Information-Theoretic Error Bounds for Source Localization in Neural Sensing

Leighton Pate Barnes
Carnegie Mellon University
leightonb@cmu.edu

Yuxin Guo
Carnegie Mellon University
yuxinguo@andrew.cmu.edu

Alex Dytso
Qualcomm Flarion Technology, Inc.
odytso2@gmail.com

Pulkit Grover
Carnegie Mellon University
pgrover@andrew.cmu.edu

Abstract

We formulate a point-source localization problem in d dimensions, where a source inside the ball of radius R emits a signal that is picked up by various sensors located at the surface of the ball. For $d = 3$, this can model problems in neural sensing, where a net of electroencephalography (EEG) or magnetoencephalography (MEG) sensors try to locate the source of a distinct neural event such as a seizure. For a power law decay model with exponent $\alpha > 0$ for the sensors, we obtain a lower bound on the minimax risk for localizing the source that is asymptotically $\frac{d^2\sigma^2 R^{2\alpha+2}}{n\alpha^2 PK}$ under mean-squared error loss, where σ^2 is the noise variance, P is the signal power, K is the number of sensors, and n is the number of independent measurements. In the case $d \leq 2(\alpha + 1)$ with uniformly distributed sensor locations, we then give a matching upper bound, including getting the exact constant correct, for the asymptotic minimax rate in a neighborhood of the origin. We show that there is a phase transition at $d = 2(\alpha + 2)$, above which a certain Fisher information quantity is minimized at the boundary of the ball, and below which it is minimized at the origin. At the critical dimension $d = 2(\alpha + 2)$, the Fisher information is constant throughout the entire parameter space. For the special case $d = 3$, we supplement and compare this information-theoretic analysis with a simulated forward EEG model that uses a realistic head model derived from population-averaged magnetic resonance imaging data.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

A ubiquitous problem in neural sensing, as well as other remote sensing applications such as global positioning, is to localize the source of an event from electromagnetic or other signals picked up by various sensors. In this paper, we formalize the problem of localizing a point-source event as a parametric statistical problem. We focus mainly on a scenario where the sensors are distributed across the surface of a d-dimensional ball, with $d = 3$ being the obvious motivating example. In the case of neural sensing, this scenario models a net of electroencephalography (EEG) or magnetoencephalography (MEG) sensors on the surface of the head, and a distinct neural event happening somewhere inside the brain. Accurately estimating the location of an event such as a neural response or a seizure onset is of great interest in both neuroscience and clinical contexts, with several techniques proposed that utilize optimization and machine-learning approaches (Hämäläinen and Ilmoniemi, 1994; Jatoi et al., 2014; Lee and Choi, 2009; Michel and He, 2019; Sun et al., 2022).

Obtaining fundamental limits to understand what localization error is possible at a given signal-to-noise ratio has received less attention, which has limited our understanding of topics such as optimal sensor placement. In a few past works, information-theoretic tools have been used to analyze the fundamental limits of similar point-source localization problems. In particular, Venkatesh and Grover (2017) focus on the EEG problem and use Fano's method for a statistical lower-bound, but limit their explicit analysis to the $d = 1$ case. They are able to get interesting lower bounds without making explicit assumptions on the shape of the "sensor model" other than some Lipschitz continuity assumptions. Their lower bounds show how the localization error decays as the number of sensors and independent measurements increase, but is only tight for these quantities up to arbitrary scaling factors. The same authors have also explored how spherical harmonics and the spatial Nyquist rate can inform an information-theoretic view of the EEG sensing problem (Grover and Venkatesh, 2017).

By contrast, in this paper we explicitly handle the general

d case, and in particular focus on $d = 3$. We bring physics into the setup, by modeling how measurement strength decays as a function of source distance to each sensor. These changes lead to a much more realistic model of the EEG problem, and allow us to characterize how the localization difficulty varies as a function of source depth from the surface of the sphere.

We take a Fisher information approach, which enables us to fully characterize the (asymptotic) minimax risk for the localization problem, including getting the correct constant. Fisher information is a powerful tool for parametric statistical lower bounds. It can be used for both unbiased estimators via the usual Cramér-Rao bound (Cramér, 1946; Radhakrishna Rao, 1945), as well as biased estimators using a Bayesian bound (Gill and Levit, 1995). It is also the correct measure for asymptotic complexity, giving both upper and lower bounds in the local asymptotic minimax theory (Hájek, 1972; LeCam, 1970; Van der Vaart, 2000). While the usual bounds focus on mean-squared error, Fisher information can be used to give lower bounds for any L^p loss as described by Chen and Özgür (2024). Other recent works that have used Fisher information to get lower bounds on nontrivial estimation problems include (Barnes and Özgür, 2021), that shows how the Fisher information can be related to other information-theoretic measures such as mutual information; and (Lee et al., 2021), that deals with estimation over an analog communication channel which is conceptually similar to neural sensing.

1.1 Relation to Prior Work

Previous work by Mosher et al. (1993) also looked at the Fisher information quantity in neural sensing to use with the Cramér-Rao lower bound. They do not analytically characterize the error rate, and instead rely only on simulation. Their simulations are based on a four-sphere head model, with four concentric spheres each modeling different conductivities for the different layers (brain, cerebrospinal fluid, skull, and scalp). In Section 5, we build on the Fisher information framework developed in the present paper, as well as that of Mosher et al., by showing that the Fisher information quantity can be computed (or approximated) in a more realistic model of a human brain built with anatomical magnetic resonance imaging (MRI) data. These simulations provide a proof of concept that the Fisher information quantities can be computed to guide real localization problems.

By handling the general d case instead of just the simplified $d = 1$ case, and by getting a matching (asymptotic) achievability result, we fundamentally advance the current state of understanding for information-theoretic limits in localization from (Grover and Venkatesh, 2017; Venkatesh and Grover, 2017; Xu and Coleman, 2020). The power-law decay model we define shortly does not previously appear in this literature, and so the analysis of its Fisher information is also new. However, our analysis does build on past work that considers how Fisher information can inform the localization problem such as (Mosher et al., 1993) (which focuses on simulation), and (Xu and Coleman, 2020) (which only considers d=1). By providing the first tight bounds that work in $d = 3$, we open up the possibility of analyzing the gap from optimality in localization algorithms used for technologies like EEG, and can help inform the design of these critical technologies. Furthermore, our analysis of the Fisher information for the power law model reveals a new and unexpected result – that the Fisher information can be minimized at the center of the ball or the outer boundary, depending on the dimension d.

In practice, localization is done via a linear forward model (such as the one described here in Section 5), which is used to solve the associated ill-posed inverse problem through regularization to obtain an estimate of the neural activity at each of the discretized locations (i.e. voxels) of the brain. The estimate of neural activity can then be used to estimate the location of a distinct event or cluster of activity. Multiple algorithms have been proposed to improve the accuracy of source localization, including classical optimization-based methods such as MNE (Hämäläinen and Ilmoniemi, 1994) and LORETA/sLORETA (Pascual-Marqui et al., 2002, 1994), to more recent deep learning-based approaches (Sun et al., 2022). The localization error reflects a combination of inverse-algorithm-related factors (e.g. optimization techniques, regularization) and sensing modality constraints (e.g. electrode placement, SNR, depth of the region of interests). These two factors cannot be disentangled in evaluation, making it difficult to interpret if the localization error is due to the limitation of the inverse algorithm or fundamentally due to the physical constraints/limitations of the sensing modality. Our work is aimed at addressing this issue by establishing the fundamental limits of estimating the location parameter theta (i.e. of the sensing modality itself), which are universal and hold regardless of the choice of inverse algorithm. Because our bounds are the first asymptotically tight bounds for the localization problem, they help understand whether poor localization is due to unavoidable physical constraints, or if there is still room for improvement with a better inverse algorithm (especially at high electrode densities, which is a problem of increasing interest). This further opens up the possibility of assessing how far from optimality any localization algorithm is, regardless of whether it uses the linear inverse or a maximum-likelihood type approach. In subsequent work, we intend to utilize these fundamental limits with matching upper bounds to determine optimal sensor placements in complex geometries, such as those obtained when sources in the region of interest lie in a sulcus. Practically, these placements and optimized algorithms can help improve clinical inferences for epileptic source localization, which is a problem of immense clinical importance

(Michel and He, 2019).

