Notes on Fisher Information-Based Measurement Design for Network Tomography

Ting He†, Chang Liu†, Ananthram Swami§, Don Towsley†, Theodoros Salonidis†, Andrei Iu. Bejan‡, and Paul Yu§

†University of Massachusetts, Amherst, MA, USA. {cliu, towsley}@cs.umass.edu
‡IBM T. J. Watson Research Center, Yorktown, NY, USA. {the, tsaloni}@us.ibm.com
§Army Research Laboratory, Adelphi, MD, USA. {ananthram.swami, paul.i.yu}@mail.mil
*Cambridge University, Cambridge, UK. aib29@cam.ac.uk

I. Problem Formulation

A. Network Model

Let \( G = (V, L) \) denote a network with nodes \( V \) and links \( L \). Each link \( l \in L \) is associated with a performance metric that varies stochastically according to a distribution with unknown parameter \( \theta_l \). Let \( P \) be a given set of candidate probing paths in \( G \). Each path \( p_y \in P \) consists of one or more pairwise adjacent links in \( G \). We assume that the monitoring system can inject probes on any selected paths in \( P \) and observe their end-to-end performance. We also introduce a matrix \( A \) defined by \( \sum_{y,l} A_{y,l} = 1 \). Let \( \phi_y \) be the vector of link success rates \( \phi_y \), and \( \phi \) be the parameter of interest.

B. Stochastic Link Metric Tomography

Given the stochastic link performance metrics of interest (e.g., delay, loss, jitter) that follow a given type of distribution with unknown parameters \( \theta := (\theta_l)_{l \in L} \), the goal of (parametric) stochastic link metric tomography is to infer \( \theta \) from observations of the corresponding path performance metrics (e.g., path delay, loss, jitter) over probed paths. In this paper, we consider probabilistic probe allocation, where each probe is sent over a randomly selected path from \( P \), with probability \( \phi_y \) of selecting path \( p_y \). Let \( f_y(x; \theta) \) denote the conditional probability of observing path metric \( x \), given that the probe is sent on path \( p_y \) and the link parameters are \( \theta \). Then the problem of stochastic link metric tomography is to infer the parameter \( \theta \) from the observations \( (x, y) := (x_t, y_t)_{t=1}^N \), where \( x_t \) is the outcome of the \( t \)-th probe and \( y_t \) the corresponding path index. Under the assumption that the performance experienced by probes is independent both across probes and across links, the observations are i.i.d., each with the following distribution:

\[
\phi_y f_y(x; \theta), \tag{1}
\]

where \( \phi := (\phi_y)_{y=1}^{|P|} \), satisfying \( \phi_y \geq 0 \) and \( \sum_{y=1}^{|P|} \phi_y = 1 \), is a design parameter controlling the distribution of probes over the candidate paths. As concrete examples, we will address in detail two representative performance metrics as follows.

1) Packet Loss Tomography: Packet loss is a typical performance metric that is multiplicative over links on a path. Packet loss tomography aims at inferring loss rates on individual links by observing end-to-end packet losses on probed paths. Let the parameter of interest \( \theta \) be the vector of link success rates (i.e., complements of loss rates), and each probe outcome \( x \) be an indicator that the corresponding probe (sent on path \( p_y \)) successfully reaches its destination. Assume that losses of the same probe on different links and of different probes on the same link are both independent. Then the observation model is:

\[
f(x, y; \theta, \phi) = \phi_y (\prod_{l \in p_y} \theta_l)^x (1 - \prod_{l \in p_y} \theta_l)^{1-x}. \tag{2}
\]

2) Delay Jitter Tomography: Delay jitter is a typical performance metric that is additive over links on a path. Delay jitter between a pair of sender and receiver is defined as the change in inter-packet delay (IPD) due to the communication, i.e., the difference between the IPD at the receiver and the IPD at the sender\(^1\), and is a crucial performance metric in streaming applications. Suppose that jitters on individual links follow the normal distribution \( N(0, \theta_l) \) with zero mean and unknown variance \( \theta_l \) (\( l \in L \)). Assume that jitters experienced by the same probe on different links and by different probes on the same link are both independent. Delay jitter tomography aims at inferring \( \theta \) from the observed end-to-end jitter \( x \) based on the following observation model:

\[
f(x, y; \theta, \phi) = \phi_y \frac{1}{\sqrt{2\pi \sum_{l \in p_y} \theta_l^2}} \exp \left( -\frac{x^2}{2 \sum_{l \in p_y} \theta_l} \right). \tag{3}
\]

C. Main Problem: Optimal Experiment Design

The focus of this paper is to develop a systematic framework for optimally allocating probes over candidate paths such that the overall error in estimating \( \theta \) can be minimized. Specifically, given an error measure \( C(\theta, \phi) \) (e.g., L2-norm) and a total number of probes \( N \), we want to design the optimal probe allocation \( \phi \), such that in conjunction with an

\(^1\)One can verify that the end-to-end delay jitter on a path equals the sum of the corresponding jitters at each link.
appropriate estimator \( \hat{\theta} \), the expected error \( \mathbb{E}[C(\hat{\theta}, \theta)] \) after taking \( N \) probes can be minimized.

D. Examples

***example of why design is needed (e.g., benefit over even distribution)***

II. PRELIMINARY

Intuitively, the optimal experiment design will depend on the estimator used and vary for different applications. We will show, however, that there exists a fundamental limit on estimation performance that allows efficient experiment design independently of the specific estimator and guarantees asymptotic optimality (when used with a particular estimator) as \( N \) becomes large.

