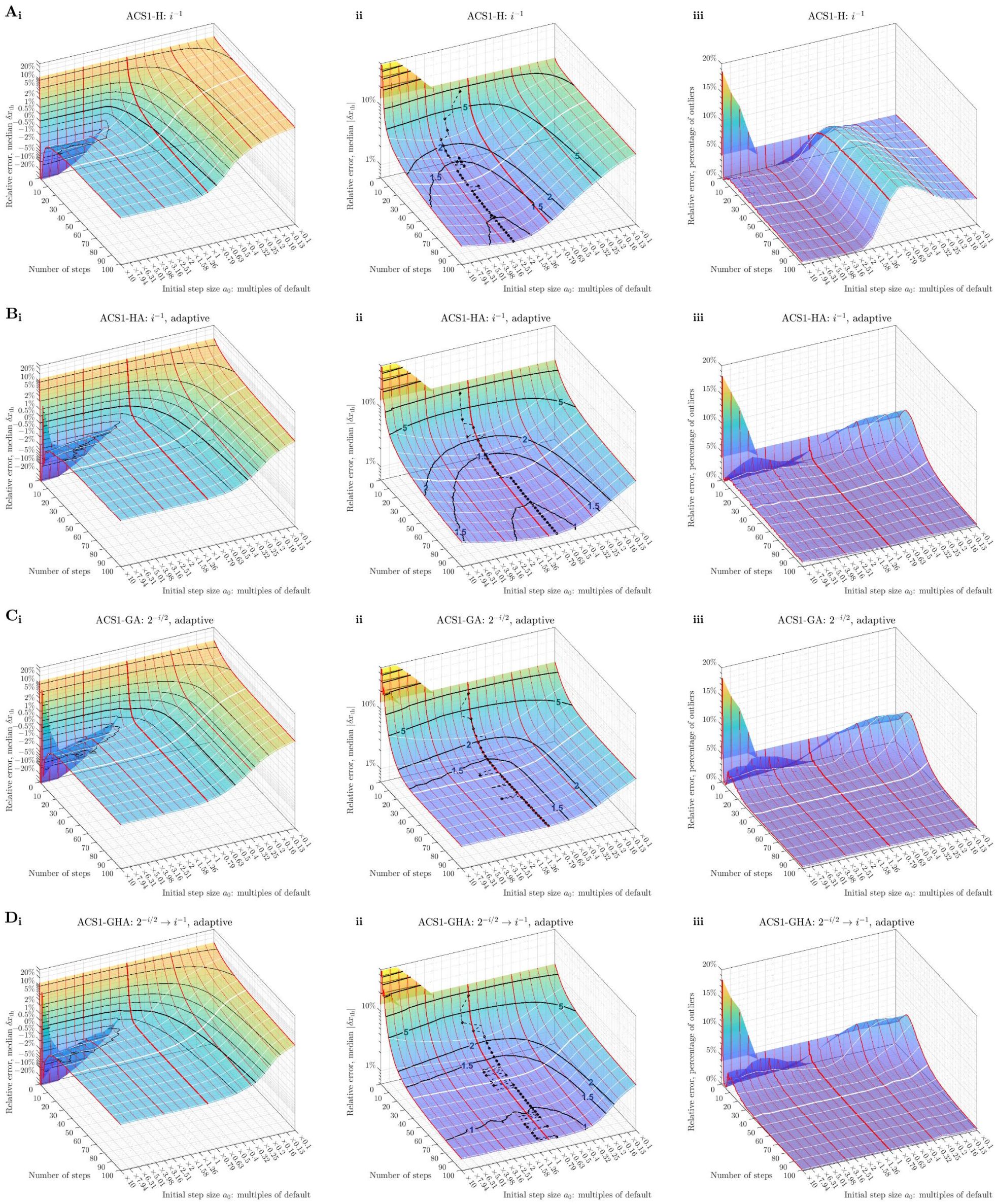

Figure S11: Relative threshold errors of root finding methods using first-order control sequence for different initial step size a_0

The red lines show results for different initial step size a_0, which are provided as multiples of the default value. The thick red line shows default value (results in Figures 4). The white lines mark every tenth step, with thick white line at the 40th step corresponding to the median number of steps of the IFCN five-out-of-ten method with 2%MSO step size. **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 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.