1.2 Paper Outline

The paper proceeds as follows. In Section 2, we mathematically formulate the source localization problem and introduce the necessary Fisher information quantities. In Section 3, we use these quantities to obtain fundamental limits (lower bounds) on the minimax risk for the localization error. In Section 4, we analyze the Fisher information matrix to obtain achievable error rates (upper bounds), attained by the maximum likelihood estimator. In Section 5, we compute the Fisher information matrix on a realistic forward model of EEG, and qualitatively compare it with the analytical theory. In Section 6, we conclude by mentioning several future directions.

2 TECHNICAL PRELIMINARIES

Suppose there are K sensors that each measure Y_i for $i = 1, \dots, K$, and their goal is to localize (i.e., estimate) an event $\theta = (\theta_1, \dots, \theta_d)$ that lies in some open set $\Theta \subseteq \mathbb{R}^d$. We assume

$$Y = x(\theta) + Z, \tag{1}$$

where $x(\theta)$ is a known deterministic function of θ that describes the sensor model, and $Z \sim \mathcal{N}(0, \sigma^2 I_K)$ is additive white Gaussian noise. The *Fisher information matrix* for estimating θ from Y is

$$I_Y(\theta) = \mathbb{E}_{Y|\theta}\left[(\nabla_\theta \log p(y|\theta))(\nabla_\theta \log p(y|\theta))^T\right] \tag{2}$$

$$= -\mathbb{E}_{Y|\theta}\left[\left(\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(y|\theta)\right)_{j,k}\right] . \tag{3}$$

where $p(y|\theta)$ is the density (with respect to the Lebesgue measure) of the random variable Y for a particular parameter value θ. The alternative form (3) holds provided that sufficient regularity conditions hold (Kay, 1993), and this will be the case throughout the paper. We will sometimes refer to the trace of the matrix (3) as *the Fisher information*. We use $\langle \cdot, \cdot \rangle$ to denote the standard inner product and $\| \cdot \|$ to denote the Euclidean norm.

Proposition 1. *For the measurement model in* (1)*, the entries of the Fisher information matrix can be written as*

$$[I_Y(\theta)]_{j,k} = \frac{1}{\sigma^2}\left\langle \frac{\partial}{\partial\theta_j} x(\theta)\,,\, \frac{\partial}{\partial\theta_k} x(\theta) \right\rangle . \tag{4}$$

Proof. The density $p(y|\theta)$ is

$$p(y|\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{K}{2}}} \exp\left(-\frac{\|x(\theta) - y\|^2}{2\sigma^2}\right) , \tag{5}$$

so that

$$\frac{\partial}{\partial\theta_j} \log p(y|\theta) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\theta_j}\|x(\theta) - y\|^2 \tag{6}$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{K}(x_i(\theta) - y_i)\frac{\partial}{\partial\theta_j} x_i(\theta) \tag{7}$$

and

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k} \log p(y|\theta) = -\frac{1}{\sigma^2}\sum_{i=1}^{K}\frac{\partial}{\partial\theta_j} x_i(\theta)\frac{\partial}{\partial\theta_k} x_i(\theta) \tag{8}$$

$$-\frac{1}{\sigma^2}\sum_{i=1}^{K}(x_i(\theta) - y_i)\frac{\partial^2}{\partial\theta_j\partial\theta_k} x_i(\theta) . \tag{9}$$

Taking the expected value, the second term becomes zero, and we get (4) as desired.

□

We emphasize that because of the potentially nonlinear function $x(\theta)$, the Fisher information matrix (which is d by d), is different from the inverse noise correlation matrix (which is K by K). It is instead related to the derivatives (or the Jacobian matrix) of $x(\theta)$, as seen in Proposition 1.

2.1 Sensor Models

Suppose that our K different sensors are at positions $s_1, \dots, s_K$ where each $s_i \in \mathbb{R}^d$, and we constrain the sensors to be on the sphere of radius R. Let the set of possible parameters Θ that we are interested in be the open ball of radius R contained in this sphere, i.e.,

$$\Theta = \{\theta \in \mathbb{R}^d \ : \ \|\theta\| < R\} . \tag{10}$$

(i) We first consider a power law decay model, as would be the case with a point-charge ($\alpha = 1$) or dipole[1] ($\alpha = 2$) (Buzsáki et al., 2012). In this model

$$x_i(\theta) = \frac{\sqrt{P}}{\|s_i - \theta\|^\alpha} \tag{11}$$

for some $\alpha > 0$ and some nominal power $P > 0$.

Proposition 2. *Under the power law decay model in* (11)*, the Fisher information is given by*

$$\mathsf{Tr}(I_Y(\theta)) = \frac{\alpha^2 P}{\sigma^2}\sum_{i=1}^{K}\frac{1}{\|s_i - \theta\|^{2\alpha+2}} . \tag{12}$$

Proof. In this case,

$$\mathsf{Tr}(I_Y(\theta)) = \frac{1}{\sigma^2}\sum_{j=1}^{d}\left\|\frac{\partial}{\partial\theta_j} x(\theta)\right\|^2 \tag{13}$$

[1] We call the $\alpha = 2$ case the "dipole" case because it decays radially at the same rate as a dipole's potential field. A true dipole also has an orientation, as well as near and far field effects, which are not accounted for in this model. A more realistic dipole model could be worked out separately from the general α framework we work with in this paper.

$$= \frac{1}{\sigma^2} \sum_{i=1}^{K} \|\nabla_\theta x_i(\theta)\|^2 \tag{14}$$

$$= \frac{\alpha^2 P}{\sigma^2} \sum_{i=1}^{K} \frac{1}{\|s_i - \theta\|^{2\alpha+2}} \, . \tag{15}$$

□

Note that if θ is the origin and since $\|s_i\| = R$,

$$\mathsf{Tr}(I_Y(0)) = \frac{\alpha^2 PK}{\sigma^2 R^{2\alpha+2}} \, . \tag{16}$$

(ii) Under other general conditions, such as uniformly bounded gradient magnitudes

$$\|\nabla_\theta x_i(\theta)\|^2 \leq G \, , \tag{17}$$

(14) gives

$$\mathsf{Tr}(I_Y(\theta)) \leq \frac{GK}{\sigma^2} \, . \tag{18}$$

If the parameter space Θ is reparameterized by a linear map $\theta = A\eta$, then the Fisher information matrix is transformed by

$$I_Y(\eta) = A^T I_Y(\theta = A\eta) A \, . \tag{19}$$

In particular, if A is a rotation matrix (i.e. $AA^T = I$), then both $\mathsf{Tr}(I_Y(\eta)) = \mathsf{Tr}(I_Y(\theta = A\eta))$ and $\mathsf{Tr}(I_Y(\eta)^{-1}) = \mathsf{Tr}(I_Y(\theta = A\eta)^{-1})$.

3 MINIMAX LOWER BOUNDS

In this section we consider lower bounds on the minimax squared-error risk in estimating θ, i.e.,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{Y^n|\theta} \|\widehat{\theta} - \theta\|^2 \, , \tag{20}$$

where $\widehat{\theta}$ is an estimate of θ from the measured data. We assume that we observe n independent measurements of Y for the same θ, and $\widehat{\theta}$ is a function of those n measurements.

