Gradient of Error Probability of M-ary Hypothesis Testing Problems under Multivariate Gaussian Noise

Minoh Jeong*, Alex Dytso†, Martina Cardone*
* University of Minnesota, Minneapolis, MN 55404, USA, Email: {jeong316, cardo089}@umn.edu
† New Jersey Institute of Technology, Newark, NJ 07102, USA Email: alex.dytso@njit.edu

***Abstract*—This letter considers an M-ary hypothesis testing problem on an n-dimensional random vector perturbed by the addition of Gaussian noise. A novel expression for the gradient of the error probability, with respect to the covariance matrix of the noise, is derived and shown to be a function of the cross-covariance matrix between the noise matrix (i.e., the matrix obtained by multiplying the noise vector by its transpose) and Bernoulli random variables associated with the correctness event.**

I. Introduction

Hypothesis testing is a well-defined problem setting for any detection or estimation problem and hence it is broadly used in various areas, such as signal processing [1], information theory [2], regression theory [3], [4] and biostatistics [5], [6].

The performance of a hypothesis testing problem is measured in terms of the error probability, which is determined by the used decision criterion. In the binary hypothesis testing problem, the *optimum* decision criterion (i.e., the one that leads to the smallest error probability) is derived in [7] and shown to be the so-called likelihood ratio test. For the case of an M-ary hypothesis testing problem, the optimum decision criterion can be obtained by minimizing the Bayes risk [8], which is the maximum a posteriori probability decision. However, in general settings characterizing the optimum decision regions is not an easy task, which in turn leads to a very few existing results on the minimum error probability. For instance, in [9], the authors characterized the minimum error probability of an M-ary hypothesis testing problem via two alternative expressions. In [10], the authors provided upper and lower bounds on the error probability of an M-ary hypothesis testing problem in terms of the Arimoto-Rényi conditional entropy.

In this letter, we consider an M-ary hypothesis testing problem on an n-dimensional random vector $\mathbf{X}$, which is perturbed by the addition of Gaussian noise. We are interested in analyzing the performance of this hypothesis testing problem in terms of the error probability under the optimal decision criterion. The main merit of our work is the derivation of a novel expression for the gradient of the error probability, which can be obtained as a function of the cross-covariance matrix between the noise matrix (i.e., the matrix obtained by multiplying the noise vector by its transpose) and Bernoulli random variables associated with the correctness event. This result can be leveraged to study the first-order behavior (in terms of the noise variance) of the error probability. For example, in practice, the first-order behavior of the error probability is often estimated by using Monte Carlo simulations, by first estimating the error probability itself, and then approximating the derivative by using finite differences [11], [12]. For such a procedure to be consistent, one needs to estimate the error probability at multiple values of the noise variance, which can lead to a large sample complexity. Our result allows for an alternative procedure that enables to directly study the derivative, and hence can potentially lead to a lower sample complexity.

The work of M. Jeong and M. Cardone was supported in part by the U.S. National Science Foundation under Grant CCF-1849757.

II. Problem Statement and Main Result

Notation. Boldface upper case letters $\mathbf{X}$ denote vector random variables; the boldface lower case letter $\mathbf{x}$ indicates a specific realization of $\mathbf{X}$; $\mathbb{E}[\cdot]$ is the expectation with respect to the joint distribution of all the (vector) random variables inside the brackets. $[n_1 : n_2]$ is the set of integers from n_1 to $n_2 \geq n_1$; 0_n is the column vector of dimension n of all zeros; $0_{n\times n}$ is the matrix of dimension $n \times n$ of all zeros; I_n is the identity matrix of dimension $n \times n$; for a square matrix A, A^{-1} is its inverse, $\mathrm{Tr}\,[\mathrm{A}]$ is its trace, and $\det(A)$ is its determinant; $\|\mathbf{x}\|$ is the ℓ_2 norm of $\mathbf{x}$, and $\mathbf{x}^T$ is the transpose of $\mathbf{x}$; $\langle\cdot,\cdot\rangle$ is the inner product operator. Calligraphic letters indicate sets; $1_{\mathcal{A}}(\mathbf{x})$ is the indicator function that has value 1 for $\mathbf{x} \in \mathcal{A}$ and value 0 for $\mathbf{x} \notin \mathcal{A}$; $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$.

We consider a framework in which an n-dimensional random vector $\mathbf{X}$ is generated according to the distribution $p_{\mathbf{X}}(\cdot)$ and then passed through an additive Gaussian noise channel. The output of the channel is denoted by $\mathbf{Y}$ and given by

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{N}$ – which is independent of $\mathbf{X}$ – is an n-dimensional Gaussian random vector with zero mean and covariance matrix $K_{\mathbf{N}}$. We also assume that $p_{\mathbf{X}}(\cdot)$ is not a function of $K_{\mathbf{N}}$.