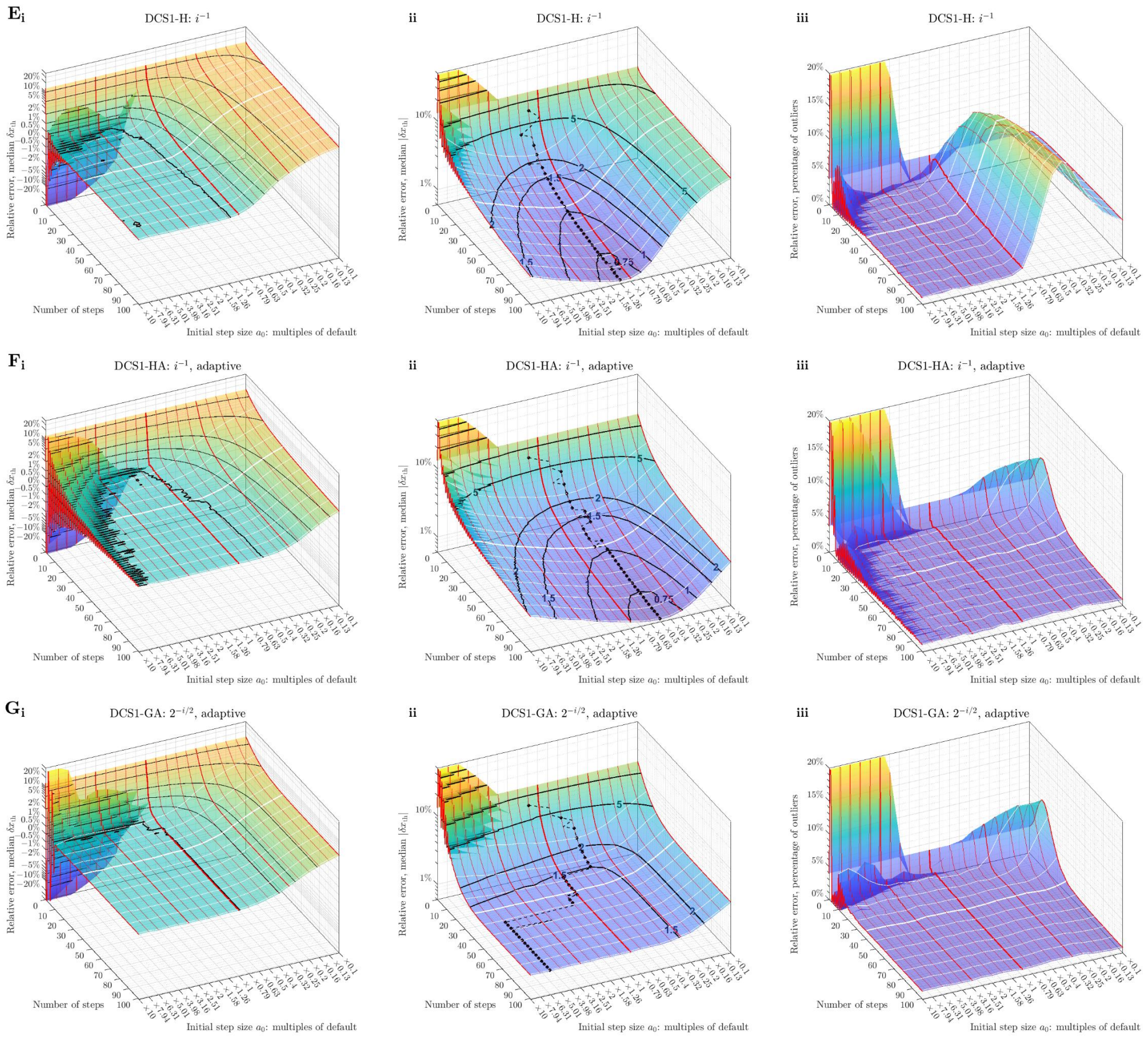

Figure S11: continued.
E.–G. Results for DCS1-H, DACS1-HA, and DCS1-GA, respectively.