In a Bayesian setting, if the parameter θ is taken from a prior distribution μ on Θ, then the Van Trees inequality (Gill and Levit, 1995) gives

$$\mathbb{E}_{\theta, Y^n} \|\widehat{\theta} - \theta\|^2 \geq \frac{d^2}{n\mathbb{E}_{\theta \sim \mu}[\mathsf{Tr}(I_Y(\theta))] + J(\mu)} \tag{21}$$

where

$$J(\mu) = \int_\Theta \frac{\|\nabla_\theta \mu(\theta)\|^2}{\mu(\theta)} d\theta \tag{22}$$

is the Fisher information from the prior.

In order to optimize the lower bound given by (21), it would be advantageous to find the prior μ on Θ that minimizes (22). For the parameter space $\Theta' = (-B, B)^d$, such a μ is known (see Tsybakov (2009); Uhrmann-Klingen (1995)), and results in

$$J(\mu) = \frac{d\pi^2}{B^2} \, . \tag{23}$$

One can show that the order $1/B^2$ is tight for the smallest ball containing Θ' by using Stam's inequality (Stam, 1959) $J(\mu)N(\mu) \geq d$, where $N(\mu) = \frac{1}{2\pi \mathrm{e}} \mathrm{e}^{\frac{2}{d} h(X)}$ is the entropy power. This yields

$$\min_{\mathrm{supp}(\mu) \subseteq \mathcal{B}_d(\sqrt{d}B)} J(\mu) \geq \frac{d}{\max_{\mathrm{supp}(\mu) \subseteq \mathcal{B}_d(\sqrt{d}B)} N(\mu)} \tag{24}$$

$$= \frac{C_d}{B^2} \tag{25}$$

where $C_d = 2\mathrm{e}\Gamma(d/2+1)^{2/d} = \Omega(d)$ with the standard gamma function Γ. Line (25) follows from uniform distributions maximizing entropy over compact supports (Cover and Thomas, 2006).

Equations (24) and (25) show that for the natural parameter space in this problem (the d dimensional ball of radius R), there are bounds on how small the Fisher information from the prior can be. This bound is roughly the same as that for the d dimensional box that is contained in the ball, so we do not lose much in our lower bound by using either the box or the ball. Either one could be used in the proof of Theorem 1 below.

There is a close connection between minimax estimation and Bayesian estimation. Bayes risk always lower bounds minimax risk, and if a Bayes estimator for a given prior has the same Bayesian risk as worst-case risk, then it must also be a minimax estimator (and the prior must be the worst-case) (Lehmann and Casella, 1998). Fisher information is a tool of interest to understand both of these quantities, and in that sense our contributions can also potentially help with the Bayesian case. While we do not explicitly focus on analyzing the Bayesian risk of this estimation problem for various priors, such as the ones discussed above, we note that the inequalities in Gill and Levit (1995) that we use here would be a useful starting point.

Theorem 1. *Under the power law decay model in* (11)*, let* Θ *be the ball of radius* R*. Then, for any measurable estimator* $\widehat{\theta}$*,*

$$\sup_{\theta \in \Theta} \mathbb{E}_{Y^n|\theta} \|\widehat{\theta} - \theta\|^2 \geq \frac{d^2 \sigma^2 R^{2\alpha+2}}{n\alpha^2 PK} \left(1 - o_{nK}(1)\right) . \tag{26}$$

If instead we have bounded gradients as in (17)*, then*

$$\sup_{\theta \in \Theta} \mathbb{E}_{Y^n|\theta} \|\widehat{\theta} - \theta\|^2 \geq \frac{d^2 \sigma^2}{nGK} \left(1 - o_{nK}(1)\right) \, . \tag{27}$$

Remark 1. The little-o terms above approach zero as the product nK goes to infinity, even if $n = 1$. The $R^{2\alpha+2}$

term is interesting in that it becomes *very* disadvantageous to have sensors further from the source.

Proof. We begin with the proof of (26). Let $B = \frac{R}{d^{\frac{1}{2}}(nK)^{\frac{1}{3}}}$ and note that $\Theta' = (-B, B)^d \subset \Theta$. When nK is large and B is small, the Fisher information (15) can be bounded *above* by $\frac{\alpha^2 PK}{\sigma^2 R^{2\alpha+2}}(1 + o_{nK}(1))$ uniformly over Θ'. This is because for $\theta \in \Theta'$,

$$\|s_i - \theta\| \geq \left\| R - \frac{R}{(nk)^{\frac{1}{3}}} \right\| \tag{28}$$

so that each term in (12)

$$\frac{1}{\|s_i - \theta\|^{2\alpha+2}} \leq \frac{1}{\left(R\left(1 - \frac{1}{(nK)^{\frac{1}{3}}}\right)\right)^{2\alpha+2}} . \tag{29}$$

Furthermore, if

$$n\mathsf{Tr}(I_Y(\theta)) \gg \frac{d\pi^2}{B^2} , \tag{30}$$

then the Fisher information from Y dominates the Fisher information from the prior in (21), and this will be the case uniformly in $\theta \in \Theta'$ when the product nK is large.

We have

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathbb{E}_{Y^n|\theta}\|\widehat{\theta} - \theta\|^2 &\geq \sup_{\theta\in\Theta'} \mathbb{E}_{Y^n|\theta}\|\widehat{\theta} - \theta\|^2 && (31)\\
&\geq \mathbb{E}_{Y^n,\theta\sim\mu}\|\widehat{\theta} - \theta\|^2 && (32)\\
&\geq \frac{d^2}{n\mathbb{E}_{\theta\sim\mu}[\mathsf{Tr}(I_Y(\theta))] + J(\mu)} && (33)\\
&\geq \frac{d^2}{n\frac{\alpha^2 PK}{\sigma^2 R^{2\alpha+2}}(1 + o_{nK}(1)) + \frac{d\pi^2}{B^2}} && (34)\\
&\geq \frac{d^2\sigma^2 R^{2\alpha+2}}{n\alpha^2 PK}\left(1 - o_{nK}(1)\right) . && (35)
\end{aligned}$$

The proof of (27) is similar, except that in step (34) we instead directly use (18). □

Remark 2. It will become clear in Section 4 that for $d < 2(\alpha + 2)$, focusing on the subset of the parameter space near the origin gives best possible lower bound from (21) – this is where the Fisher information is minimized.

4 ACHIEVABLE MINIMAX RATES

In order to get a matching asymptotic upper bound, we need to consider how our sensors are positioned, i.e., the values of $s = (s_1, \ldots, s_K)$. For example, if all of the sensors are positioned on one side of the ball, then the worst case θ that maximizes the expected error would intuitively be on the other side of the ball. To simplify this problem, we will assume that the sensors are uniformly placed around the surface of the sphere. This may be a reasonable assumption in the EEG example, where a net of sensors (with as high density as possible) is placed with roughly uniform placement about the scalp. This assumption can break down as the source approaches the boundary of the ball, where it becomes important whether the source is close to the exact location of a sensor or not – since signal strength can get arbitrarily large close to a sensor. However, this is not a concern with noninvasive sensing such as EEG.

Formally, the uniform sensor assumption is realized by choosing the sensor locations to be independent and identically drawn random variables S_i taken from the uniform distribution (i.e., the Haar measure) on the sphere. We still assume that the locations are known, so really we consider the Fisher information from the parametric model

$$p(s, y|\theta) = p(s)p(y|\theta, s) . \tag{36}$$

This does not change the score vector in that

$$\nabla_\theta \log p(y, s|\theta) = \nabla_\theta \log p(y|\theta, s) . \tag{37}$$

Note, however, that the Fisher information entries $[I_Y(\theta)]_{j,k}$ are now averaged over the uniformly distributed locations S_i. This along with (19) also means that $\mathsf{Tr}(I_Y(\theta))$ and $\mathsf{Tr}(I_Y(\theta)^{-1})$ are invariant to rotations, i.e, they are spherically symmetric.