Given the observation of $\mathbf{Y}$ in (1), we are interested in analyzing the probability of error of an M-ary hypothesis testing problem under the optimal decision rules [7], [8]. A standard M-ary hypothesis testing problem consists of the following:

1) A collection of *hypothesis regions* to be denoted by
$$\mathcal{H}_i \subset \mathbb{R}^n,\ i \in [1 : M];\ \text{and} \tag{2}$$
2) A collection of *decision regions* to be denoted by
$$\mathcal{R}_{i,K_{\mathbf{N}}} \subset \mathbb{R}^n,\ i \in [1 : M], \tag{3}$$
where the region $\mathcal{R}_{i,K_{\mathbf{N}}}$ corresponds to declaring $\mathcal{H}_i$.

Remark 1. *We assume that both hypothesis and decision regions are partitions. Moreover, $\mathcal{R}_{i,K_{\mathbf{N}}}, i \in [1:M]$ in (3) highlights the fact that the M decision regions can be a function of the noise covariance matrix $K_{\mathbf{N}}$.*

For a given collection of hypothesis and decision regions, the probability of error is given by

$$\begin{aligned} p_e(n, K_{\mathbf{N}}) &= \sum_{i\in[1:M]} \Pr\left(\mathbf{Y} \notin \mathcal{R}_{i,K_{\mathbf{N}}} | \mathcal{H}_i\right) \Pr(\mathcal{H}_i) \\ &= 1 - \sum_{i\in[1:M]} \Pr\left(\mathbf{Y} \in \mathcal{R}_{i,K_{\mathbf{N}}} | \mathcal{H}_i\right) \Pr(\mathcal{H}_i). \quad (4) \end{aligned}$$

The next theorem (the proof of which is provided in Section III) is the main result of this work. The theorem provides an expression for the gradient of the error probability $p_e(n, K_{\mathbf{N}})$ of an M-ary hypothesis testing problem.

Theorem 1. *Given an observation $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where $\mathbf{N} \sim \mathcal{N}(0_n, K_{\mathbf{N}})$ with positive definite $K_{\mathbf{N}}$, consider a hypothesis testing problem with M hypotheses $\mathcal{H}_i, i \in [1:M]$ on $\mathbf{X}$, and corresponding decision regions $\mathcal{R}_{i,K_{\mathbf{N}}}$. Then,*

$$\nabla_{K_{\mathbf{N}}} p_e(n, K_{\mathbf{N}}) = -\frac{K_{\mathbf{N}}^{-1} \mathrm{Cov}\left\{\mathbf{N}\mathbf{N}^{\mathrm{T}}, 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\} K_{\mathbf{N}}^{-1}}{2}, \quad (5)$$

where

$$\mathcal{A} = \bigcup_{i\in[1:M]} \{(\mathbf{X}, \mathbf{Y}) | \mathbf{X} \in \mathcal{H}_i, \mathbf{Y} \in \mathcal{R}_{i,K_{\mathbf{N}}}\} \quad (6)$$

is the event of correctness and where $\mathrm{Cov}\left\{\mathbf{N}\mathbf{N}^{\mathrm{T}}, 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\}$ is the cross-covariance matrix between $\mathbf{N}\mathbf{N}^{\mathrm{T}}$ and $1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})$.

By applying the fundamental theorem of calculus on (5), we readily obtain the next corollary.

Corollary 1. *Assume positive definite $K_{\mathbf{N},1}$ and $K_{\mathbf{N},2}$. Let $\tilde{\mathbf{N}} \sim \mathcal{N}(0_n, \tilde{K})$ and $\tilde{\mathbf{Y}} = \mathbf{X} + \tilde{\mathbf{N}}$. Then,*

$$\begin{aligned} &p_e(n, K_{\mathbf{N},2}) - p_e(n, K_{\mathbf{N},1}) \\ &= -\int_{\mathcal{C}} \left\langle \frac{\tilde{K}^{-1} \mathrm{Cov}\left\{\tilde{\mathbf{N}}\tilde{\mathbf{N}}^{\mathrm{T}}, 1_{\mathcal{A}}(\mathbf{X}, \tilde{\mathbf{Y}})\right\} \tilde{K}^{-1}}{2}, \mathrm{d}\tilde{K} \right\rangle, \quad (7) \end{aligned}$$

where $\mathcal{C}$ is an arbitrary path from $K_{\mathbf{N},1}$ to $K_{\mathbf{N},2}$ that preserves the positive definite property.

The following lemma simplifies (5) when $K_{\mathbf{N}} = \sigma^2 I_n$.