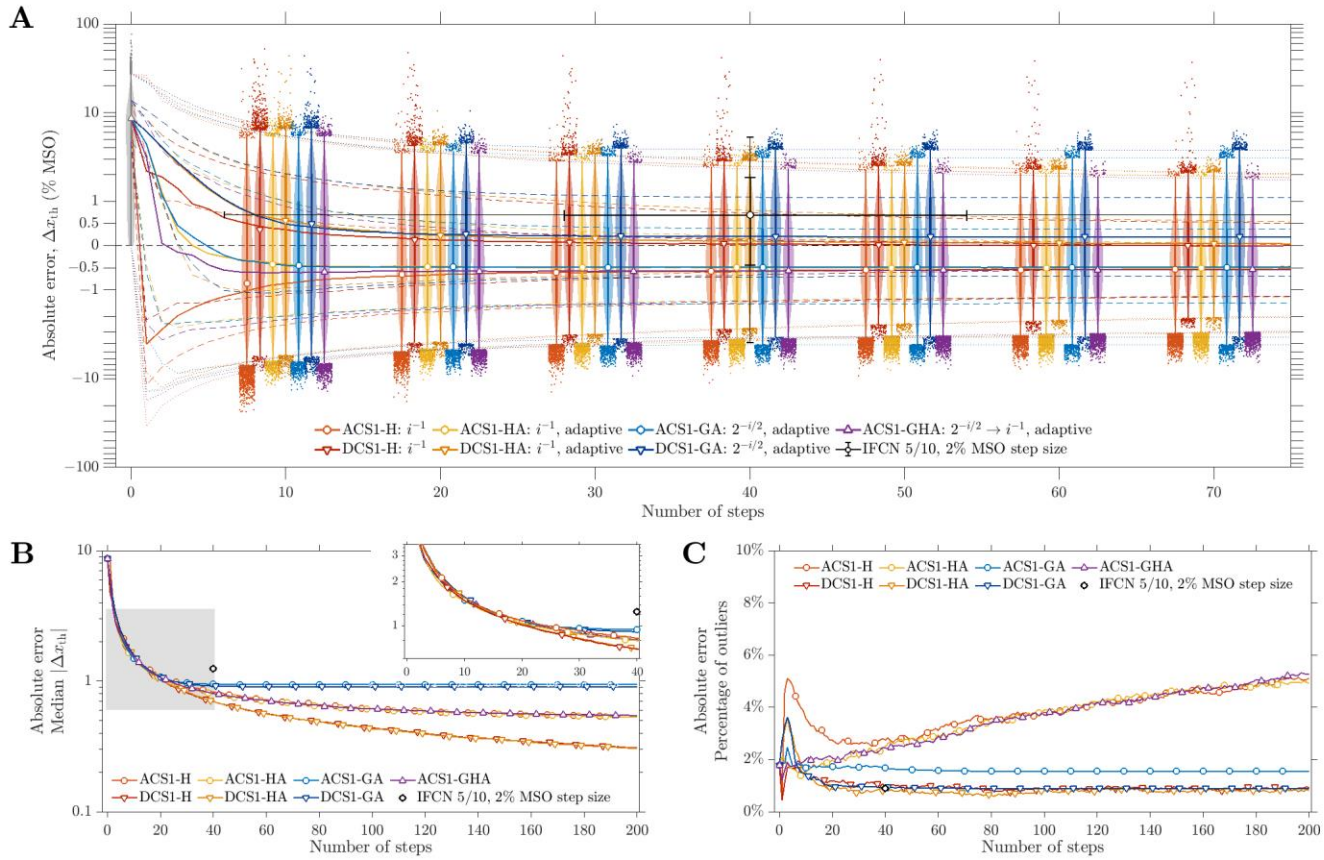

Figure S12: Absolute threshold errors of root finding methods using first-order control sequence with optimal initial step size a_0

Shown similarly as the relative threshold error in Figure 5.

