Error exponents for composite hypothesis testing of Markov forest distributions

Error Exponents for Composite Hypothesis Testing of Markov Forest Distributions

Vincent Y. F. Tan, Animashree Anandkumar and Alan S. Willsky

Stochastic Systems Group, LIDS, MIT, Cambridge, MA 02139. Email: {vtan,animakum,willsky}@mit.edu

***Abstract*—The problem of composite binary hypothesis testing of Markov forest (or tree) distributions is considered. The worst-case type-II error exponent is derived under the Neyman-Pearson formulation. Under simple null hypothesis, the error exponent is derived in closed-form and is characterized in terms of the so-called bottleneck edge of the forest distribution. The least favorable distribution for detection is shown to be Markov on the second-best max-weight spanning tree with mutual information edge weights. A necessary and sufficient condition to have positive error exponent is derived.**

***Index Terms*—Worst-case error exponent, Markov forests, Least favorable distribution, Neyman-Pearson formulation.**

I. Introduction

Binary composite hypothesis testing is a classical problem where one is required to decide if a set of samples is drawn from a distribution in the set Λ_0 or the set Λ_1, corresponding to the null and alternative hypotheses respectively [1]. Under the Neyman-Pearson formulation, the goal is to minimize the type-II (misdetection) error probability, where the alternative hypothesis is mistaken as the null, under the constraint that the type-I (false alarm) error probability is below a fixed size.

As the number of samples available for detection increases, the type-II error probability typically decays exponentially and the rate of decay is known as the *error exponent*. For composite hypothesis testing, the *worst-case error exponent* is the slowest decay rate of the type-II error probability for any distribution in Λ_1. It serves as a performance benchmark for detection in the large-sample regime. The distribution that achieves the worst-case exponent is said to be the *least favorable* for detection.

In this paper, we derive the worst-case error exponent as well as the least favorable distribution when Λ_0 and Λ_1 are sets of multivariate distributions (called *graphical models*) which are Markov on trees[1] and, more generally, forests. We consider both discrete and Gaussian graphical models. We also simplify the generalized likelihood ratio test (GLRT), which is commonly used in composite hypothesis testing.

We first consider the special case when the null hypothesis is simple, *i.e.*, $\Lambda_0 = \{p\}$, for a fixed distribution p Markov on a tree T_0, and Λ_1 is a set of distributions Markov on all other trees except T_0. A brute force computation of the worst-case error exponent requires a search over all the trees and is thus computationally prohibitive. We derive the worst-case error exponent which is characterized by the mutual information on the so-called *bottleneck edge* of T_0. Moreover, we prove that the least favorable distribution is Markov on the second-best max-weight spanning tree (MWST) with mutual information edge weights (based on p). We generalize these results to forest distributions and composite null hypotheses, and provide conditions on Λ_0 and Λ_1 to ensure that the error exponent is positive.

We now describe some related work in the areas of composite hypothesis testing [1] and learning graphical models. Hoeffding first proposed an asymptotically optimal test for composite hypothesis testing assuming that null hypothesis is simple [2]. The test was subsequently generalized to the case when the null hypothesis is also composite [3]. The GLRT [1] is another test for composite hypothesis testing, which is only optimal under certain conditions [4]. The error exponent is easily characterized when the source is i.i.d. [2]–[4] and in this case, it is usually available in closed-form.

In [5], [6], we derived the error exponent for maximum-likelihood (ML) learning of the structure of tree-structured graphical models using the Chow-Liu algorithm [7]. In [8], a learning algorithm for tree distributions was proposed for a the specific purpose of hypothesis testing. In contrast, composite hypothesis testing of forest distributions is considered in this paper. In [9], the authors derived the error exponent for binary hypothesis testing of Markov forest distributions on randomly placed nodes where number of nodes (and hence, the number of variables) goes to infinity and each node has one independent sample from the forest distribution. In contrast, in this paper, we fix the number of nodes in the forest and draw large number of samples from the graphical model.