The Fisher information matrix characterizes the complexity of the parameter estimation problem asymptotically as the number of independent samples n goes to infinity, and locally around the particular θ used to evaluate $I_Y(\theta)$. More concretely, we consider the *maximum likelihood estimator*

$$\begin{aligned}
\widehat{\theta}_n(\mathbf{y}, \mathbf{s}) &= \arg\max_\theta p\Big(\mathbf{y} = (y^1, \ldots, y^n), \\
&\qquad \mathbf{s} = (s^1, \ldots, s^n)\Big|\theta\Big) && (38)\\
&= \arg\max_\theta \sum_{i=1}^n \log p(y^i, s^i|\theta) && (39)\\
&= \arg\min_\theta \sum_{i=1}^n \|x(\theta, s^i) - y^i\|^2 . && (40)
\end{aligned}$$

If the statistical model $p(y, s|\theta)$ satisfies a technical condition known as *differentiability in quadratic mean*, then this estimator is known to be asymptotically normal in the sense that $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges to $\mathcal{N}(0, I_Y(\theta)^{-1})$ in distribution (see Van der Vaart (2000) Thm 5.39). In our case, with additive Gaussian noise, the required technical conditions do hold provided that $x(\theta)$ is a smooth function. In general, this convergence in distribution does not also imply convergence in moments, however, in our scenario, we find that $\sqrt{n}(\widehat{\theta}_n - \theta)$ does have a convergent L^2 norm and

$$n\mathbb{E}_{Y^n,S^n|\theta}\|\widehat{\theta}_n - \theta\|^2 \to \mathsf{Tr}(I_Y(\theta)^{-1}) . \tag{41}$$

See the appendix for details on the convergence of the mean-squared error.

In light of (41), we would like to understand $\mathsf{Tr}\left(I_Y(\theta)^{-1}\right)$, and we have the following characterizations of the Fisher information matrix.

Proposition 3. *With uniformly random sensor locations and the point-source model in* (11)*, the Fisher information matrix $I_Y(\theta_0)$ is diagonal at the point $\theta_0 = (c, 0, \ldots, 0)$.*

Proof. For $j \neq k$, consider an off-diagonal entry $[I_Y(\theta_0)]_{j,k}$.

$$[I_Y(\theta_0)]_{j,k} = \frac{\alpha^2 P}{\sigma^2} \int \sum_{i=1}^{K} \frac{(s_{i,j} - \theta_{0,j})(s_{i,k} - \theta_{0,k})}{\|s_i - \theta_0\|^{2\alpha+4}} \mathsf{d}s \tag{42}$$

$$= \frac{\alpha^2 PK}{\sigma^2} \int \frac{(s_{1,j} - \theta_{0,j})(s_{1,k} - \theta_{0,k})}{\|s_1 - \theta_0\|^{2\alpha+4}} \mathsf{d}s_1 \tag{43}$$

$$= \frac{\alpha^2 PK}{\sigma^2} \begin{cases} \int \frac{(s_{1,j}-c)s_{1,k}}{\|s_1-\theta_0\|^{2\alpha+4}} \mathsf{d}s_1 & , \ j \text{ or } k = 1 \\ \int \frac{s_{1,j}s_{1,k}}{\|s_1-\theta_0\|^{2\alpha+4}} \mathsf{d}s_1 & , j, k \neq 1 \end{cases} \tag{44}$$

$$= 0 \ . \tag{45}$$

Here we use $\mathsf{d}s$ to denote the product integral $\prod_i \mathsf{d}s_i$ where each $\mathsf{d}s_i$ denotes integration with respect to the Haar measure. Equation (42) follows from Proposition 1 and (11). The final step follows from the oddness of the function with respect to $s_{1,k}$ along with the symmetry of the Haar measure. The symmetry we use is the fact that the Haar measure is invariant to the change of variables $s_{1,k} \mapsto -s_{1,k}$ (i.e., the reflection of s_1 across axis k). So we can break the integral into the space with positive $s_{1,k}$ values and the space with negative $s_{1,k}$ values, and the two terms cancel.

We do not need to deal with entry $[I_Y(\theta_0)]_{1,1}$, since we are only trying to show the matrix is diagonal, i.e., the off-diagonal entries are zero. □

The next lemma shows that $\mathsf{Tr}(I_Y(\theta))$ exhibits striking behavior as the dimension d changes. For $d < 2(\alpha + 2)$, the trace of the Fisher information is minimized at the origin, suggesting that it is the most difficult part of the parameter space to localize. At what we term the *critical dimension*, $d = 2(\alpha+2)$, the trace of the Fisher information is exactly constant throughout the ball. For $d > 2(\alpha + 2)$, the trace of the Fisher information is minimized at the boundary of the ball. Note that in the practically relevant $d = 3$ case, d is smaller than the critical dimension for all $\alpha > 0$. So for our main application (and, indeed, most applications) the Fisher information at the center of the ball is key to characterizing the minimax rate. However, in high dimensions we have the counterintuitive result that this is no longer the case. While it is more difficult to work with $\mathsf{Tr}(I_Y(\theta)^{-1})$ instead of $\mathsf{Tr}(I_Y(\theta))$, we are nonetheless able to show that for $d \leq 2(\alpha + 1)$ the quantity $\mathsf{Tr}(I_Y(\theta)^{-1})$ is maximized at the origin over a neighborhood of the origin, and that for $d > 2(\alpha + 2)$, it is not maximized at the origin.

Lemma 1. *The sum of the diagonal entries of the form* (42)*,*

$$\mathsf{Tr}(I_Y(\theta)) = \frac{\alpha^2 PK}{\sigma^2} \sum_{j=1}^{d} \int \frac{(s_{1,j} - \theta_j)^2}{\|s_1 - \theta\|^{2\alpha+4}} \mathsf{d}s_1 \tag{46}$$

$$= \frac{\alpha^2 PK}{\sigma^2} \int \frac{1}{\|s_1 - \theta\|^{2\alpha+2}} \mathsf{d}s_1 \ , \tag{47}$$

has the following properties.

(i) If $d < 2(\alpha + 2)$, then $\mathsf{Tr}(I_Y(\theta))$ achieves its unique global minimum at $\theta = 0$.

(ii) If $d = 2(\alpha + 2)$, then $\mathsf{Tr}(I_Y(\theta))$ is constant over all $\theta \in \Theta$.

(iii) If $d > 2(\alpha + 2)$, then $\mathsf{Tr}(I_Y(\theta))$ approaches its minimum only at the boundary, i.e., as $\|\theta\| \to R$.

Proof. The function in (47) is a superposition of shifted copies of the map $\theta \mapsto \frac{1}{\|\theta\|^{2\alpha+2}}$. The trace of the Hessian (i.e., the Laplacian) of this map is

$$\nabla^2 \left(\frac{1}{(\theta_1^2 + \ldots + \theta_d^2)^{\alpha+1}} \right) = \frac{-2(\alpha+1)(d - 2(\alpha+2))}{(\theta_1^2 + \ldots + \theta_d^2)^{\alpha+2}} \tag{48}$$

which is positive, zero, or negative everywhere it is defined (everywhere but at the origin) depending on whether we are in case (i), (ii), or (iii), respectively. Using dominated convergence to interchange limiting operations, the trace of the Hessian of (47) must also be positive, zero, or negative, respectively, for all $\|\theta\| < R$. Since (47) is the convolution of a spherically symmetric function with a spherically symmetric measure, it is also spherically symmetric. We can therefore parametrize (47) by a function $g(r = \|\theta\|)$. Again using dominated convergence, we have that g is infinitely differentiable for $r < R$. For spherical symmetric functions, note (Taylor, 2020)

$$\mathsf{Tr}(\mathsf{Hessian}_g(r = \|\theta\|)) = g''(r) + \frac{d-1}{r} g'(r) \ . \tag{49}$$

(Case (i)): By symmetry $g'(0) = 0$ and therefore (49) implies $g''(0) > 0$ and 0 is a local minimum. Supposing, for contradiction, that g is not strictly increasing in r, then there must exist an $0 < r_0 < R$ with $g'(r_0) = 0$ and $g''(r_0) \leq 0$. At r_0, (49) must be less than or equal to zero. But this contradicts the trace of the Hessian being positive for $\|\theta\| < R$, and therefore g must be increasing in r.