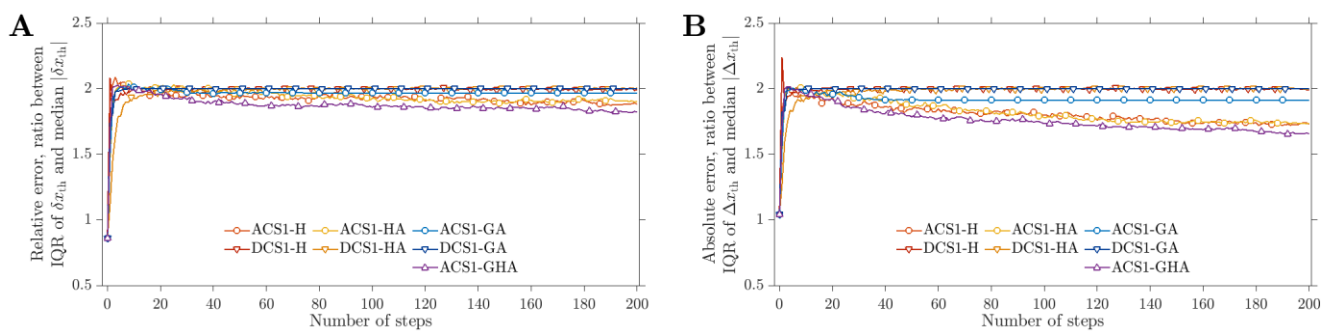

Figure S13: Ratio of median of absolute value of errors and interquartile range (IQR) of errors of first-order root finding methods using optimal initial step size a_0

A. Ratio for relative threshold errors (related to Figures 5). **B.** Ratio for absolute threshold errors (related to Figures S12).

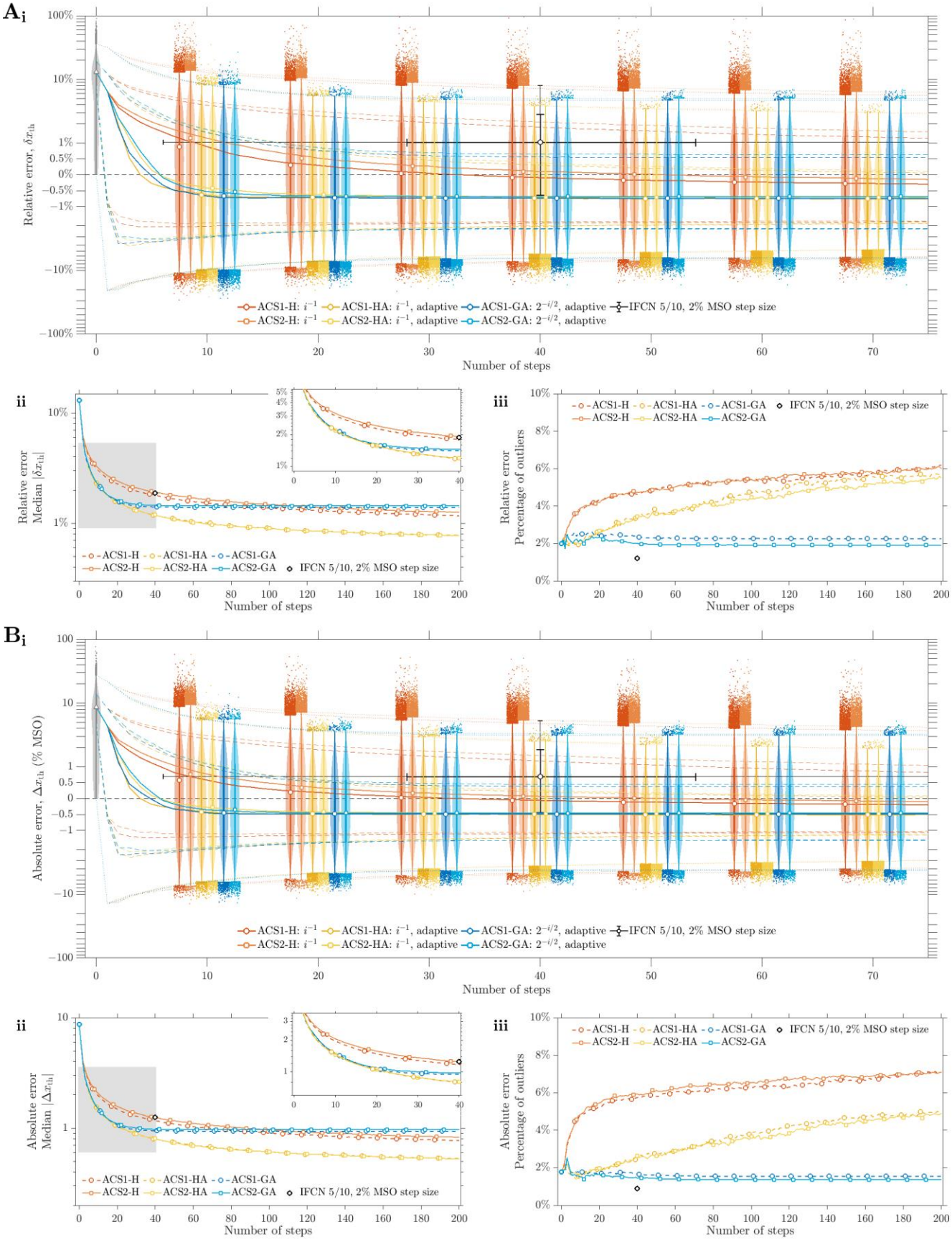

Figure S14: Threshold errors of root finding methods using second-order analog control sequence with default parameters

