Analysis of Uncertain Data:
Evaluation of Given Hypotheses

Anatole Gershman
anatoleg@cs.cmu.edu
www.cs.cmu.edu/~anatoleg

Eugene Fink
e.fink@cs.cmu.edu
www.cs.cmu.edu/~eugene

Bin Fu
binf@cs.cmu.edu
www.cs.cmu.edu/~binf

Jaime G. Carbonell
jgc@cs.cmu.edu
www.cs.cmu.edu/~jgc

Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, United States

Abstract—**We consider the problem of heuristic evaluation of given hypotheses based on limited observations, in situations when available data are insufficient for rigorous statistical analysis.**

Keywords—**Uncertainty, hypothesis evaluation, Bayesian reasoning, artificial intelligence.**

I. INTRODUCTION

WHEN estimating the likelihood of given hypotheses, we sometimes need to base our analysis on very limited data. Such situations are common in the fields where collection of statistically significant evidence is prohibitively expensive or impossible. For example, when an astrophysicist compares alternative cosmological theories, she may have data on a limited number of related astronomical objects, as there may be no other such objects within the observable range. Similarly, when a physician diagnoses a patient's illness, she may have a limited number of available tests.

Researchers have long realized the importance of techniques for hypothesis evaluation based on limited data, and accumulated much evidence that the humans apply such techniques in their reasoning. Bayes studied this problem back in the eighteenth century and derived his famous rule for updating probabilities [Bayes, 1763]. Modern scientists have extended his results and developed Bayesian-network tools [Pearl, 1988]. Recently, researchers have investigated related challenges in business and military intelligence [Heuer, 1999; Elsaesser and Stech, 2006].

We have addressed several related problems and built a suite of tools for representation of insufficient and partially inaccurate data [Bardak et al., 2006a; Fu et al., 2008]; reasoning under uncertainty [Fink et al., 2006; Gardiner et al., 2008]; and targeted data gathering [Bardak et al., 2006b].

We report one of the developed approaches to evaluating hypotheses based on limited evidence. We give an example that illustrated this problem (Section III), formalize the general problem (Section IV), describe a technique for solving it (Section IV–VIII), and give initial experiments (Section IX).

A closely related complementary problem is to construct

plans for targeted information gathering when available data are insufficient. Examples of such situations include collecting astronomical observations and ordering medical tests. We have addressed this complementary problem in another publication [Gershman et al., 2009].

II. EXAMPLE

We consider the task of a business analyst who is observing a small company and trying to infer the plans of its management. The company has recently released a new product and announced that its main focus is on expanding its sales. The analyst however suspects that it is also working on a second new product, which has a potential for greater sales. If she is right, the company is developing the second product in secrecy, as it does not want to reveal its plans to competitors. The analyst has to distinguish between two hypotheses:

- The company focuses exclusively on expanding the sales of the announced product.
- The company puts significant resources into the development of another product.

She also has to account for the possibility that neither of her hypotheses is correct, and something entirely different is in the offing. For instance, the company may be sold to a competitor or file for a bankruptcy. While the chances of such unexpected outcomes are low, they are not negligible.

The analyst has some idea of prior probabilities from her past experience with similar situations. For instance, she may believe that the prior for the first hypothesis is 0.6, for the second is 0.3, and for unexpected developments is 0.1. If she had no other information, she would use these probabilities; however, she has other data, which include public accounting records provided by the company, statements of its president, announced contracts, and so on. While a lot of these data are irrelevant, some may be "gold nuggets" that greatly help the analyst in her task. She has to (1) identify relevant data and (2) evaluate posterior probabilities of her hypotheses.

While the last two tasks may sound similar to the standard Bayesian reasoning, there are two important differences.

- Available data may correlate in complex ways, and the analyst may not know about these correlations. For instance, several news articles may come from different sources (making them near-independent) or from the same

source (making them highly dependent), and the analyst may not know which is the case.

- The analyst has to evaluate the chances that none of her hypotheses is correct, which cannot be directly addressed by the Bayesian rule.