II. System Model and Problem Formulation

A. Notations and Mathematical Preliminaries

Let $G = (V, E)$ be an undirected graph with vertex set $V := \{1, \ldots, d\}$ and edge set $E \subset \binom{V}{2}$ and let $\mathrm{nbd}(i) := \{j \in V : (i, j) \in E\}$ be the set of neighbors of vertex i. Let $\mathcal{F}^d$ denote the set of forests, *i.e.*, undirected acyclic graphs with d nodes and let $\mathcal{T}^d \subset \mathcal{F}^d$ denote the set of spanning trees (also called trees) with d nodes. For a fixed forest $F \in \mathcal{F}^d$, let the set of *supergraphs* of F (with d nodes) be $\mathcal{S}(F) \subset \mathcal{F}^d$, *i.e.*, F is a subgraph of any element of $\mathcal{S}(F)$ and by definition $F \in \mathcal{S}(F)$. The *type* or empirical distribution of a sequence $\mathbf{x}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is denoted as $\widehat{\mu}^n := \widehat{\mu}^n(\mathbf{x}^n)$.

This work is supported by a AFOSR funded through Grant FA9559-08-1-1080, a MURI funded through ARO Grant W911NF-06-1-0076 and a MURI funded through AFOSR Grant FA9550-06-1-0324. V. Tan is also funded by A*STAR, Singapore.

Due to space constraints, the proofs of the results are not included here. The reader is encouraged to view all the proofs at http://web.mit.edu/vtan/www/isit10b.

[1] A *tree* is a connected acyclic graph, a *forest* is any acyclic graph and a *strict forest* is an acyclic graph which is not a tree, *i.e.*, disconnected.

B. Graphical Models: Markov Tree/Forest Distributions

An *undirected graphical model* [10] is a probability distribution that factorizes according to the structure of a given undirected graph $G = (V, E)$, where each random variable is associated to a vertex (or node) in G. More precisely, for an alphabet $\mathcal{X}$, a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ distributed according to $p \in \mathcal{P}(\mathcal{X}^d)$ is said to be *Markov* on a graph G if $p(x_i|x_{V\setminus\{i\}}) = p(x_i|x_{\mathrm{nbd}(i)}), \forall\, i \in V$. We also say that G is the *graph* of p. In this paper, we assume that G is a *minimal representation* for p, *i.e.*, G has the smallest number of edges for the conditional independence relations to hold.

The results in this paper apply to both discrete ($\mathcal{X}$ finite and p is the probability mass function of $\mathbf{X}$) and Gaussian graphical models ($\mathcal{X} = \mathbb{R}$ and p is the probability density function). In the Gaussian case, the random vector $\mathbf{X} \sim p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is a zero-mean Gaussian with covariance matrix $\mathbf{\Sigma}$. It is known [10] that the inverse covariance matrix (or precision matrix) $\mathbf{K} := \mathbf{\Sigma}^{-1}$ encodes the structure of $G = (V, E)$, *i.e.*, $\mathbf{K}(i, j) = 0$ if and only if $(i, j) \notin E$.

In this paper, we focus on the class of positive[2] distributions Markov on some forest $F = (V, E)$. Such a distribution admits the following factorization into node and pairwise marginals:

$$p(\mathbf{x}) = \prod_{i\in V} p_i(x_i) \prod_{(i,j)\in E} \frac{p_{i,j}(x_i, x_j)}{p_i(x_i)p_j(x_j)}. \tag{1}$$

Denote the set of positive d-dimensional distributions Markov on forests with alphabet $\mathcal{X}^d$ as $\mathcal{D}(\mathcal{X}^d, \mathcal{F}^d)$ and the set of distributions Markov on a particular forest F as $\mathcal{D}(\mathcal{X}^d, F)$. A similar set of notation applies for the set of trees $\mathcal{T}^d$.

C. General Hypothesis Testing of Forest Distributions

Let $\mathbf{x}^n$ be a set of i.i.d. samples drawn from a positive distribution with support $\mathcal{X}^d$. Consider the following binary composite hypothesis testing problem [1]:

$$\begin{aligned} H_0 &: \mathbf{x}^n \sim \{p : p \in \Lambda_0 \subset \mathcal{D}(\mathcal{X}^d, \mathcal{F}^d)\}, \\ H_1 &: \mathbf{x}^n \sim \{q : q \in \Lambda_1 \subset \mathcal{D}(\mathcal{X}^d, \mathcal{F}^d)\}, \end{aligned} \tag{2}$$

where $\Lambda_0 \cap \Lambda_1 = \emptyset$ for identifiability and Λ_0 is closed. For the error exponents to be positive, additional constraints on Λ_0 and Λ_1 need to be imposed and we discuss these constraints for specific examples of (2) in the sequel.

