A Distance Measure for Bounding Probabilistic Belief Change

Hei Chan and Adnan Darwiche

Computer Science Department
University of California, Los Angeles, CA 90095
{*hei,darwiche*}*@cs.ucla.edu*

Abstract

We propose a distance measure between two probability distributions, which allows one to bound the amount of belief change that occurs when moving from one distribution to another. We contrast the proposed measure with some well known measures, including KL–divergence, showing some theoretical properties on its ability to bound belief changes. We then present two practical applications of the proposed distance measure: sensitivity analysis in belief networks and probabilistic belief revision. We show how the distance measure can be easily computed in these applications, and then use it to bound global belief changes that result from either the perturbation of local conditional beliefs or the accommodation of soft evidence. Finally, we show that two well known techniques in sensitivity analysis and belief revision correspond to the minimization of our proposed distance measure and, hence, can be shown to be optimal from that viewpoint.

Key words: probabilistic reasoning, uncertainty, belief revision

1 Introduction

We propose in this paper a distance measure which allows one to bound the amount of belief change that results from transforming one probabilistic state of belief into another. Specifically, given a probability distribution Pr representing an initial state of belief, and a distribution Pr' representing a new state of belief, we define a distance measure which allows us to tightly bound belief change as follows: $1/k \leq O'(\alpha \mid \beta)/O(\alpha \mid \beta) \leq k$. Here, k is a constant

[⋆] A shorter version of this paper appeared in the *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, pages 539-545.

that depends on the proposed distance, α and β are arbitrary events, $O(\alpha \mid \beta)$ is the odds of event α given β with respect to Pr, and $O'(\alpha \mid \beta)$ is the odds of event α given β with respect to Pr'. We show a number of theoretical results about the proposed measure and then present two of its key applications.

On the theoretical side, we prove that our proposed measure satisfies the three properties of distance. We also contrast our distance measure with classical measures, including KL–divergence [13] where we present some results on its ability to bound belief changes. Specifically, we show that belief change between two states of belief can be unbounded, even when their KL–divergence tends to zero. We show, however, that KL–divergence can be used to bound the average change in beliefs as opposed to the worst–case change in beliefs.

On the practical side, we present two main applications of our proposed distance measure. The first application is sensitivity analysis in belief networks, an area which concerns itself with bounding global belief change that results from applying a local perturbation to a belief network [14,1,4,3,12,2,16]. We show three key results here. First, we show that if Pr is the distribution induced by a belief network N, and if Pr' is the distribution induced by a belief network N' that results from changing some conditional probability table (CPT) in N, then the distance between Pr and Pr' can be computed locally by only examining the changed CPT. Second, we use our distance measure to provide a bound on global belief change that results from a local CPT change, and show that our bound generalizes and provides more insights into the bound given by Chan and Darwiche recently [2]. Finally, we use our proposed distance measure to prove the optimality of a prevalent, but formally unjustified, technique in the literature on sensitivity analysis relating to changing the CPTs of multivalued variables [14,12,4].

The second application we consider for our distance measure is in belief revision [6]. Here, we show how our distance measure can be used to bound belief change that results from incorporating uncertain evidence according to both Jeffrey's rule [9] and Pearl's method of virtual evidence [15]. We actually prove the optimality of Jeffrey's rule with regards to minimizing belief change and, finally, consider the application of our distance measure to quantifying the strength of evidence, as measured by the amount of belief change it induces.

Proofs of all theorems in this paper can be found in Appendix A.

2 A probabilistic distance measure

Our proposed measure is defined between two probability distributions.

Definition 2.1 *Let Pr and Pr' be two probability distributions over the same set of worlds w. We define a measure $D(Pr, Pr')$ as follows:*

$$D(Pr, Pr') \overset{def}{=} \ln \max_{w} \frac{Pr'(w)}{Pr(w)} - \ln \min_{w} \frac{Pr'(w)}{Pr(w)},$$

where we will define, $0/0 \overset{def}{=} 1$ and $\infty/\infty \overset{def}{=} 1$ in this paper.

We will say that two probability distributions Pr and Pr' *have the same support*, if for every world w, $Pr(w) = 0$ iff $Pr'(w) = 0$. Note that if two distributions Pr and Pr' do not have the same support, $D(Pr, Pr') = \infty$.

Our first result on the defined measure is that it satisfies the three properties of distance, hence, it is a *distance measure.*

Theorem 2.1 *Let Pr, Pr' and Pr'' be three probability distributions over the same set of worlds. The distance measure given in Definition 2.1 satisfies these three properties:*

Positiveness: $D(Pr, Pr') \geq 0$, and $D(Pr, Pr') = 0$ iff $Pr = Pr'$;
Symmetry: $D(Pr, Pr') = D(Pr', Pr)$;
Triangle Inequality: $D(Pr, Pr') + D(Pr', Pr'') \geq D(Pr, Pr'')$.

The interest in the defined distance measure stems from two reasons. First, it can be easily computed in a number of practical situations which we discuss in later sections. Second, it allows us to bound the difference in beliefs captured by two probability distributions.

Theorem 2.2 *Let Pr and Pr' be two probability distributions over the same set of worlds. Let α and β be two events. We then have:*

$$e^{-D(Pr,Pr')} \leq \frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \leq e^{D(Pr,Pr')},$$

where $O(\alpha \mid \beta) = Pr(\alpha \mid \beta)/Pr(\overline{\alpha} \mid \beta)$ is the odds of event α given β with respect to Pr, and $O'(\alpha \mid \beta) = Pr'(\alpha \mid \beta)/Pr'(\overline{\alpha} \mid \beta)$ is the odds of event α given β with respect to Pr'.[1] *The bound is tight in the sense that for every pair of distributions Pr and Pr', there are events α and β such that:*

$$\frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} = e^{D(Pr,Pr')}; \; \frac{O'(\overline{\alpha} \mid \beta)}{O(\overline{\alpha} \mid \beta)} = e^{-D(Pr,Pr')}.$$

[1] Of course, we must have $Pr(\beta) \neq 0$ and $Pr'(\beta) \neq 0$ for the odds to be defined.

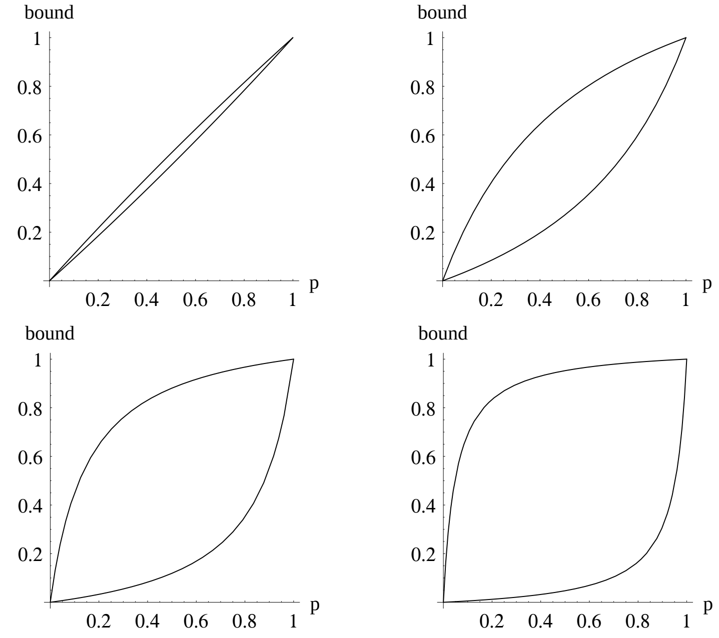

Fig. 1. The bounds of $Pr'(\alpha \mid \beta)$, as given by Inequality 2, plotted against the initial belief $p = Pr(\alpha \mid \beta)$ for several different values of distance $d = D(Pr, Pr')$: $d = .1$ (top left), $d = 1$ (top right), $d = 2$ (bottom left), and $d = 3$ (bottom right).

We can express the bound of Theorem 2.2 in two other useful forms. First, we can use logarithms:

$$|\ln O'(\alpha \mid \beta) - \ln O(\alpha \mid \beta)| \leq D(Pr, Pr'). \tag{1}$$

Second, we can use probabilities instead of odds:

$$\frac{p\, e^{-d}}{p\,(e^{-d} - 1) + 1} \leq Pr'(\alpha \mid \beta) \leq \frac{p\, e^{d}}{p\,(e^{d} - 1) + 1}, \tag{2}$$

where $p = Pr(\alpha \mid \beta)$ is the initial belief in α given β, and $d = D(Pr, Pr')$ is the distance. The bounds of $Pr'(\alpha \mid \beta)$ are plotted against p for several different values of d in Figure 1.

In the applications we shall discuss next, Pr is a distribution which represents some initial state of belief, and Pr' is a distribution which represents a new state of belief. The new state of belief results form applying some kind of local change to the initial state. Examples include the change in some conditional belief or the incorporation of new evidence. Our goal is then to assess the global impact of such local belief changes. According to Theorem 2.2, if we are able to compute the distance measure $D(Pr, Pr')$, then we can bound global belief change in a very precise sense. For example, we can use Inequality 2 to compute the bound on any query $Pr'(\alpha \mid \beta)$. We will later show two applications from sensitivity analysis and belief revision where the distance measure can be computed efficiently.

3 Comparisons with existing measures

Before we discuss the applications of our proposed distance measure, we first need to settle a major question: Can we bound belief change in the sense given above using one of the classical probabilistic measures? We show next that this is not possible using at least two of the most commonly used measures: KL–divergence and Euclidean distance. We show, however, that KL–divergence can be used to provide an average–case bound on belief changes and we also provide a relationship between that bound and ours.

We start first with Kullback–Leibler (KL) divergence, which is one of the most common measures for comparing probability distributions [13].

Definition 3.1 *Let Pr and Pr' be two probability distributions over the same set of worlds w. The KL–divergence between Pr and Pr' is defined as:*

$$KL(Pr, Pr') \overset{def}{=} -\sum_{w} Pr(w) \ln \frac{Pr'(w)}{Pr(w)}.$$

The first thing to note about KL–divergence [2] is that it is incomparable with our distance measure.