Lemma 1. *For the case of independent and identically distributed Gaussian noise, i.e., $\mathbf{N} \sim \mathcal{N}(0_n, \sigma^2 I_n)$, we have*

$$\frac{\partial}{\partial \sigma^2} p_e(n, \sigma^2 I_n) = -\frac{1}{2\sigma^4} \mathrm{Cov}\left\{||\mathbf{N}||^2, 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\}. \quad (8)$$

Proof: When $K_{\mathbf{N}} = \sigma^2 I_n$, the expression in (5) becomes

$$\nabla_{K_{\mathbf{N}}} p_e(n, \sigma^2 I_n) = -\frac{1}{2\sigma^4} \mathrm{Cov}\left\{\mathbf{N}\mathbf{N}^T, 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\}. \quad (9)$$

Moreover, we have that

$$\frac{\partial}{\partial \sigma^2} p_e(n, \sigma^2 I_n) = \mathrm{Tr}\left[\left[\nabla_{K_{\mathbf{N}}} p_e(n, \sigma^2 I_n)\right]^T \frac{\partial \sigma^2 I_n}{\partial \sigma^2}\right]$$

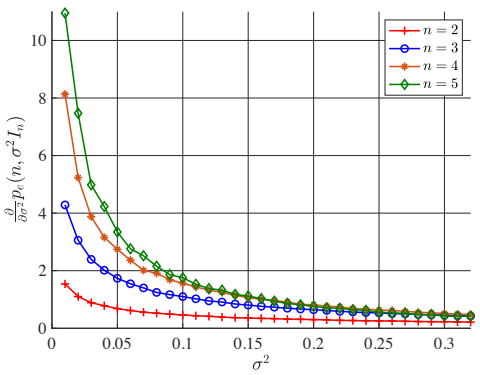

Fig. 1: Derivative of the error probability for the permutation recovery problem versus σ^2 for different values of n.

$$\begin{aligned} &= -\frac{1}{2\sigma^4} \mathrm{Tr}\left[\mathrm{Cov}\left\{\mathbf{N}\mathbf{N}^{\mathrm{T}}, 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\}\right] \\ &= -\frac{1}{2\sigma^4} \mathrm{Cov}\left\{\mathrm{Tr}\left[\mathbf{N}\mathbf{N}^T\right], 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\} \\ &= -\frac{1}{2\sigma^4} \mathrm{Cov}\left\{||\mathbf{N}||^2, 1_{\mathcal{A}}(\mathbf{X}, \mathbf{Y})\right\}, \quad (10) \end{aligned}$$

where the first equality follows by using the chain rule in [13, eq.(137)], and the second equality follows by using (9). ■

We next present an example that showcases how our result can be used to study limiting behaviors of the error probability.

Example. Let $\mathbf{X} \sim \mathcal{N}(0_n, I_n)$ and $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ with $\mathbf{N} \sim \mathcal{N}(0_n, \sigma^2 I_n)$. Given the noisy observation $\mathbf{Y}$, according to which permutation (among the $n!$ possible ones) was $\mathbf{X}$ sorted? This question falls within the topic of recovering the structure (i.e., permutation) of noisy data, as recently analyzed in [14] and references therein. In [15], the authors characterized the structure of the optimal decision regions by showing that $\mathcal{R}_{i,K_{\mathbf{N}}} = \mathcal{H}_i, i \in [1:n!]$. However, in [15] a question remained open: How does the error probability behave in the low noise regime? Using the result in (5), we can gain insights helpful to answer this question, as we describe next. Fig. 1 empirically shows the behavior of $\frac{\partial}{\partial \sigma^2} p_e(n, \sigma^2 I_n)$ versus different values of σ^2 and for several values of n. From Fig. 1, we observe that the error probability has a steep behavior with respect to σ^2, hence suggesting that even a slight increase of the noise variance results in large increases of the error probability. We also observe that this noisy behavior becomes more remarkable as n increases and that the slope seems to be infinite when $\sigma \to 0$. Consequently, we have obtained some insights into the behavior of the minimum probability of error. Note that obtaining similar insights (e.g., slope of the error probability) by computing the probability of error using the Monte Carlo method would have been a more computationally involved task. □

III. Proof of Theorem 1

We start by noting that the probability of correctness associated with the hypothesis $\mathcal{H}_i, i \in [1:M]$ is

$$p_c(n, K_{\mathbf{N}}, \mathcal{H}_i) = \int_{\mathbf{y}\in\mathbb{R}^n} 1_{\mathcal{R}_{i,K_{\mathbf{N}}}}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}, \mathcal{H}_i)\, \mathrm{d}\mathbf{y},$$

where $f_{\mathbf{Y}}(\mathbf{y}, \mathcal{H}_i)$ is defined as

$$f_{\mathbf{Y}}(\mathbf{y}, \mathcal{H}_i) = f_{\mathbf{Y}}(\mathbf{y}|\mathcal{H}_i) \Pr(\mathcal{H}_i),\ i \in [1:M], \quad (11)$$

and $\mathcal{R}_{i,K_{\mathbf{N}}}$ is the optimal decision region for $\mathcal{H}_i$ in (3) when the noise covariance is $K_{\mathbf{N}}$. Then, the probability of correctness of the hypothesis testing problem can be computed as