A. Fundamental Performance Limit

***vector-form CRB and definition of FIM***

B. Identifiability and Invertibility of \( I(\theta; \phi) \)

The CRB has an implicit assumption that the FIM is indeed invertible. In our problem we will show that the invertibility of \( I(\theta; \phi) \) directly follows from identifiability of the inference problem. We say that unknown parameters \( \theta \) are identifiable from observations \( x \) related to \( \theta \) via observation model \( f(x; \theta) \) if and only if \( f(x; \theta) \neq f(x; \theta') \) (at some \( x \)) for any \( \theta \neq \theta' \). We show the following relationships.

Theorem 1. Suppose that the stochastic link metric tomography problem can be cast as a linear system \( A \hat{z}(\theta) = w \), where \( A \) is the measurement matrix, \( \hat{z}(\theta) \) is a bijection of \( \theta \), and \( w \) is a vector of path performance parameters such that the probe outcomes depend on \( \theta \) only through \( w \). Moreover, suppose that \( w \) can be estimated consistently\(^2\) from probes. Then the following statements are equivalent:

1) \( \theta \) is identifiable;
2) \( \theta \) has full column rank;
3) the FIM \( I(\theta; \phi) \) is invertible.

Proof. 1) \( \Leftrightarrow \) 2): It is easy to see that the notion of identifiability is equivalent to the existence of a consistent estimator, i.e., an estimator whose error goes to zero as the sample size goes to infinity. If \( A \) has full column rank, then \( z^{-1}((A^T A)^{-1} A^T w) \) is a consistent estimator of \( \theta \) (\( w \) is a consistent estimator of \( \theta \)), and hence \( \theta \) is identifiable. If \( A \) does not have full column rank, then \( \exists z_1 \neq z_2 \) such that \( A z_1 = A z_2 = w \). Since probe outcomes are independent of \( \theta \) given \( w \), the observation models are identical for \( \theta_1 = z^{-1}(z_1) \) and \( \theta_2 = z^{-1}(z_2) \) (\( \theta_1 \neq \theta_2 \)), and hence \( \theta \) is unidentifiable.

1) \( \Leftrightarrow \) 3): If \( I(\theta; \phi) \) is invertible, then the maximum likelihood estimator (MLE; see Section III-A) must have a mean squared error (MSE) of order \( I^{-1}(\theta; \phi)/N \) in estimating \( \theta_1 \) for each \( l \in L \), and thus it must be a consistent estimator of \( \theta \), implying the identifiability of \( \theta \). If \( \theta \) is identifiable, then the MLE must be consistent. Meanwhile, as \( I(\theta; \phi) \) becomes singular, one of its eigenvalues must go to zero, which causes \( I^{-1}(\theta; \phi) \rightarrow \infty \) for some \( l \in L \), contradicting the consistency of MLE. Hence, \( I(\theta; \phi) \) must be invertible.

Both loss tomography and jitter tomography admit a linear system model \( Ax = w \), where \( z_l = \log \theta_l \), \( w_y = \log (\prod_{l \in p_y} \theta_l) \) for loss tomography, and \( z_l = \theta_l \), \( w_u = \sum_{l \in p_u} \theta_l \) for jitter tomography. Intuitively, identifiability is a baseline requirement that guarantees the inference problem is solvable with infinite measurements, and experiment design is a fine-tuning step that aims at maximizing the accuracy of inference with finite measurements. Therefore, we will assume in the sequel that our problem is identifiable (and hence \( A \) has full column rank, \( I(\theta; \phi) \) is invertible) to focus on the issue of experiment design.

III. MAXIMUM LIKELIHOOD ESTIMATOR (MLE) FOR STOCHASTIC LINK METRIC TOMOGRAPHY

Before going into experiment design, we first review a widely-adopted estimator and its special properties in our problem.

A. Background on MLE

Given observations, the MLE solves for the parameter value that maximizes the likelihood of these observations and uses this value as an estimate of the parameter. Although one can directly apply this definition to solve for MLE by searching for maximum of the likelihood function (e.g., by solving the likelihood equation), this approach can introduce complications in network tomography problems because the maximum may not be unique. Instead, we leverage a unique property of MLE, that it is invariant under one-to-one parameter transformations. Specifically, if \( \hat{\theta} \) is an MLE of \( \theta \) and \( \alpha = g(\theta) \) is a one-to-one transformation, then \( \hat{\alpha} = g(\hat{\theta}) \) is an MLE of \( \alpha \). We will see later that this approach allows us to borrow known results on MLE for classic (non-tomographic) estimation problems and greatly simplifies the derivation.

The MLE plays a significant role in FIM-based experiment design. According to the CRB (Section II-A), using the FIM (specifically \( \frac{1}{N} I^{-1}(\theta; \phi) \)) to represent the estimation error in experiment design implies an implicit assumption that the adopted estimator is efficient (i.e., unbiased and achieving the CRB). In this regard, the MLE has a superior property that it is the only candidate for efficient estimator, i.e., if an efficient estimator exists, it must equal the MLE (see Lemma 3.4 of [1])\(^3\). Moreover, although efficient estimators do not always exist for finite sample sizes, the MLE is asymptotically efficient under mild regularity conditions, i.e., its expectation converges to the true parameter at a rate approximating the CRB as the sample size \( N \) goes to infinity. ***Ting: verify the regularity conditions*** Therefore, adopting the MLE as the estimator guarantees that our FIM-based experiment design will optimize the decoding rate of the given error measure, as we send more and more probes.

\(^2\)That is, the error in estimating \( w \) does to zero as the number of probes goes to infinity.