Example 3.1 *Consider the following distributions, Pr, Pr' and Pr'', over worlds w_1, w_2 and w_3:*

$$\begin{aligned}
Pr(w_1) = .50,\ \ Pr(w_2) = .25,\ \ Pr(w_3) = .25;\\
Pr'(w_1) = .50,\ Pr'(w_2) = .30,\ Pr'(w_3) = .20;\\
Pr''(w_1) = .43,\ Pr''(w_2) = .32,\ Pr''(w_3) = .25.
\end{aligned}$$

Computing the KL–divergence and our distance measure gives us:

$$\begin{aligned}
KL(Pr, Pr') = .0102,\ KL(Pr, Pr'') = .0137;\\
D(Pr, Pr') = .405, \quad D(Pr, Pr'') = .398.
\end{aligned}$$

[2] Note that KL–divergence is asymmetric, and is thus technically not a distance measure.

Therefore, according to KL–divergence, Pr' is closer to Pr than Pr'', while according to our distance measure, Pr'' is closer to Pr than Pr'.

The next example shows that we can make the KL–divergence arbitrarily close to 0, while keeping some odds ratio arbitrarily close to some constant k. Hence, KL–divergence cannot be used to bound belief changes as permitted by our proposed measure.

Example 3.2 *Consider the following distributions, Pr and Pr', over worlds w_1, w_2 and w_3:*

$$\begin{aligned} Pr(w_1) = p, \quad Pr(w_2) = q - p, \quad Pr(w_3) = 1 - q; \\ Pr'(w_1) = kp, \; Pr'(w_2) = q - kp, \; Pr'(w_3) = 1 - q; \end{aligned}$$

where p, q and k are constants, with $0 \leq p \leq q \leq 1$ and $0 \leq k \leq q/p$. The KL–divergence between Pr and Pr' is:

$$KL(Pr, Pr') = -p \ln k - (q - p) \ln \frac{q - kp}{q - p}.$$

Assume we have events $\alpha = w_1$ and $\beta = w_1, w_2$. The odds ratio of α given β between Pr and Pr' is:

$$\frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} = \frac{k(q - p)}{q - kp}.$$

We can see that as p approaches 0, the KL–divergence also approaches 0, while the odds ratio $O'(\alpha \mid \beta)/O(\alpha \mid \beta)$ approaches k.

This example shows that we can make the KL–divergence arbitrarily close to 0, while keeping some odds ratio arbitrarily close to some constant k. In this example, we condition on event β, which has a probability of q that can be arbitrarily large. However, the probability of α, which is p according to Pr and kp according to Pr', is very small. Hence, although we have $Pr'(\alpha)/Pr(\alpha) = k$, this ratio is ignored by KL–divergence because the term $-p \ln k$ is very small as p approaches 0. More generally, the "contribution" of a world w to KL–divergence is equal to $-Pr(w) \ln(Pr'(w)/Pr(w))$. Therefore for a fixed ratio $Pr'(w)/Pr(w)$, this "contribution" becomes closer to 0 as $Pr(w)$ decreases, and becomes infinitesimal when $Pr(w)$ approaches 0.

Another popular measure to compare two probability distributions is the Euclidean distance.

Definition 3.2 *Let Pr and Pr' be two probability distributions over the same set of worlds w. The Euclidean distance between Pr and Pr' is defined as:*

$$ED(Pr, Pr') \overset{def}{=} \sum_{w} \sqrt{(Pr'(w) - Pr(w))^2}.$$

That is, when computing the Euclidean distance, we add up the squared differences between pairs of probability values. Therefore, this measure has the same problem as KL–divergence: even if there is a large relative difference for the probability of a world with respect to Pr and Pr', it will be ignored if this probability is very small. Consequently, we cannot provide any guarantee on the ratio $O'(\alpha \mid \beta)/O(\alpha \mid \beta)$, no matter how small the Euclidean distance is (unless it is zero).

To summarize, neither KL–divergence nor Euclidean distance can be used to provide guarantees on the ratio $O'(\alpha \mid \beta)/O(\alpha \mid \beta)$, as we did in Theorem 2.2 using our distance measure.

Finally, we note that our distance measure is an improvement over computing the L-infinity metric $L_\infty(\ln Pr, \ln Pr') \overset{def}{=} \max_w |\ln Pr'(w) - \ln Pr(w)|$, because our distance measure is computed from both the maximum and minimum values of $\ln Pr'(w) - \ln Pr(w)$. If $L_\infty(\ln Pr, \ln Pr') = x$, we can conclude that $-x \leq \ln Pr'(w) - \ln Pr(w) \leq x$ for any w, and prove the following two guarantees. First, if α and β are two events, we have $e^{-2x} \leq O'(\alpha \mid \beta)/O(\alpha \mid \beta) \leq e^{2x}$. Second, our distance measure is bounded such that $D(Pr, Pr') \leq 2x$. However, the first guarantee is also a direct result of the second guarantee, and by computing the exact value of the distance measure $D(Pr, Pr')$, we can obtain a bound that is no worse than the bound obtained from $L_\infty(\ln Pr, \ln Pr')$. Therefore, our distance measure should always be preferred.

3.1 KL-divergence as an average–case bound

Even though KL–divergence cannot be used to bound belief changes as discussed above, it can still be used to offer a bound but on the average change in beliefs. This is given by the following theorem.

Theorem 3.1 *Let Pr and Pr' be two probability distributions over the same set of worlds. Let α and β be two events. We then have:*

$$\begin{aligned} &KL(Pr, Pr') \\ &\geq -Pr(\beta)\left(Pr(\alpha \mid \beta) \ln \frac{Pr'(\alpha \mid \beta)}{Pr(\alpha \mid \beta)} + (1 - Pr(\alpha \mid \beta)) \ln \frac{1 - Pr'(\alpha \mid \beta)}{1 - Pr(\alpha \mid \beta)}\right), \end{aligned}$$

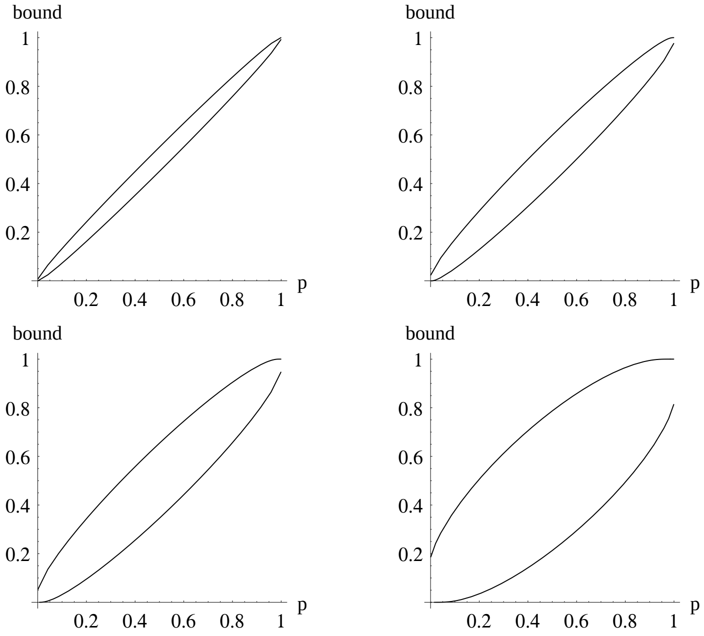

Fig. 2. The bounds of $Pr'(\alpha \mid \beta)$, as given by Theorem 3.1, plotted against the initial belief $p = Pr(\alpha \mid \beta)$, for different values of $KL(Pr, Pr')$ and $Pr(\beta)$: $KL(Pr, Pr') = .005$ and $Pr(\beta) = 1$ (top left), $KL(Pr, Pr') = .02$ and $Pr(\beta) = 1$ (top right), $KL(Pr, Pr') = .005$ and $Pr(\beta) = .1$ (bottom left), and $KL(Pr, Pr') = .02$ and $Pr(\beta) = .1$ (bottom right).

or alternatively, in odds:

$$KL(Pr, Pr') \geq Pr(\beta) \left(\ln \frac{O'(\alpha \mid \beta) + 1}{O(\alpha \mid \beta) + 1} - \frac{O(\alpha \mid \beta)}{O(\alpha \mid \beta) + 1} \ln \frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \right),$$

where $O(\alpha \mid \beta) = Pr(\alpha \mid \beta)/Pr(\overline{\alpha} \mid \beta)$ is the odds of event α given β with respect to Pr, and $O'(\alpha \mid \beta) = Pr'(\alpha \mid \beta)/Pr'(\overline{\alpha} \mid \beta)$ is the odds of event α given β with respect to Pr'.

According to this theorem, the KL–divergence can be used to provide a guarantee on the new odds value $O'(\alpha \mid \beta)$ in terms of the original odds value $O(\alpha \mid \beta)$. However, the provided guarantee depends on the probability of event β, where the quality of the guarantee degrades as the probability decreases. This echoes Example 3.2, where we can get a constant log–odds change even when the KL–divergence approaches 0, because the probability $Pr(\beta)$ also approaches 0. To give better insights into the guarantee offered by Theorem 3.1, we plot in Figure 2 the bounds it provides for different values of $KL(Pr, Pr')$ and $Pr(\beta)$.

One can also provide a bound on the change in odds of $\alpha \mid \beta$ which is independent of the probability of β by taking an average over all possible probabilities of β, but that would be an average–case bound as opposed to the worst–case bound provided by our measure.

3.2 Bayes factor

One useful term that can help us further understand our distance measure and its relation to KL–divergence is the *Bayes factor* [7,8,10], defined as follows.

Definition 3.3 *If Pr and Pr' are two probability distributions, and α and β are two events, the Bayes factor (or odds factor), $F_{Pr',Pr}(\alpha : \beta)$, is defined as the ratio of new-to-old odds:*

$$F_{Pr',Pr}(\alpha : \beta) \overset{def}{=} \frac{Pr'(\alpha)/Pr'(\beta)}{Pr(\alpha)/Pr(\beta)}.$$

Our distance measure given in Definition 2.1 can be expressed using the Bayes factor:

$$D(Pr, Pr') = \ln \max_{w_i, w_j} F_{Pr',Pr}(w_i : w_j).$$

Therefore, our distance measure can be regarded as the logarithm of the maximum Bayes factor of any two worlds between the two distributions. Consequently, it can be used to bound the Bayes factor of any two events between the two distributions.