D. Definition of Worst-Case Type-II Error Exponent

For a test ϕ, let $A_n(\phi) \subset (\mathcal{X}^d)^n$ be an *acceptance region* for H_0, *i.e.*, $\mathbf{x}^n \in A_n(\phi)$ represents a decision in favor of H_0. Then, for fixed $p \in \Lambda_0$ and $q \in \Lambda_1$, the type-I and type-II error probabilities[3] are given by $p^n(A_n^c(\phi))$ and $q^n(A_n(\phi))$ respectively. When the type-II error probability decays exponentially with the sample size n under a distribution $q \in \Lambda_1$, the *type-II error exponent* under q is defined as

$$J(\Lambda_0, q; \phi) := \liminf_{n\to\infty} -\frac{1}{n} \log q^n(A_n(\phi)). \tag{3}$$

[2]We say a distribution p is *positive* if $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$.

[3]For a Borel measurable set $B \subset (\mathcal{X}^d)^n$, $p^n(B) := \int_B p^n \, d\nu$ (ν is either the Lebesgue or counting measure) and $p^n(\mathbf{x}^n) = \prod_{k=1}^n p(\mathbf{x}_k)$.

The *optimal type-II error exponent* for composite hypothesis testing is the supremum of $J(\Lambda_0, q; \phi)$ over all tests subject to the type-I error being below a fixed size α:

$$J^*(\Lambda_0, q) := \sup_{\phi} \{J(\Lambda_0, q; \phi) : \forall p \in \Lambda_0, p^n(A_n^c(\phi)) \le \alpha\}. \tag{4}$$

The corresponding (universal) detector ϕ^* is said to be *asymptotically optimal*. Also, if this universal detector has the same type-II error exponent as the Neyman-Pearson detector for the corresponding simple hypotheses, it is said to be *asymptotically uniformly most powerful*.

The *worst-case type-II error exponent* is now defined as

$$J^*(\Lambda_0, \Lambda_1) := \inf_{q\in\Lambda_1} J^*(\Lambda_0, q). \tag{5}$$

Furthermore, if there exists a distribution $q^* \in \Lambda_1$ that attains the infimum[4] above, then q^* is known as the *least favorable distribution* (LFD). The worst-case error exponent serves as a performance benchmark for composite hypothesis testing.

In this paper, we focus on finding closed-form (i.e., simple) solutions to the worst-case type-II error exponent $J^*(\Lambda_0, \Lambda_1)$ for the hypothesis testing problem in (2). Moreover, we show that these error exponents can be computed efficiently. Note that an exhaustive search for the optimization in (5) is intractable since there are d^{d-2} undirected trees with d nodes [11]. However, we are able to express the error exponent as relatively elementary functions of the parameters of the distribution p. We also provide intuitive interpretations of these results.

III. Tests for Composite Hypothesis Testing

A. The Hoeffding Test for Composite Hypothesis Testing

The *Hoeffding test* [2] produces acceptance regions

$$A_n(\mathrm{HT}) := \left\{ \mathbf{x}^n : \inf_{p\in\Lambda_0} D(\widehat{\mu}^n \,||\, p) \le \epsilon_n \right\}, \tag{6}$$

where $\epsilon_n = O(\frac{\log n}{n})$ and $D(\cdot||\cdot)$ is the KL-divergence.

Proposition 1 (Worst-Case Type-II Error Exponent [12]): The worst-case type-II error exponent in (5) is

$$J^*(\Lambda_0, \Lambda_1) = \inf_{p\in\Lambda_0, q\in\Lambda_1} D(p \,||\, q). \tag{7}$$

and is achieved by the Hoeffding test in (6). Furthermore, if the null hypothesis H_0 is simple, then the Hoeffding test is asymptotically uniformly most powerful.

The type-II error exponent for the Hoeffding test in (6) is $\inf_{p\in\Lambda_0} D(p||q)$ for every $q \in \Lambda_1$. In addition, if $\Lambda_0 = \{p\}$, the Hoeffding test is asymptotically uniformly most powerful. For the proof of the above result, see [12, Thm. 2.3]. Hence, the Hoeffding test is worst-case asymptotically optimal [2], *i.e.*, there does not exist a test with a better worst-case type-II error exponent. This is because the Neyman-Pearson test, which is the optimal (most powerful) test, for corresponding simple hypotheses in the worst case has the same error exponent as the Hoeffding test.