III. GENERAL PROBLEM

To formalize this problem, suppose that the analyst has to distinguish among n mutually exclusive hypotheses, denoted $H_1, H_2, \ldots, H_n$. She bases her analysis on m observable features of the company, denoted $obs_1, obs_2, \ldots, obs_m$, where each observation is a variable that takes one of several discrete values. Suppose further that she has the following information about prior probabilities, likely observations under different hypotheses, and actual observed values.

- **Prior probabilities:** For each hypothesis, the analyst knows its prior probability; thus, she has an array of n priors, denoted $prior[1..n]$. The sum of its elements may be less than 1.0, which means that none of the hypotheses may be correct. Intuitively, the "no correct hypothesis" situation is a surprise that does not match the analyst's expectations.

- **Possible observations:** For every observation obs_a, the analyst knows the number of its possible values, denoted $num[a]$. Thus, she has the array $num[1..m]$ that represents the number of possible values for each of the m observations. We assume without loss of generality that the values of obs_a are the integers from 1 to $num[a]$. In real-world situations, they may be nominals, strings, and so on, but of course we can always enumerate them with integers.

- **Observation distributions:** The likelihood of specific observed values depends on which hypothesis is correct. For every hypothesis, the analyst knows the related probability distribution of each observation. Thus, she has a matrix with $n \cdot m$ elements, denoted $chance[1..n, 1..m]$, where each element is a probability-density function of possible observed values. Every element $chance[i, a]$ of this array is itself a one-dimensional array with $num[a]$ elements, representing the probabilities of possible values of obs_a. That is, $chance[i, a][b]$ is the likelihood of observing the bth value of obs_a in case if hypothesis H_i is correct.

- **Actual observations:** The analyst also knows a specific value of each observation, which represents the available data about the company. Thus, she has an array of m observed values, denoted $val[1..m]$.

The analyst has to use this knowledge in evaluating the posterior probabilities of the given hypotheses.

IV. USE OF ONE OBSERVATION

We begin with a simple special case, in which (1) the analyst can observe only one feature obs_a of the company, and (2) the given hypotheses completely cover all possibilities, leaving no room for any surprises.

- **Problem 1:** *Suppose that we have one observed value, $val[a]$, and that the sum of the prior$[1..n]$ probabilities is*

exactly 1.0. We have to determine the posterior probabilities of the n given hypotheses.

In this case, the analyst can directly apply the Bayesian rule. First, she finds the *integrated likelihood* (also called *marginal likelihood*) of observing $val[a]$, by summing the chances of observing it under all hypotheses:

$$likelihood(val[a])$$
$$= chance[1, a][val[a]] \cdot prior[1]$$
$$+ \ldots + chance[n, a][val[a]] \cdot prior[n].$$

Then, she calculates the posterior for each hypothesis H_i:

$$post[i] = prob(H_i \mid val[a])$$
$$= chance[i, a][val[a]] \cdot prior[i] / likelihood(val[a]).$$

Note that the resulting posteriors depend on the value of obs_a. Formally, it means that the array $post[1..n]$ is a function of $val[a]$, and we can write this function as $post(val[a])$.

V. REJECTION OF ALL HYPOTHESES

We next consider the situation when the given hypotheses do *not* cover all possibilities, and the analyst has to evaluate the chances of a surprising situation that fits no hypothesis.

- **Problem 2:** *Suppose that we have one observed value, $val[a]$, and that the sum of the prior$[1..n]$ probabilities is less than 1.0. We have to estimate the posterior probability that none of the hypotheses is correct.*

The analyst thus considers an additional hypothesis H_0, which represents the belief that all n given hypotheses are incorrect; its prior is

$$prior[0] = 1.0 - prior[1] - \ldots - prior[n].$$

If we reuse the expression for $likelihood(val[a])$ from Section IV, we can write the following Bayesian rule for the H_0 posterior:

$$post[0] = prior[0] \cdot prob(val[a] \mid H_0)$$
$$/ \big(prior[0] \cdot prob(val[a] \mid H_0)$$
$$+ likelihood(val[a])\big).$$

Unfortunately, the analyst cannot directly calculate this posterior because she has no knowledge of $prob(val[a] \mid H_0)$.