$$\begin{aligned} p_c(n, K_{\mathbf{N}}) &= \sum_{i\in[1:M]} p_c(n, K_{\mathbf{N}}, \mathcal{H}_i) \\ &= \sum_{i\in[1:M]} \int_{\mathbf{y}\in\mathbb{R}^n} 1_{\mathcal{R}_{i,K_{\mathbf{N}}}}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}, \mathcal{H}_i)\ \mathrm{d}\mathbf{y} \\ &= \int_{\mathbf{y}\in\mathbb{R}^n} \sum_{i\in[1:M]} 1_{\mathcal{R}_{i,K_{\mathbf{N}}}}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}, \mathcal{H}_i)\ \mathrm{d}\mathbf{y}, \end{aligned}$$

where the last equality follows by using the Fubini-Tonelli theorem [16]. Since $p_e(n, K_{\mathbf{N}}) = 1 - p_c(n, K_{\mathbf{N}})$, we obtain

$$p_e(n, K_{\mathbf{N}}) = 1 - \int_{\mathbf{y}\in\mathbb{R}^n} \sum_{i=1}^{M} 1_{\mathcal{R}_{i,K_{\mathbf{N}}}}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}, \mathcal{H}_i) \mathrm{d}\mathbf{y}. \quad (12)$$

We now claim that the gradient of the probability of error with respect to $K_{\mathbf{N}}$ is given by (5). In order to verify this claim, it is sufficient to show the following equation, which follows from the fundamental theorem of calculus for line integral,

$$\begin{aligned} \int_{\mathcal{C}} \langle \nabla_{\mathbf{r}} p_e(n, \mathbf{r}), \mathrm{d}\mathbf{r} \rangle &= p_e(n, K_{\mathbf{N}}) - p_e(n, 0_{n\times n}) \\ &= p_e(n, K_{\mathbf{N}}), \end{aligned} \quad (13)$$

where $\mathcal{C}$ is an arbitrary path from $0_{n\times n}$ to $K_{\mathbf{N}}$ that preserves the positive definite property, and where we let the probability of error be zero for the noiseless case, i.e., when $K_{\mathbf{N}} = 0_{n\times n}$.

By defining $\mathbf{r}(t) = K_{\mathbf{N}_t} = tK_{\mathbf{N}}$, we obtain

$$\begin{aligned} &\int_{\mathcal{C}} \langle \nabla_{\mathbf{r}} p_e(n, \mathbf{r}), \mathrm{d}\mathbf{r} \rangle = \int_0^1 \left\langle \nabla_{\mathbf{r}(t)} p_e(n, \mathbf{r}(t)), \mathbf{r}'(t) \right\rangle \mathrm{d}t \\ &= \int_0^1 \mathrm{Tr}\left[\left[\nabla_{\mathbf{r}(t)} p_e(n, \mathbf{r}(t)) \right]^T \mathbf{r}'(t) \right] \mathrm{d}t, \end{aligned} \quad (14)$$

where the last equality follows since $\mathrm{Tr}[\cdot]$ is the inner product operator over the space of matrices, particularly, $\langle A, B \rangle = \mathrm{Tr}[A^T B] = \mathrm{Tr}[AB^T]$.

With the goal to show that (13) holds, we now substitute (5) inside (14), where remember that $\mathbf{r}(t) = K_{\mathbf{N}_t} = tK_{\mathbf{N}}$, and hence $\mathbf{r}'(t) = K_{\mathbf{N}}$. In order to highlight the fact that the gradient depends on t, we use the notation $\mathbf{N}_t, \mathbf{Y}_t, \mathcal{A}_t$ when the noise covariance $K_{\mathbf{N}_t}$ is used. With this, and by using symmetry of the covariance matrix, we obtain

$$\begin{aligned} &\int_{\mathcal{C}} \langle \nabla_{\mathbf{r}} p_e(n, \mathbf{r}), \mathrm{d}\mathbf{r} \rangle \\ &= \int_0^1 \mathrm{Tr}\left[-\frac{1}{2} K_{\mathbf{N}_t}^{-1} \mathrm{Cov}\left\{ \mathbf{N}_t \mathbf{N}_t^T, 1_{\mathcal{A}_t}(\mathbf{X}, \mathbf{Y}_t) \right\} K_{\mathbf{N}_t}^{-1} K_{\mathbf{N}} \right] \mathrm{d}t \\ &\overset{(a)}{=} \int_0^1 \mathrm{Tr}\left[-\frac{1}{2} \mathbb{E}\left[\Omega(\mathbf{N}_t) 1_{\mathcal{A}_t}(\mathbf{X}, \mathbf{Y}_t) \right] K_{\mathbf{N}} \right] \mathrm{d}t \\ &\overset{(b)}{=} \int_0^1 \mathrm{Tr}\left[-\frac{1}{2} \mathbb{E}\left[\Omega(\mathbf{N}_t) \sum_{i\in[1:M]} 1_{\mathcal{H}_i}(\mathbf{X}) 1_{\mathcal{R}_{i,t}}(\mathbf{Y}_t) \right] K_{\mathbf{N}} \right] \mathrm{d}t \\ &\overset{(c)}{=} \int_0^1 \sum_{i\in[1:M]} \mathrm{Tr}\left[-\frac{1}{2} \mathbb{E}\left[\Omega(\mathbf{N}_t) 1_{\mathcal{H}_i}(\mathbf{X}) 1_{\mathcal{R}_{i,t}}(\mathbf{Y}_t) \right] K_{\mathbf{N}} \right] \mathrm{d}t \\ &\overset{(d)}{=} -\int_0^1 \sum_{i\in[1:M]} \mathrm{Tr}\left[\Pr[\mathcal{H}_i] \mathbb{E}\left[\frac{\Omega(\mathbf{N}_t) 1_{\mathcal{R}_{i,t}}(\mathbf{Y}_t)}{2} \,\middle|\, \mathcal{H}_i \right] K_{\mathbf{N}} \right] \mathrm{d}t, \end{aligned} \quad (15)$$