[4]The closedness of Λ_1 ensures the existence of such a q^*.

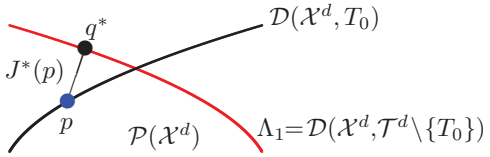

Fig. 1. The geometry of the hypothesis testing problem in (9). The simple null hypothesis (in blue) involves p, which is Markov on T_0. The alternative hypothesis (in red) involves the set of distributions Markov on some other tree not equal to T_0. The worst-case type-II error exponent $J^*(p)$ is the minimum KL-divergence from p to Λ_1. q^*, the LFD, is the projection of p onto Λ_1.

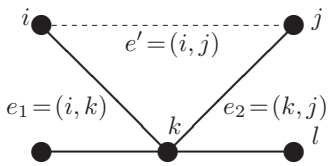

Fig. 2. Illustration of Thm. 3. The unique path associated to non-edge e' is $\mathrm{Path}(e'; E_0) = \{(i,k), (k,j)\}$. The non-edge $e' = (i,j)$ (of length $L(i,j) = 2$) replaces the edge $e_1 = (i,k)$ if $I(p_{e_1}) \geq I(p_{e_2})$ and e_1 is the bottleneck edge. Otherwise e' replaces $e_2 = (k,j)$.

However, it is not possible to simplify the Hoeffding test (6) further for Markov forest distributions. Thus, we now turn our attention to another test which is easier to implement for forest distributions but does not have similar optimality guarantees.

B. The Generalized Likelihood Ratio Test (GLRT)

The GLRT [1] is a test that produces acceptance regions

$$A_n(\mathrm{GLRT}) := \left\{ \mathbf{x}^n : \frac{1}{n} \log \frac{\sup_{q \in \Lambda_1} q^n(\mathbf{x}^n)}{\sup_{p \in \Lambda_0} p^n(\mathbf{x}^n)} \in \mathcal{R}_n(\alpha) \right\}, \quad (8)$$

where the region $\mathcal{R}_n(\alpha) \subset \mathbb{R}$ is chosen to satisfy the false alarm constraint in (3). That is, given observations $\mathbf{x}^n$, the GLRT is the likelihood ratio test between the ML distributions under H_0 and H_1. Note that the Hoeffding test is independent of Λ_1 but the GLRT depends on Λ_1.

For simple null and alternative hypotheses, the GLRT reduces to the canonical likelihood ratio test, which is optimal under the Neyman-Pearson formulation. However, for composite hypotheses, the GLRT is in general not guaranteed to be uniformly optimal for all null and alternative hypotheses but is nevertheless employed ubiquitously due to its simplicity. In this paper, we simplify the GLRT for a few special cases of the hypothesis test in (2).

IV. Results Under Simple Null Hypothesis

In this section, we consider a special case of (2), given as

$$\begin{aligned} H_0 &: \mathbf{x}^n \sim \{p\}, \\ H_1 &: \mathbf{x}^n \sim \{q : q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d \setminus \{T_0\})\}, \end{aligned} \quad (9)$$

with a given distribution $p \in \mathcal{D}(\mathcal{X}^d, T_0)$ Markov on a fixed tree $T_0 = (V, E_0)$. Hence, H_0 is a simple hypothesis, *i.e.*, $\Lambda_0 = \{p\}$. The set in the alternative $\Lambda_1 = \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d \setminus \{T_0\})$ is the set of all distributions Markov on all trees other than T_0. See Fig. 1 for an geometric illustration of the problem in (9).

This problem is related to the classical universal hypothesis testing problem, where one has to decide whether the samples are drawn from a fixed distribution [1]; the difference here is that we limit the class of distributions under H_1 to tree-structured distributions. This problem is also known in various contexts as *anomaly detection* where the (nominal) null hypothesis is known while the (anomalous) alternative hypothesis is unknown.

A. Simplification of the GLRT for Tree Distributions

We now simplify the GLRT for the problem in (9) given samples $\mathbf{x}^n$. With an abuse of notation, we denote the mutual information between random variables X_i and X_j under p as $I(p_e) := I(X_i; X_j)$, where p_e is the marginal on $e = (i,j)$.