The corresponding methods using first-order analog control sequence are also shown for comparison (as in Figures 4 and S9). **A.** Relative threshold error. **B.** Absolute threshold error.

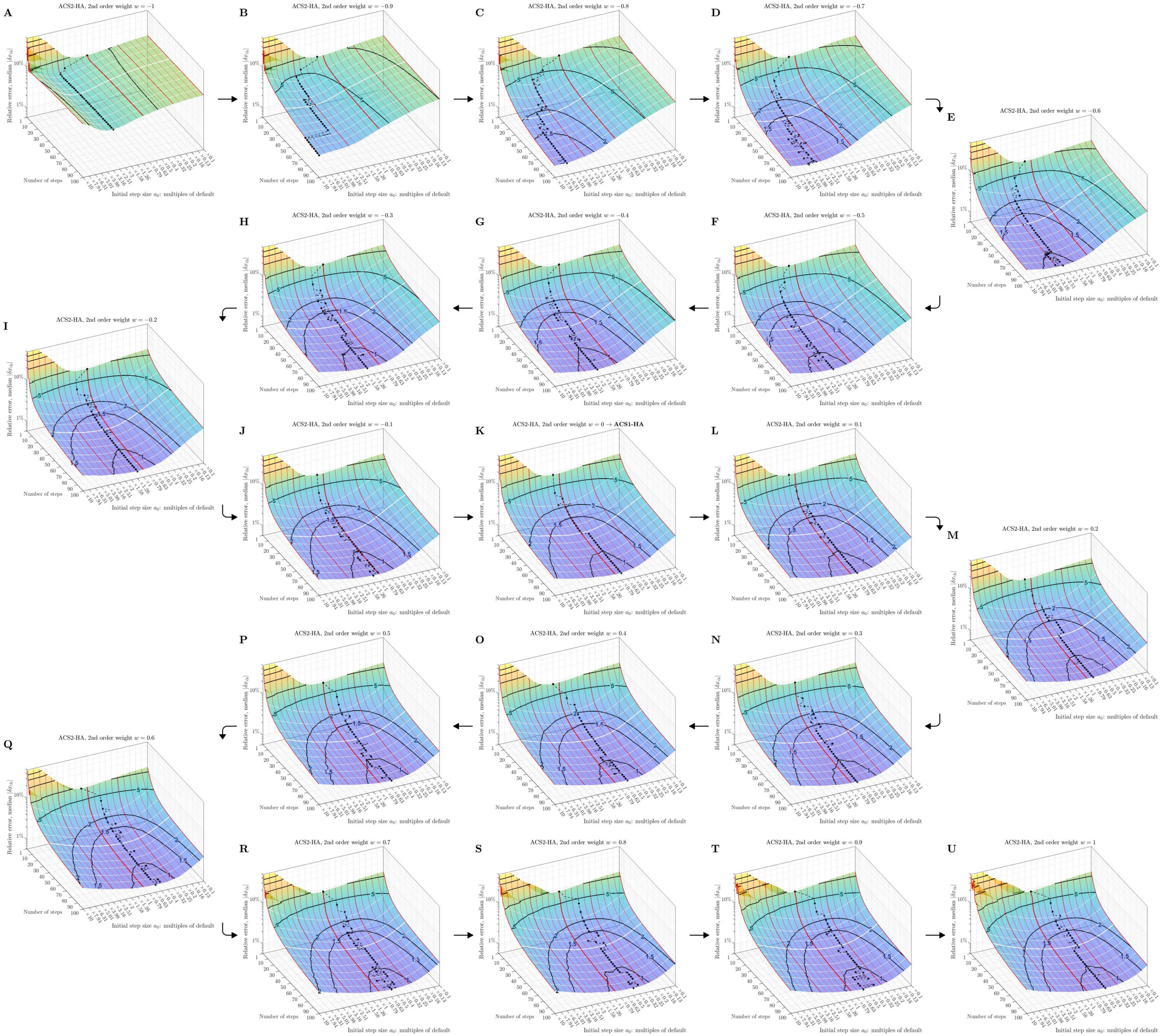

Figure S15: Relative threshold errors of root finding methods using second-order analog control sequence for different initial step size a_0 and given second order weights w
A.–U. Results for second order weights from −1 to 1 in steps of 0.1, shown similarly as in Figure S11A$_{ii}$.

