A Unified Information Criterion for Evaluating Probe and Test Selection

Juan Liu, Johan de Kleer, Lukas Kuhn, Bob Price, and Rong Zhou
Palo Alto Research Center,
3333 Coyote Hill Rd, Palo Alto, CA94304.
Email: {juan.liu, dekleer, lkuhn, bprice, rzhou}@parc.com

Abstract—**Diagnostic tasks often need to make the decision of what measurement to make or what action to take in order to resolve ambiguities in diagnosis. Intuitively one would like to seek the most "informative" choice. In the paper, we formalize this intuition and propose an information criterion for evaluating and comparing measurement/action choices based on their information contribution. The criterion is mutual information, an information-theoretic concept measuring statistical dependence. The information criterion gives a precise quantitative metric to differentiate the quality of measurement/action choices. We use a few concrete example in two separate paradigms, probe selection in circuit diagnosis and test generation in production plants, to illustrate the mutual information criterion. Despite the apparent differences of the two paradigms, the information criterion works coherently. We demonstrate how different probing actions or test plans vary in their information values.**

Index Terms—**Mutual Information, Diagnostics, Information Theory.**

Fig. 1. Scope of this paper: we will only focus on the information criterion (the block with thick border).

I. INTRODUCTION

A significant challenge in diagnostic tasks is to identify what new measurements to make next or what new experiments to try next in order to resolve ambiguities quickly. These two tasks are often termed "active probing" and "test generation". While their goals are clear, the actual practice of selecting probing locations and designing test is more art than science. In this paper, we outline a general conceptual framework which unifies the tasks of selecting probing locations and test plans based on the concept of mutual information. The basic idea is simple: not all probing and test choices are equal; some are more informative than others. In this paper, we formalize the intuition using an information theoretic concept, mutual information. It provides a single metric to precisely evaluate the amount of information that a probing measurement or a test plan are expected to bring to the diagnostic problem. Using this metric, different choices can be compared fairly.

In this paper, we focus on the information evaluation criterion rather than the overall diagnosis. Figure 1 shows the general flowchart of sequential diagnosis. Based on a current diagnosis belief, one may decide where to make observations. This is the top block in the diagram marked as "probe selection or test generation". The measurement is then taken, and used to update the belief via some inference mechanism such as a GDE (General Diagnostic Engine) [1] or a Bayesian inference engine [2]. With the updated belief, the probe selection/test generation process may repeat to find
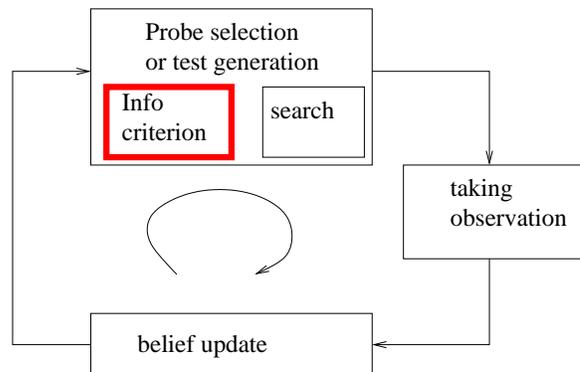
the next suitable action. In this paper, we will not address the inference or the measurement process, but only focus on the probe selection/test generation part. Furthermore, this part involves an information evaluation criterion to measure the quality of different choices and a search strategy to find the optimal choice. The search is a sophisticated problem by itself. We will not address the search problem in this paper, but only show that the information criterion provides accurate heuristics to guide the search. The focus of this paper is on the information criterion (the block with thick border in Figure 1).

The information criterion for evaluating probe and test selections is most useful if active probing and test execution incurs a non-negligible cost. This is often the case as special equipments or technician labor are needed. On the other hand, if the probing and test execution is nearly effortless, then evaluation and selection would not be a problem. The burden is shifted to the inference part: being able to update the diagnosis in the presence of a large amount of test evidence. In this paper, we restrict our discussion in the former case; furthermore, we assume that all probing actions or test plan executions incur a uniform cost. If the cost is non-uniform, we can devise the selection strategy to strike a balance between two competing goals: finding an action with the best information content, and yet keeping the action cost low. Action cost is typically known prior to taking the action, and our information criterion provides a metric for the information content.

In this paper, we first explain in Sec. II, on an abstraction level, how the information criterion is formulated, computed,

and used to guide the decision of which measurement to take. Then the information criterion is instantiated in two separate paradigms: (1) probe selection in circuit diagnosis, and (2) test plan generation in production plants. These two paradigms are conceptually different and have been addressed via quite different techniques, but we show that they share the common need for informative measurements, and can be unified under the information criterion. In Sec. III, we show the information criterion for active probing in circuit diagnosis with two concrete examples: a simple linear cascade of inverter and a full adder. Our analysis shows that different probing locations vary a lot in terms of their information content, and the best probe location depends on the fault assumptions. In Sec. IV, we explain the information criterion for the plan selection in modular production plants. In this case, the information criterion takes a simple form and can be evaluated efficiently. In Sec. V, we discuss a few possible extensions of the information criterion. The paper concludes with Sec. VI.