\(^3\)The original proof is for scalar parameter, but analogous argument yields the same conclusion for vector parameter as long as the FIM is invertible.
B. MLE for Packet Loss Tomography

1) Log-likelihood Function and Sufficient Statistics: Based on the observation model (2), we can write down the log-likelihood function for packet loss tomography after $N$ probes as:

$$
\mathcal{L}(\theta) = \sum_{t=1}^{N} \log \phi_{y_t} + \sum_{t=1}^{N} x_t \log f(x = 1|y_t, \theta) \\
+ \sum_{t=1}^{N} (1 - x_t) \log f(x = 0|y_t, \theta),
$$

where $f(x = 1|y, \theta) := \prod_{t \in P_y} \theta_t$ and $f(x = 0|y, \theta) := 1 - \prod_{t \in P_y} \theta_t$ are the conditional success/loss probabilities given that the probing path is $p_y$. For each $y \in \{1, \ldots, |P|\}$, define $n_{1,y}$ as the number of probes sent on path $p_y$ and successfully received, and $n_{0,y}$ the number of probes sent on path $p_y$ but lost, both among the first $N$ probes. We can rewrite (4) as a function of $(n_{0,y}, n_{1,y})_{y=1}^{P}$ as:

$$
\mathcal{L}(\theta) = \sum_{y=1}^{|P|} (n_{0,y} + n_{1,y}) \log \phi_y + \sum_{y=1}^{|P|} \left[ n_{1,y} \log f(x = 1|y, \theta) \\
+ n_{0,y} \log f(x = 0|y, \theta) \right].
$$

Besides being a more compact form of the log-likelihood function, this equation also implies that $n := (n_{0,y}, n_{1,y})_{y=1}^{P}$ are sufficient statistics for the problem of packet loss tomography.

2) MLE: As mentioned before, a typical method to compute the MLE for loss tomography is by directly maximizing the log-likelihood function (5) over $\theta$. Instead of taking this approach, we notice that our problem is closely related to the problem of inferring path success rates. Given observations $(n_{0,y}, n_{1,y})$ about path $p_y$, we can view $n_{1,y}$ as the sum of $n_{0,y} + n_{1,y}$ i.i.d. Bernoulli random variables with success probability $\alpha_y$ (i.e., the path success rate for $p_y$), whose MLE is known to be $\hat{\alpha}_y = n_{1,y}/(n_{1,y} + n_{0,y})$. If we can establish a one-to-one transformation between $\alpha_y$’s and $\theta_i$’s, we can obtain the MLE of $\theta$ using the invariance property of MLE. We now formalize this idea. Without loss of generality, we assume that $n_{1,y} + n_{0,y} > 0$ for $y = 1, \ldots, |P|$, as $p_y$ can be ignored from $P$ otherwise. For ease of presentation, we use $g(z)$ to denote the vector obtained by applying a scalar function $g(\cdot)$ to each element of a vector $z$.

**Proposition 2.** If the measurement matrix $A$ has full column rank and there is at least one successful probe per path (i.e., $n_{1,y} > 0$ for $y = 1, \ldots, |P|$), then the MLE for loss tomography equals:

$$
\hat{\theta} = \exp \left( (A^T A)^{-1} A^T \log \hat{\alpha}(n) \right),
$$

where $\hat{\alpha}(n) := (\hat{\alpha}_y(n))_{y=1}^{|P|}$ for $\hat{\alpha}_y(n) := n_{1,y}/(n_{1,y} + n_{0,y})$ is the vector of empirical path success rates.

**Proof.** Since $A$ has full column rank, the link success rates $\theta$ and the path success rates $\alpha$ form a one-to-one mapping $\log \theta = (A^T A)^{-1} A^T \log \alpha$ for $\alpha > 0$. Since the MLE of $\alpha$ is known to be $\hat{\alpha}(n)$ (by classic results on Bernoulli random variables), the MLE of $\theta$ follows from its invariance property.

3) Example: Consider a simple 2-link network as illustrated in Fig... Applying the MLE formula in (6) yields that the MLE of $(\hat{\theta}_1, \hat{\theta}_2)^T$ is

$$
\hat{\theta}_1 = \frac{n_{1,1} + n_{1,3}}{n_{1,1} + n_{1,3} + n_{1,2}}, \\
\hat{\theta}_2 = \frac{n_{1,2}(n_{1,1} + n_{1,3})}{n_{1,1} + (n_{1,2} + n_{1,3})}.
$$

C. MLE for Delay Jitter Tomography

1) Log-likelihood Function and Sufficient Statistics: For delay jitter tomography, we have the following log-likelihood function for a total of $N$ probes based on (3):

$$
\mathcal{L}(\theta) = \sum_{t=1}^{N} \log \phi_{y_t} - \frac{1}{2} \sum_{t=1}^{N} \log (2\pi \sum_{l \in p_y} \theta_l) - \frac{1}{2} \sum_{t=1}^{N} \frac{x_{t}^2}{\theta_l}.
$$

For each $y \in \{1, \ldots, |P|\}$, define $n_y$ as the number of probes sent on path $p_y$ and $x_{y,k}$ the end-to-end jitter for the $k$-th probe on $p_y$. We can rewrite (8) as:

$$
\mathcal{L}(\theta) = \sum_{y=1}^{|P|} n_y \log \phi_y - \frac{1}{2} \sum_{y=1}^{|P|} n_y \log (2\pi \sum_{l \in p_y} \theta_l) - \frac{1}{2} \sum_{y=1}^{|P|} \sum_{k=1}^{n_y} x_{y,k}^2 \theta_l.
$$

The final expression (9) implies that $(n_y, \sum_{k=1}^{n_y} x_{y,k}^2)_{y=1}^{|P|}$ are sufficient statistics for the problem of delay jitter tomography.

2) MLE: Following the same approach as in Proposition 2, we can show the following for delay jitter tomography. Again, we assume without loss of generality that $n_y > 0$ for $y = 1, \ldots, |P|$.