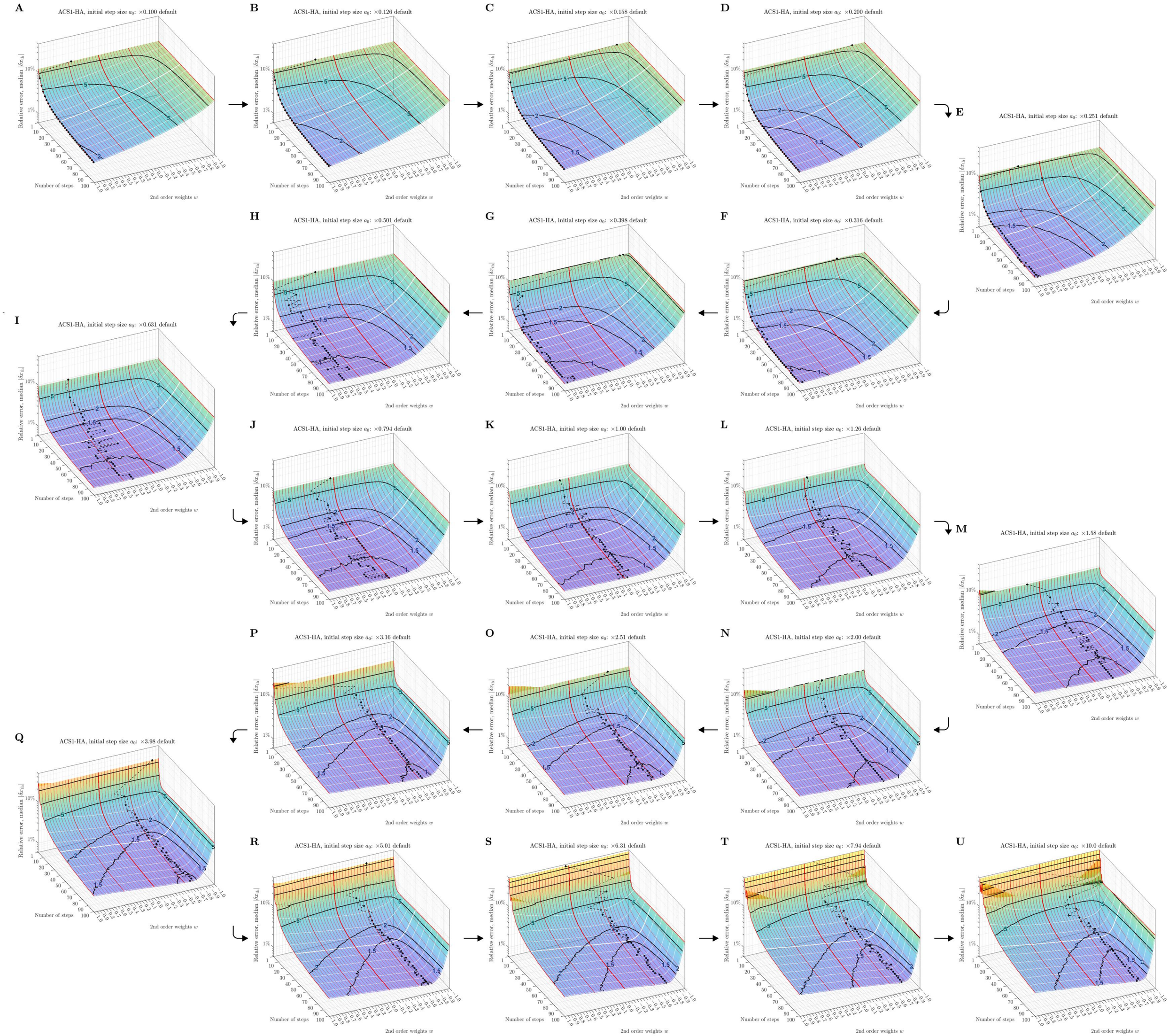

Figure S16: Relative threshold errors of root finding methods using second-order analog control sequence for different second order weights *w* and given initial step size a_0

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 S11A$_{ii}$ but with second-order weight *w* as the parameter. The thick red line indicates *w*=0, corresponding to the first-order version.