Theorem 3.2 *Let Pr and Pr' be two probability distributions over the same set of worlds. Let γ_1 and γ_2 be two events. We then have:*

$$e^{-D(Pr,Pr')} \leq F_{Pr',Pr}(\gamma_1 : \gamma_2) \leq e^{D(Pr,Pr')}.$$

Theorem 2.2 is a special case of Theorem 3.2, when we substitute the following terms: $\gamma_1 = \alpha \mid \beta$ and $\gamma_2 = \overline{\alpha} \mid \beta$.

While our distance measure provides us a worst–case bound of Bayes factors, the KL–divergence can be perceived as an average–case bound of Bayes factors, as shown by the next theorem.

Theorem 3.3 *If α is an arbitrary event, and $\gamma_1, \ldots, \gamma_n$ is a set of mutually exclusive and exhaustive events, we then have:*

$$0 \leq \sum_i Pr(\gamma_i) \ln F_{Pr',Pr}(\alpha : \gamma_i) - \ln \frac{Pr'(\alpha)}{Pr(\alpha)} \leq KL(Pr, Pr').$$

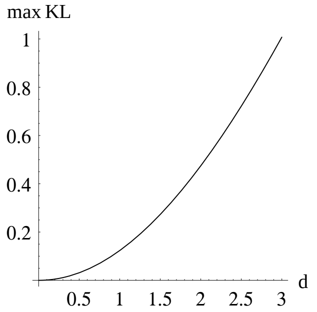

Fig. 3. The bound of KL–divergence $KL(Pr, Pr')$ as a function of the distance measure $d = D(Pr, Pr')$, as given in Theorem 3.4.

If instead of an arbitrary partition $\gamma_1, \ldots, \gamma_n$, we are given the set of worlds w, then we will get the following equality relation:

$$KL(Pr, Pr') = \sum_w Pr(w) \ln F_{Pr',Pr}(\alpha : w) - \ln \frac{Pr'(\alpha)}{Pr(\alpha)}.$$

Theorem 3.3 gives us a bound on the weighted sum of the logarithms of the Bayes factors using KL–divergence.

To further relate the two bounds, we ask an interesting question: If we are given the value of our distance measure between two distributions, can we put a bound on the KL–divergence between them? The following theorem provides us the answer.

Theorem 3.4 *Given two distributions Pr and Pr', where $D(Pr, Pr') = d > 0$, we have:*

$$KL(Pr, Pr') \leq -1 - \ln \frac{d}{e^d - 1} + \frac{d}{e^d - 1}.$$

The plot of the bound of Theorem 3.4 against d is shown in Figure 3. Hence, the smaller d is, the more we can say about the KL–divergence as we can get a tighter bound. If we view our distance measure as providing a worst–case bound on belief changes, and KL–divergence as providing an average–case bound on such changes, it is then not surprising that we can say more about the average–case bound (KL–divergence) as the worst–case bound (our distance measure) is less dramatic.

We close this section with some further comments on the suitability of the average–case versus worst–case bounds:

(1) Average–case bounds have proved to be useful in learning algorithms, while worst–case bounds are more important in common-sense reasoning, where conditioning on unlikely events is not uncommon.

(2) Even in learning algorithms, worst–case analysis can be useful if it gives us tight results.

4 Applications to sensitivity analysis

We now consider a major application of our distance measure to sensitivity analysis in belief networks [14,1,12,4,2]. A belief network is a graphical probabilistic model, composed of two parts: a directed acyclic graph where nodes represent variables, and a set of conditional probability tables (CPTs), one for each variable [15,11]. The CPT for variable X with parents $\mathbf{U}$ defines a set of conditional beliefs of the form $\theta_{x|\mathbf{u}} = Pr(x \mid \mathbf{u})$, where x is a value of variable X, $\mathbf{u}$ is an instantiation of parents $\mathbf{U}$, and $\theta_{x|\mathbf{u}}$ is the probability value of x given $\mathbf{u}$, and is called a network parameter.

One of the key questions with respect to belief networks is this: what can we say about the global effect of changing some parameter $\theta_{x|\mathbf{u}}$ to a new value $\theta'_{x|\mathbf{u}}$? That is, what is the effect of such a local parameter change on the value of some arbitrary query $Pr(\alpha \mid \beta)$?

Our earlier results [2] have provided a partial answer to this question, for the case where: variable X is binary (it has only two values, x and $\overline{x}$); α is the value y of some variable Y; β is the instantiation $\mathbf{e}$ of some variables $\mathbf{E}$, and neither $\theta_{x|\mathbf{u}}$ nor $\theta'_{x|\mathbf{u}}$ is extreme (equal to 0 or 1). Specifically under these conditions, we have shown that:

$$|\ln O'(y \mid \mathbf{e}) - \ln O(y \mid \mathbf{e})| \leq \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta'_{\overline{x}|\mathbf{u}}} - \ln \frac{\theta_{x|\mathbf{u}}}{\theta_{\overline{x}|\mathbf{u}}} \right|. \tag{3}$$

Using the above bound, we provided a formalization of a number of intuitions relating to the sensitivity of probabilistic queries to changes in network parameters. We will now show how our distance measure can be used to derive a generalization of the above bound, which applies without any of the previously mentioned restrictions.

Suppose that our initial belief network is N and it induces a probability distribution Pr. By changing the CPT for variable X, we produce a new belief network N' that induces a probability distribution Pr'. If we are able to compute the distance between Pr and Pr', $D(Pr, Pr')$, we can then use Theorem 2.2 to provide a guarantee on the global effect of the local CPT change. As it turns out, the distance can be computed locally as given by the following theorem.

Theorem 4.1 *Let N and N' be belief networks which induce distributions Pr and Pr' respectively, and let X be a variable with parents $\mathbf{U}$ in network N.*

Suppose that N' is obtained from N by changing the conditional probability distribution of variable X given parent instantiation $\mathbf{u}$ from $\Theta_{X|\mathbf{u}}$ to $\Theta'_{X|\mathbf{u}}$, i.e. we change parameter $\theta_{x|\mathbf{u}}$ to $\theta'_{x|\mathbf{u}}$ for every value x. If $Pr(\mathbf{u}) > 0$, then:

$$D(Pr, Pr') = D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}).$$

The above theorem shows that the distance between the global probability distributions induced by networks N and N' is exactly the distance between the local distributions of X given $\mathbf{u}$, assuming that all other local distributions in N and N' are the same.

Theorem 4.1 is of great practical importance as it allows us to invoke Theorem 2.2 to provide a generalized sensitivity analysis formula for belief networks.

Corollary 4.1 *Let N and N' be belief networks which induce distributions Pr and Pr' respectively, and let X be a variable with parents $\mathbf{U}$ in network N. Suppose that N' is obtained from N by changing the conditional probability distribution of variable X given parent instantiation $\mathbf{u}$ from $\Theta_{X|\mathbf{u}}$ to $\Theta'_{X|\mathbf{u}}$, i.e. we change parameter $\theta_{x|\mathbf{u}}$ to $\theta'_{x|\mathbf{u}}$ for every value x. If $Pr(\mathbf{u}) > 0$, then:*

$$e^{-D(\Theta_{X|\mathbf{u}},\Theta'_{X|\mathbf{u}})} \leq \frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \leq e^{D(\Theta_{X|\mathbf{u}},\Theta'_{X|\mathbf{u}})}.$$

The bound given by Inequality 3 is a special case of Corollary 4.1, when X has only two values x and $\overline{x}$. In this case, the distance $D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}})$ is equal to:

$$\begin{aligned} D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}) &= \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta_{x|\mathbf{u}}} - \ln \frac{\theta'_{\overline{x}|\mathbf{u}}}{\theta_{\overline{x}|\mathbf{u}}} \right| \\ &= \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta'_{\overline{x}|\mathbf{u}}} - \ln \frac{\theta_{x|\mathbf{u}}}{\theta_{\overline{x}|\mathbf{u}}} \right|. \end{aligned}$$

We have therefore generalized their results on sensitivity analysis to arbitrary events and belief networks. We have also relaxed the condition that neither $\theta_{x|\mathbf{u}}$ nor $\theta'_{x|\mathbf{u}}$ can be extreme.

4.1 Comparison with KL–divergence

Suppose now that we want to compute the KL–divergence between two belief networks under the same circumstances. We have:

$$KL(Pr, Pr') = Pr(\mathbf{u})KL(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}).$$

We can see that to compute the KL–divergence between two belief networks due to a single parameter change, we need to know the probability $Pr(\mathbf{u})$. Hence, the KL–divergence cannot be obtained locally as it would require a global computation to obtain $Pr(\mathbf{u})$. This is not necessary when we compute our distance measure as given in Theorem 4.1. Hence, we have two differences between our distance measure and KL–divergence in this regard. One is semantical in which our measure can be used to bound worst–case belief changes, while KL–divergence can be used to bound average–case belief changes. The second difference is computational where our measure can be computed in constant time under local changes, while the KL–divergence cannot.

4.2 Changing parameters of multi–valued variables

We now close this section with a final application of our distance measure. Suppose X is a variable with parents $\mathbf{U}$, values x_1, x_2 and x_3, and parameters $\theta_{x_1|\mathbf{u}} = .6$, $\theta_{x_2|\mathbf{u}} = .3$ and $\theta_{x_3|\mathbf{u}} = .1$. Suppose further that we want to change the parameter $\theta_{x_1|\mathbf{u}} = .6$ to $\theta'_{x_1|\mathbf{u}} = .8$. As a result, we will need to change the other parameters $\theta_{x_2|\mathbf{u}}$ and $\theta_{x_3|\mathbf{u}}$ so that the sum of all three parameters remains to be 1. Because X is multivalued, there are infinitely many ways to change the other two parameters and the question is: which one of them should we choose? One popular scheme, which we will call the *proportional scheme,* distributes the mass $1 - \theta'_{x_1|\mathbf{u}} = 1 - .8 = .2$ among the other two parameters proportionally to their initial values. That is, the new parameters will be $\theta'_{x_2|\mathbf{u}} = .2(.3/.4) = .15$ and $\theta'_{x_3|\mathbf{u}} = .2(.1/.4) = .05$. This scheme has been used in all approaches to sensitivity analysis we are familiar with [14,12,4], yet without justification. As it turns out, we can use our distance measure to prove the optimality of this scheme in a very precise sense.