**Proposition 3.** If the measurement matrix $A$ has full column rank, then the MLE for jitter tomography equals:

$$
\hat{\theta} = (A^T A)^{-1} A^T \hat{\beta},
$$

where $\hat{\beta} := (x_{y,k}^2)_{y=1}^{|P|}$ for $x_{y,k}^2 := \frac{1}{n_y} \sum_{k=1}^{n_y} x_{y,k}^2$ is the vector of empirical variances of path jitter.

**Proof.** When $A$ has full column rank, the link jitter variances $\theta$ and the path jitter variances $\beta$ form a one-to-one mapping $\theta = (A^T A)^{-1} A^T \beta$. Moreover, as the variance of a zero-mean normal random variable (i.e., path jitter), it is known that the MLE of $\beta_y$ equals the empirical variance $x_{y}^2$. The MLE of $\theta$ then follows from the invariance property of MLE.

3) Example: We use the same example as illustrated in Fig... to demonstrate the MLE for jitter tomography. By (10), the MLE in this case is

$$
\hat{\theta}_1 = x_1^2 = \frac{1}{n_2} \sum_{k=1}^{n_2} x_{1,k}^2, \\
\hat{\theta}_2 = x_2^2 - x_1^2 = \frac{1}{n_1} \sum_{k=1}^{n_1} x_{1,k}^2.
$$

**D. Implication on Experiment Design**

Applying the MLE formulas in Propositions 2 and 3 requires special treatment in the probing process. Besides requiring the probed paths to form a full-column-rank measurement matrix, loss tomography also requires at least one successful probe per path. Both requirements can be satisfied by introducing an initialization phase, where instead of randomly selecting probing paths according to a distribution \( \phi \), we sequentially probe a set of paths that contain a basis of \( \mathbb{R}^{|L|} \) (hence yielding a full-column-rank measurement matrix) until the requirements for applying MLE are satisfied. For jitter tomography, it suffices to send one probe per path; for loss tomography, the initialization may take multiple probes per path until obtaining a success.

**IV. Optimal Experiment Design based on FIM**

The essence of FIM-based experiment design is to treat the CRB as an approximate of the estimation error matrix and select the design parameter \( \phi \) to optimize a given objective function based on the CRB. Given an estimated value of \( \theta \), the FIM (and hence the CRB) only depends on \( \phi \), which in theory allows us to optimize over \( \phi \). In practice, however, solving this optimization is highly nontrivial as it requires the optimization of a \(|P|\)-dimensional vector, which renders numerical solutions inefficient and error-prone for large \(|P|\). In this section, we will show that under certain conditions, this optimization can be solved in closed-form for some well-known design objectives, and the conditions are satisfied by both loss and jitter tomography.

**A. D-Optimal Design**

In D-optimal experiment design, we seek to minimize the determinant of inverse FIM \( \det(I^{-1}(\theta; \phi)) \), or equivalently maximize \( \det(I(\theta; \phi)) \). The CRB implies that this design will minimize the volume of the error ellipsoid.

1) Structure of \( \det(I(\theta; \phi)) \): We begin by establishing a special structure of the determinant of the FIM that holds for any network topology and any set of probing paths, under certain conditions on the observation model. We first show a general property of the FIM as follows.

**Lemma 4.** The FIM for the observation model (1) equals a convex combination of per-path FIMs:

\[
I(\theta; \phi) = \sum_{y=1}^{|P|} \phi_y I^{(y)}(\theta),
\]

where \( I^{(y)}(\theta) \) is the FIM for path \( p_y \) based on the observation model \( f_y(x; \theta) \). Note that \( I^{(y)}(\theta) \) is independent of \( \phi \) and is only a function of \( \theta \).

**Proof.** Let \( \mathcal{L}(\theta) \) and \( \mathcal{L}_y(\theta) \) be the log-likelihood functions for the overall experiment and path \( p_y \), respectively. Since \( \mathcal{L}(\theta) = \log \phi_y + \mathcal{L}_y(\theta) \), applying the definition of FIM (REFER TO DEFINITION in Section II-A) yields:

\[
I_{i,j}(\theta; \phi) = -\mathbb{E}_{x,y}[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \mathcal{L}(\theta)]
= \sum_{y=1}^{|P|} \phi_y \left[ -\mathbb{E}_{x,y}[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \mathcal{L}_y(\theta)] \right],
\]

and \( -\mathbb{E}_{x,y}[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \mathcal{L}_y(\theta)] \) equals \( I^{(y)}_{i,j}(\theta) \) by definition.

Based on this decomposition, we can show that the determinant of the FIM has a particular structure as follows.

**Theorem 5.** Assume that the number of paths is not less than the number of links \(|P| \geq |L|\), and let \( S_{|L|} \) be the collection of all size-\(|L|\) subsets of \( \{1, \ldots, |P|\} \). If the per-path FIM satisfies

\[
I^{(y)}_{i,j}(\theta) I^{(y)}_{j,i}(\theta) = I^{(y)}_{i,i}(\theta) I^{(y)}_{j,j}(\theta)
\]

for any \( y \in \{1, \ldots, |P|\} \) and any \( i, j, k, l \in \{1, \ldots, |L|\} \) such that \( i \neq j, k \neq l \), then there exist functions \( B_C(\theta) \) indexed by the elements of \( S_{|L|} \) such that

\[
\det(I(\theta; \phi)) = \sum_{C \in S_{|L|}} B_C(\theta) \prod_{i \in C} \phi_i.
\]

Functions \( B_C(\theta) \) (\( C \in S_{|L|} \)) do not depend on \( \phi \) and are not all zero.