She can get its upper bound by observing that the above expression for $post[0]$ monotonically increases on $prob(val[a] \mid H_0)$; hence, it reaches its maximal possible value when $prob(val[a] \mid H_0) = 1.0$, which implies that

$$post[0] \leq prior[0] / \big(prior[0] + likelihood(val[a])\big).$$

We next derive a lower bound for $post[0]$, which is based on the following intuitive principle.

- **General plausibility principle:** *If a theoretical model indicates that some event is very unlikely, but we have observed this event in the real world, then we should reject the model or at least strongly doubt its validity.*

Informally, it states that very unlikely events normally do not happen, and we should explain the world based on the assumption that the observed events were in fact likely to happen. Its justification may be analogous to that for the famous Ockham's razor. That is, we may be unable to prove it

formally, and we can find plenty of counterexamples, but the common sense and experience suggest that it usually gives good results.

We use a more specific version of that principle, which is directly applicable to the problem stated in Section III.

- **Specific plausibility principle:** *If the analyst has observed* val[a]*, then* prob(val[a]) *must not be very low.*

To derive an estimate for $post[0]$, we quantify the notion of "very low" by introducing the following constant.

- **Plausibility threshold:** *We use a global constant, denoted "plaus," which must be between* 0.0 *and* 1.0*. We assume that, if the analyst has observed* val[a]*, then*

$$prob(val[a]) \geq plaus / num[a].$$

Note that we can represent $prob(val[a])$ as the sum of the chances of $val[a]$ under all hypotheses, including H_0:

$prob(val[a])$
  $= prob(val[a] \mid H_0) \cdot prior[0]$
    $+ chance[1, a][val[a]] \cdot prior[1]$
    $+ \dots + chance[n, a][val[a]] \cdot prior[n]$
  $= prob(val[a] \mid H_0) \cdot prior[0] + likelihood(val[a]).$

We substitute this representation into the plausibility-threshold inequality:

$prob(val[a] \mid H_0) \cdot prior[0] + likelihood(val[a])$
  $\geq plaus / num[a],$

which implies that

$prob(val[a] \mid H_0)$
  $\geq (plaus / num[a] - likelihood(val[a])) / prior[0].$

Since the dependency of $post[0]$ on $prob(val[a] \mid H_0)$ is monotonic, the above inequality immediately leads to a lower bound for $post[0]$:

$post[0] = prior[0] \cdot prob(val[a] \mid H_0)$
    $/ (prior[0] \cdot prob(val[a] \mid H_0)$
      $+ likelihood(val[a]))$
  $\geq (plaus / num[a] - likelihood(val[a]))$
      $/ (plaus / num[a] - likelihood(val[a])$
        $+ likelihood(val[a]))$
  $= 1.0 - likelihood(val[a]) \cdot num[a] / plaus.$

We have thus obtained both lower and upper bounds on the posterior probability $post[0]$ of H_0.

- **Probability bounds:** *The posterior probability that none of the n given hypotheses is correct has the following bounds.*

  Lower: $1.0 - likelihood(val[a]) \cdot num[a] / plaus.$
  Upper: $prior[0] / (prior[0] + likelihood(val[a])).$

VI. JUDGMENT CALLS

When the analyst uses the approach of Section V to estimate the chances that all given hypotheses are incorrect, she has to make two judgment calls. First, she must select a specific plausibility threshold, *plaus*. Second, she needs to decide whether to use the lower or the upper bound as the final probability estimate.

The choice of *plaus* depends on how conservative she wants to be. Reducing this threshold leads to more reliable conclusions at the expense of getting a looser lower bound. The preliminary experiments in several domains suggest that setting *plaus* to 0.1 usually gives good practical results; however, we have not investigated the dependency of appropriate thresholds on domain properties, which is an interesting direction for the future work.

The choice between the lower and the upper bounds depends on what constitutes the worse-case scenario; that is, the analyst should usually err on the pessimistic side. For example, if the n given hypotheses represent positive developments, whereas H_0 means the absence of any positive outcome, the analyst should use the upper bound as her pessimistic estimate. On the other hand, if H_0 would be a pleasant surprise, the conservative approach is to use the lower bound.

After the analyst estimates the posterior of H_0, she should "prorate" the posteriors of the n given hypotheses, $post[1..n]$, obtained using the technique of Section IV. That is, she should multiply all $post[1..n]$ values by the same constant so that their sum becomes $(1 - post[0])$. She thus obtains the array $post[0..n]$ of estimates that sum to 1.0.