Theorem 4.2 *When changing a parameter $\theta_{x|\mathbf{u}}$ to $\theta'_{x|\mathbf{u}}$ for a multivalued variable X, the proportional scheme, i.e. the one that sets $\theta'_{x_i|\mathbf{u}} = (1-\theta'_{x|\mathbf{u}})(\theta_{x_i|\mathbf{u}}/(1-\theta_{x|\mathbf{u}}))$ for all $x_i \neq x$, leads to the smallest distance between the original and new distributions of X, which is given by:*

$$\begin{aligned} D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}) &= \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta_{x|\mathbf{u}}} - \ln \frac{\theta'_{\overline{x}|\mathbf{u}}}{\theta_{\overline{x}|\mathbf{u}}} \right| \\ &= \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta'_{\overline{x}|\mathbf{u}}} - \ln \frac{\theta_{x|\mathbf{u}}}{\theta_{\overline{x}|\mathbf{u}}} \right|, \end{aligned}$$

where we define $\theta'_{\overline{x}|\mathbf{u}} = 1 - \theta'_{x|\mathbf{u}}$ and $\theta_{\overline{x}|\mathbf{u}} = 1 - \theta_{x|\mathbf{u}}$.

Theorem 4.2 thus justifies the use of the proportional scheme on the grounds that it leads to the tightest bound on the amount of associated belief change.

5 Applications to belief revision

The problem of probabilistic belief revision can be defined as follows. We are given a probability distribution Pr, which captures a state of belief and assigns a probability p to some event γ. We then obtain evidence suggesting a probability of $q \neq p$ for γ. Our goal is to change the distribution Pr to a new distribution Pr' such that $Pr'(\gamma) = q$. There are two problems here. First, usually there are many choices for Pr'. Which one should we adopt? Second, if we decide to choose the new state of belief Pr' according to some specific method, can we provide any guarantee on the amount of belief change that will be undergone as a result of moving from Pr to Pr'?

As for the first question, we will consider two methods for updating a probability distribution in the face of new evidence: Jeffrey's rule [9] and Pearl's method of virtual evidence [15]. As for the second question, we will show next that we can indeed provide interesting guarantees on the amount of belief change induced by both methods. We present the guarantees first and then some of their applications.

5.1 Jeffrey's rule

We start with Jeffrey's rule for accommodating uncertain evidence.

Definition 5.1 *Let Pr be a probability distribution over worlds w, and let $\gamma_1, \ldots, \gamma_n$ be a set of mutually exclusive and exhaustive events that are assigned probabilities $p_1, \ldots, p_n$, respectively, by Pr. Suppose we want to change Pr to a new distribution Pr' such that the probabilities of $\gamma_1, \ldots, \gamma_n$ become $q_1, \ldots, q_n$, respectively. Jeffrey's rule defines the new distribution Pr' as follows:*

$$Pr'(w) \overset{def}{=} Pr(w)\frac{q_i}{p_i}, \quad \textit{if } w \models \gamma_i.$$

The main result we have about Jeffrey's rule is that the distance between probability distributions Pr and Pr' can be computed directly from the old and new probabilities of $\gamma_1, \ldots, \gamma_n$. This immediately allows us to invoke Theorem 2.2 as we show next.

Theorem 5.1 *Let Pr and Pr' be two distributions, where Pr' is obtained by applying Jeffrey's rule to Pr as given in Definition 5.1. We then have:*

$$D(Pr, Pr') = \ln \max_i \frac{q_i}{p_i} - \ln \min_i \frac{q_i}{p_i}.$$

We immediately get the following bound.

Corollary 5.1 *If O and O' are the odds functions before and after applying Jeffrey's rule as given in Definition 5.1, then:*

$$e^{-d} \leq \frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \leq e^{d},$$

where $d = \ln \max_i (q_i/p_i) - \ln \min_i (q_i/p_i)$.

To consider an example application of Corollary 5.1, we use a simple example from Jeffrey [9].

Example 5.1 *Assume that we are given a piece of cloth, where its color can be one of: green (c_g), blue (c_b), or violet (c_v). We also want to know whether in the next day, the cloth will be sold (s), or not sold ($\overline{s}$). Our original state of belief is given by the probability distribution of the worlds Pr:*

$$Pr(s, c_g) = .12,\ Pr(s, c_b) = .12,\ Pr(s, c_v) = .32,$$
$$Pr(\overline{s}, c_g) = .18,\ Pr(\overline{s}, c_b) = .18,\ Pr(\overline{s}, c_v) = .08.$$

Therefore, our original state of belief on the color of the cloth (c_g, c_b, c_v) is given by the distribution $(.3, .3, .4)$. Assume that we now inspect the cloth by candlelight, and we want to revise our state of belief on the color of the cloth to the new distribution $(.7, .25, .05)$ using Jeffrey's rule. The distance between the original and new distributions of the worlds can be computed by simply examining the original and new distributions on the color variable as given by Theorem 5.1. Specifically, the distance between the two distributions is $\ln(.7/.3) - \ln(.05/.4) = 2.93$. We can now use this distance to provide a bound on the change in any of our beliefs. Consider for example our belief that the cloth is green given that it is sold tomorrow, $Pr(c_g \mid s)$, which is initially $.214$. Suppose we want to find the bound on the change in this belief induced by the new evidence. Given Corollary 5.1 and Inequality 2, we have:

$$.0144 \leq Pr'(c_g \mid s) \leq .836,$$

which suggests that a dramatic change in belief is possible in this case. If we actually apply Jeffrey's rule, we get the new distribution Pr':

$$Pr'(s, c_g) = .28,\ Pr'(s, c_b) = .10,\ Pr'(s, c_v) = .04,$$
$$Pr'(\overline{s}, c_g) = .42,\ Pr'(\overline{s}, c_b) = .15,\ Pr'(\overline{s}, c_v) = .01,$$

according to which $Pr'(c_g \mid s) = .667$, which does suggest a dramatic change. On the other hand, if the new evidence on the color of the cloth is given by the distribution $(.25, .25, .50)$ instead, the distance between the old and new distributions will be $.406$, and our bound will be: $.153 \leq Pr'(c_g \mid s) \leq .290$, which is obviously much tighter as this evidence is much weaker.

Alternatively, we can compute the KL–divergence after applying Jeffrey's rule:

$$KL(Pr, Pr') = \sum_i -p_i \ln \frac{q_i}{p_i}.$$

Hence for Jeffrey's rule, one can obtain locally both the average–case bounds and the worst–case bounds on the amount of belief change.

We close this section by showing that Jeffrey's rule commits to a probability distribution which minimizes our distance measure. Hence, Jeffrey's rule leads to the strongest bound on the amount of belief change.[3]

Theorem 5.2 *The new distribution Pr' obtained by applying Jeffrey's rule to an initial distribution Pr is optimal in the following sense. Among all possible distributions that assign probabilities $q_1, \ldots, q_n$ to events $\gamma_1, \ldots, \gamma_n$, Pr' minimizes the distance from Pr, according to the measure defined in Definition 2.1.*[4]

5.2 Pearl's method of virtual evidence

We now consider Pearl's method of virtual evidence. According to this method, we also have a new evidence η that bears on a set of mutually exclusive and exhaustive events $\gamma_1, \ldots, \gamma_n$, but the evidence is not specified as a set of new probabilities for these events. Instead, for each γ_i, $i \neq 1$, we are given a number λ_i which is interpreted as the ratio $Pr(\eta \mid \gamma_i)/Pr(\eta \mid \gamma_1)$. That is, λ_i represents the likelihood ratio that we would obtain evidence η given γ_i, compared with given γ_1. Note that under this interpretation, we must have $\lambda_1 = 1$.

Definition 5.2 *Let Pr be a probability distribution over worlds w, and let $\gamma_1, \ldots, \gamma_n$ be a set of mutually exclusive and exhaustive events that are assigned probabilities $p_1, \ldots, p_n$, respectively, by Pr. Suppose we want to change Pr to a new distribution Pr' to incorporate virtual evidence η, specified by $\lambda_1, \ldots, \lambda_n$,*

[3] It has been previously proven that the probability distribution obtained by Jeffrey's rule also minimizes the KL–divergence [5]. Hence, our result strengthens this previously known result as it shows that Jeffery's rule lead to the strongest worst–case and average–case bounds on belief changes.

[4] Note that this distribution Pr' is not necessarily unique.

with $\lambda_1 = 1$ *and* $\lambda_i = Pr(\eta \mid \gamma_i)/Pr(\eta \mid \gamma_1)$ *if* $i \neq 1$*. Pearl's method of virtual evidence defines the new distribution* Pr' *as follows:*

$$Pr'(w) \overset{def}{=} Pr(w)\frac{\lambda_i}{\sum_j p_j \lambda_j}, \quad \textit{if } w \models \gamma_i.$$

Again, we can easily compute the distance between distributions Pr and Pr' using only local information.

Theorem 5.3 *Let* Pr *and* Pr' *be two distributions, where* Pr' *is obtained from* Pr *by accommodating virtual evidence as given by Definition 5.2. We then have:*

$$D(Pr, Pr') = \ln \max_i \lambda_i - \ln \min_i \lambda_i.$$

This immediately gives us the following bound.

Corollary 5.2 *If* O *and* O' *are the odds functions before and after applying Pearl's method as given in Definition 5.2, then:*

$$e^{-d} \leq \frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \leq e^{d},$$

where $d = \ln \max_i \lambda_i - \ln \min_i \lambda_i$.

For the special case where our evidence η bears only on $\neg\gamma$ versus γ, with $\lambda = Pr(\eta \mid \gamma)/Pr(\eta \mid \neg\gamma)$, the above bound reduces to $|\ln O'(\alpha \mid \beta) - \ln O(\alpha \mid \beta)| \leq |\ln \lambda|$. Therefore, the bound is tighter when λ is closer to 1. Clearly, when $\lambda = 1$, the evidence is trivial and the two distributions are the same.