Proposition 2 (GLRT for Tree Distributions): The acceptance region for the GLRT in (8) is given as

$$A_n(\mathrm{GLRT}) = \left\{ \mathbf{x}^n : \sum_{(i,j) \in E^*} I(\widehat{\mu}^n_{i,j}) - \sum_{(i,j) \in E_0} I(\widehat{\mu}^n_{i,j}) \in \mathcal{R}_n(\alpha) \right\},$$

where E^* is the edge set given by the optimization problem:

$$E^* := \operatorname*{argmax}_{E : T = (V,E) \in \mathcal{T}^d \setminus \{T_0\}} \sum_{(i,j) \in E} I(\widehat{\mu}^n_{i,j}). \quad (10)$$

We observe from (10) that E^* is the solution to a constrained MWST problem where the edge weights given by the empirical mutual information quantities $\{I(\widehat{\mu}^n_{i,j})\}$ and the resulting tree is not allowed to be equal to T_0. We propose the following simple procedure to find the edge set E^* for the GLRT.

1) Run a MWST algorithm (*e.g.*, Kruskal's [11]) to get the edge set $\widehat{E} := \operatorname{argmax}_{E : T = (V,E) \in \mathcal{T}^d} \sum_{(i,j) \in E} I(\widehat{\mu}^n_{i,j})$.
2) If $\widehat{E} \neq E_0$, then $E^* = \widehat{E}$ and end.
3) Otherwise if $\widehat{E} = E_0$, run the second-best MWST algorithm [11] with the same set of edge weights and set E^* to be the output of the algorithm.

Thus, once the empirical mutual informations have been computed, the tree structure E^* of the ML distribution in Λ_1 can be determined efficiently.[5] Subsequently, the simplified test as stated in Theorem 2 can be used to determine whether the set of observations $\mathbf{x}^n$ is drawn from p or some other distribution q whose graph is a tree not equal to T_0.

B. Worst-Case Type-II Error Exponent

We now derive the error exponent $J^*(p)$ at which the probability of misdetection decays for the hypothesis testing problem in (9). Let $L(i,j)$ denote the *graph distance* between nodes i and j, *i.e.*, $L(i,j)$ is the number of edges between nodes i and j in T_0. We abbreviate $J^*(\{p\}, \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d \setminus \{T_0\}))$ in (5) as $J^*(p)$. Let $\mathrm{Path}(e'; E_0)$ be the unique path between i and j in tree T_0. See Fig. 2. We now state our main result.

Theorem 3 (Worst-Case Error Exponent for Trees): The worst-case type-II error exponent J^*, defined in (5), for the hypothesis testing problem in (9) is

$$J^*(p) = \min_{\substack{e' = (i,j) \notin E_0 \\ L(i,j) = 2}} \min_{e \in \mathrm{Path}(e'; E_0)} \{I(p_e) - I(p_{e'})\}, \quad (11)$$

[5] The MWST and second-best MWST have time complexity $O(d^2)$ [11].

and can be found with $d-1$ computations of $I(p_e) - I(p_{e'})$. Moreover, the LFD for detection in (5) is attained by $q^* \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d \backslash \{T_0\})$ which is Markov on the second-best MWST

$$E_{q^*} = \operatorname*{argmax}_{E:T=(V,E)\in\mathcal{T}^d\backslash\{T_0\}} \sum_{e\in E} I(p_e), \tag{12}$$

with parameters $q^*_i = p_i, \forall\, i \in V$ and $q^*_{i,j} = p_{i,j}, \forall\, (i,j) \in E_{q^*}$.

The closed-form expression in terms of mutual information quantities implies that the computation of $J^*(p)$ is simple and there is no need to perform an exhaustive search over all d^{d-2} trees for the LFD. Observe that the LFD q^* in (12) that achieves the worst-case type-II error exponent is the "closest" tree distribution to p (in the KL-divergence sense) which is not Markov on T_0. Moreover, q^* is Markov on the second-best MWST with mutual information edge weights. Recall that the MWST is achieved by T_0, and hence, the second-best MWST is the "closest" tree to T_0. Moreover, the two trees differ in exactly one edge, and we call the edge in T_0 the *bottleneck edge*, since it is replaced by a non-edge such that difference in mutual information quantities is minimized. In addition, the constraint $L(i,j) = 2$ in (11) comes from the the data-processing lemma which says that the mutual information of pairs of variables farther apart are smaller than pairs of variables that are closer. Thus, shorter non-edges are more likely to replace a true edge than longer non-edges. Consequently, to compute $J^*(p)$, we only have to consider those non-edges with distance 2.