where the labeled equalities follow from: (a) the definition of covariance and by defining

$$\Omega(\boldsymbol{\omega}) = K_{\mathbf{N}_t}^{-1} \boldsymbol{\omega}\boldsymbol{\omega}^{\mathrm{T}} K_{\mathbf{N}_t}^{-1} - K_{\mathbf{N}_t}^{-1}; \quad (16)$$

(b) letting $\mathcal{R}_{i,t} = \mathcal{R}_{i,K_{\mathbf{N}_t}}$ for shorthand, and using $\mathcal{A}$ defined in (6) by recalling that the hypothesis and decision regions are partitions; (c) the fact that $\mathrm{Tr}[\cdot]$ and $\mathbb{E}[\cdot]$ are linear operators; and (d) the fact that $\mathbb{E}\left[f(\cdot) 1_{\mathcal{S}}(\mathbf{X})\right] = \Pr(\mathbf{X} \in \mathcal{S}) \mathbb{E}\left[f(\cdot) \mid \mathbf{X} \in \mathcal{S}\right]$.

Since $\mathbf{N}_t = \mathbf{Y}_t - \mathbf{X}$, the trace in (15) can be expressed as

$$\begin{aligned} &\mathrm{Tr}\left[\Pr[\mathcal{H}_i] \mathbb{E}\left[\frac{\Omega(\mathbf{Y}_t - \mathbf{X}) 1_{\mathcal{R}_{i,t}}(\mathbf{Y}_t)}{2} \,\middle|\, \mathcal{H}_i \right] K_{\mathbf{N}} \right] \\ &\overset{(a)}{=} \mathrm{Tr}\left[\Pr[\mathcal{H}_i] \mathbb{E}\left[\int_{\mathbf{y}\in\mathcal{R}_{i,t}} \frac{\Omega(\mathbf{y} - \mathbf{X}) f_{\mathbf{N}_t}(\mathbf{y} - \mathbf{X})}{2} \mathrm{d}\mathbf{y} \,\middle|\, \mathcal{H}_i \right] K_{\mathbf{N}} \right] \\ &\overset{(b)}{=} \int_{\mathbf{y}\in\mathcal{R}_{i,t}} \mathrm{Tr}\left[\Pr[\mathcal{H}_i] \mathbb{E}\left[\frac{\Omega(\mathbf{y} - \mathbf{X}) f_{\mathbf{N}_t}(\mathbf{y} - \mathbf{X})}{2} \,\middle|\, \mathcal{H}_i \right] K_{\mathbf{N}} \right] \mathrm{d}\mathbf{y} \\ &\overset{(c)}{=} \int_{\mathbf{y}\in\mathcal{R}_{i,t}} \mathrm{Tr}\left[\nabla_{K_{\mathbf{N}_t}} f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i) K_{\mathbf{N}} \right] \mathrm{d}\mathbf{y} \\ &\overset{(d)}{=} \int_{\mathbf{y}\in\mathcal{R}_{i,t}} \frac{\partial f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i)}{\partial t}\ \mathrm{d}\mathbf{y}, \end{aligned} \quad (17)$$

where the labeled equalities follow from: (a) the definition of expected value; (b) using Fubini-Tonelli theorem [16], which is verified from the fact that $\mathbb{E}\left[\frac{1}{2} \Omega(\mathbf{Y}_t - \mathbf{X}) 1_{\mathcal{R}_{i,t}}(\mathbf{Y}_t) \mid \mathcal{H}_i \right]$ is a finite matrix and the fact that $\mathrm{Tr}[\cdot]$ is a linear operator; (c) using Lemma 2 (below); and (d) using the chain rule for the derivative of structured matrices [13, eq.(137)].