II. MUTUAL INFORMATION CRITERION

To measure the information content, we consider mutual information, a concept rooted in information theory measuring statistical dependence. It is a commonly used metric for characterizing the performance of data compression and classification [3].

For illustration, we use the following notation: let X be the underlying diagnostic state and Y be the observation. For example, X is the the bit-vector 011000 if the only second and third module have fault, and Y is the measurement obtained at a probing location, or the outcome of a test plan execution. Note that X and Y are both random variables. The mutual information between X and Y is defined as:

$$I(X;Y) \overset{\triangle}{=} \sum_{x,y} \left[p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \right].$$ (1)

It measures the amount of information (in bits if logarithm is in base 2) the observation Y tells about the underlying diagnosis state X. It has the following properties:

- Symmetry: $I(X;Y) = I(Y;X)$ — the amount of information that X tells about Y is the same as the amount that Y tells about X.
- Non-negativity: $I(X;Y) \geq 0$, with equality only when X and Y are independent.
- Connection with entropy:

$$I(X;Y) = H(X) - H(X|Y),$$ (2)

where $H(X) \overset{\triangle}{=} \sum_x \left[p(x) \log \frac{1}{p(x)} \right]$ is the entropy of X, and $H(X|Y) \overset{\triangle}{=} \sum_{x,y} \left[p(x,y) \log \frac{1}{p(x|y)} \right]$ is the entropy of X conditioning on observing Y, i.e., the "remaining uncertainty" after the observation.

In diagnostic tasks, observations are made from probing locations or test executions. Hence Y is actually implicitly parametrized by the probing location or the test plan. To emphasize this, we use Y_m, with m denoting the measuring
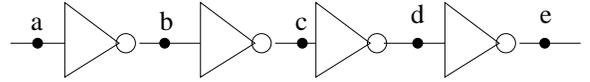

Fig. 2. Cascaded Inverters.

action. The worst possible choice is when $I(X;Y_m)$ is zero, in which case, measuring any value of Y_m has no implication on refining the underlying diagnosis X. In practice, we should avoid such an irrelevant observation, but rather make an observation that reveals as much information as possible regarding X. The goal for probe/plan selection is to find the m such that $I(X;Y_m)$ is maximized. Maximizing $I(X;Y_m)$ is equivalent to minimizing $H(X|Y_m)$. This is equivalent to say: we would like to select the best probe/plan, which leaves as little uncertainty as possible. This intuition of minimizing conditional entropy is used for example in the General Diagnostic Engine (GDE) [1].

Search for the best probe/plan maximizing $I(X;Y_m)$ involves search over all possible plans; we will not address the search problem in this paper. Rather, it is straight-forward to compare probes/plans. For example, given two choices m_1 and m_2, we say m_1 is more informative and preferable than m_2 if,

$$I(X;Y_{m_1}) > I(X;Y_{m_2}).$$ (3)

To calculate mutual information, we take advantage of the symmetry: exchanging X and Y in (2), we have

$$I(X;Y) = H(Y) - H(Y|X).$$ (4)

Although (2) and (4) are equivalent, the latter is often much easier to compute, since it uses the observation likelihood $p(y|\mathbf{x})$ which is often known a priori. In contrast, (2) uses the posterior belief $p(\mathbf{x}|y)$, which is a lot harder to compute.

Mutual information has been used as an evaluation and selection criterion in a number of applications. For example, [4] uses mutual information to decide which sensors to activate in the context of tracking a moving target. Similarly, [5] uses an information criterion to control a fleet of robots, sending robots to most advantageous locations. Medical researchers have used mutual information to guide feature selection to diagnose human lung cancer [6]. Related to machine diagnostics, [7] used a similar criterion in process diagnosis based on Bayesian networks. Its information criterion is derived for continuous health states. In this paper, we extend this framework to the discrete diagnostics domain and make connections between high-level intuitions and precise information evaluation.

III. PROBE SELECTION

Circuit diagnosis is a canonical example of model-based reasoning. In this paradigm, a number of techniques have been proposed. The early GDE work [1] proposes a minimum entropy criterion that determines what measurements to make next: the one minimizing the conditional entropy of candidate probabilities resulting from the measurement. This is well aligned with our information criterion. In this section, we extend this early work to a variety of fault assumptions.

Here we use a simple example of a cascade of inverters (Figure 2). The same circuit was also used as a working example for illustration in [1]. An inverter gives $y_{output} = \neg y_{input}$ if it is working properly. An inverter that has a fault may produce wrong output. In the strong fault assumption, we assume that faulty inverter operates in a known way. Or one may retreat to a weak fault assumption, assuming that a faulty inverter operates in an unknown way and produces incorrect output for some inputs. In practice, intermittency of fault may add further complications. For instance, a faulty component may not always malfunction. Its malfunctioning is a random event with probability q due to unknown environmental conditions. The diagnosis of intermittent fault has been addressed for example in [8]. In the analysis below, we consider four combinations: strong and weak faults, persistent ($q = 1$) and intermittent ($0 < q < 1$), and show how the fault assumption affects the optimal probing locations.