As in other related results such as Thm. 8 in [5], observe that as the difference between the mutual information quantities $I(p_e)$ and $I(p_{e'})$ increases, the worst-case type-II error exponent also increases, resulting in better detection performance in the large-sample regime. This is intuitive in light of the Chow-Liu algorithm [7] for learning the ML tree structure from a set of i.i.d. samples; if the true mutual information quantities are far apart, then the true edges in T_0 can be distinguished more easily from the non-edges and it is less likely for the ML estimator to mistake a non-edge as a true edge.

We now specialize Thm. 3 for Gaussians. Let the correlation coefficient between X_i and X_j (under p) be denoted as $\rho_{i,j}$.

Corollary 4 (Error Exponent for Gaussian Trees): If $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ under both H_0 and H_1 in (9), then the error exponent for the hypothesis test in (2) is

$$J^*(p) = \min_{e_1,e_2\in E_0 : e_1\sim e_2} \frac{1}{2}\log\left[\frac{1-\rho_{e_1}^2\rho_{e_2}^2}{1-\rho_{e_1}^2}\right], \tag{13}$$

where the notation $e_1 \sim e_2$ means that the edges e_1 and e_2 share a common node.

C. Extension to Forest Distributions

We now generalize the results in Section IV-B to forest distributions. In this case, the hypothesis testing problem is

$$\begin{aligned} H_0 &: \mathbf{x}^n \sim \{p\}, \\ H_1 &: \mathbf{x}^n \sim \{q : q \in \mathcal{D}(\mathcal{X}^d, \mathcal{F}^d \setminus \mathcal{S}(F_0))\}, \end{aligned} \tag{14}$$

where $p \in \mathcal{D}(\mathcal{X}^d, F_0)$ for a forest $F_0 \in \mathcal{F}^d$. Notice that the uncertainty set in the alternative hypothesis $\mathcal{D}(\mathcal{X}^d, \mathcal{F}^d \backslash \mathcal{S}(F_0))$ excludes distributions which are Markov on supergraphs of F_0. If this were not so (*i.e.*, $\Lambda_1 = \mathcal{D}(\mathcal{X}^d, \mathcal{F}^d \setminus F_0)$) and F_0 is a strict forest, then the error exponent $J^*(p)$ will necessarily be zero. This is because the infimum in (5) will be achieved by a sequence of distributions, each Markov on a strict supergraph of F_0, but with arbitrarily weak potentials on an additional edge that respects the tree constraint. For example, in Fig. 3, the non-edge (i,j) with an arbitrarily weak potential will be added in this scenario, resulting in $J^*(p) = 0$.

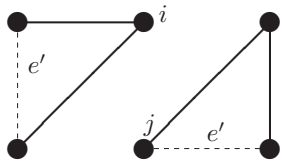

Fig. 3. Illustration of Thm. 5. In this 6-node forest, there are two possible non-edges e' that can replace a true edge along its path. The replacement resulting in the minimum difference in mutual information quantities $I(p_e) - I(p_{e'})$ gives the error exponent $J^*(p)$.

Theorem 5 (Worst-Case Error Exponent for Forests): The error exponent for the (forest) hypothesis problem in (14) is

$$J^*(p) = \min_{\substack{e'=(i,j)\notin E_0 \\ L(i,j)=2}} \min_{e\in \mathrm{Path}(e';E_0)} \{I(p_e) - I(p_{e'})\}. \tag{15}$$

Furthermore, the LFD q^* exists and its edge set and parameters are the same as in Thm. 3. In particular, $|E_{q^*}| = |E_0|$.

The error exponent for the forest hypothesis test in (14) mirrors that for trees. Also, the GLRT for the forests can be implemented in exactly the same fashion as in Thm. 2 with the caveat that E^* in (10) is only allowed to have $|E_0|$ edges.