By substituting (17) into (15), we obtain

$$\begin{aligned} &\int_{\mathcal{C}} \langle \nabla_{\mathbf{r}} p_e(n, \mathbf{r}), \mathrm{d}\mathbf{r} \rangle \\ &= -\int_0^1 \sum_{i\in[1:M]} \int_{\mathbf{y}\in\mathcal{R}_{i,t}} \frac{\partial f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i)}{\partial t}\ \mathrm{d}\mathbf{y}\ \mathrm{d}t \\ &= -\int_{\mathbf{y}\in\mathbb{R}^n} \int_0^1 \sum_{i\in[1:M]} \frac{\partial f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i)}{\partial t} 1_{\mathcal{R}_{i,t}}(\mathbf{y})\ \mathrm{d}t\ \mathrm{d}\mathbf{y}, \end{aligned} \quad (18)$$

where the last equality follows by the Fubini-Tonelli theorem [16].

We now analyze the integrand in (18). By using the result in Lemma 3 (below), we can restrict our attention to $\mathbf{y}$'s that belong to the *interior* of $\mathcal{R}_{i,K_{\mathbf{N}}}$. Thus, we have

$$\begin{aligned} &\frac{\partial f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i) 1_{\mathcal{R}_{i,t}}(\mathbf{y})}{\partial t} \\ &= \frac{\partial f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i)}{\partial t} 1_{\mathcal{R}_{i,t}}(\mathbf{y}) + f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i) \frac{\partial 1_{\mathcal{R}_{i,t}}(\mathbf{y})}{\partial t} \\ &= \frac{\partial f_{\mathbf{Y}_t}(\mathbf{y}, \mathcal{H}_i)}{\partial t} 1_{\mathcal{R}_{i,t}}(\mathbf{y}), \end{aligned} \quad (19)$$

where the second equality follows by leveraging the result in Lemma 4 (below), i.e., the continuity property of $1_{\mathcal{R}_{i,t}}(\mathbf{y})$ implies that it is also differentiable (because of the property of indicator function), and hence $\frac{\partial 1_{\mathcal{R}_{i,t}}(\mathbf{y})}{\partial t}=0$. Consequently, we can rewrite (18) as

$$\begin{aligned}
&\int_{\mathcal{C}}\langle\nabla_{\mathbf{r}}p_e(n,\mathbf{r}),\mathrm{d}\mathbf{r}\rangle\\
&=-\int_{\mathbf{y}\in\mathbb{R}^n}\int_0^1\sum_{i\in[1:M]}\frac{\partial f_{\mathbf{Y}_t}(\mathbf{y},\mathcal{H}_i)1_{\mathcal{R}_{i,t}}(\mathbf{y})}{\partial t}\,\mathrm{d}t\,\mathrm{d}\mathbf{y}\\
&=-\int_{\mathbf{y}\in\mathbb{R}^n}\sum_{i\in[1:M]}f_{\mathbf{Y}_1}(\mathbf{y},\mathcal{H}_i)1_{\mathcal{R}_{i,1}}(\mathbf{y})-f_{\mathbf{Y}_0}(\mathbf{y},\mathcal{H}_i)1_{\mathcal{R}_{i,0}}(\mathbf{y})\,\mathrm{d}\mathbf{y}\\
&\overset{(a)}{=}p_e(n,K_{\mathbf{N}})-1-p_e(n,0_{n\times n})+1\\
&=p_e(n,K_{\mathbf{N}}),
\end{aligned}\tag{20}$$

where the equality in (a) follows by using (12).

The expression in (20) is equivalent to (13), and hence the proof of Theorem 1 is concluded.

A. Ancillary Results

We here state and prove three lemmas, which we have used in the proof of Theorem 1.

Lemma 2. *Let $\widetilde{\mathbf{X}}=\mathbf{y}-\mathbf{X}$. Then,*

$$\begin{aligned}
&\frac{1}{\Pr(\mathcal{H}_i)}\nabla_{K_{\mathbf{N}}}f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_i)\\
&=\frac{1}{2}\mathbb{E}\left[f_{\mathbf{N}}(\widetilde{\mathbf{X}})\left(-K_{\mathbf{N}}^{-1}+K_{\mathbf{N}}^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}K_{\mathbf{N}}^{-1}\right)\,\middle|\,\mathcal{H}_i\right].
\end{aligned}\tag{21}$$

Proof: We start by noting that, by using the chain rule, we obtain

$$\nabla_{K_{\mathbf{N}}}\ln f_{\mathbf{N}}(\mathbf{n})=\frac{1}{f_{\mathbf{N}}(\mathbf{n})}\nabla_{K_{\mathbf{N}}}f_{\mathbf{N}}(\mathbf{n}).\tag{22}$$