**Proof.** Applying the Leibniz formula for determinant calculation to the decomposed FIM in (12) shows that

\[
\det(I(\theta; \phi)) = \sum_{\pi \in \Pi_{|L|}} \text{sgn}(\pi) \prod_{i=1}^{|L|} I_{i,\pi_i}(\theta; \phi)
= \sum_{\pi \in \Pi_{|L|}} \text{sgn}(\pi) \left( \prod_{y_1=1}^{|L|} \cdots \prod_{y_{|L|}=1}^{|L|} \phi_{y_1} I^{(y_1)}_{1,\pi_1}(\theta) \right),
\]

where \( \pi \) is a permutation of \( \{1, \ldots, |L|\} \) (\( \Pi_{|L|} \) is the set of all permutations), and \( \text{sgn}(\pi) \) is a sign function that equals 1 if \( \pi \) is achievable by an even number of pairwise swaps, and \(-1\) if it is achievable by an odd number of swaps. Equation (17) shows that the determinant of the FIM can be written as a sum of order-\(|L|\) terms of \( \phi \) (i.e., \( \prod_{i=1}^{|L|} \phi_{y_i} \)), weighted by functions of \( \theta \). Each term in the summation is uniquely determined by \( \pi \) and \( y \).

The key to the proof is to show that after combining these order-\(|L|\) terms, the remaining terms only contain product of \(|L|\) distinct \( \phi_y \)’s, i.e., terms containing duplicate variables \( (y_i = y_j \text{ for } i \neq j) \) all disappear. We prove this by showing that terms with duplicate variables will combine to zero.

For each term with at least one duplicate variable, i.e., the corresponding \( \pi \) and \( y \) satisfy: \( \exists i, j \in \{1, \ldots, |L|\} \) \((i \neq j)\) such that \( y_i = y_j = y_0 \) for some \( y_0 \in \{1, \ldots, |P|\} \), there must exist a corresponding term, referred to as the opposite term, for the same \( y \) and a slightly different permutation \( \pi' \) that is identical as \( \pi \) except that \( \pi'_i = \pi_j \) and \( \pi'_j = \pi_i \). The
absolute value of this opposite term equals
\[
\prod_{k \neq i,j} \phi_{y_k} I_{k,i}^{(y_k)}(\theta) \phi_{y_j}^2 I_{j,i}^{(y_j)}(\theta) I_{j,j}^{(y_j)}(\theta),
\]
which equals the absolute value of the first term
\[
\prod_{k \neq i,j} \phi_{y_k} I_{k,i}^{(y_k)}(\theta) \phi_{y_j} I_{j,i}^{(y_j)}(\theta) I_{j,j}^{(y_j)}(\theta)
\]
because \(I_{i,j}^{(y_i)}(\theta)I_{j,i}^{(y_j)}(\theta) = I_{i,j}^{(y_j)}(\theta)I_{j,i}^{(y_i)}(\theta)\). Meanwhile, \(\text{sgn}(\pi)\) and \(\text{sgn}(\pi')\) must differ as the permutations differ by one pairwise swap. Therefore, the two terms sum up to zero.

Moreover, if we define the opposite term of a term containing duplicate variables as the term obtained by swapping \(\pi_i\) and \(\pi_j\) for the first two duplicate variables (i.e., for the smallest \(i, j\) with \(y_i = y_j\)), then it is easy to see that the opposite term of the opposite term is the original term, and thus no two different terms can have the same opposite. Therefore, after combining terms, only terms consisting of a product of \(|L|\) distinct \(\phi_i\)'s remain, implying the formula (15).

Remark: The essence of this theorem is that under the condition (14), the determinant of the FIM, when viewed as a function of \(\phi\), can be written as a weighted sum of order-\(|L|\) terms, where each term is a product of \(|L|\) distinct \(\phi_i\)'s. We will show later that this property can help to simplify our FIM-based experiment design.

In fact, similar arguments can be used to show an analogous formula for any minor\(^4\) of the FIM as follows.

**Corollary 6.** Let \(M\) be an \(n\)-dimensional submatrix of \(I(\theta; \phi)\) after removing \(|L| - n\) rows and columns, and \(S_n\) be the collection of all size-\(n\) subsets of \([1, \ldots, |P|]\). Then under the condition (14), there exist functions \(B_C(\theta)\) indexed by the elements of \(S_n\), such that the determinant of \(M\) (i.e., a minor of \(I(\theta; \phi)\)) admits the following decomposition:

\[
\det(M(\theta; \phi)) = \sum_{C \in S_n} B_C(\theta) \prod_{i \in C} \phi_i.
\]

Functions \(B_C(\theta) (C \in S_n)\) do not depend on \(\phi\) and are not all zero.

**Proof.** The proof is analogous to that of Theorem 5. Let \(\{r_1, \ldots, r_n\}\) and \(\{c_1, \ldots, c_n\}\) denote the rows and columns of \(I(\theta; \phi)\) that remain in \(M\). From the proof of Theorem 5, we see that the minor can be decomposed as:

\[
\det(M(\theta; \phi)) = \sum_{\pi \in \Pi_n} \text{sgn}(\pi) \left( \prod_{y_1=1}^{|P|} \prod_{y_n=1}^{|P|} \phi_{y_i} I_{y_i,y_i}^{(y_i)}(\theta) \right),
\]

where \(\Pi_n\) is the set of all permutations of \(\{1, \ldots, n\}\), and \(I_{y_i,y_j}^{(y_i)}(\theta)\) is the \((i, j)\)-th entry of the per-path FIM for path \(y_i\). Then by similar arguments as in Theorem 5, we can show that each term containing duplicate \(\phi_i\) variables must have a (unique) opposite term that is equal in absolute value but opposite in sign. Therefore, after combining the terms, the only remaining terms are those corresponding to products of \(n\) distinct \(\phi_i\)'s, which proves the result.