(Case (ii)): When the Laplacian is zero, $g(r)$ must satisfy the differential equation

$$g''(r) + \frac{d-1}{r} g'(r) = 0 \tag{50}$$

with initial conditions $g'(0) = 0$ and $g(0) = g_0$. The unique solution is the constant function $g(r) = g_0$.

(Case (iii): By the argument in Case (i), except negating the quantities in question, we see that g must be decreasing in r. Note also that (47) is bounded below by zero, so we can extend g by continuity to the closure $[0, R]$, and its value on the boundary $g(R)$ will be its minimum. □

One way of thinking about Lemma 1 is that the trace of the Fisher information is either subharmonic, harmonic, or superharmonic, along with being spherically symmetric, depending on which of the three cases is true.

We now turn to analyzing $\mathsf{Tr}(I_Y(\theta)^{-1})$, which does not uniformly satisfy the same (sub)harmonicity properties as in Lemma 1. However, we can still build on the techniques from Lemma 1 to show the following properties. In Lemma 2 below, it is more challenging to show optimality of the origin than in Lemma 1, and we only show it for a slightly smaller subset of d values (for a fixed α).

Lemma 2. *If $d \leq 2(\alpha+1)$, then $\mathsf{Tr}(I_Y(\theta)^{-1})$ is maximized at $\theta = 0$ over a neighborhood $U \subseteq \Theta$ containing the origin.*

Proof. We consider only a single term of the summation in (46),

$$[I_Y(\theta)]_{j,j} = \frac{\alpha^2 PK}{\sigma^2} \int \frac{(s_{1,j} - \theta_j)^2}{\|s_1 - \theta\|^{2\alpha+4}} \mathsf{d}s_1 \tag{51}$$

which is a superposition of shifted copies of $\theta \mapsto \frac{\theta_j^2}{\|\theta\|^{2\alpha+4}}$. The trace of the Hessian of this map is

$$\nabla^2 \left(\frac{\theta_j^2}{(\theta_1^2 + \ldots + \theta_d^2)^{\alpha+2}} \right) = \frac{2}{(\theta_1^2 + \ldots + \theta_d^2)^{\alpha+2}} - \frac{(\alpha+2)(10 + 2(d-1) - 4(\alpha+3))}{(\theta_1^2 + \ldots + \theta_d^2)^{\alpha+3}} \tag{52}$$

which is strictly positive if $d \leq 2(\alpha+1)$. Using dominated convergence, the trace of the Hessian of (51) is also strictly positive for $\|\theta\| < R$.

Note that

$$\nabla^2 \left(\frac{1}{f(\theta)} \right) = \sum_{j=1}^{d} \frac{1}{f(\theta)^2} \left(\frac{2 \frac{\partial f}{\partial \theta_j}(\theta)^2}{f(\theta)} - \frac{\partial^2 f}{\partial \theta_j^2}(\theta) \right) \tag{53}$$

$$= \frac{1}{f(\theta)^2} \left(\frac{2\|\nabla f(\theta)\|^2}{f(\theta)} - \nabla^2 f(\theta) \right) . \tag{54}$$

Setting $f(\theta) = [I_Y(\theta)]_{j,j}$ we have $\nabla f(0) = 0$ by symmetry, so that $\nabla^2 \left(\frac{1}{f} \right)_{\theta=0} < 0$. Therefore, the summation

$$\varphi(\theta) = \sum_{j=1}^{d} \frac{1}{[I_Y(\theta)]_{j,j}} \tag{55}$$

also has $\nabla^2 \varphi(0) < 0$, and by the symmetry with respect to different coordinates $i = 1, \ldots, d$ we have $\frac{\partial^2 \varphi}{\partial \theta_i^2}(0) < 0$. This means φ has a local maximum along the line $\theta_0 = (c, 0, \ldots, 0)$ at the origin. By Proposition 3, $\varphi(\theta_0)$ matches $\mathsf{Tr}(I_Y(\theta_0)^{-1})$ along this line. By the rotational symmetry of $\mathsf{Tr}(I_Y(\theta)^{-1})$, this establishes $\theta = 0$ as its local maximum. □

Lemma 3. *If $d > 2(\alpha+2)$, then $\mathsf{Tr}(I_Y(\theta)^{-1})$ is minimized at the origin over all $\theta \in \Theta$.*

Proof. By Jensen's inequality and a singular value decomposition,

$$\frac{d^2}{\mathsf{Tr}(I_Y(\theta))} \leq \mathsf{Tr}(I_Y(\theta)^{-1}) , \tag{56}$$

with equality at $\theta = 0$. In the case $d > 2(\alpha+2)$, the proof of Lemma 1 shows that

$$\frac{d^2}{\mathsf{Tr}(I_Y(0))} \leq \frac{d^2}{\mathsf{Tr}(I_Y(\theta))} , \tag{57}$$

with equality only at $\theta = 0$. Putting (56) and (57) together, $\mathsf{Tr}(I_Y(\theta)^{-1})$ is lower bounded by the constant $\frac{d^2}{\mathsf{Tr}(I_Y(0))}$ and there is equality only at $\theta = 0$. □

Finally, using these observations about the Fisher information, we can write down an upper bound on the minimax rate that matches the lower bound from Theorem 1.

Theorem 2. *For the power law decay model in* (11) *with uniformly distributed sensor locations, if $d \leq 2(\alpha+1)$, then*

$$\inf_{\{\widehat{\theta}_n\}} \sup_{\theta \in U} \lim_{n \to \infty} n \mathbb{E}_{Y^n, S^n | \theta} \left\| \widehat{\theta}_n - \theta \right\|^2 \leq \frac{d^2 \sigma^2 R^{2\alpha+2}}{\alpha^2 PK} \tag{58}$$

where the outer infimum is over all sequences of measurable estimators $\{\widehat{\theta}_n\}$, and the supremum is over all $\theta \in U$ as defined in Lemma 2.

Proof. The result follows immediately from (41) along with Lemma 2. □

The minimax rate $\frac{d^2 \sigma^2 R^{2\alpha+2}}{n\alpha^2 PK}$ suggests the following "rule of thumb" for the point-source localization problem: if sensors are twice as far away from a dipole that you are trying localize, you can expect the worst case mean-squared localization error to increase by a factor of $2^{2\alpha+2} = 2^6$, and this corresponds to a factor of $\sqrt{2^6} = 8$ in the Euclidean distance localization error.

While we only formally establish the origin as the worst case locally, we can numerically verify how the Fisher information behaves as we move radially from the surface of the sphere for any individual choice of d, α. In Figure 1, we numerically compute the curve of how $\mathsf{Tr}(I_Y(\theta)^{-1})$

varies across depth for the point-charge ($\alpha = 1$) and dipole ($\alpha = 2$) cases across different numbers of dimensions d. The trace of the inverse matrix behaves similarly to the trace of the original matrix as described in Lemma 1, in that there is a change in behavior exactly at the critical $d = 2(\alpha + 2)$.