In this example, with four cascaded inverters, the underlying hypothesis space contains 16 hypotheses: $X = \{0000, 0001, \ldots, 1111\}$. Active probing in this case is to compare the probing locations $m = \{a, b, c, d, e\}$ and compute the respective mutual information values $I(X; Y_m)$.

A. Strong fault models, persistent

In the strong fault case, a faulty inverter operates in a known way, e.g., $y_{output} = y_{input}$. In practice, the input/output relationship of a faulty component can be learned via diagnostic inference. To evaluate the mutual information, we first need to specify the observation likelihood model $p(y_m|x)$. Note that in this linear cascaded inverter example, we have,

$$p(y_m|\mathbf{x}) = \sum_{(y_1, \ldots, y_{m-1}, y_{m+1}, \ldots, y_K)} p(y_1, y_2, \ldots, y_K|\mathbf{x}), \quad (5)$$

where K is the total number of components in the linear cascade ($K = 4$), and

$$p(y_1, y_2, \ldots, y_K|\mathbf{x}) = p(y_1) \prod_i p(y_{i+1}|y_i, x_i). \quad (6)$$

The individual term $p(y_{i+1}|y_i, x_i)$ is the property of the i-th inverter module, with y_i as its input and y_{i+1} as the output. The variable x_i is 0 if the i-th inverter has no fault and 1 otherwise.

The component-wise likelihood function is the following:

$$p(y_{i+1}|y_i, x_i) = \begin{cases} 1 & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 0; \\ 0 & \text{if } y_{i+1} = y_i \text{ and } x_i = 0; \\ 1 & \text{if } y_{i+1} = y_i \text{ and } x_i = 1; \\ 0 & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 1; \end{cases} \quad (7)$$

Plugging the component-wise likelihood into (6) to get the joint distribution $p(y_1, y_2, \ldots, y_K|\mathbf{x})$ and marginalizing (as in (5)) to obtain $p(y_m|\mathbf{x})$, we can evaluate the mutual information.

Under the strong fault assumption, given any hypothesis x and initial input at point a, there is no uncertainty in the output of any inverter, hence the second term $H(Y_m|X) = 0$. Therefore, we simply have $I(X; Y_m) = H(Y_m)$.

TABLE I
THE MUTUAL INFORMATION $I(X; Y_m)$ FOR DIFFERENT PROBING LOCATIONS $m = \{a, b, c, d, e\}$ IN THE CASCADED INVERTERS EXAMPLE. THE BOLD FONTS MARK THE BEST PROBING LOCATION (WITH MAXIMAL $I(X; Y_m)$). THE PRIOR KNOWLEDGE IS THAT FOUR INVERTERS MAY HAVE FAULT WITH PROBABILITY $0.2, 0.1, 0.1, 0.1$ RESPECTIVELY. THE TABLE IS EVALUATED AFTER THE OBSERVATION $a = 1, e = 0$.

	Persistent strong	Intermittent strong $q = 0.1$	Persistent weak	Intermittent weak $q = 0.1$
a	0	0	0	0
b	**0.988**	0.7648	**0.849**	0.755
c	0.961	**0.7652**	0.839	**0.757**
d	0.721	0.585	0.633	0.579
e	0	0	0	0

To illustrate the information criterion, let us consider a concrete example. For the cascaded inverters shown in Figure 2, with input $a = 1$ and output $e = 0$, there must be something wrong with this circuit, and the diagnostic task needs to decide which location to probe. The mutual information $I(X; Y_m)$ for different probing locations $m = \{a, b, c, d, e\}$ is listed in Table 1 (second column) under the strong persistent fault assumption. The initial condition is that the inverters A, B, C, D are independently faulty with probability $0.2, 0.1, 0.1, 0.1$ respectively. The best probing location in this case is b, immediately after the first inverter. This is intuitive, given that inverter A is most likely to have fault than others. The two ends a and e have zero information value, since we already know their values $a = 1$ and $e = 0$.

The preceeding likelihood function (7) was derived by direct inspection. Model-based diagnosis algorithms such as GDE compute $p(y_m|\mathbf{x})$ through first principles reasoning from a description of the system. For example, the 1st line of equation (7) is inferred from the fact that a correctly working inverter ($x_i = 0$) always ($p(y_{i+1}|y_i, x_i) = 0$) inverts its output ($y_{i+1} = \neg y_i$). For more details of such algorithms see [1].