D. Comparison to ML Structure Learning

In this section, we compare and contrast the above results (Thm. 3 and Thm. 5) to the line of research that concerns learning of tree-structured graphical models from data [5], [6]. In [5], [6], the learner is given samples $\mathbf{x}^n$ drawn from a fixed tree distribution p and the error exponent for ML structure learning using the Chow-Liu algorithm [7] was derived. However, in this paper, we instead consider the Neyman-Pearson formulation where the probability of false alarm $p^n(A_n^c(\phi))$ is kept below a fixed size α and we quantify the worst-case (smallest) exponential decay of the probability of misdetection. Furthermore, consistent learning (which implies positive error exponent) of strict forest-structured distributions is not possible because the ML estimate will always be a connected tree. However, we do have a positive error exponent in Thm. 5 because we have explicitly excluded all distributions Markov on supergraphs of the graph of p from the set Λ_1 in (14).

V. Generalization to Composite Null Hypothesis

A. Positivity of Error Exponent

We now analyze the conditions under which the probability of misdetection $q^n(A_n(\phi))$ decays to zero exponentially fast for every q in the uncertainty set Λ_1. Since we assumed that the graph of any distribution is a minimal representation, for

the special case where the null hypothesis is simple, the worst-case type-II error exponents $J^*(p)$ in (11) and (15) are positive as minimality implies that the mutual information quantities $I(p_e) > 0$ for all $e \in E_0$ and so the difference $I(p_e) - I(p_{e'}) > 0$.

In this section, we derive a necessary and sufficient condition to ensure the positivity of the worst-case type-II error exponent in the general hypothesis testing problem in (2). In (2), both the null and alternative hypotheses are composite. Let $\mathcal{E}_0$ and $\mathcal{E}_1$ be the set of edge sets of the distributions in Λ_0 and Λ_1 respectively, *i.e.*, every distribution $p \in \Lambda_0$ is Markov on some tree with edge set $E_0 \in \mathcal{E}_0$ and similiarly for Λ_1.

Proposition 6 (Positivity of Error Exponent): Assume the general composite hypothesis test for forest-structured distributions in (2). Suppose there exists a $\delta > 0$ such that

$$\min_{p \in \Lambda_0} \min_{(E_0, E_1) \in \mathcal{E}_0 \times \mathcal{E}_1} \min_{e' \in E_1} \min_{e \in \mathrm{Path}(e'; E_0)} I(p_e) \geq \delta, \qquad (16)$$

then the worst-case type-II error exponent $J^*(\Lambda_0, \Lambda_1) > 0$. This result says that the mutual information quantities on all edges $e \in E_0$ along the path of some non-edge $e' \in E_1$ has to be bounded away from zero. We illustrate these results with the following example.

B. A Specific Example: Gaussian Distributions

We now consider a special case of (2) in which the sets Λ_0 and Λ_1 are defined parametrically. Note that this formulation is in contrast to robust hypothesis testing [1, Ch. 11]. Specifically, to obtain a closed-form solution to $J^*(p)$, we consider the Gaussian random vector, $\mathbf{X} \sim p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma}(i,i) = 1$ for all $i = 1, \ldots, d$. By the Markov property, the specification of the correlations along the edges $e \in E_0$ suffices for a complete characterization of a Gaussian tree distribution [6]. For $\eta_1, \eta_2 \in (0, 0.5)$, define the sets of Gaussian tree models

$$\Lambda_0 := \{p \in \mathcal{D}(\mathcal{X}^d, T_0) : \eta_1 \leq |\rho_e| \leq 1 - \eta_2, \forall e \in E_0\}, \quad (17a)$$

$$\Lambda_1 := \{q : q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d \setminus \{T_0\})\}. \quad (17b)$$

That is, the (closed) uncertainty set Λ_0 is parameterized by (η_1, η_2) and comprises zero-mean, unit-variance Gaussians Markov on T_0 and for which the magnitude of the correlation coefficients along the edges are between η_1 and $1 - \eta_2$.

Proposition 7 (Error Exponent for Bounded Correlations): Assume that the uncertainty sets in the null and alternative hypotheses for the problem in (2) are given in (17), for some $\eta_1, \eta_2 \in (0, 0.5)$. Then the worst-case type-II error exponent, defined in (5), is given as

$$J^*(\Lambda_0, \Lambda_1) = \frac{1}{2} \log \left[\frac{1 - \eta_1^2 (1 - \eta_2)^2}{1 - \eta_1^2} \right]. \qquad (18)$$

Furthermore, if $\eta_1 = \eta_2 = \eta$, then $\lim_{\eta \downarrow 0} J^*(\Lambda_0, \Lambda_1)/\eta^3 = 1$.