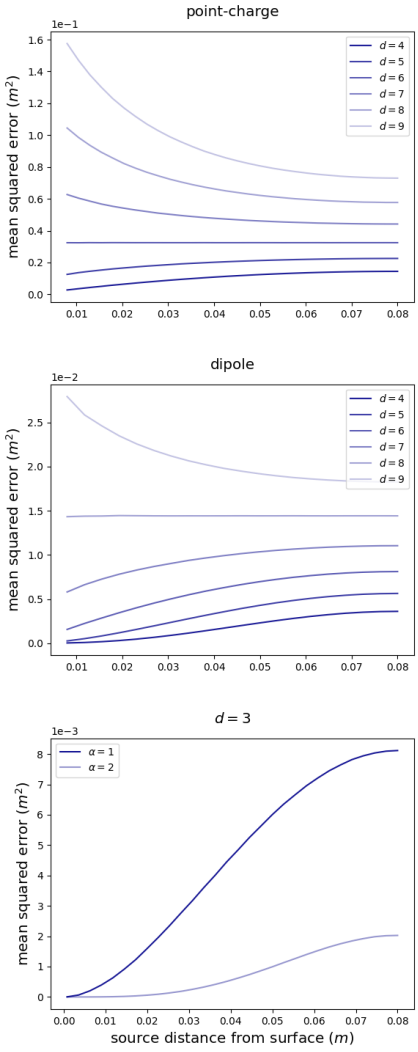

Figure 1: The trace of the inverse Fisher information matrix as a function of source distance from center in $\alpha = 1$ (top) and $\alpha = 2$ (middle) cases for various dimensions. We use $n = 1$, $K = 71$, $R = .08m$, and $\frac{P}{\sigma^2}$ set so that signal power to noise variance is 0.1 at $\theta = 0$. The bottom plot compares this same quantity for $\alpha = 1, 2$ in the case $d = 3$.

5 SIMULATIONS

To investigate how the Fisher information varies with source depth in EEG, we utilize the high-resolution New York head model from Huang et al. (2016). This standardized finite-element head model is based on the averaged anatomy of 152 MRI scans, with the goal of making its geometry and tissue segmentation representative of the general population. In our analysis, this head model provides the biophysical basis for EEG forward modeling, which links source activities inside the brain to scalp potentials recorded by EEG sensors. The resulting model is linear and is represented by a matrix multiplication

$$Y = Ls_\theta + Z. \tag{59}$$

The vector $s_\theta \in \mathbb{R}^M$ encodes the single-timepoint activities in a discretized volume of M brain sources, with $s_{\theta,i}$ specifying the dipole strength at position θ_i. The forward matrix $L \in \mathbb{R}^{K \times M}$ maps source space to sensor space, where each column characterizes the scalp potential vector resulting from a unit activation at the corresponding source location. $Y \in \mathbb{R}^K$ thus denotes the vector of EEG measurements at a single time point across K scalp sensors, and $Z \in \mathbb{R}^K$ models additive Gaussian noise. This formulation can be easily extended to model time series data.

The forward matrix L encodes the geometric and conductivity properties of the head. In this work, we use the New York head model segmented into six tissue types, each with isotropic conductivity (indicated in brackets): white matter (0.126 S/m), gray matter (0.276 S/m), cerebrospinal fluid (1.65 S/m), skull (0.01 S/m), scalp (0.465 S/m), and air (2.5e-14 S/m)—with a spatial resolution of 0.5mm. The computation of L is performed using the FieldTrip toolbox (Oostenveld et al., 2011), which discretizes the head volume into meshes using the finite element method and solves the quasi-static Poisson equation to determine the potential at the scalp sensors. We use $K = 71$ sensing channels on the scalp (Fig. 2). A total of $M = 13,318$ sources are defined within the brain. Computations were run on a 64-core CPU (AMD Ryzen Threadripper Pro 3995WX) with 512 GB of RAM. Construction of L took three days to run, which dominated the rest of the compute time needed.

To investigate how the Fisher information quantity varies with source depth, we construct source arrays originating from the cortical surface and extending toward the brain center. For each array, sources are sampled at 1 cm intervals, starting from a cortical surface source inward to the brain center, up to a depth of 6 cm, corresponding to the minimum distance from any surface point to the brain center. Cortical surface points are sampled to provide approximately even coverage of array orientations across the head. For simplicity, all source dipoles are assumed to have a radial orientation, directed outward from the brain center.

We calculate the Fisher information quantity according to (4). Spatial derivatives are approximated using finite differences with neighboring sources along the RAS anatomical coordinate system (R: right, A: anterior, S: superior, see Figure 2), up to 1.5 cm away. The noise variance σ^2

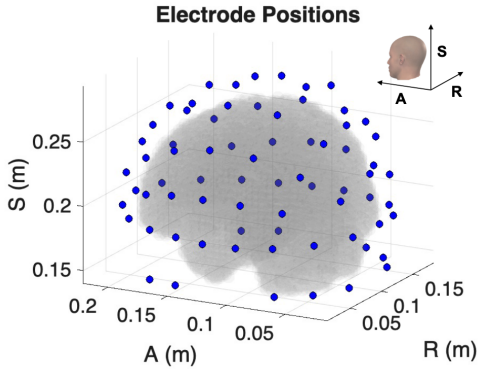

Figure 2: Scalp sensor positions (blue) and brain sources (gray) in the RAS coordinate system, with axes R (right), A (anterior), and S (superior). The head schematic (courtesy of Complete Anatomy software) illustrates the orientation of these axes relative to the brain.

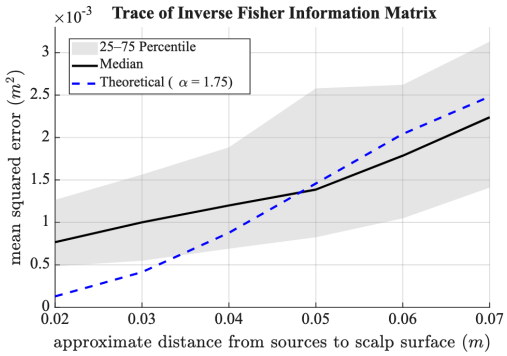

Figure 3: Trace of the inverse Fisher information matrix as a function of approximate source distance to the scalp surface. Sources at each depth are spaced by approximately 1 cm, and the x-axis does not start at 0 due to skull thickness separating brain tissue from the scalp.

is chosen such that resulting average signal-to-noise ratio is 0.1 when the brain-center source is activated with unit strength. Figure 3 shows the trace of the inverse Fisher information matrix, with the median and interquartile range across 2,649 sampled sources at six depth levels. It also shows the theoretical curve $\mathsf{Tr}(I_Y(\theta)^{-1})$, like from Figure 1, that fits the median best (least squares) with $d = 3$, which is at $\alpha = 1.75$. Figure 4 shows the locations of the sampled sources colored by the trace of inverse Fisher information matrix. Sources located near the cortical surface generally have smaller values. In contrast, sources closer to the brain center and those near the bottom of the brain exhibit much larger values. This is because medial sources are far from all sensors and therefore difficult to record from, while ventral sources are especially hard to detect since there are no sensors placed below the skull base.

The trace of the inverse Fisher information matrix in Figure 1 exhibits a qualitatively similar trend to those of Fig-

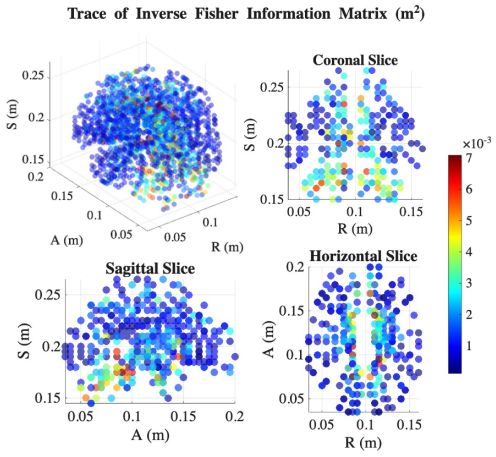

Figure 4: Trace of the inverse Fisher information matrix across sampled brain sources. The color scale is truncated at the 99th percentile due to outliers.

ure 1 (for $d = 3$ and $1 \leq \alpha \leq 2$), and the magnitudes of the quantities are comparable. We would not necessarily expect perfect correspondence – the non-uniform nature of the realistic head model should account for local variations in geometry and conductances throughout the brain. This also manifests as the considerable variability of the quantities in Figure 3 at each given depth.