B. Strong fault models, intermittent

The component-wise likelihood function is:

$$p(y_{i+1}|y_i, x_i) = \begin{cases} 1 & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 0; \\ 0 & \text{if } y_{i+1} = y_i \text{ and } x_i = 0; \\ q & \text{if } y_{i+1} = y_i \text{ and } x_i = 1; \\ (1-q) & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 1; \end{cases} \quad (8)$$

To evaluate mutual information, we need to compute $p(y_m|\mathbf{x})$. Starting from point a, the probability of $p(y_{i+1}|\mathbf{x})$ can be evaluated recursively from $p(y_i|\mathbf{x})$, until reaching the probe location m. We use the shorthand notation $p_i(b)$ for $p(y_i = b|\mathbf{x})$ with $b = 0, 1$. The evaluation is recursive:

$$p_{i+1}(b) = \begin{cases} p_i(\neg b) \cdot q + p_i(b) \cdot (1-q) & \text{if } x_i = 1 \\ p_i(\neg b) & \text{if } x_i = 0 \end{cases} \quad (9)$$

Through the recursion, we can push to the probe location m to evaluate the outcome probability $p(y_m = b|\mathbf{x})$, and then compute the mutual information $I(X; Y_m)$ as in (4).

The analysis above does not assume that all inverters have the same intermittent parameter q. On the other hand, if they

have the same q value, it further simplifies into

$$p(y_m|\mathbf{x}) = \begin{cases} \sum_{\text{even } k} \left[C_N^k q^k (1-q)^{N-k} \right] & \text{for } y_m = a \\ \sum_{\text{odd } k} \left[C_N^k q^k (1-q)^{N-k} \right] & \text{for } y_m \neq a \end{cases}$$

(10)

Here N is the total number of modules that have fault before the probing location m in the hypothesis $\mathbf{x}$, i.e., $N = \sum_{i=1,\ldots,m} x_i$. The notation C_N^k denotes N-choose-k. This result is also easy to understand: if the outcome at probing location m is the same as a, then there must be an even number of malfunctioning inverters between a and m, which can be chosen randomly from a total of N possible faults. Hence the probability has the C_N^k term and the polynomial term with even k. Same for the case of $y_m \neq a$, which must have an odd number of faults. From $p(y_m|\mathbf{x})$, we can evaluate $H(Y_m|X)$.

The third column of Table I shows the comparison of probing locations under the intermittent strong fault assumption. The best probing location is c instead of b as in the previous two cases. On the conceptual level, the output at a good probing location should have a decent probability of observing an actual malfunctioning unit. If that probability is too low, one cannot learn much from the measurement. In the intermittent fault case, the small q value means that the malfunctioning is rarely occurring. This causes the best probing location to shift towards the middle. In this example, if we observe $c = 0$, then there must be a fault in only of the first two inverters; if $c = 1$, then one of the last two inverters must have a fault.

C. Weak fault models, persistent

In the weak fault model, the observation likelihood is the following:

$$p(y_{i+1}|y_i, x_i) = \begin{cases} 1 & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 0; \\ 0 & \text{if } y_{i+1} = y_i \text{ and } x_i = 0; \\ 0.5 & \text{if } y_{i+1} = 0 \text{ and } x_i = 1; \\ 0.5 & \text{if } y_{i+1} = 1 \text{ and } x_i = 1; \end{cases}$$

(11)

In this case, the mutual information criterion is the following:

$$I(X; Y_m) = H(Y_m) - p_m^U,$$

(12)

where p_m^U denotes the probability that the outcome at m is unknown (0 or 1). The derivation is straight forward: at any given measuring location m, with any particular hypothesis $\mathbf{x} \in \mathcal{X}$, its outcome can be deterministically 0, deterministically 1, or unknown. For instance, with $a = 1, e = 0$, the no-fault hypothesis $\{0000\}$ is ruled out. With the remaining hypotheses, the measurement at location b has a few possible values: (i) 0 under the hypotheses $\mathcal{X}_0 = \{0001, 0010, 0011, 0100, 0101, 0110, 0111\}$, since the first module has no fault; (ii) 1 under the hypothesis $\mathcal{X}_1 = 1000$ since the last three modules have no fault; and (iii) unknown under all other remaining hypotheses $\mathcal{X}_{\mathcal{U}}$. The conditional entropy

$$H(Y|X) = \sum_{\mathbf{x} \in \mathcal{X}} H(Y|X = \mathbf{x}) p(\mathbf{x})$$

(13)

This sum can be broken down into three sets: over the set $\mathcal{X}_1$, the conditional entropy is 0 since the outcome is deterministic

with value 1; same for the set $\mathcal{X}_0$. The only remaining set is $\mathcal{X}_{\mathcal{U}}$, in which each hypothesis $\mathbf{x}$ has a corresponding conditional entropy $H(Y|X = \mathbf{x}) = 1$ bit from the equal probability 0/1 outcome, and the whole set has an accumulated probability of p_m^U. Putting them altogether, we have (12).

The fourth column of Table I shows the $I(X; Y_m)$ values. Similar to the persistent strong fault case, the best probing location is b.