VII. INFORMATION UTILITY

If the analyst has multiple observations, she may need to decide which of them are most relevant, which poses the problem of quantifying their information utility.

- **Problem 3:** *Suppose that we know the* prior[1..n] *probabilities of the given hypotheses. Suppose further that we have used some observation to estimate their posteriors,* post[0..n]*, including the probability* post[0] *that none of them is correct. We have to evaluate the utility of this estimate, which represents the value of the used observation.*

This problem is somewhat "philosophical"; it has no specific mathematical solution and the choice of an appropriate utility measure may depend on an application domain [Lin, 1991; Grünwald and Dawid, 2004].

We review two classical measures; in Section IX, we give empirical results of using them. While they are appropriate in most practical situations, they are not perfect and may occasionally lead the evaluation procedure astray. We have suggested a more general framework for constructing utility functions and analyzed its relation to these two classical measures in another publication [Gershman et al., 2009].

Shannon's entropy: The first measure of the observation utility is the negation of Shannon's entropy of the posterior probabilities [Shannon, 1948]:

$entropy\text{-}util(post)$
  $= post[0] \cdot \log post[0] + \dots + post[n] \cdot \log post[n].$

This measure rewards "high certainty," that is, situations in which the posteriors clearly favor one hypothesis over all others. In other words, the utility is high when the probability of one specific hypothesis is close to 1.0; it is low when all hypotheses are about equally likely.

Its main drawback is that it may reward unwarranted certainty. If an application domain is highly nondeterministic and the true probabilities of all hypotheses are about the same,

this measure may give preference to observations that falsely create the impression of certainty. Intuitively, the evaluation procedure may prefer the illusion of certainty over the truth.

Kullback-Leibler divergence: The second measure is the KL-divergence between the priors and the posteriors [Kullback and Leibler, 1951; Kullback, 1987]:

$$KL\text{-}util(post)$$
$$= post[0] \cdot \log \left(post[0] / (1.0 - prior[1] - \ldots - prior[n])\right)$$
$$+ post[1] \cdot \log \left(post[1] / prior[1]\right)$$
$$+ \ldots + post[n] \cdot \log \left(post[n] / prior[n]\right).$$

This measure rewards situations in which the posteriors are very different from the priors. Thus, it gives preference to observations that have the potential for "paradigm shifts."

Its main drawback is that it may encourage unwarranted departure from the right conclusions. If the priors are already accurate, it may give preference to observations that falsely change probabilities. Intuitively, the evaluation procedure may prefer sensation seeking to fact checking.

VIII. Use of multiple observations

We now consider the general problem of evaluating hypotheses based on multiple observations.

- **Problem 4:** *Suppose that we have multiple observed values, val[1..m]. We have to determine the posterior probabilities of the n given hypotheses.*

If the analyst were to use all available observations in the standard Bayesian reasoning, she would need data on their *joint* distribution. Unfortunately, obtaining even rough estimates of joint probabilities is impractically difficult in most real-world situations. Furthermore, the analyst usually cannot assume that her observations are independent. For instance, when she hears similar facts from multiple sources, she cannot reliably determine whether these sources corroborate or just repeat each other. The history of business analysis and military intelligence is full of spectacular mistakes due to inappropriate uses of the independence assumption.

To avoid this pitfall, we use a conservative approach that *never* overestimates the value of available data. Specifically, we identify the highest-utility observation and do not use other observations to corroborate it. In other words, we assume the worst-case correlation among available data. On the upside, it never leads to excessive confidence; on the downside, it may underestimate the value of observations.

To formalize it, suppose that the analyst has selected an appropriate utility function, *util(post)*, which may be Shannon's entropy, KL-divergence, or some other measure [Grünwald and Dawid, 2004; Gershman et al., 2009].

For each available observation, *val[a]*, the analyst can compute the respective posteriors, *post(val[a])*, using the technique in Sections IV and V, and then evaluate its utility, *util(post(val[a]))*. She selects the single "most relevant" observation that maximizes this utility. She uses it in computing the posteriors and ignores all other observations.