Alternatively, we can compute the KL–divergence due to applying virtual evidence:

$$KL(Pr, Pr') = \sum_i -p_i \ln \frac{\lambda_i}{\sum_j p_j \lambda_j}.$$

Note, however, that to compute the KL–divergence, we need to know $p_i = Pr(\gamma_i)$, which is not available locally from simply examining the virtual evidence. Hence, although we can use KL–divergence to offer average–case bounds on belief changes induced by the method of virtual evidence, we cannot do this in constant time as is possible using our proposed distance measure.

Consider the following example from Pearl [15].

Example 5.2 *On any given day, there is a burglary on any given house with probability* $Pr(b) = 10^{-4}$*, while the alarm of Mr. Holmes' house will go off if there is a burglary with probability* $Pr(a \mid b) = .95$*, and go off if there is no burglary with probability* $Pr(a \mid \bar{b}) = .01$*. One day, Mr. Holmes' receives a call from his neighbor, Mrs. Gibbons, saying she may have heard the alarm of his house going off. Mr. Holmes concludes that there is an 80% chance that Mrs. Gibbons did hear the alarm going off. According to Pearl's method, this evidence can be interpreted as:* $\lambda = Pr(\eta \mid a)/Pr(\eta \mid \overline{a}) = 4$*. Therefore, the distance between the original distribution* Pr*, and the new distribution* Pr' *which results from incorporating the virtual evidence, is* $|\ln \lambda| = |\ln 4| = 1.386$*. We can use this distance to bound the change in any of our beliefs. In particular, we may want to bound the new probability that there was a burglary at Mr. Holmes' house. Inequality 2 gives us:*

$$2.50 \times 10^{-5} \leq Pr'(b) \leq 4.00 \times 10^{-4}.$$

If we actually apply Pearl's method, we get $Pr'(b) = 3.85 \times 10^{-4}$*.*

Our distance measure is then useful for approximate reasoning given *soft evidence*, as we can use the bound to approximate the probability of any event after the accommodation of such evidence. The approximation itself takes constant time to compute since we only need to compute the distance measure and apply Inequality 2. We stress, however, that the bound becomes trivial in the case of *hard evidence* since the initial and new distributions no longer have the same support in this case, making the distance between them infinitely large.

We close this section by a final application of our distance measure, relating to the notion of *evidence strength.*

Example 5.3 *Going back to Example 5.1, we ask: What kind of evidence will assure us that our belief in the cloth being green given that it is sold tomorrow, which is now at .214, would not exceed .3? Inequality 2 can be used in this case to obtain a sufficient condition on the strength of evidence which will ensure this. Specifically, Inequality 2 gives us:*

$$\frac{.214\, e^{-d}}{.214\,(e^{-d}-1)+1} \leq Pr'(c_g \mid s) \leq \frac{.214\, e^{d}}{.214\,(e^{d}-1)+1}.$$

To ensure that $Pr'(c_g \mid s) \leq .3$*, we must find a distance* d *that equates the above upper bound to* $.3$*. A value of* $d = .454$ *has this property. Hence, any piece of evidence which has a distance of no more than* $.454$ *from the current distribution on color,* $(.3, .3, .4)$*, would guarantee that* $Pr'(c_g \mid s)$ *does not exceed* $.3$*. Following are some pieces of evidence which satisfy this condition:* $(.25, .25, .5)$*,* $(.25, .3, .45)$ *and* $(.35, .3, .35)$*.*

6 Conclusion

We proposed a distance measure between two probability distributions, which allows one to bound the amount of belief change that occurs when moving from one distribution to the other. We also contrasted the proposed measure with some well known measures, including KL–divergence. We then presented two practical applications of the proposed distance measure: sensitivity analysis in belief networks and probabilistic belief revision. We showed how the distance measure can be easily computed in these applications, and then used it to bound global belief changes that result from either the perturbation of local conditional beliefs or the accommodation of soft evidence. Finally, we showed that two well known techniques in sensitivity analysis and belief revision correspond to the minimization of our proposed distance measure and, hence, can be shown to be optimal from that viewpoint.

Acknowledgments

This work has been partially supported by NSF grant IIS-9988543, MURI grant N00014-00-1-0617, and by DiMI grant 00-10065.

References

[1] E. Castillo and J. M. Gutiérrez and A. S. Hadi. Sensitivity Analysis in Discrete Bayesian Networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part A (Systems and Humans)*, volume 27, pages 412-423, 1997.

[2] H. Chan and A. Darwiche. When do Numbers Really Matter? In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 65-74. Morgan Kaufmann Publishers, San Francisco, California, 2001.

[3] V. M. H. Coupé, F. V. Jensen, U. Kjærulff and L. C. van der Gaag. *A Computational Architecture for N-way Sensitivity Analysis of Bayesian Networks.* Technical Report, 2000.

[4] A. Darwiche. A Differential Approach to Inference in Bayesian Networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 123-132. Morgan Kaufmann Publishers, San Francisco, California, 2000.

[5] P. Diaconis and S. L. Zabell. Updating Subjective Probability. *Journal of the American Statistical Association*, volume 77, pages 822-830, 1982.

[6] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT press, Cambridge, Massachusetts, 1988.

[7] I. J. Good. *Probability and the Weighing of Evidence.* Charles Griffin, London, 1950.

[8] I. J. Good. *Good Thinking: the Foundations of Probability and Its Applications.* University of Minnesota Press, Minneapolis, 1983.

[9] R. C. Jeffrey. *The Logic of Decision.* McGraw-Hill, New York, 1965. (2nd edition: University of Chicago Press, Chicago, 1983; Paperback correction: 1990.)

[10] R. C. Jeffrey. *Probability and the Art of Judgement.* Cambridge University Press, Cambridge, 1992.

[11] F. V. Jensen. *Bayesian Networks and Decision Graphs.* Springer–Verglag, New York, 2001.

[12] U. Kjærulff and L. C. van der Gaag. Making Sensitivity Analysis Computationally Efficient. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 317-325. Morgan Kaufmann Publishers, San Francisco, California, 2000.

[13] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, volume 22, pages 79-86, 1951.

[14] K. B. Laskey. Sensitivity Analysis for Probability Assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics*, volume 25, pages 901-909, 1995.

[15] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Francisco, 1988.

[16] L. C. van der Gaag and S. Renooij. Analysing Sensitivity Data from Probabilistic Networks. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 530-537. Morgan Kaufmann Publishers, San Francisco, California, 2001.

A Proofs

Proof of Theorem 2.1 The distance measure given in Definition 2.1 satisfies these three properties:

Positiveness: Obviously, we have $D(Pr, Pr') \geq 0$ and $D(Pr, Pr) = 0$. We can easily see that $\max_w (Pr'(w)/Pr(w)) \geq 1$ and $\min_w (Pr'(w)/Pr(w)) \leq 1$, since $\sum_w Pr'(w) = \sum_w Pr(w) = 1$. When $D(Pr, Pr') = 0$, we must have $\max_w (Pr'(w)/Pr(w)) = \min_w (Pr'(w)/Pr(w)) = 1$. Therefore, $Pr(w) = Pr'(w)$ for all w, and thus $Pr = Pr'$.

Symmetry: Since $\max_w (Pr'(w)/Pr(w)) = \min_w (Pr(w)/Pr'(w))$ and $\min_w (Pr'(w)/Pr(w)) = \max_w (Pr(w)/Pr'(w))$, we can easily see that $D(Pr, Pr') = D(Pr', Pr)$.

Triangle inequality: Let $Pr''(w_i)/Pr(w_i) = \max_w (Pr''(w)/Pr(w))$ and $Pr''(w_j)/Pr(w_j) = \min_w (Pr''(w)/Pr(w))$. Since $\ln(Pr''(w_i)/Pr(w_i)) = \ln(Pr''(w_i)/Pr'(w_i)) - \ln(Pr'(w_i)/Pr(w_i))$, we have:

$$
\begin{aligned}
& D(Pr, Pr'') \\
&= \ln \max_w \frac{Pr''(w)}{Pr(w)} - \ln \min_w \frac{Pr''(w)}{Pr(w)} \\
&= \ln \frac{Pr''(w_i)}{Pr(w_i)} - \ln \frac{Pr''(w_j)}{Pr(w_j)} \\
&= \left(\ln \frac{Pr''(w_i)}{Pr'(w_i)} - \ln \frac{Pr'(w_i)}{Pr(w_i)}\right) - \left(\ln \frac{Pr''(w_j)}{Pr'(w_j)} - \ln \frac{Pr'(w_j)}{Pr(w_j)}\right) \\
&= \left(\ln \frac{Pr'(w_j)}{Pr(w_j)} - \ln \frac{Pr'(w_i)}{Pr(w_i)}\right) + \left(\ln \frac{Pr''(w_i)}{Pr'(w_i)} - \ln \frac{Pr''(w_j)}{Pr'(w_j)}\right) \\
&\leq \left(\ln \max_w \frac{Pr'(w)}{Pr(w)} - \ln \min_w \frac{Pr'(w)}{Pr(w)}\right) + \left(\ln \max_w \frac{Pr''(w)}{Pr'(w)} - \ln \min_w \frac{Pr''(w)}{Pr'(w)}\right) \\
&= D(Pr, Pr') + D(Pr', Pr''). \square
\end{aligned}
$$

Proof of Theorem 2.2 If Pr and Pr' do not have the same support, we have $D(Pr, Pr') = \infty$, and thus $-\infty = e^{-D(Pr,Pr')} \leq O'(\alpha \mid \beta)/O(\alpha \mid \beta) \leq e^{D(Pr,Pr')} = \infty$. If they have the same support, let $r_w = Pr'(w)/Pr(w)$. The odds ratio $O'(\alpha \mid \beta)/O(\alpha \mid \beta)$ can be expressed as:

$$
\begin{aligned}
\frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} &= \frac{Pr'(\alpha \mid \beta)/Pr'(\overline{\alpha} \mid \beta)}{Pr(\alpha \mid \beta)/Pr(\overline{\alpha} \mid \beta)} \\
&= \frac{Pr'(\alpha, \beta)/Pr'(\overline{\alpha}, \beta)}{Pr(\alpha, \beta)/Pr(\overline{\alpha}, \beta)} \\
&= \frac{(\sum_{w \models \alpha,\beta} Pr'(w))/(\sum_{w \models \overline{\alpha},\beta} Pr'(w))}{(\sum_{w \models \alpha,\beta} Pr(w))/(\sum_{w \models \overline{\alpha},\beta} Pr(w))} \\
&= \frac{(\sum_{w \models \alpha,\beta} r_w Pr(w))/(\sum_{w \models \overline{\alpha},\beta} r_w Pr(w))}{(\sum_{w \models \alpha,\beta} Pr(w))/(\sum_{w \models \overline{\alpha},\beta} Pr(w))}.
\end{aligned}
$$