The prior work [1] proposes a minimum entropy criterion of selecting m to minimize $H(X|Y)$. Through a lengthy derivation, its entropy criterion is $-H(Y_m) + p_m^U \log M$, where M is the number of distinct values that Y_m can take. This is exactly the same as in (12). The mutual information derivation is much simpler, and can be readily generalized to a variety of fault assumptions.

D. Weak fault models, intermittent

For the weak fault, intermittency will make the output less random. The component-wise likelihood function is:

$$p(y_{i+1}|y_i, x_i) = \begin{cases} 1 & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 0; \\ 0 & \text{if } y_{i+1} = y_i \text{ and } x_i = 0; \\ q/2 & \text{if } y_{i+1} = y_i \text{ and } x_i = 1; \\ 1 - q/2 & \text{if } y_{i+1} = \neg y_i \text{ and } x_i = 1; \end{cases}$$

(14)

The $q/2$ in the third line comes from the fact that a faulty module malfunctions with a probability of q and in that situation, the probability of observing a wrong output is 0.5 due to the weak fault assumption. In this particular cascaded inverter example, the intermittent weak fault case is identical to the strong intermittent fault case, except with a new q value. The mutual information criterion can be computed in the same way.

The last column of Table I enumerates the mutual information values under the intermittent weak fault assumption. The best probe location is c. Another thing to note in Table I is that the information content decreases when the faults become intermittent. The persistent strong fault column has the highest values. This is because the observation likelihood model is very informative: observing the input and output of any inverter, one can immediately say whether the inverter has fault. In the intermittent weak fault assumption, the input-output observation is hardly conclusive: observing a correct input-output pair does not mean the inverter is good. One can conclude the inverter is bad only when a malfunctioning shows up and the output happens to be wrong, but this is a very rare event by the intermittency nature. Hence we see this information value decreasing as we retreat to weaker and weaker assumptions.

E. Another example: full adder

Another example is the diagnosis of a full adder, shown in Figure 3. It exhibits a level of sophistication lacking in the previous example, but typical in real circuits: the gates are diverse, and the connections are non-linear. We show that the mutual information criterion can be used to guide the probe selection in this representative circuit.
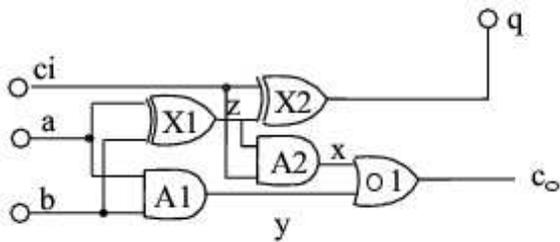
Fig. 3. Full adder.

TABLE II
THE MUTUAL INFORMATION $I(X; Y_m)$ FOR DIFFERENT PROBING
LOCATIONS $m = \{x, y, z\}$ IN THE FULL ADDER CIRCUIT. THE BOLD
FONTS MARK THE BEST PROBING LOCATION (WITH MAXIMAL $I(X; Y_m)$).
EACH MODULE IS ASSUMED TO HAVE FAULT WITH PRIOR PROBABILITY
0.2. THE TABLE IS EVALUATED FOR INPUTS $\{a = 1, b = 1, c_i = 0\}$ AND
OUTPUTS $\{q = 1, c_o = 1\}$.

	Persistent strong	Intermittent strong $q = 0.1$	Persistent weak	Intermittent weak $q = 0.1$
x	**0.773**	**0.049**	0.368	0.042
y	0.440	0.003	0.100	0.001
z	0	0	**0.889**	**0.808**

The full adder takes three inputs: a, b, c_i (carry-in), and produces two outputs q and c_o (carry-out). With inputs $\{a = 1, b = 1, c_i = 0\}$ and outputs $\{q = 1, c_o = 1\}$, it is clear that the circuit is incorrect and needs diagnosis. The possible probe locations are x, y, and z. Which one to probe is the choice that need to be made. We again consider four fault combinations. For the strong fault model, we assume the following: (1) any faulty XOR gate operates like an OR gate; (2) any faulty AND gate operates like a NAND gate; and (3) any faulty OR gate operates like a NOR gate. These are assumptions for illustration in this paper, and should be adjusted accordingly in real diagnosis problems. For the weak fault model, we assume any malfunctioning gate produces 0 or 1 with equal probability.

Table II lists the results under a variety of fault assumptions. Notice a few things from the table:

- Not all probing locations are equal, and some could even be useless. For instance, probing the location z under the strong fault model (the second column, the last row) has zero information. Due to the strong fault assumption, only a handful of fault combinations could have produced the observed input/output relationship. Concatenating the gates in the order (X1,A1,X2,A2,O1), the only possible fault combinations are: $X = \{10000, 10010, 10100, 10110, 11001, 11010, 11101, 11110\}$. Under all these combinations, we will have $z = 1$. Hence probing at z does not contribute to the diagnosis.
- The best probe location changes as the fault assumption varies. For instance, the best probe location under the strong fault model (the second and third columns) are x, while the best probe location under the weak fault model (the last two columns) is z.