\(\Box\)

2) **Optimal Experiment Design:** Theorem 5 implies that when \(|P| = |L|\), the determinant of the FIM is proportional to the product of \(\phi_i\)'s:

\[
\det(I(\theta; \phi)) = B(\theta) \prod_{i=1}^{|L|} \phi_i.
\]

Since \(\sum_{i=1}^{|L|} \phi_i = 1\), by the inequality of arithmetic and geometric means, we see that (21) is maximized by setting \(\phi_i = 1/|L|\) for all \(i = 1, \ldots, |L|\), i.e., uniform probing is D-optimal (when \(|P| = |L|\)).

**B. A-Optimal Design**

In A-optimal experiment design, we seek to minimize the trace of inverse FIM \(\text{Tr}(I^{-1}(\theta; \phi))\). The CRB states that this design will minimize the mean squared error (MSE) for estimating \(\theta\).

1) **Structure of \(\text{Tr}(I^{-1}(\theta; \phi))\):** Theorem 5 implies, in particular, that when the numbers of paths and links coincide (i.e., \(|P| = |L|\)), the following multiplicative decomposition for the determinant of the FIM holds:

\[
\det(I(\theta; \phi)) = B(\theta) \prod_{k=1}^{|L|} \phi_k.
\]

This fact, together with Corollary 6, can be used to prove the following structure of \(\text{Tr}(I^{-1}(\theta; \phi))\).

**Theorem 7.** Suppose \(|P| = |L|\) and the FIM is invertible. If the condition (14) holds, then the trace of the inverse FIM \(\text{Tr}(I^{-1}(\theta; \phi))\) admits the following representation:

\[
\text{Tr}(I^{-1}(\theta; \phi)) = \sum_{i=1}^{|L|} \frac{1}{\phi_i} A_i(\theta),
\]

where \(A_1(\theta), \ldots, A_{|L|}(\theta)\) are only functions of \(\theta\).

**Proof.** Let us denote the \((i, j)\)-element of \(I^{-1}(\theta; \phi)\) by \(I_{i,j}^{-1}(\theta; \phi)\). Applying Cramer’s rule of calculating the inverse of a matrix, we can write

\[
I_{i,j}^{-1}(\theta; \phi) = (-1)^{i+j} \frac{\det(M_{ji}(\theta; \phi))}{\det(I(\theta; \phi))},
\]

where \(\det(M_{ji}(\theta; \phi))\) is the minor of element \((j, i)\) of \(I(\theta; \phi)\) (i.e., the determinant of the submatrix after removing row \(j\) and column \(i\)). In particular, the diagonal elements of \(I^{-1}(\theta; \phi)\) have the following form:

\[
I_{k,k}^{-1}(\theta; \phi) = \frac{\det(M_{kk}(\theta; \phi))}{\det(I(\theta; \phi))}, \quad k = 1, \ldots, |L|.
\]

Corollary 6 implies the following representation of the numer-
The trace of $I^{-1} (\theta; \phi)$ is then equal to

\[
\text{Tr}(I^{-1} (\theta; \phi)) = \sum_{k=1}^{[L]} I_{k,k}^{-1} (\theta; \phi)
\]

\[
= \frac{1}{\det(I(\theta; \phi))} \sum_{k=1}^{[L]} \sum_{C \in S_{[L]-1}} B_{C,k} (\theta) \prod_{i \in C} \phi_i
\]

\[
= \sum_{k=1}^{[L]} \sum_{C \in S_{[L]-1}} \frac{B_{C,k} (\theta)}{B(\theta)} \prod_{i \in C} \phi_i
\]

\[
= \sum_{k=1}^{[L]} \prod_{i \in C} \phi_i \frac{B_{C,k} (\theta)}{B(\theta)}
\]

where we have used the multiplicative representation (21) for $\det(I(\theta; \phi))$. Next, we observe that $S_{[L]-1}$ has exactly $[L]$ members $C_1, \ldots, C_{[L]}$, where each $C_i$ is the subset of $\{1, \ldots, [L]\}$ that excludes $i$. Thus,

\[
\text{Tr}(I^{-1} (\theta; \phi)) = \sum_{i=1}^{[L]} \frac{1}{\phi_i} \left( \sum_{k=1}^{[L]} B_{C_i,k} (\theta) \right) / B(\theta)
\]

\[
= \sum_{i=1}^{[L]} \frac{1}{\phi_i} A_i (\theta),
\]

where $A_i (\theta) = \sum_{k=1}^{[L]} B_{C_i,k} (\theta) / B(\theta)$. \hfill \Box

**Remark:***Ting: perhaps a counterexample that the formula does not hold when there are more than $[L]$ paths***

**2) Optimal Experiment Design:*** Andrei: proof of $A_i(\theta) > 0$***

Using the Lagrange Multiplier method, we have a closed-form solution for minimizing (22) wrt $\phi$ as the following:

\[
\phi_i = \frac{\sqrt{A_i (\theta)}}{\sum_{j=1}^{[L]} \sqrt{A_j (\theta)}}
\]

for $i = 1, 2, \ldots, [L]$.

**C. Weighted A-Optimal Design**

In practice, different links may be of different values to the application. To model this aspect, we extend the A-optimal design criterion to a new criterion, called weighted A-optimal design, that differentiates the importance of monitoring different links by a weight vector $\omega := (\omega_k)_{k=1}^{[L]}$, where $\omega_k$ denotes the weight of link $l_k$. Adding weights changes the objective from minimizing $\text{Tr}(I^{-1} (\theta; \phi))$ to minimizing the weighted sum of the diagonal elements of $I^{-1} (\theta; \phi)$:

\[
\sum_{k=1}^{[L]} \omega_k I_{k,k}^{-1} (\theta; \phi).
\]

By the CRB, this design will minimize the weighted average MSE for estimating $\{\theta_i\}_{i \in L}$.