We now introduce $\max_w r_w$ and $\min_w r_w$ to get the upper bound of the odds ratio:

$$
\frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \leq \frac{((\max_w r_w) \sum_{w \models \alpha,\beta} Pr(w))/((\min_w r_w) \sum_{w \models \overline{\alpha},\beta} Pr(w))}{(\sum_{w \models \alpha,\beta} Pr(w))/(\sum_{w \models \overline{\alpha},\beta} Pr'(w))}
$$

$$= \frac{\max_w r_w}{\min_w r_w}.$$

Similarly, we can also get the lower bound on the odds ratio:

$$\frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \geq \frac{((\min_w r_w) \sum_{w \models \alpha,\beta} Pr(w))/((\max_w r_w) \sum_{w \models \overline{\alpha},\beta} Pr(w))}{(\sum_{w \models \alpha,\beta} Pr(w))/(\sum_{w \models \overline{\alpha},\beta} Pr'(w))}$$
$$= \frac{\min_w r_w}{\max_w r_w}.$$

We note that the bounds can be expressed using our distance measure:

$$e^{D(Pr,Pr')} = \frac{\max_w Pr'(w)/Pr(w)}{\min_w Pr'(w)/Pr(w)} = \frac{\max_w r_w}{\min_w r_w};$$
$$e^{-D(Pr,Pr')} = \frac{\min_w Pr'(w)/Pr(w)}{\max_w Pr'(w)/Pr(w)} = \frac{\min_w r_w}{\max_w r_w}.$$

Therefore, we have $e^{-D(Pr,Pr')} \leq O'(\alpha \mid \beta)/O(\alpha \mid \beta) \leq e^{D(Pr,Pr')}$. If both $O'(\alpha \mid \beta)$ and $O(\alpha \mid \beta)$ takes on either 0 or ∞, the theorem still holds because $0/0 \overset{def}{=} 1$ and $\infty/\infty \overset{def}{=} 1$.

The bound is tight in the sense that for every pair of distributions Pr and Pr', there are events $\alpha = w_i$ and $\beta = w_i$ or w_j, where $r_{w_i} = \max_w r_w$ and $r_{w_j} = \min_w r_w$, such that:

$$O'(\alpha \mid \beta) = \frac{\sum_{w \models \alpha,\beta} r_w Pr(w)}{\sum_{w \models \overline{\alpha},\beta} r_w Pr(w)}$$
$$= \frac{r_{w_i} Pr(w_i)}{r_{w_j} Pr(w_j)}.$$

Since $O(\alpha \mid \beta) = Pr(w_i)/Pr(w_j)$ and $e^{D(Pr,Pr')} = (\max_w r_w)/(\min_w r_w) = r_{w_i}/r_{w_j}$, we have $O'(\alpha \mid \beta)/O(\alpha \mid \beta) = e^{D(Pr,Pr')}$. Similarly, we can get $O'(\overline{\alpha} \mid \beta)/O(\overline{\alpha} \mid \beta) = e^{-D(Pr,Pr')}$.□

Proof of Theorem 3.1 To obtain the largest probability change for the conditional event $\alpha \mid \beta$ given some KL–divergence, the new probability distribution Pr' needs to satisfy the following three properties:

- For all $w \models \alpha, \beta$, $Pr'(w) = Pr(w)(Pr'(\alpha \mid \beta)/Pr(\alpha \mid \beta))$.
- For all $w \models \overline{\alpha}, \beta$, $Pr'(w) = Pr(w)(Pr'(\overline{\alpha} \mid \beta)/Pr(\overline{\alpha} \mid \beta))$.
- For all $w \models \overline{\beta}$, $Pr'(w) = Pr(w)$, i.e., $Pr'(\overline{\beta}) = Pr(\overline{\beta})$.

Under these three conditions, we have the following equality:

$$
\begin{aligned}
&KL(Pr, Pr') \\
&= -Pr(\alpha, \beta) \ln \frac{Pr'(\alpha, \beta)}{Pr(\alpha, \beta)} - Pr(\overline{\alpha}, \beta) \ln \frac{Pr'(\overline{\alpha}, \beta)}{Pr(\overline{\alpha}, \beta)} \\
&= -Pr(\beta) \left(Pr(\alpha \mid \beta) \ln \frac{Pr'(\alpha \mid \beta)}{Pr(\alpha \mid \beta)} + Pr(\overline{\alpha} \mid \beta) \ln \frac{Pr'(\overline{\alpha} \mid \beta)}{Pr(\overline{\alpha} \mid \beta)} \right).
\end{aligned}
$$

Since this is the case where we get the largest probability change, the following inequality must be satisfied:

$$
\begin{aligned}
&KL(Pr, Pr') \\
&\geq -Pr(\beta) \left(Pr(\alpha \mid \beta) \ln \frac{Pr'(\alpha \mid \beta)}{Pr(\alpha \mid \beta)} + (1 - Pr(\alpha \mid \beta)) \ln \frac{1 - Pr'(\alpha \mid \beta)}{1 - Pr(\alpha \mid \beta)} \right).
\end{aligned}
$$

We can now plug in $Pr'(\alpha \mid \beta) = O'(\alpha \mid \beta)/(O'(\alpha \mid \beta) + 1)$ and $Pr(\alpha \mid \beta) = O(\alpha \mid \beta)/(O(\alpha \mid \beta) + 1)$, and after simplification, we get:

$$
KL(Pr, Pr') \geq Pr(\beta) \left(\ln \frac{O'(\alpha \mid \beta) + 1}{O(\alpha \mid \beta) + 1} - \frac{O(\alpha \mid \beta)}{O(\alpha \mid \beta) + 1} \ln \frac{O'(\alpha \mid \beta)}{O(\alpha \mid \beta)} \right) . \square
$$

Proof of Theorem 3.2 Similar to the proof of Theorem 2.2, by replacing α, β with γ_1, and $\overline{\alpha}, \beta$ with γ_2.□

Proof of Theorem 3.3 If $\gamma_1, \ldots, \gamma_n$ is a set of mutually exclusive and exhaustive events, we have:

$$
\begin{aligned}
\sum_i Pr(\gamma_i) \ln F_{Pr',Pr}(\alpha : \gamma_i) &= \sum_i Pr(\gamma_i) \ln \frac{Pr'(\alpha)/Pr(\alpha)}{Pr'(\gamma_i)/Pr(\gamma_i)} \\
&= \sum_i Pr(\gamma_i) \ln \frac{Pr'(\alpha)}{Pr(\alpha)} - \sum_i Pr(\gamma_i) \ln \frac{Pr'(\gamma_i)}{Pr(\gamma_i)} \\
&= \ln \frac{Pr'(\alpha)}{Pr(\alpha)} \sum_i Pr(\gamma_i) - \sum_i Pr(\gamma_i) \ln \frac{Pr'(\gamma_i)}{Pr(\gamma_i)} \\
&= \ln \frac{Pr'(\alpha)}{Pr(\alpha)} - \sum_i Pr(\gamma_i) \ln \frac{Pr'(\gamma_i)}{Pr(\gamma_i)}.
\end{aligned}
$$

We also note the following inequality about KL–divergence:

$$
KL(Pr, Pr') = \sum_w -Pr(w) \ln \frac{Pr'(w)}{Pr(w)}
$$

$$\begin{aligned}&\geq \sum_i -Pr(\gamma_i) \ln \frac{Pr'(\gamma_i)}{Pr(\gamma_i)}\\ &\geq 0.\end{aligned}$$

Therefore, we have:

$$\ln \frac{Pr'(\alpha)}{Pr(\alpha)} \leq \sum_i Pr(\gamma_i) \ln F_{Pr',Pr}(\alpha : \gamma_i) \leq \ln \frac{Pr'(\alpha)}{Pr(\alpha)} + KL(Pr, Pr'),$$

or:

$$0 \leq \sum_i Pr(\gamma_i) \ln F_{Pr',Pr}(\alpha : \gamma_i) - \ln \frac{Pr'(\alpha)}{Pr(\alpha)} \leq KL(Pr, Pr').$$

If instead of an arbitrary partition $\gamma_1, \ldots, \gamma_n$, we are given the set of worlds w, then we will get the following equality relation:

$$\sum_w Pr(w) \ln F_{Pr',Pr}(\alpha : w) - \ln \frac{Pr'(\alpha)}{Pr(\alpha)} = KL(Pr, Pr').\square$$

Proof of Theorem 3.4 To prove this theorem, we first establish two lemmas:

Lemma A.1 *Given two distributions Pr and Pr', and a set of worlds S_+, where $\forall_{w\in S_+} Pr'(w) > Pr(w)$, let $r_{max} = \max_{w\in S_+}(Pr'(w)/Pr(w))$. We have the following inequality:*

$$\sum_{w\in S_+} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} \leq -\frac{Pr'(S_+) - Pr(S_+)}{r_{max} - 1} \ln r_{max}. \tag{A.1}$$

Lemma A.2 *Given two distributions Pr and Pr', and a set of worlds S_-, where $\forall_{w\in S_-} Pr'(w) < Pr(w)$, let $r_{min} = \min_{w\in S_-}(Pr'(w)/Pr(w))$. We have the following inequality:*

$$\sum_{w\in S_-} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} \leq -\frac{Pr'(S_-) - Pr(S_-)}{r_{min} - 1} \ln r_{min}. \tag{A.2}$$

We use induction to prove Lemma A.1. Lemma A.2 can be proved similarly.

Assume $S = \{w\}$, where $Pr'(w) > Pr(w)$. We have $r_{max} = Pr'(w)/Pr(w)$. Therefore:

$$-Pr(w) \ln \frac{Pr'(w)}{Pr(w)} = -\frac{r_{max}Pr(w) - Pr(w)}{r_{max} - 1} \ln r_{max}$$

$$= -\frac{Pr'(w) - Pr(w)}{r_{max} - 1} \ln r_{max}.$$

Therefore, Inequality A.1 is satisfied for the set S.