6 CONCLUSIONS

In this paper, we framed the point-source localization problem as a parametric statistical problem, and derived the minimax rates using Fisher information. We then computed similar Fisher information quantities in a realistic head model for EEG to build understanding of the fundamental limits for real neural sensing problems. There are several ways to extend our results in future research, e.g.: (i) Incorporating spatial correlation models of noise based on realistic models/data (see "brain noise" in Grover and Venkatesh (2017)). The idea would be to add noise at each sensor with correlation that depends on how close the sensors are to each other. This would be particularly relevant in the high-K regime to understand the limits there; (ii) Utilizing these bounds to optimize sensor placement, and getting as close as possible to the fundamental limits with few sensors; (iii) Examining which of the existing EEG/MEG source-localization algorithms, if any, approach fundamental limits in the limit of high sensor density.

References

Barnes, L. P. and Özgür, A. (2021). Fisher information and mutual information constraints. In *Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2179–2184.

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature reviews neuroscience*, 13(6):407–420.

Chen, W.-N. and Özgür, A. (2024). Lq lower bounds on distributed estimation via fisher information. In *Proceedings of the 2024 IEEE International Symposium on Information Theory (ISIT)*, pages 91–96.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton Univ. Press.

Gill, R. D. and Levit, B. Y. (1995). Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, 1(1–2):59–79.

Grover, P. and Venkatesh, P. (2017). An information-theoretic view of EEG sensing. *Proceedings of the IEEE*, 105(2):367–384.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194.

Hämäläinen, M. S. and Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32(1):35–42.

Huang, Y., Parra, L. C., and Haufe, S. (2016). The new york head—a precise standardized volume conductor model for EEG source localization and tES targeting. *NeuroImage*, 140:150–162.

Jatoi, M. A., Kamel, N., Malik, A. S., Faye, I., and Begum, T. (2014). A survey of methods used for source localization using EEG signals. *Biomedical Signal Processing and Control*, 11:42–52.

Kay, S. M. (1993). *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc.

LeCam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828.

Lee, C.-Z., Barnes, L. P., and Özgür, A. (2021). Over-the-air statistical estimation. *IEEE Journal on Selected Areas in Communications*, 40(2):548–561.

Lee, H. and Choi, S. (2009). Group nonnegative matrix factorization for EEG classification. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5, pages 320–327. PMLR.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer.

Michel, C. M. and He, B. (2019). EEG source localization. *Handbook of clinical neurology*, 160:85–101.

Mosher, J. C., Spencer, M. E., Leahy, R. M., and Lewis, P. S. (1993). Error bounds for EEG and MEG dipole source localization. *Electroencephalography and clinical Neurophysiology*, 86(5):303–321.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational intelligence and neuroscience*, 2011:1–9.

Pascual-Marqui, R. D. et al. (2002). Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. *Methods find exp clin pharmacol*, 24(Suppl D):5–12.

Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. (1994). Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of psychophysiology*, 18(1):49–65.

Radhakrishna Rao, C. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91.

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112.

Sun, R., Sohrabpour, A., Worrell, G. A., and He, B. (2022). Deep neural networks constrained by neural mass models improve electrophysiological source imaging of spatiotemporal brain dynamics. *Proceedings of the National Academy of Sciences*, 119(31):1–12.

Taylor, M. E. (2020). *Introduction to Analysis in Several Variables: Advanced Calculus*, volume 46. American Mathematical Soc. `https://mtaylor.web.unc.edu/wp-content/uploads/sites/16915/2018/04/analmv.pdf`.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Uhrmann-Klingen, E. (1995). Minimal fisher information distributions with compact-supports. *Sankhyā: The Indian Journal of Statistics, Series A*, 57(3):360–374.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press.

Venkatesh, P. and Grover, P. (2017). Lower bounds on the minimax risk for the source localization problem. In *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3080–3084.

Xu, A. and Coleman, T. (2020). Minimax lower bounds for circular source localization. In *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1242–1247.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] **Yes.** In Section 2 we describe the mathematical setting for our point source localization problem.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] **Yes.** Sections 3 and 4 are devoted to analyzing the sample complexity of the estimation/localization problem.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] **Yes.** The code for our plots and simulations is provided with the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] **Yes.**
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] **Yes.** All proofs are provided in the main text, except for a detail regarding the convergence of the mean-squared error for the maximum likelihood estimator, which is provided in the appendix.
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] **Yes.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] **Yes.** The code and instructions needed for our plots and simulations is provided with the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] **Not Applicable**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] **Yes.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] **Yes.** We describe the compute used for the simluations in Section 5.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] **Yes.** We use a realistic head model by Huang et al. (2016) as well as the FieldTrip toolbox (Oostenveld et al., 2011) and cite both of them in the text.
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] **Not Applicable**

Appendix

A LIMITING MEAN-SQUARED ERROR RATE OF THE MLE

Recall that we have the maximum likelihood estimator

$$\widehat{\theta}_n \in \arg\max_{\theta \in \bar{U}} p(\mathbf{y}, \mathbf{s}|\theta) \ . \tag{60}$$

By $\bar{U}$ we mean the closure of U, and by continuity the likelihood will achieve its maximum on the compact set $\bar{U}$.

First note that the model $p(y, s|\theta)$ is identifiable in the sense that $p(y, s|\theta)$ is not identical to $p(y, s|\theta^*)$ for any $\theta \neq \theta^*$. This can be seen by fixing a collection of s values and considering the conditional distributions $p(y|\theta, s)$ which are Gaussian with a particular mean. Each sensor position s_i considered specifies a distance $\|s_i - \theta\|$ via the mean $x_i(\theta)$. Once sufficiently many sensor positions have been considered, θ is fully specified (i.e., its position is "triangulated"). The identifiability property means that $\widehat{\theta}_n$ is a consistent estimator of θ (see Lemma 5.35 and Theorem 5.14 in Van der Vaart (2000)).

Consider a first order Taylor expansion of

$$\ell_{\mathbf{y},\mathbf{s}}(\widehat{\theta}_n) = \nabla_\theta \log p(\mathbf{y}, \mathbf{s}|\theta)\big|_{\theta=\widehat{\theta}_n} \tag{61}$$

at the true θ,

$$\ell_{\mathbf{y},\mathbf{s}}(\theta) + \mathsf{Hess}_\theta(\log p(\mathbf{y}, \mathbf{s}|\theta))(\widehat{\theta}_n - \theta) + R = \ell_{\mathbf{y},\mathbf{s}}(\widehat{\theta}_n) \tag{62}$$

where each component of the remainder vector R can be written as

$$R_i = \frac{1}{2}(\widehat{\theta}_n - \theta)^T \mathsf{Hess}_\theta \left(\frac{\partial}{\partial \theta_i} \log p(\mathbf{y}, \mathbf{s}|\theta) \right) \Big|_{\theta=\theta^*} (\widehat{\theta}_n - \theta) \tag{63}$$

and θ^* is a point on the line between θ and $\widehat{\theta}_n$ (with possibly a different θ^* for each component i, j, k). Multiplying (62) by $1/n$,

$$\frac{1}{n}\ell_{\mathbf{y},\mathbf{s}}(\theta) + \frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y}, \mathbf{s}|\theta))(\widehat{\theta}_n - \theta) + \frac{1}{n}R = \frac{1}{n}\ell_{\mathbf{y},\mathbf{s}}(\widehat{\theta}_n) \ . \tag{64}$$

Taking the third partial derivatives of $\log p(\mathbf{y}, \mathbf{s}|\theta)$, we get

$$\left| \frac{1}{n} \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log p(\mathbf{y}, \mathbf{s}|\theta) \right|_{\theta=\theta^*} \leq C_0 + C_1 \frac{1}{n} \sum_{i=1}^{n} \|y^i\|_1 \tag{65}$$

where the constants C_0, C_1 depend on the maximum possible magnitude of some derivatives of $x_j(\theta)$ over all $\theta \in \bar{U}$. The random vectors Y^i are sub-Gaussian under the true probability distribution $p(y, s|\theta) = p(y|s, \theta)p(s)$ since $p(y|s, \theta)$ is Gaussian with means and variances bounded across all s values. Therefore, each entry of $\frac{1}{n}\mathsf{Hess}_\theta \left(\frac{\partial}{\partial \theta_i} \log p(\mathbf{y}, \mathbf{s}|\theta) \right)_{\theta=\theta^*}$ has a sub-Gaussian tail bound with

$$\mathsf{Pr}_\theta \left(\left| \frac{1}{n} \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \log p(\mathbf{y}, \mathbf{s}|\theta) \right|_{\theta=\theta^*} > C_0' + t \right) \leq e^{-ct^2 n} \ . \tag{66}$$

The above holds no matter what θ^* is (provided it is in $\bar{U}$), and the probability is over the distribution with respect to the true parameter θ.