1) **Structure of $\sum_{k=1}^{[L]} \omega_k I_{k,k}^{-1} (\theta; \phi)$:** Based on analogous arguments, we can easily extend Theorem 7 to the following result.

**Corollary 8.** Under the conditions in Theorem 7, the weighted sum of the diagonal elements of the inverse FIM admits the following representation:

\[
\sum_{k=1}^{[L]} \omega_k I_{k,k}^{-1} (\theta; \phi) = \sum_{i=1}^{[L]} \frac{1}{\phi_i} A_i (\theta),
\]

where $A_1 (\theta), \ldots, A_{[L]} (\theta)$ are only functions of $\theta$.

**Remark:** Using the notations in the proof of Theorem 7, we can explicitly write out the expression for $A_i (\theta)$ as:

\[
A_i (\theta) = \sum_{k=1}^{[L]} \omega_k B_{C_i,k} (\theta) / B(\theta).
\]

**2) Optimal Experiment Design:** ***Modified from design under A-optimal criterion***

**D. Application to Loss and Jitter Tomography**

We are now ready to apply our generic results on FIM-based experiment design to the problems of loss and jitter tomography. The first step is to verify condition (14), which is the basis for Theorem 5, Theorem 7, and Corollary 8. Moreover, to use these formulas in experiment design, we need an efficient way of evaluating the coefficients ($B_{C_i}(\theta), A_i(\theta)$, and $A_1(\theta)$) for a given (estimated) value of $\theta$.

**1) Application to Packet Loss Tomography:** Based on the observation model (2), we can obtain the per-packet FIM $I^{(y)}(\theta)$ for loss tomography, whose $(i, j)$-th entry equals

\[
I^{(y)}_{i,j}(\theta) = f(x = 1|y, \theta) \theta_i \theta_j \sum_{y \in \mathbb{Y}} f(x = 0|y, \theta) \mathbb{1}\{i, j \in p_y\},
\]

where $\mathbb{1}\{\cdot\}$ is the indicator function. It is easy to verify that this FIM satisfies the condition (14), and thus the formulas in Theorem 5, Theorem 7, and Corollary 8 hold.

To derive explicit expressions for their coefficients, we take a detailed look at the FIM, which leads to a decomposition into a product of matrices with special structures. Plugging (20) into (12) gives the $(i, j)$-th entry of the FIM as

\[
I_{i,j}(\theta) = \sum_{y=1}^{P} \phi_y f(x = 1|y, \theta) f(x = 0|y, \theta) \mathbb{1}\{i, j \in p_y\}.
\]

We introduce two auxiliary matrices$^5$: $D = \text{diag} \left( (d_y)_{y=1}^{P} \right)$ for $d_y := \phi_y f(x = 1|y, \theta) / f(x = 0|y, \theta)$, and $\Theta = \text{diag} (\theta)$. Then the above FIM can be written in matrix form as

\[
I(\theta; \phi) = \Theta^{-1} A^T D A \Theta^{-1},
\]

where $A$ is the measurement matrix.

Based on this decomposition, we can evaluate its determinant and trace of inverse as functions of $\Theta$, $A$, and $D$, which lead to the following results.

**Lemma 9.** Let $A_C$ denote a $[L] \times [L]$ submatrix of the measurement matrix $A$ formed by rows with indices in $C$.

$^5$Here $\text{diag} (\mathbf{d})$ denotes a diagonal matrix with the main diagonal $\mathbf{d}$. 
Then $\det(I(\theta; \phi))$ for loss tomography can be expressed as (15) with coefficients

$$B_C(\theta) = \frac{\det(A_C)^2}{\prod_{l \in L} \theta_l^2} \prod_{i \in C} \frac{f(x = 1|i, \theta)}{f(x = 0|i, \theta)}$$

for each $C \in S_{|L|}$.

Moreover, if $|P| = |L|$ and $I(\theta; \phi)$ is invertible, then $\text{Tr}(I^{-1}(\theta; \phi))$ can be expressed as (22) with coefficients

$$A_i(\theta) = \frac{f(x = 0|i, \theta)}{f(x = 1|i, \theta)} \sum_{k=1}^{[L]} \theta_k^2 b_{k,i}^2$$

for $i = 1, \ldots, |L|$, where $b_{k,i}$ is the $(k, i)$-th entry of $A^{-1}$. Similarly, the weighted sum of the diagonal elements of $I^{-1}(\theta; \phi)$ can be expressed as (28) with coefficients

$$\tilde{A}_i(\theta) = \frac{f(x = 0|i, \theta)}{f(x = 1|i, \theta)} \sum_{k=1}^{[L]} \omega_k \theta_k^2 b_{k,i}^2,$$

where $\omega_k$ is the weight of link $l_k$.

**Proof.** To derive $B_C(\theta)$, we evaluate the determinant of the FIM by $\det((\theta^{-1})^2 \det(A^T DA))$. Applying the Cauchy-Binet formula to $\det(A^T DA)$ gives

$$\det(I(\theta; \phi)) = \prod_{l \in L} \theta_l^2 \sum_{C \subseteq S_{|L|}} \det(A_C) \det((DA)_C),$$

where similar to $A_C, (DA)_C$ is a $[L] \times [L]$ submatrix of $DA$ formed by rows with indices in $C$. Since $D$ is diagonal, we can further decompose $\det((DA)_C)$ into $\det(D_C) \det(A_C)$, where $D_C = \text{diag}((d_y)_{y \in C})$. Since the only term depending on $\phi$ is $\det(D_C)$, we can rewrite (35) as a function of $\phi$ as

$$\det(I(\theta; \phi)) = \sum_{C \subseteq S_{|L|}} \left[ \frac{\det(A_C)^2}{\prod_{l \in L} \theta_l^2} \prod_{i \in C} \frac{f(x = 1|i, \theta)}{f(x = 0|i, \theta)} \prod_{i \in C} \phi_i \right],$$

which matches formula (15) with $B_C(\theta)$ defined as in (32).