Now we are given sets $S_1, \ldots, S_n$, where $S_i \cap S_j = \emptyset$ for all S_i, S_j, $i \neq j$, and $\forall_{w \in S_i} Pr'(w) > Pr(w)$ for all S_i. Assume that Inequality A.1 is true for all S_i, i.e. if $r_i = \max_{w \in S_i} (Pr'(w)/Pr(w))$, we have:

$$\sum_{w \in S_i} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} \leq -\frac{Pr'(S_i) - Pr(S_i)}{r_i - 1} \ln r_i.$$

We want to prove that Inequality A.1 is also true for $S_+ = \bigcup_{i=1}^{n} S_i$. First, we note that if $p \geq q > 1$, we get $-(\ln p)/(p-1) \geq -(\ln q)/(q-1)$. Now let $r_{max} = \max_{w \in S_+} (Pr'(w)/Pr(w))$. Since $r_{max} \geq r_i > 1$ for all i, we have:

$$\begin{aligned}
\sum_{w \in S_+} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} &= \sum_{i=1}^{n} \sum_{w \in S_i} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} \\
&\leq \sum_{i=1}^{n} -\frac{Pr'(S_i) - Pr(S_i)}{r_i - 1} \ln r_i \\
&\leq \sum_{i=1}^{n} -\frac{Pr'(S_i) - Pr(S_i)}{r_{max} - 1} \ln r_{max} \\
&= -\frac{\sum_{i=1}^{n} (Pr'(S_i) - Pr(S_i))}{r_{max} - 1} \ln r_{max} \\
&= -\frac{Pr'(S_+) - Pr(S_+)}{r_{max} - 1} \ln r_{max}.
\end{aligned}$$

This proves that Inequality A.1 is also true for S_+. Therefore, by induction, Lemma A.1 is true.

We now proceed with the proof of Theorem 3.4. Given distributions Pr and Pr', we divide all worlds into three subsets: S_+, where $\forall_{w \in S_+} Pr'(w) > Pr(w)$; S_-, where $\forall_{w \in S_-} Pr'(w) < Pr(w)$; and S_0, where $\forall_{w \in S_0} Pr'(w) = Pr(w)$. It is obvious that we must have $Pr'(S_+) - Pr(S_+) = -(Pr'(S_-) - Pr(S_-))$. If we are given $r_{max} = \max_w (Pr'(w)/Pr(w))$ and $r_{min} = \min_w (Pr'(w)/Pr(w))$, then the KL–divergence between Pr and Pr' is bounded by:

$$\begin{aligned}
KL(Pr, Pr') &= \sum_{w} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} \\
&= \sum_{w \in S_+} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} + \sum_{w \in S_-} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{w \in S_0} -Pr(w) \ln \frac{Pr'(w)}{Pr(w)} \\
\leq & -\frac{Pr'(S_+) - Pr(S_+)}{r_{max} - 1} \ln r_{max} - \frac{Pr'(S_-) - Pr(S_-)}{r_{min} - 1} \ln r_{min} \\
= & (Pr'(S_+) - Pr(S_+)) \left(\frac{\ln r_{min}}{r_{min} - 1} - \frac{\ln r_{max}}{r_{max} - 1} \right).
\end{aligned}
$$

We first maximize $Pr'(S_+) - Pr(S_+)$ with respect to r_{max} and r_{min}. The constraints we have are $Pr'(S_+)/Pr(S_+) \leq r_{max}$, $Pr'(S_-)/Pr(S_-) \geq r_{min}$, and $Pr(S_+) + Pr(S_-) = Pr'(S_+) + Pr'(S_-) \leq 1$. The solution we get is:

$$
Pr'(S_+) - Pr(S_+) \leq \frac{(r_{max} - 1)(1 - r_{min})}{r_{max} - r_{min}}.
$$

Therefore:

$$
\begin{aligned}
KL(Pr, Pr') \leq & \left(\frac{(r_{max} - 1)(1 - r_{min})}{r_{max} - r_{min}} \right) \left(\frac{\ln r_{min}}{r_{min} - 1} - \frac{\ln r_{max}}{r_{max} - 1} \right) \\
= & \frac{(1 - r_{max}) \ln r_{min} - (1 - r_{min}) \ln r_{max}}{r_{max} - r_{min}}.
\end{aligned}
$$

We now maximize this term with respect to $d = \ln r_{max} - \ln r_{min}$. The maximum value is attained at $r_{max} = e^d d/(e^d - 1)$ and $r_{min} = d/(e^d - 1)$. The maximum KL–divergence value is thus given by:

$$
\begin{aligned}
KL(Pr, Pr') \leq & \frac{(1 - \frac{e^d d}{e^d - 1}) \ln \frac{d}{e^d - 1} - (1 - \frac{d}{e^d - 1}) \ln \frac{e^d d}{e^d - 1}}{\frac{e^d d}{e^d - 1} - \frac{d}{e^d - 1}} \\
= & \frac{(e^d - 1 - e^d d) \ln \frac{d}{e^d - 1} - (e^d - 1 - d)(d + \ln \frac{d}{e^d - 1})}{(e^d - 1)d} \\
= & \frac{-(e^d - 1)d \ln \frac{d}{e^d - 1} - (e^d - 1 - d)d}{(e^d - 1)d} \\
= & -\ln \frac{d}{e^d - 1} - \frac{e^d - 1 - d}{e^d - 1} \\
= & -1 - \ln \frac{d}{e^d - 1} + \frac{d}{e^d - 1}. \square
\end{aligned}
$$

Proof of Theorem 4.1 We first prove the following lemma.

Lemma A.3 *Assume that we change parameter $\theta_{x|\mathbf{u}}$ to $\theta'_{x|\mathbf{u}}$ for every value x, and $Pr(\mathbf{u}) > 0$. For all x_i where $\theta'_{x_i|\mathbf{u}} > 0$ or $\theta_{x_i|\mathbf{u}} > 0$, there must exist some $w \models x_i, \mathbf{u}$ such that it satisfies the condition $Pr'(w)/Pr(w) = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$. For* *all other worlds w that do not satisfy this condition, we must have $Pr'(w) = Pr(w)$, and thus $Pr'(w)/Pr(w) = 1$.*[5]

First we note that $Pr'(\mathbf{u}) = Pr(\mathbf{u}) > 0$. For any world w, either $w \models \bar{\mathbf{u}}$, or $w \models x_i, \mathbf{u}$ for some x_i. We consider the different cases of w.

- If $w \models \bar{\mathbf{u}}$, we must have $Pr'(w) = Pr(w)$, because we are only changing parameters $\theta_{x|\mathbf{u}}$.
- If $w \models x_i, \mathbf{u}$, we consider four cases of x_i:
  - · If $\theta'_{x_i|\mathbf{u}} = \theta_{x_i|\mathbf{u}} = 0$, $Pr'(x_i, \mathbf{u}) = Pr(x_i, \mathbf{u}) = 0$. Therefore, for all worlds $w \models x_i, \mathbf{u}$, $Pr'(w) = Pr(w) = 0$.
  - · If $\theta'_{x_i|\mathbf{u}} = 0$ and $\theta_{x_i|\mathbf{u}} > 0$, $Pr'(x_i, \mathbf{u}) = 0$ and $Pr(x_i, \mathbf{u}) > 0$. Therefore, for all worlds $w \models x_i, \mathbf{u}$, either $Pr'(w) = Pr(w) = 0$; or $Pr'(w) = 0$ and $Pr(w) > 0$, giving us $Pr'(w)/Pr(w) = 0 = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$. Moreover, because $Pr(x_i, \mathbf{u}) > 0$, there must exist some $w \models x_i, \mathbf{u}$ such that $Pr(w) > 0$, and thus satisfying the condition $Pr'(w)/Pr(w) = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$.
  - · If $\theta'_{x_i|\mathbf{u}} > 0$ and $\theta_{x_i|\mathbf{u}} = 0$, $Pr'(x_i, \mathbf{u}) > 0$ and $Pr(x_i, \mathbf{u}) = 0$. Therefore, for all worlds $w \models x_i, \mathbf{u}$, either $Pr'(w) = Pr(w) = 0$; or $Pr'(w) > 0$ and $Pr(w) = 0$, giving us $Pr'(w)/Pr(w) = \infty = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$. Moreover, because $Pr'(x_i, \mathbf{u}) > 0$, there must exist some $w \models x_i, \mathbf{u}$ such that $Pr'(w) > 0$, and thus satisfying the condition $Pr'(w)/Pr(w) = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$.
  - · If $\theta'_{x_i|\mathbf{u}} > 0$ and $\theta_{x_i|\mathbf{u}} > 0$, $Pr'(x_i, \mathbf{u}) > 0$ and $Pr(x_i, \mathbf{u}) > 0$. Therefore, for all worlds $w \models x_i, \mathbf{u}$, either $Pr'(w) = Pr(w) = 0$; or $Pr'(w) > 0$ and $Pr(w) > 0$, giving us $Pr'(w)/Pr(w) = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$. Moreover, because $Pr'(x_i, \mathbf{u}) > 0$ and $Pr(x_i, \mathbf{u}) = 0$, there must exist some $w \models x_i, \mathbf{u}$ such that $Pr'(w) > 0$ and $Pr(w) > 0$, and thus satisfying the condition $Pr'(w)/Pr(w) = \theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}$.

Therefore Lemma A.3 is correct, and we can conclude $\max_w(Pr'(w)/Pr(w)) = \max_{x_i}(\theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}})$ and $\min_w(Pr'(w)/Pr(w)) = \min_{x_i}(\theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}})$. Consequently, we have $D(Pr, Pr') = D(\Theta_{X,\mathbf{u}}, \Theta'_{X,\mathbf{u}})$.$\square$

Proof of Theorem 4.2 Let $\Theta'_{X|\mathbf{u}}$ be the distribution generated by the proportional scheme, where $\theta'_{x_i|\mathbf{u}} = (1 - \theta'_{x|\mathbf{u}})(\theta_{x_i|\mathbf{u}}/(1 - \theta_{x|\mathbf{u}}))$ for all $x_i \neq x$, and $\Theta''_{X|\mathbf{u}}$ be an arbitrary distribution with $\theta''_{x|\mathbf{u}} = \theta'_{x|\mathbf{u}}$. We want to prove that $D(\Theta_{X|\mathbf{u}}, \Theta''_{X|\mathbf{u}}) \geq D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}})$.