- With fault intermittency, the probing action becomes less informative. This is consistent with what we observed in the cascaded inverters example.

F. Extension: test vector generation

One diagnosis strategy is to choose suitable test vectors. By varying the input to the circuits, for instance, the input $\{a, b, c\}$ to the full adder, one may isolate faults and help the overall diagnostic task. Similar information criteria can be extended to test vector selection: to choose the most informative test vector. The same mutual information evaluation mechanism can be used. Instead of evaluating $I(X; Y_m)$ and choose the best probe location m, we can change the variable m to be the input test vectors, or even the combination of probe locations and test vectors. The detailed evaluation would be different, but the general idea of using mutual information to differentiate the quality of test choices remains the same. One direct approach to find the next best test vector is to apply GDE's approach to every possible input vector.

IV. PLAN SELECTION

A common diagnostic problem is the diagnosis of a production plant. A product often goes through many steps or modules in the manufacturing process. When the outcome is unsatisfactory, one needs to diagnose which step or steps have caused the problem. In this section, we use PARC's prototype modular printer (shown in Figure 4) for illustration. This printer has over 170 independently controlled modules and many possible paper paths; the redundancy enables high-speed high-throughput printing [9]. The product in this case is a paper sheet, which enters from the left and exits on the right. It may go through paper path modules (dark black edges with small rollers in the figure) and printer modules (the 4 large rectangles). At the output, we may observe a fault; the most commonly observed is a damaged paper (wrinkled, ripped, or dog-eared). Unlike active probing in circuit diagnosis, we cannot make observations at arbitrary modules before a paper sheet exits. What can be leveraged is to control which modules the paper sheet goes through. For example, if we suspect the top-right printer module to have a fault but not the other three printer modules, we may control the path so that (1) it avoids the top-right printer if we want to maintain a working system, or (2) it passes the top-right printer, if our goal is to diagnose the suspected fault.

Although we use the prototype printer as our illustration vehicle, the analysis in this section is general and can be extended to the diagnosis of many production plants. At the abstraction level, the problem of plan selection is to choose the most informative production plan P to maximize the information content $I(X; Y_P)$, where X is the underlying diagnosis state, and Y_P is the outcome of the production plan, which is binary-valued: 0 for success plan, and 1 for unsuccessful (e.g., damaged paper sheet).

A. Single intermittent fault

In practical situations, the number of potential faults is small. A simplification of the diagnostic problem is to assume
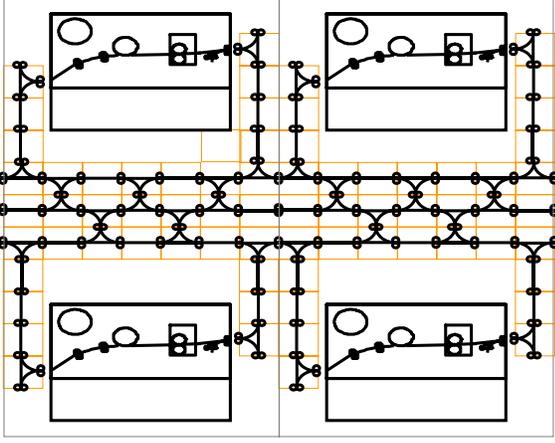
Fig. 4. PARC's prototype printer. It consists of two towers each containing 2 printers (large rectangles). Sheets enter on the left and exit on the right. Dark black edges with small rollers represent possible paper paths.

that the whole system has at most one fault, known as the single fault assumption. It reduces the diagnosis space from exponential 2^M to M, where M is the total number of modules, i.e., $X = \{1, 2, \ldots, M\}$. The prior knowledge is that module i has fault with probability p_i. We further assume that if a module i has fault, it has an intermittency probability of q_i.

The question is how to diagnose this modular printer when a fault has been observed? A common divide-and-conquer scheme is to devise a plan P to go through only half of the modules. If the plan observes a fault, that means P contains the fault, and the other half that P excludes is cleared of suspicion. If the plan is successful, then P is cleared, and the fault must be in the other half. In this way, every plan dissects the diagnosis space by half. This divide-and-conquer strategy is used for example in [10].

The strategy can be generalized via the mutual information criterion. In the production plant abstraction, each faulty module can damage the product with an intermittent probability q if it is included in the production plan P, i.e., with the observation likelihood:

$$p(y_P|X = i) = \begin{cases} 0 & \text{if } y = 1 \text{ and } i \notin P; \\ 1 & \text{if } y = 0 \text{ and } i \notin P; \\ q_i & \text{if } y = 1 \text{ and } i \in P; \\ 1 - q_i & \text{if } y = 0 \text{ and } i \in P; \end{cases} \quad (15)$$

Define $H_q^{(i)}$ as the entropy corresponding to the binomial distribution q_i, i.e., $H_q^{(i)} \triangleq -[q_i \log q_i + (1 - q_i) \log(1 - q_i)]$. The mutual information can be evaluated in the following form

$$I(X; Y_P) = [-y_0 \log y_0 - y_1 \log y_1] - \sum_{i \in P} p_i H_q^{(i)}, \quad (16)$$

where y_0 is the probability of observing a success, y_1 is the probability of observing a failure. The derivation follows from (4). The first term (the bracketed term) is $H(Y_P)$, and one can easily verify the second term is $H(Y_P|X) = \sum_{i \in P} p_i H_q^{(i)}$.