Then, we have that

$$\begin{aligned}
&\frac{1}{\Pr(\mathcal{H}_i)}\nabla_{K_{\mathbf{N}}}f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_i)\\
&\overset{(a)}{=}\nabla_{K_{\mathbf{N}}}\mathbb{E}\left[f_{\mathbf{N}}(\mathbf{y}-\mathbf{X})\mid\mathcal{H}_i\right]\overset{(b)}{=}\mathbb{E}\left[\nabla_{K_{\mathbf{N}}}f_{\mathbf{N}}(\mathbf{y}-\mathbf{X})\mid\mathcal{H}_i\right]\\
&\overset{(c)}{=}\mathbb{E}\left[f_{\mathbf{N}}(\mathbf{y}-\mathbf{X})\nabla_{K_{\mathbf{N}}}\ln f_{\mathbf{N}}(\mathbf{y}-\mathbf{X})\mid\mathcal{H}_i\right]\\
&\overset{(d)}{=}\frac{1}{2}\mathbb{E}\left[f_{\mathbf{N}}(\widetilde{\mathbf{X}})\left(-K_{\mathbf{N}}^{-1}+K_{\mathbf{N}}^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}K_{\mathbf{N}}^{-1}\right)\,\middle|\,\mathcal{H}_i\right],
\end{aligned}\tag{23}$$

where the labeled equalities follow from: (a) using (11) and $f_{\mathbf{Y}}(\mathbf{y}|\mathcal{H}_i)=\mathbb{E}\left[f_{\mathbf{N}}(\mathbf{y}-\mathbf{X})|\mathcal{H}_i\right]$; (b) the Leibniz rule which is here trivially verifiable [16]; (c) using (22); and (d) using the gradient rules (see [13] for details). This concludes the proof of Lemma 2, which we have used in step (c) in (17). ∎

Lemma 3. *Let $\partial\mathcal{R}_{i,K_{\mathbf{N}}}$, $i\in[1:M]$ be the boundary of the set $\mathcal{R}_{i,K_{\mathbf{N}}}$. Then, $\partial\mathcal{R}_{i,K_{\mathbf{N}}}$ is a set of Lebesgue measure zero for all $i\in[1:M]$.*

Proof: The proof is by contradiction, i.e., we assume that $\partial\mathcal{R}_{i,K_{\mathbf{N}}}$ is a set of positive measure, and we show that this leads to a contradiction. Note that, since $\partial\mathcal{R}_{i,K_{\mathbf{N}}}$ is the boundary of a decision region, then by the optimal decision criterion in [8, Appendix 3C], there must exist $j\neq i$ such that $f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_i)=f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_j)$, for all $\mathbf{y}\in\partial\mathcal{R}_{i,K_{\mathbf{N}}}$. We now leverage two well-known properties: (i) a convolution with Gaussian preserves analyticity [17], and hence $f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_i)$ is an analytic function for every $i\in[1:M]$; and (ii) two analytic functions in $\mathbb{R}^n$ that agree on a set of a positive measure must be equal everywhere [18], and hence, for all $\mathbf{y}\in\mathbb{R}^n$, we have

$$f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_i)=f_{\mathbf{Y}}(\mathbf{y},\mathcal{H}_j).\tag{24}$$

By using characteristic functions, it is a standard exercise to show that if the outputs of the probability density function are the same, so are the input distributions [16]. Therefore, for all measurable sets $\mathcal{B}\subseteq\mathbb{R}^n$, we have

$$\Pr(\mathbf{X}\in\mathcal{B},\mathbf{X}\in\mathcal{H}_i)=\Pr(\mathbf{X}\in\mathcal{B},\mathbf{X}\in\mathcal{H}_j).\tag{25}$$

However, since $\mathcal{H}_i$ and $\mathcal{H}_j$ are disjoint (Remark 1), choosing $\mathcal{B}=\mathcal{H}_i$ leads to a contradiction since $\Pr(\mathbf{X}\in\mathcal{H}_i,\mathbf{X}\in\mathcal{H}_i)\neq\Pr(\mathbf{X}\in\mathcal{H}_i,\mathbf{X}\in\mathcal{H}_j)=0$. This concludes the proof of Lemma 3 (used to analyze the integrand in (18)). ∎

Lemma 4. *For any $(i,j)\in[1:M]^2$ if $\mathbf{y}$ belongs to the interior of $\mathcal{R}_{i,t}$, then $1_{\mathcal{R}_{j,t}}(\mathbf{y})$ is continuous in t.*

Proof: The proof is by contradiction, i.e., we assume that $1_{\mathcal{R}_{j,t}}(\mathbf{y})$ is discontinuous[1] in t and we show that this leads to a contradiction. Let $\tilde{\mathbf{y}}\in\mathcal{R}_{i,\tilde{t}}$ be a fixed point. Assume that $1_{\mathcal{R}_{i,t}}(\tilde{\mathbf{y}})$ is discontinuous at $t=\tilde{t}>0$. Without loss of generality, we can assume that $\tilde{\mathbf{y}}\in\mathcal{R}_{i,t}$ when $t\to\tilde{t}^+$ and that $\tilde{\mathbf{y}}\in\mathcal{R}_{j,t}$, $j\neq i$ when $t\to\tilde{t}^-$. From the optimal decision rule [8], we then get the following two inequalities

$$\lim_{t\to\tilde{t}^+}\left(f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_i)-f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_j)\right)>0,\tag{26}$$

$$\lim_{t\to\tilde{t}^-}\left(f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_i)-f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_j)\right)<0.\tag{27}$$