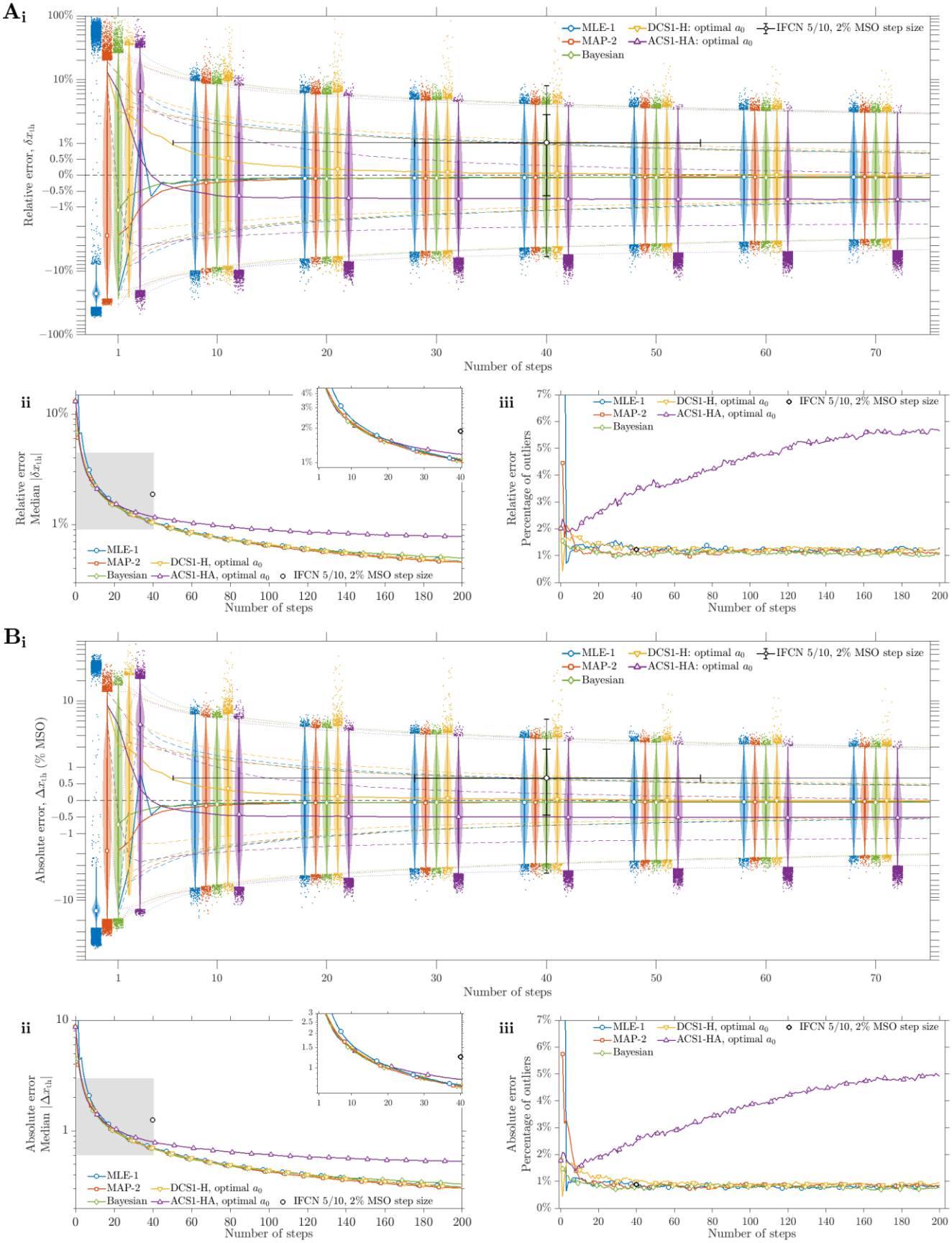

Figure S17: Relative threshold errors of the best performing methods in each category
Similar to Figures 3 and S5. **A.** Relative threshold error. **B.** Absolute threshold error.