Note the following additional continuity and concentration conditions, which all hold for our statistical model $p(y, s|\theta)$:

i. each entry of $\frac{1}{\sqrt{n}}\ell_{\mathbf{y},\mathbf{s}}(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta \log p(y^i, s^i|\theta)$ is sub-Gaussian with variance $O(1)$

ii. each entry of $\frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y},\mathbf{s}|\theta)) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{Hess}_\theta(\log p(y^i, s^i|\theta))$ is sub-Gaussian with variance $O(1/n)$

iii. $\mathbb{E}_{Y,S|\theta}[\mathsf{Hess}_\theta(\log p(Y,S|\theta))] = -I_Y(\theta)$ is continuous in θ

The first two properties follow from the first two derivatives (with respect to θ) of the function $\log p(y,s|\theta)$ being sub-Gaussian. This is because they are Gaussian under the law of y conditioned on s, and are therefore a mixture of Gaussians (with bounded means and variances). The last property follows by our equivalent form for $I_Y(\theta)$ and using dominated convergence to take limits inside the integral with respect to $\mathsf{d}s$.

We write

$$\frac{1}{n}R = M(\widehat{\theta}_n - \theta) \tag{67}$$

and formally manipulate (64) to get

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \left(\frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y},\mathbf{s}|\theta)) + M\right)^{-1}\frac{1}{\sqrt{n}}\left(\ell_{\mathbf{y},\mathbf{s}}(\widehat{\theta}_n) - \ell_{\mathbf{y},\mathbf{s}}(\theta)\right) . \tag{68}$$

We are going to break the probability space into three disjoint events, the first of which is

$$E_1 = \Bigg\{\mathbf{y},\mathbf{s} : \exists \text{ an entry of } \frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y},\mathbf{s}|\theta)) \text{ that varies by more than } \frac{1}{n^{1/3}} \text{from its mean or} \tag{69}$$

$$\exists \text{ an entry of } \frac{1}{n}\mathsf{Hess}_\theta\left(\frac{\partial}{\partial\theta_i}\log p(\mathbf{y},\mathbf{s}|\theta)\right)\Bigg|_{\theta=\theta^*} \text{with magnitude greater than } C_0' + \frac{1}{n^{1/3}}\Bigg\} . \tag{70}$$

The key properties of event E_1 are that it has exponentially decreasing probability and that for sufficiently large n, outside of E_1 the matrix

$$\left(\frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y},\mathbf{s}|\theta)) + M\right) \tag{71}$$

has a minimum singular value $\sigma_0 > 0$. For the minimum singular value property, consider the following. $I_Y(0)$ is invertible with some nonzero minimal singular value and $I_Y(\theta)$ is continuous. Shrinking U if necessary, $I_Y(\theta)$ has a nonzero minimal singular value across all $\theta \in U$, and then for sufficiently large n we have $\frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y},\mathbf{s}|\theta))$ has a nonzero minimal singular value outside of E_1. Furthermore, the entries of M are bounded outside of E_1 and can be made small enough by shrinking U.

Because the maximum likelihood estimator is consistent, $\mathsf{Pr}(\|\widehat{\theta}_n - \theta\| > \varepsilon_n) \leq \varepsilon_n$ for some $\varepsilon_n \to 0$. We define

$$E_2 = E_1^C \cap \{\mathbf{y},\mathbf{s} : \|\widehat{\theta}_n - \theta\| > \varepsilon_n\} , \tag{72}$$

and

$$E_3 = E_1^C \cap E_2^C . \tag{73}$$

We have

$$n\mathbb{E}_{Y^n,S^n|\theta}\|\widehat{\theta}_n - \theta\|^2 = \int_{E_1} n\|\widehat{\theta}_n - \theta\|^2 \mathsf{d}P_{Y^n,S^n|\theta} \tag{74}$$

$$+ \int_{E_2} n\|\widehat{\theta}_n - \theta\|^2 \mathsf{d}P_{Y^n,S^n|\theta} \tag{75}$$

$$+ \int_{E_3} n\|\widehat{\theta}_n - \theta\|^2 \mathsf{d}P_{Y^n,S^n|\theta} . \tag{76}$$

Using property (ii) and (66), and a union bound over entries, the probability of E_1 is less than $2(d^2+d^3)e^{-cn^{\frac{1}{3}}}$ for some $c > 0$ related to the sub-Gaussian constants. Therefore (74) can be bounded by

$$\int_{E_1} n\|\widehat{\theta}_n - \theta\|^2 \mathsf{d}P_{Y^n,S^n|\theta} \leq 8R^2 n(d^2+d^3)e^{-cn^{\frac{1}{3}}} \to 0 . \tag{77}$$

By Cauchy-Schwarz, consistency, and (68), we can bound (75) by,

$$\int 1_{E_2} \cdot n\|\widehat{\theta}_n - \theta\|^2 \mathsf{d}P_{Y^n,S^n|\theta} \leq \frac{\sqrt{\varepsilon_n}}{\sigma_0^2}\sqrt{\mathbb{E}_{Y^n,S^n|\theta}\left\|\frac{1}{\sqrt{n}}\left(\ell_{\mathbf{y},\mathbf{s}}(\widehat{\theta}_n) - \ell_{\mathbf{y},\mathbf{s}}(\theta)\right)\right\|^4} \to 0 \ , \tag{78}$$

where the convergence to zero is due to property (i) bounding the fourth moment term by a constant. Finally we are left with the "nice" set E_3. In this set, for sufficiently large n, we have that $\widehat{\theta}_n$ is in the interior of U and therefore must satisfy $\ell_{\mathbf{y},\mathbf{s}}(\widehat{\theta}_n) = 0$. We also have

$$\left(\frac{1}{n}\mathsf{Hess}_\theta(\log p(\mathbf{y},\mathbf{s}|\theta)) + M\right) \to -I_Y(\theta) \ . \tag{79}$$

Therefore

$$\begin{aligned}
\int_{E_3} n\|\widehat{\theta}_n - \theta\|^2 \mathsf{d}P_{Y^n,S^n|\theta} &\to \int \left\|I_Y(\theta)^{-1}\frac{1}{\sqrt{n}}\ell_{\mathbf{y},\mathbf{s}}(\theta)\right\|^2 \mathsf{d}P_{Y^n,S^n|\theta} && (80)\\
&= \int \mathsf{Tr}\left(I_Y(\theta)^{-1}\frac{1}{n}\ell_{\mathbf{y},\mathbf{s}}(\theta)\ell_{\mathbf{y},\mathbf{s}}(\theta)^T I_Y(\theta)^{-T}\right)\mathsf{d}P_{Y^n,S^n|\theta} && (81)\\
&= \mathsf{Tr}\left(I_Y(\theta)^{-1}\int \frac{1}{n}\ell_{\mathbf{y},\mathbf{s}}(\theta)\ell_{\mathbf{y},\mathbf{s}}(\theta)^T \mathsf{d}P_{Y^n,S^n|\theta}\ I_Y(\theta)^{-T}\right) && (82)\\
&= \mathsf{Tr}(I_Y(\theta)^{-1}) \ . && (83)
\end{aligned}$$

We can justify the interchange of limits to get (80) by using the minimum singular value σ_0 outside of E_1 and property (i).