If the analyst happens to have a joint distribution for some subset of observations, she can improve her evaluation,

by using this subset as a single "unified" observation and each entry in the joint-distribution table as its possible value.

IX. Experiments

We have evaluated the described technique in the context of the PAINT architecture, built by teams from several institutions under the PAINT program of IARPA. This architecture is a suite of tools for modeling a partially observable organization; evaluating the accuracy of models and related observations; analyzing possible future developments; and planning the collection of additional data. It supports modeling of the organization's management, decision processes, projects, and resources. We have been responsible for uncertainty-analysis tools in PAINT, including hypothesis evaluation.

We have experimented with tasks of determining whether an observed company is secretly working on a new product, similar to the example in Section II. We describe the hypotheses and observations in these experiments, and give the results of applying the system to several specific tasks.

- **Hypotheses:** We have considered four hypotheses, which give rise to four different organizational models in PAINT.
  - **H_1:** The company focuses exclusively on the sales of its current product and has no secret agenda.
  - **H_2:** The company puts limited resources into its secret development of a new product, but its main focus is the sales of the current product.
  - **H_3:** The company puts major resources into its new-product development, but also focuses on the current sales.
  - **H_4:** The company focuses almost exclusively on its new product; its current sales are just a smokescreen to mislead its competitors.
- **Observations:** We have considered eighty-four observable features of the modeled company, which represent public data about its sales and known projects. The observables do not include any direct data about the new product, since these data are "invisible" to the public.
- **Observation distributions:** To obtain probability distributions for observations, we have simulated the company behavior under different scenarios. For every hypothesis, we have run the PAINT models to predict the likelihood of each possible observed value.

We have applied the system to distinguishing between H_1 and *one* other hypothesis at a time. We have run this experiment with two different priors of H_1, specifically, 0.5 and 0.7. We have tested the following three techniques for selecting the most relevant observation.

- **Shannon's entropy:** We select the observation that minimizes the expected entropy of the posteriors, as described in Section VII.
- **Kullback-Leibler divergence:** We select the observation that maximizes the expected divergence between the priors and the posteriors, which is also described in Section VII.
- **Randomly chosen observation:** We pick a "relevant" observation *at random* among the *n* available observations. Clearly, it is a very ineffective method; we have included it

in order to show that the developed technique is at all useful, that is, far better than a trivial approach.

Since we have *three* techniques to be evaluated, *two* different priors, and *three* hypotheses that can be alternatives to H_1, we have experimented with all $3 \cdot 2 \cdot 3 = 18$ possible settings. For every setting, we have run 2000 Monte Carlo simulations of the PAINT models and applied the system to evaluate hypothesis H_1 based on the simulated observations. The hypothesis evaluation has included the following steps.

- Pick an observation using the specified selection strategy.
- Evaluate the posterior of H_1 based on this observation.
- If it is at least 0.5, report that H_1 is correct; else, H_1 is wrong.

We have computed the percentage of the right answers among the 2000 runs, which is the final performance measure.

In Table I, we show this percentage for each of the eighteen settings. Unsurprisingly, the performance based on the random choice of an observation is poor. It is slightly better than the direct random guessing of a hypothesis, which would be 50% accurate, because the system occasionally picks an observation that is indeed relevant, and then uses it to identify the correct hypothesis.

The other two techniques give much better results. The accuracy of distinguishing H_1 from H_2 or H_3 ranges from 75% to 80%, whereas the accuracy of distinguishing H_1 from H_4 is 100%. The reason is that the observed sales are similar under hypotheses H_1, H_2, and H_3, which makes them hard to differentiate based on available data. On the other hand, H_1 and H_4 are very different, and thus they are easier to distinguish.

If the prior of H_1 is 0.5, the Kullback-Leibler divergence is mathematically equivalent to Shannon's entropy of the posteriors, so the two techniques give the same results. On the other hand, when the prior is 0.7, the Kullback-Leibler divergence gives slightly better results, although the difference is not statistically significant.

TABLE I: DIFFERENTIATION BETWEEN HYPOTHESES. WE HAVE EXPERIMENTED WITH TWO DIFFERENT PRIORS AND THREE TECHNIQUES FOR SELECTING RELEVANT OBSERVATIONS. FOR EACH PRIOR AND EACH TECHNIQUE, WE HAVE RUN 2000 EXPERIMENTS, AND WE GIVE THE PERCENTAGES OF CORRECT HYPOTHESIS EVALUATIONS.