Letting $g(t,\tilde{\mathbf{y}})=f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_i)-f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_j)$, we hence have that $g(t,\tilde{\mathbf{y}})$ is a discontinuous function in t, where the discontinuity occurs at $t=\tilde{t}$. However, for $\tilde{t}>0$ we have

$$\begin{aligned}
&\lim_{t\to\tilde{t}}f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_i)=\Pr(\mathcal{H}_i)\lim_{t\to\tilde{t}}\mathbb{E}\left[f_{\mathbf{N}_t}(\tilde{\mathbf{y}}-\mathbf{X})|\mathcal{H}_i\right]\\
&\overset{(a)}{=}\Pr(\mathcal{H}_i)\mathbb{E}\left[\lim_{t\to\tilde{t}}f_{\mathbf{N}_t}(\tilde{\mathbf{y}}-\mathbf{X})|\mathcal{H}_i\right]\\
&\overset{(b)}{=}\Pr(\mathcal{H}_i)\mathbb{E}\left[f_{\mathbf{N}_{\tilde{t}}}(\tilde{\mathbf{y}}-\mathbf{X})|\mathcal{H}_i\right]=f_{\mathbf{Y}_{\tilde{t}}}(\tilde{\mathbf{y}},\mathcal{H}_i),
\end{aligned}\tag{28}$$

where the labeled equalities follow from: (a) using the dominated convergence theorem, which is verified since

$$\mathbb{E}\left[f_{\mathbf{N}_t}(\mathbf{y}-\mathbf{X})|\mathcal{H}_i\right]<\mathbb{E}\left[\left(2\pi\det\left(K_{\mathbf{N}_t}\right)\right)^{-\frac{n}{2}}\right]<\infty,\tag{29}$$

where the last inequality follows by recalling that $K_{\mathbf{N}_t}=tK_{\mathbf{N}}$ where $K_{\mathbf{N}}$ is positive definite, and $t\to\tilde{t}$ with $\tilde{t}>0$; and (b) since $f_{\mathbf{N}_t}(\mathbf{y}-\mathbf{X})$ is continuous. The equation (28) shows that $f_{\mathbf{Y}_t}(\tilde{\mathbf{y}},\mathcal{H}_i)$ is continuous in t, which implies that $g(t,\tilde{\mathbf{y}})$ is continuous in t. This contradicts the assumption and concludes the proof of Lemma 4 (used in the proof of (19)). ∎

[1]To prove this claim, it is sufficient to consider only *jump* discontinuity since the *removable* discontinuity cannot happen.

REFERENCES

[1] S. Kritchman and B. Nadler, "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3930–3941, Oct 2009.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[3] A. Javanmard and A. Montanari, "Confidence intervals and hypothesis testing for high-dimensional regression," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2869–2909, 2014.

[4] ——, "Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6522–6554, July 2014.

[5] Y. Xia and J. Sun, "Hypothesis testing and statistical analysis of microbiome," *Genes & Diseases*, vol. 4, no. 3, pp. 138–148, 2017.

[6] M. D. Crisp, S. A. Trewick, and L. G. Cook, "Hypothesis testing in biogeography," *Trends in Ecology & Evolution*, vol. 26, no. 2, pp. 66–72, 2011.

[7] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933. [Online]. Available: http://www.jstor.org/stable/91247

[8] S. Kay, *Fundamentals of Statistical Signal Processing: Detection theory*. Prentice-Hall PTR, 1998.

[9] G. Vazquez-Vilar, A. Tauste Campo, A. Guillén i Fàbregas, and A. Martinez, "Bayesian M-ary hypothesis testing: The meta-converse and Verdú-Han bounds are tight," *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2324–2333, May 2016.

[10] I. Sason and S. Verdú, "Arimoto–Rényi conditional entropy and Bayesian M-ary hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, Jan 2018.

[11] R. E. Melchers and M. Ahammed, "Gradient estimation for applied Monte Carlo analyses," *Reliability Engineering & System Safety*, vol. 78, no. 3, pp. 283–288, 2002.

[12] M. Ahammed and R. E. Melchers, "Gradient and parameter sensitivity estimation for systems evaluated using Monte Carlo analysis," *Reliability Engineering & System Safety*, vol. 91, no. 5, pp. 594–601, 2006.

[13] K. Petersen and M. Pedersen, *The Matrix Cookbook*. Technical Univ. Denmark, Kongens Lyngby, Denmark, Tech. Rep, 2012, vol. 3274.

[14] A. Dytso, M. Cardone, M. S. Veedu, and H. Vincent Poor, "On estimation under noisy order statistics," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 36–40.

[15] M. Jeong, A. Dytso, M. Cardone, and H. V. Poor, "Recovering structure of noisy data through hypothesis testing," in *2020 IEEE International Symposium on Information Theory (ISIT)*, June 2020, pp. 1307–1312.

[16] S. I. Resnick, *A Probability Path*. Springer, 2003.

[17] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 1999, vol. 40.

[18] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*. Springer Science & Business Media, 2002.