- If $\theta_{x|\mathbf{u}} = 1$ and $\theta''_{x|\mathbf{u}} = \theta'_{x|\mathbf{u}} < 1$, we must have $D(\Theta_{X|\mathbf{u}}, \Theta''_{X|\mathbf{u}}) = \infty$.[6]

[5] Either $Pr'(w) = Pr(w) > 0$; or $Pr'(w) = Pr(w) = 0$, and thus $Pr'(w)/Pr(w) \overset{def}{=} 1$.

[6] Here, we will define the proportional scheme as making $\theta'_{x_i|\mathbf{u}} = \theta'_{x_j|\mathbf{u}}$ for all $x_i, x_j \neq x$, since we have $0/0 \overset{def}{=} 1$.

- If $\theta_{x|\mathbf{u}} = 0$ and $\theta''_{x|\mathbf{u}} = \theta'_{x|\mathbf{u}} > 0$, we must have both $D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}) = \infty$ and $D(\Theta_{X|\mathbf{u}}, \Theta''_{X|\mathbf{u}}) = \infty$.
- We now consider the case where $\theta_{x|\mathbf{u}} \neq 0$ and $\theta_{x|\mathbf{u}} \neq 1$. We first assume that $\theta''_{x|\mathbf{u}} = \theta'_{x|\mathbf{u}} > \theta_{x|\mathbf{u}}$. For $\Theta'_{X|\mathbf{u}}$, we have $\max_{x_i}(\theta'_{x_i|\mathbf{u}}/(\theta_{x_i|\mathbf{u}}) = \theta'_{x|\mathbf{u}}/\theta_{x|\mathbf{u}}$ and $\min_{x_i}(\theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}) = (1-\theta'_{x|\mathbf{u}})/(1-\theta_{x|\mathbf{u}})$. If $\Theta_{X|\mathbf{u}}$ and $\Theta''_{X|\mathbf{u}}$ do not have the same support, we have $D(\Theta_{X|\mathbf{u}}, \Theta''_{X|\mathbf{u}}) = \infty \geq D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}})$. If they have the same support, we have the following inequality:

$$\begin{aligned}
\frac{1-\theta''_{x|\mathbf{u}}}{1-\theta_{x|\mathbf{u}}} &= \frac{\sum_{x_i \neq x} \theta''_{x_i|\mathbf{u}}}{\sum_{x_i \neq x} \theta_{x_i|\mathbf{u}}} \\
&= \frac{\sum_{x_i \neq x} \theta_{x_i|\mathbf{u}}(\theta''_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}})}{\sum_{x_i \neq x} \theta_{x_i|\mathbf{u}}} \\
&\geq \frac{\sum_{x_i \neq x} \theta_{x_i|\mathbf{u}}(\min_i\,(\theta''_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}))}{\sum_{x_i \neq x} \theta_{x_i|\mathbf{u}}} \\
&= \min_i \frac{\theta''_{x_i|\mathbf{u}}}{\theta_{x_i|\mathbf{u}}}.
\end{aligned}$$

  Thus for $\Theta''_{X|\mathbf{u}}$, we have $\max_{x_i}(\theta''_{x|\mathbf{u}}/\theta_{x_i|\mathbf{u}}) \geq \theta''_{x|\mathbf{u}}/\theta_{x|\mathbf{u}} = \theta'_{x|\mathbf{u}}/\theta_{x|\mathbf{u}} = \max_{x_i}(\theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}})$ and $\min_{x_i}(\theta''_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}}) \leq (1-\theta''_{x|\mathbf{u}})/(1-\theta_{x|\mathbf{u}}) = (1-\theta'_{x|\mathbf{u}})/(1-\theta_{x|\mathbf{u}}) = \min_{x_i}(\theta'_{x_i|\mathbf{u}}/\theta_{x_i|\mathbf{u}})$. Therefore, we have:

$$\begin{aligned}
D(\Theta_{X|\mathbf{u}}, \Theta''_{X|\mathbf{u}}) &= \ln \max_{x_i} \frac{\theta''_{x_i|\mathbf{u}}}{\theta_{x_i|\mathbf{u}}} - \ln \min_{x_i} \frac{\theta''_{x_i|\mathbf{u}}}{\theta_{x_i|\mathbf{u}}} \\
&\geq \ln \max_{x_i} \frac{\theta'_{x_i|\mathbf{u}}}{\theta_{x_i|\mathbf{u}}} - \ln \min_{x_i} \frac{\theta'_{x_i|\mathbf{u}}}{\theta_{x_i|\mathbf{u}}} \\
&= D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}).
\end{aligned}$$

  We can prove the above result for the similar case of $\theta''_{x|\mathbf{u}} = \theta'_{x|\mathbf{u}} < \theta_{x|\mathbf{u}}$.

Therefore, the proportional scheme gives us the smallest distance, and this distance is equal to:

$$\begin{aligned}
D(\Theta_{X|\mathbf{u}}, \Theta'_{X|\mathbf{u}}) &= \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta_{x|\mathbf{u}}} - \ln \frac{1-\theta'_{x|\mathbf{u}}}{1-\theta_{x|\mathbf{u}}} \right| \\
&= \left| \ln \frac{\theta'_{x|\mathbf{u}}}{\theta'_{\overline{x}|\mathbf{u}}} - \ln \frac{\theta_{x|\mathbf{u}}}{\theta_{\overline{x}|\mathbf{u}}} \right|,
\end{aligned}$$

where $\theta_{\overline{x}|\mathbf{u}} = 1 - \theta_{x|\mathbf{u}}$ and $\theta'_{\overline{x}|\mathbf{u}} = 1 - \theta'_{x|\mathbf{u}}$.□

Proof of Theorem 5.1 If $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ do not have the same support, we can easily prove that the distributions Pr and Pr' also do not have the same support, and thus $D(Pr, Pr') = \ln \max_i (q_i/p_i) - \ln \min_i (q_i/p_i) = \infty$. Otherwise, from Jeffrey's rule, we have $Pr'(w) \overset{def}{=} Pr(w)(q_i/p_i)$, if $w \models \gamma_i$. Therefore, we have $Pr'(w)/Pr(w) = q_i/p_i$, if $w \models \gamma_i$. Consequently, we have:

$$\begin{aligned} D(Pr, Pr') &= \ln \max_w \frac{Pr'(w)}{Pr(w)} - \ln \min_w \frac{Pr'(w)}{Pr(w)} \\ &= \ln \max_i \frac{q_i}{p_i} - \ln \min_i \frac{q_i}{p_i}.\square \end{aligned}$$

Proof of Theorem 5.2 Let Pr' be the distribution generated by Jeffrey's rule, while Pr'' be any distribution that satisfies the constraint, $Pr''(\gamma_i) = Pr'(\gamma_i) = q_i$ for all $i = 1, \ldots, n$. We want to prove that $D(Pr, Pr'') \geq D(Pr, Pr')$.

If Pr and Pr'' do not have the same support, we have $D(Pr, Pr'') = \infty \geq D(Pr, Pr')$. If they have the same support, let $q_j/p_j = \max_i (q_i/p_i)$ and $q_k/p_k = \min_i (q_i/p_i)$. We now introduce $r_{max} = \max_w (Pr''(w)/Pr(w))$, and write the following inequality:

$$\begin{aligned} r_{max} p_j &= r_{max} Pr(\gamma_j) \\ &= \sum_{w \models \gamma_j} r_{max} Pr(w) \\ &\geq \sum_{w \models \gamma_j} \frac{Pr''(w)}{Pr(w)} Pr(w) \\ &= \sum_{w \models \gamma_j} Pr''(w) \\ &= Pr''(\gamma_j) \\ &= q_j. \end{aligned}$$

This gives us $r_{max} \geq q_j/p_j$. We can similarly introduce $r_{min} = \min_w (Pr''(w)/Pr(w))$, and write the following inequality:

$$\begin{aligned} r_{min} p_k &= r_{min} Pr(\gamma_k) \\ &= \sum_{w \models \gamma_k} r_{min} Pr(w) \\ &\leq \sum_{w \models \gamma_k} \frac{Pr''(w)}{Pr(w)} Pr(w) \\ &= \sum_{w \models \gamma_k} Pr''(w) \\ &= Pr''(\gamma_k) \end{aligned}$$

$$= q_k.$$

This gives us $r_{min} \leq q_k/p_k$. Therefore, the distance between Pr and Pr'' is:

$$\begin{aligned} D(Pr, Pr'') &= \ln \max_w \frac{Pr''(w)}{Pr(w)} - \ln \min_w \frac{Pr''(w)}{Pr(w)} \\ &= \ln r_{max} - \ln r_{min} \\ &\geq \ln \frac{q_j}{p_j} - \ln \frac{q_k}{p_k} \\ &= \ln \max_i \frac{q_i}{p_i} - \ln \min_i \frac{q_i}{p_i} \\ &= D(Pr, Pr'). \end{aligned}$$

Therefore, the distribution Pr' gives us the smallest distance.□

Proof of Theorem 5.3 From Pearl's method of virtual evidence, we have $Pr'(w) \stackrel{def}{=} Pr(w)(\lambda_i/(\sum_j p_j\lambda_j))$, if $w \models \gamma_i$. Therefore, we have $Pr'(w)/Pr(w) = \lambda_i/(\sum_j p_j\lambda_j)$, if $w \models \gamma_i$. Consequently, we have:

$$\begin{aligned} D(Pr, Pr') &= \ln \max_w \frac{Pr'(w)}{Pr(w)} - \ln \min_w \frac{Pr'(w)}{Pr(w)} \\ &= \ln \max_i \frac{\lambda_i}{\sum_j p_j\lambda_j} - \ln \min_i \frac{\lambda_i}{\sum_j p_j\lambda_j} \\ &= \ln \max_i \lambda_i - \ln \min_i \lambda_i.\square \end{aligned}$$