An interesting special case is when all faults are persistent, i.e., $q_i = 1$ for all i. In this case, all $H_q^{(i)} = 0$, and the second term in (16) vanishes. The mutual information is hence only $-y_0 \log y_0 - y_1 \log y_1$, maximized when $y_0 = y_1 = 0.5$. This means the path P should go through half of the probability mass, i.e., $\sum_{i \in P} p_i = 0.5$. This is a generalization of the divide-and-conquer strategy above.

In the intermittent fault case, the second term is non-zero and can be considered as a "correction" term due to the intermittency. When all the modules have the same q_i value (this is likely in the paper path modules which all have the identical design), the mutual information can be further simplified. It can be evaluated as the function of a single variable $w = \sum_{i \in P} p_i$:

$$I(X; Y_P) = -[wq \log wq + (1 - wq) \log(1 - wq)] - wH_q \quad (17)$$

The last term H_q is the entropy of the Bernoulli distribution with parameter q. Note the mutual information is very easy to evaluate: given any plan P, we can obtain w as the summation of the probability mass along the plan, then plug in to obtain the corresponding mutual information value. The computation only involves several additions and multiplications.

Now consider the more complicated problem of plan generation. The optimal plan should maximize $I(X; Y_P)$. Since w is the only variable, the optimal value of w can be derived from simple calculus, and is

$$w = \frac{1}{q(2^{H_q/q} + 1)} \quad (18)$$

When $q \to 0$, w is asymptotically approaching $\frac{1}{e}$. For $q \in (0, 1]$, w takes value from $\frac{1}{e}$ to $\frac{1}{2}$. This is an interesting result: as the faults become less likely to show, we should include less probability mass in the plan P. If the plan comes out with a damaged product, the rest of the probability mass not included in P is ruled out by the single fault assumption. On the other hand, if the plan comes out without showing any damage, the modules in P cannot be ruled out due to intermittency, but have to be tested further.

Implication to the search: The original path search problem, i.e., searching for the best path P which has the maximal information gain, has been reduced to a much simpler yet equivalent problem of searching for a path P, which has an accumulative probability closest to a target value w. The latter problem is simple because the accumulative probability of the path is additive by nature. Handling a cost function that is additive helps the search problem tremendously, because (1) the order of modules in the path does not matter, (2) sub-paths can be summarized from their contribution to the overall cost function, and (3) it enables tree pruning.

Given the target value w as in (18), efficient search algorithm can be used to find the best plan. The target-value search strategy proposed in [10] tackles this problem, It starts from the product entrance (or exit), grows the search tree, and prunes it by establishing upper- and lower-bounds on the deviation of the accumulated probability to the target value w.

The search is very efficient. Interested readers may refer to [10] for details.

B. Multiple faults

We extend the analysis to multiple faults. For simplicity, we assume all modules have identical q values. The diagnosis space is $\mathbf{x} = (0/1, \ldots, 0/1)$, where the i-th element value is an indicator function regarding whether this module has fault. For each possible diagnosis, we have a probability $p(\mathbf{x})$.

To evaluate the mutual information $I(X;Y) = H(Y) - H(Y|X)$, we compute the two terms as follows.

- The first term $H(Y)$ is the entropy corresponding to the binomial distribution (y_0, y_1), with

$$
\begin{aligned}
y_0 = p(y = 0) &= \sum_{\mathbf{x}} p(y = 0|\mathbf{x})p(\mathbf{x}) \quad (19) \\
&= \sum_{\mathbf{x}} (1 - q)^{k(\mathbf{x}, P)} p(\mathbf{x}) \quad (20)
\end{aligned}
$$

here $k(\mathbf{x}, P)$ is the number of faulty modules that P goes through. : $k(\mathbf{x}, P) = \sum_{i \in P} x_i$. k is a random variable with a distribution $p(k)$. The distribution can be derived from the diagnosis belief $p(\mathbf{x})$ and the plan P. With $p(k)$ computed, we have $y_0 = \sum_k (1 - q)^k p(k)$.

- For any given value of k, the outcome is 0 with probability $(1 - q)^k$ and 1 with probability $1 - (1 - q)^k$, and we use H_k to denote the entropy corresponding to this distribution. The conditional entropy is

$$
H(Y|X) = H(Y|K) = \sum_k p(k) H_k \quad (21)
$$

A potential path P affects the cost function only over k and its distribution, i.e., the number of *faulty modules* it passes. If two different paths offer the same distribution of k, they are essentially the same from the mutual information perspective. Given any plan P, we can evaluate $p(k)$ and H_k.