To derive $A_i(\theta)$, we evaluate the inverse of the FIM by $\Theta A^{-1} D^{-1} A^{-T} \Theta$. Denoting $A^{-1}$ as $[b_{i,j}]_{i,j=1}^{[L]}$, we can evaluate the $k$-th diagonal entry as $I_{k,k}^{-1}(\theta; \phi) = \theta_k^2 \sum_{i=1}^{[L]} b_{k,i}^2 d_i^{-1}$ since $\Theta$ and $D^{-1}$ are diagonal. Plugging in $d_i^{-1} = f(x = 0|i, \theta)/[\phi_i f(x = 1|i, \theta)]$ yields

$$\text{Tr}(I^{-1}(\theta; \phi)) = \sum_{k=1}^{[L]} \frac{\theta_k^2}{\phi_i} \frac{b_{k,i}^2 f(x = 0|i, \theta)}{f(x = 1|i, \theta)} \cdot \phi_i$$

$$= \sum_{i=1}^{[L]} \frac{1}{\phi_i} \left[ \frac{f(x = 0|i, \theta)}{f(x = 1|i, \theta)} \sum_{k=1}^{[L]} \theta_k^2 b_{k,i}^2 \right],$$

which matches formula (22) with $A_i(\theta)$ defined as in (33). The same derivation will give the expression for $\tilde{A}_i(\theta)$.

**Remark:** For the case of $|P| > |L|$ and invertible $I(\theta; \phi)$, we can also give an explicit expression for $\text{Tr}(I^{-1}(\theta; \phi))$.

The key is to plug the decomposed $I(\theta; \phi)$ into Cramer’s formula of calculating $I_{k,k}^{-1}(\theta; \phi)$ (see (24)). Let $A^{(k)}$ denote the submatrix of $A$ by removing the $k$-th column and $A^{(k)}_{C_C}$ the submatrix of $A^{(k)}$ formed by rows with indices in $C$. A derivation similar to the proof of Lemma 9 shows that

$$\text{Tr}(I^{-1}(\theta; \phi)) = \sum_{C \subseteq S_{|L|-1}} \left[ \sum_{k=1}^{[L]} \theta_k^2 \det(A^{(k)}_{C_C})^2 \prod_{i \in C} d_i \right],$$

which is a rational expression of $\phi$. A similar expression holds for the weighted sum of the diagonal elements of $I^{-1}(\theta; \phi)$.

2) **Application to Delay Jitter Tomography:** Similarly, from the observation model (3), we can derive the per-path FIM for jitter tomography as

$$f_{i,j}^{(y)}(\theta) = \frac{1}{2\left( \sum_{l \in p_y} \phi_l \right)^2} \prod_{i \neq j \in p_y},$$

which also satisfies the condition (14).

Applying (39) to (12) gives an expression for individual entries of the FIM for jitter tomography. Observing its similarity to the FIM for loss tomography, we again write it in matrix form by introducing another auxiliary matrix $E = \text{diag}((e_y)_{y \in P})$, where $e_y = \phi_y/[2(\sum_{l \in p_y} \phi_l)^2]$. It can be verified that the FIM for jitter tomography satisfies $I(\theta; \phi) = A^T EA$. This decomposition immediately leads to the following results.

**Lemma 10.** The $B_C(\theta)$ for jitter tomography can be expressed as (15) with coefficients

$$B_C(\theta) = \frac{\det(A_C)^2}{2^{[L]} \prod_{l \in C} \left( \sum_{l \in p_l} \phi_l \right)^2}$$

for each $C \in S_{|L|}$ ($AC$ defined as in Lemma 9).

Moreover, if $|P| = |L|$ and $I(\theta; \phi)$ is invertible, then $\text{Tr}(I^{-1}(\theta; \phi))$ can be expressed as (22) with coefficients

$$A_i(\theta) = 2 \left( \sum_{l \in p_i} \phi_l \right)^2 \sum_{k=1}^{[L]} \theta_k^2 b_{k,i}^2,$$

for $i = 1, \ldots, |L|$ ($b_{k,i}$ defined as in Lemma 9). Similarly, the weighted sum of the diagonal elements of $I^{-1}(\theta; \phi)$ can be expressed as (28) with coefficients

$$\tilde{A}_i(\theta) = 2 \left( \sum_{l \in p_i} \phi_l \right)^2 \sum_{k=1}^{[L]} \omega_k \theta_k^2 b_{k,i}^2.$$

**Proof.** The proof is analogous to that of Lemma 9 based on evaluating $\det(A^T EA)$ and $A^{-1} E^{-1} A^{-T}$.

**Remark:** Similar to loss tomography, for the case of $|P| > |L|$ and invertible $I(\theta; \phi)$, we can explicitly write
Tr(I^{-1}(\theta; \phi)) for jitter tomography as

\[
\text{Tr}(I^{-1}(\theta; \phi)) = \frac{\sum_{C \in S_{|L|-1}} \left[ \sum_{k=1}^{|L|} \det(A_C^{(k)})^2 \right] \prod_{i \in C'} e_i}{\sum_{C \in S_{|L|}} \det(A_C)^2 \prod_{i \in C} e_i},
\]

which is again a rational expression of \( \phi \), and a similar expression holds for the weighted variation. ***Ting: But the question is: does this rational expression make it easier to optimize \( \phi \)?***

REFERENCES