This function $J^*(\Lambda_0, \Lambda_1)$ is plotted in Fig. 4. We observe that as either η_1 or η_2 tends to 0, the exponent $J^* \to 0$. This is in line with intuition because if $\eta_1 \approx 0$, then Λ_0 includes those Gaussian distributions which have very weak correlations on all edges, and hence small mutual information values. If $\eta_2 \approx 0$, then Λ_0 includes those Gaussian distributions which have almost perfectly correlated variables. Conversely, the exponent is monotonically increasing in η_1 and η_2, indicating that as the minimum distance between the sets Λ_0 and Λ_1 increases, the exponent also increases. In fact, if $\eta \approx 0$, J^* behaves like η^3.

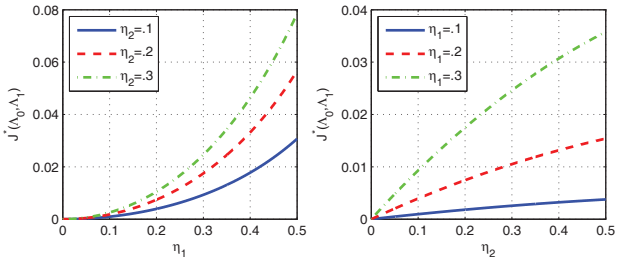

Fig. 4. Plot of $J^*(\Lambda_0, \Lambda_1)$ in (18) against (η_1, η_2).

VI. Conclusion

In this paper, we analyzed the worst-case type-II error exponent for composite hypothesis testing of Markov forest distributions under the Neyman-Pearson formulation. We characterized the error exponent in terms of the bottleneck edge when the null hypothesis was assumed to be simple. We also provided conditions for the error exponent to be positive, which ensures that there the type-II error probability decays exponentially fast. Given the nature of the results in this paper, a natural question that arises is the following: Can we derive similar closed-form (and thus interpretable) expressions for the error exponents of hypothesis testing problems which involve a more general class of graphical models, *e.g.*, decomposable models [10]?

References

[1] E. Lehmann, *Testing Statistical Hypotheses*. New York, NY: John Wiley & Sons, Inc., 1959.

[2] W. Hoeffding, "Asymptotically Optimal Tests for Multinomial Distributions," *Ann. of Math. Stats.*, vol. 36, no. 2, pp. 369–401, 1965.

[3] O. Zeitouni and M. Gutman, "On Universal Hypotheses Testing via Large Deviations," *IEEE Trans. on Info. Th.*, vol. 37, no. 2, pp. 285–290, 1991.

[4] O. Zeitouni, J. Ziv, and N. Merhav, "When is the Generalized Likelihood Ratio Test Optimal?" *IEEE Trans. on Info. Th.*, vol. 38, no. 5, pp. 1597–1602, 1992.

[5] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A Large-Deviation Analysis for the Maximum Likelihood Learning of Markov Tree Structures," *submitted to IEEE Trans. on Info. Th., Arxiv 0905.0940*, May 2009.

[6] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures," *accepted to IEEE Trans. on Sig. Proc., Arxiv 0909.5216*, Jan 2010.

[7] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees." *IEEE Trans. on Info. Th.*, vol. 14, no. 3, pp. 462–467, May 1968.

[8] S. Sanghavi, V. Y. F. Tan, and A. S. Willsky, "Learning Graphical Models for Hypothesis Testing," in *IEEE Workshop on Statistical Signal Processing*, Madison, WI, Aug 2007.

[9] A. Anandkumar, L. Tong, and A. S. Willsky, "Detection Error Exponent for Spatially Dependent Samples in Random Networks," in *Intl. Symp. Info. Th.*, Seoul, S. Korea, July 2009.

[10] S. Lauritzen, *Graphical Models*. Oxford University Press, USA, 1996.

[11] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. McGraw-Hill Science/Engineering/Math, 2003.

[12] I. Csiszar and P. Shields, *Information Theory and Statistics: A Tutorial*. Now Publishers Inc, 2004.