Prior of H_1	Accuracy of hypothesis selection (%)		
	Shannon's Entropy	Kullback–Leibler	Random Observ.
Differentiating H_1 from H_2			
0.5	79	79	51
0.7	74	76	50
Differentiating H_1 from H_3			
0.5	81	81	52
0.7	74	78	51
Differentiating H_1 from H_4			
0.5	100	100	59
0.7	100	100	54

X. CONCLUDING REMARKS

We have investigated the problem of (1) evaluating hypotheses based on limited data and (2) estimating the chances that all given hypotheses are incorrect. We have developed a heuristic solution, which combines the Bayesian reasoning with the *plausibility principle* and a method for selecting the most relevant observation.

The initial experiments have demonstrated a reasonable accuracy of this technique, which ranges from 75% when hypotheses are hard to distinguish to 100% in easier cases; however, the context of these experiments has been limited. We plan to test the system with more complex scenarios, which will include a large number of alternative hypotheses.

We have not yet done a rigorous comparison of the manual hypothesis selection with the automated evaluation, but informal pilot tests suggest that the system performs much better than human subjects recruited among the project members. This result is unsurprising, since people are known to be poor Bayesian reasoners [Heuer, 1999], and automated "number crunchers" usually outperform human experts on the tasks that require probabilistic analysis of complex evidence. We aim to run a formal comparison of the manual and the automated analysis as part of the future work.

ACKNOWLEDGMENTS

REFERENCES

[Bardak et al., 2006a] Ulas Bardak, Eugene Fink, and Jaime G. Carbonell. Scheduling with uncertain resources: Representation and utility function. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 1486–1492, 2006.

[Bardak et al., 2006b] Ulas Bardak, Eugene Fink, Chris R. Martens, and Jaime G. Carbonell. Scheduling with uncertain resources: Elicitation of additional data. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 1493–1498, 2006.

[Bayes, 1763] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. Communicated by Richard Price in a letter to John Canton, 1763.

[Elsaesser and Stech, 2006] Christopher Elsaesser and Frank J. Stech. Detecting deception. In Alexander Kott and William M. McEneaney, editors, Adversarial Reasoning: Computational Approaches to Reading the Opponent's Mind, pages 101–124. Chapman and Hall / CRC, Boca Raton, FL, 2006.

[Fink et al., 2006] Eugene Fink, P. Matthew Jennings, Ulas Bardak, Jean Oh, Stephen F. Smith, and Jaime G. Carbonell. Scheduling with uncertain resources: Search for a near-optimal solution. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 137–144, 2006.

[Fu et al., 2008] Bin Fu, Eugene Fink, and Jaime G. Carbonell. Analysis of uncertain data: Tools for representation and processing. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 3256–3260, 2008.

[Gardiner et al., 2008] Steven Gardiner, Eugene Fink, and Jaime G. Carbonell. Scheduling with uncertain resources: Learning to make reasonable assumptions. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pages 2554–2559, 2008.

[Gershman et al., 2009] Anatole Gershman, Eugene Fink, Bin Fu, and Jaime G. Carbonell. Analysis of uncertain data: Selection of probes for information gathering. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 2009.

[Grünwald and Dawid, 2004] Peter D. Grünwald and A. Philip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Annals of Statistics, 32(4), pages 1367–1433, 2004.

[Heuer, 1999] Richards J. Heuer, Jr. Psychology of Intelligence Analysis. Center for the Study of Intelligence, Central Intelligence Agency, Washington, DC, 1999.

[Kullback, 1987] The Kullback–Leibler distance. The American Statistician, 41, pages 340–341, 1987.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A. Leibler. On information and sufficiency. Annals of Mathematical Statistics, 22, pages 79–86, 1951.

[Lin, 1991] Jianhua Lin. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), pages 145–151, 1991.

[Pearl, 1988] Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA, 1988.

[Shannon, 1948] Claude E. Shannon. A mathematical theory of communication. The Bell System Technical Journal, 27, pages 379–423 and 623–656, 1948.