Note the following a few special cases:

- If $q = 1$, the second term is 0, and the optimal for the first term is $y_0 = y_1 = 0.5$ (selecting a path which has equal chance of observing a damaged/undamaged paper). This is the same as $p(k = 0) = 0.5$, i.e., with a probability of one half, all the modules in P are good.
- Consider the initial condition that all modules are independently faulty with probability s ($s < 1$). In this case, we can choose the optimal plan length(M), and which ones to include in the path does not matter since all modules are identical in their faulty probability. With $s = 0.5$, the optional path P goes through only one module. With smaller s values, the optimal length is longer. To see this, note that $p(k) = C_M^k \cdot s^k (1 - s)^{M-k}$, and $y_0 = (1 - qs)^M$. For y_0 to get close to 0.5, M increases as s decreases.

The mutual information criterion can be used to guide the search for optimal production path. The exact search problem is difficult (the number of possible paths is exponential) and is out of the scope of this paper. However, the evaluation of mutual information is easy. This evaluation can be used to compare a few plan candidates, or make local adjustment to an existing plan — e.g., adding a new module or deleting an existing module, in order to obtain an informative measurement. This serves the diagnostics goal and minimizes the number of further tests.

V. DISCUSSION

The two paradigms illustrated above, probe selection in circuit diagnosis and test plan generation in production plant diagnosis, both use a greedy strategy. At any step, the selection process uses the information criterion to find the most informative measurement to make for the time being. This greedy strategy works well for the diagnosis of static system, where the underlying ground truth of component fault does not change over time. However, there is no guarantee of optimality. On the other hand, the information criterion, $I(X;Y_m)$ can be re-formulated with a look-ahead horizon, i.e., instead of computing the mutual information between X and the immediate probing action Y_m, we can compute the mutual information $I(X; Y_m^{(t=1)}, Y_m^{(t=2)}, \ldots, Y_m^{(t=T)})$, where T is the look-ahead horizon. Using this criterion, one would be able to compare choices on their relatively long term contribution. The drawback is that the new criterion $I(X; Y_m^{(t=1)}, Y_m^{(t=2)}, \ldots, Y_m^{(t=T)})$ is much harder to evaluate. The state space grows exponentially. Various approximation techniques can be used, see [5] for an example. Another strategy is to use greedy strategy most of the time, but switch to the look-ahead strategy only occasionally to avoid getting trapped in local optimum.

The search for optimal test sequence is known to be NP-hard. The work in [11] proposes a roll-out algorithm, inspired by policy iteration of dynamic programming, to search for a suboptimal solution. Our information criterion with single and multi-step lookahead can be readily combined with this roll-out strategy.

VI. CONCLUSION

This paper proposes an information criterion for evaluating and selecting which measurement to make to help diagnosis. The criterion is based on mutual information, rooted in information theory to measure the dependence of random variables. The information criterion can be used in a variety of diagnostic problems, such as active probing in troubleshooting circuits and test plan generation in production plant diagnosis. From the analysis we can see that different action choices vary in their information contribution, and thus it is essential to be able to evaluate and compare them. The information criterion can further guide the search for optimal actions.

REFERENCES

[1] J. de Kleer and B. C. Williams, "Diagnosing multiple faults," *Artificial Intelligence*, no. 32, pp. 97–130, 1987.

[2] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer Verlay, New York, 1995.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY: John Wiley and Sons, Inc., 1991.

[4] J. Liu, J. E. Reich, and F. Zhao, "Collaborative in-network processing for target tracking," *EURASIP, Journal on Applied Signal Processing*, vol. 2003, pp. 378–391, Mar. 2003.

[5] G. M. Hoffmann, S. L. Waslander, and C. J. Tomlin, "Mutual information methods with particle filters for mobile sensor network control," in *Proceedings of the 45th IEEE Conference on Decision and Control (CDC)*, (San Diego, California, USA), Dec. 2006.

[6] G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Floyd, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, pp. 2394–2402, Dec. 2001.

[7] S. Verron, T. Tiplica, and A. Kobi, "Procedure based on mutual information and Bayeisan networks for the fault diagnosis of industrial systems," in *Proceedings of the 2007 American Control Conference*, (New York City, USA), July 2007.

[8] J. de Kleer, "Diagnosing intermittent fault," in *Proceedings of the 18th International Workshop on Principles of Diagnosis (DX-07)*, (Nashville, TN, USA), May 2007.

[9] W. Ruml, M. Do, and M. Fromherz, "On-line planning and scheduling for high-speed manufacturing," in *Proceedings of ICAPS*, pp. 30–39, 2005.

[10] L. Kuhn, B. Price, J. de Kleer, M. Do, and R. Zhou, "Heuristic search for target-value path problem," in *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-08)*, 2008. to appear.

[11] F. Tu and K. R. Pattipati, "Rollout strategies for sequential fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 33, no. 1, pp. 86–99, 2003.