Challenge Paper: Marginal Probabilities for Instances and Classes (Poster Presentation SRL Workshop)

Oliver Schulte OSCHULTE@CS.SFU.CA
School of Computing Science, Simon Fraser University, Vancouver-Burnaby, Canada

Abstract

In classic AI research on combining logic and probability, Halpern introduced an inference principle for *marginal probabilities* of ground atoms: If the corresponding population frequency is known, the marginal probability should be equal to it. For instance, if the only thing we know about Tweety is that Tweety is a bird, then the probability that Tweety flies should be the frequency of flyers in the class of birds. We provide several arguments for why a statistical-relational inference system should satisfy Halpern's principle. If the principle is accepted, the technical challenge is then to construct inference models to meet this specification. The technical part of the paper gives examples of structures and parameterizations that do and do not satisfy the constraint on marginal probabilities, using Parametrized Bayes nets.

1. Introduction

Classic AI research established a fundamental distinction between two types of probabilities associated with a relational structure (Halpern, 1990; Bacchus, 1990). *Class-level probabilities*, also called type 1 probabilities, are assigned to the rates, statistics, or frequencies of events in a database. These concern classes of entities (e.g., students, courses, users) rather than specific entities. Examples of class-level queries would be "what is the percentage of birds that fly?" and "what is the percentage of male users that have given high ratings to an action movie?". Getoor, Taskar and Koller presented Statistical-Relational Models for representing class-level database statistics (Getoor et al., 2001). In Halpern's probabilistic logic, class-level probabilities are associated with formulas that contain 1st-order variables(e.g., $\mathit{Flies}(X) = \mathit{90}\%$). The type 1 probability of such a formula is the number of groundings that satisfy the formula, divided by the number of possible groundings. For example, assuming that the domain of 1st-order variable X is the class of birds, the expression $P(\mathit{Flies}(X)) = \mathit{90}\%$ denotes that 90% of birds fly.

Instance-level probabilities, also called type 2 probabilities, are assigned to specific, non-repeatable events or the properties of specific entities. Examples of instance-level queries would be "what is the probability that Tweety flies?" and "what is the probability that Jack highly rates Captain America?". Most statistical-relational learning has been concerned with type 2 instance probabilities; models like Probabilistic Relational Models (PRMs) (Getoor et al., 2007), Parametrized Bayes Nets (Poole, 2003), and Markov Logic Networks (MLNs) (Domingos & Richardson, 2007) define type 2 probabilities for ground instances using a grounding semantics.

Marginal Equivalence of Type 1 and Type 2 Probabilities. The SRL research into type 1 and type 2 probabilities has been largely independent of each other. In contrast, AI researchers directed sustained effort towards principled connections between the two types of probabilistic reasoning (Halpern, 1990; Bacchus, 1990; Bacchus et al., 1992; Halpern, 2006). A basic principle they discovered is what we shall refer to as *the marginal equivalence principle*: If we have no particular knowledge about an individual, other than that the individual belongs to a certain class, then the probabilities associated with that individual should be the class-level probabilities. To illustrate, if the only thing we know about Tweety is that Tweety is a bird, then the probability that Tweety flies should be the frequency of flyers in the bird population.

For illustration, consider the toy database instance of Figure 1. For the sake of the example, let us suppose that this data is representative of the entire population. Then the class-level probability that a course is difficult is 1/2, since half the known courses are difficult. Now suppose we run an inference system on the query $P(\mathit{diff}(\mathit{250})) = \mathit{hi}$? which asks for the marginal

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

probability that a new course 250 is difficult. The marginal equivalence principle requires that the answer to this query be 1/2.

Students

Name	intelligence
Anna	lo
Bob	hi

Courses

ID	difficulty
100	lo
200	hi
300	hi
400	lo

Registered

Name	ID
Anna	100
Anna	300
Bob	200
Bob	300

Figure 1. A small database instance. The class-level probability of a difficult course is the frequency in the course population, which is 1/2.

Formalization. In Halpern's probabilistic logic, the equivalence principle can be written as an *instantiation schema.* By this we mean that the principle amounts to instantiating a rule schema containing first-order variables with constants, as is done with universally quantified first-order variables in logic. For example, using first-order predicate notation, an example given by Halpern is

$$P(\mathit{Flies}(X)|\mathit{Bird}(X)) = 90\%$$
$$\rightarrow P(\mathit{Flies}(\mathit{Tweety})|\mathit{Bird}(\mathit{Tweety})) = 90\%$$

where X is instantiated with the constant *Tweety*. In this formula X is interpreted as denoting a randomly selected member of the domain, and $P(\mathit{Flies}(X)|\mathit{Bird}(X))$ expresses the proportion of flyers in the bird population. (For a full description of the random selection semantics and syntax, please see (Halpern, 1990; Bacchus, 1990).) In many SRL formalisms, first-order variables and constants are typed, meaning that background knowledge assigns them to a known class; we assume typing in the remainder of the paper. If *Tweety* is known to belong to the class of birds and X is known to range over the class of birds, a simpler form of the instantiation schema is

$$P(\mathit{Flies}(X)) = 90\% \rightarrow P(\mathit{Flies}(\mathit{Tweety})) = 90\%.$$

This is an instance of the *marginal equivalence principle*: Marginal probabilities for ground atoms should be the same as marginal probabilities at the class level.

Challenge. The challenge to the SRL community that we see arising from Halpern's equivalence principle is this: If we accept the principle as a desirable feature of an SRL inference system, how can we design/learn systems so that they satisfy it? We believe that this challenge is a constructive one, in that meeting it would lead to progress in statistical-relational learning. First, with respect to marginal probabilities, a system that satisfies the equivalence principle, will achieve high accuracy on marginal queries, and likely can compute the answers to marginal queries in closed form. (More details below.) Second, the principle adds a well-motivated constraint on the features and parameters of an SRL system. Such constraints on learning and/or inference can be leveraged to obtain more efficient and effective algorithms than is possible in an unconstrained general design space.

The paper proceeds as follows. First, we provide several motivations for the equivalence principle. Then we give examples of Parametrized Bayes net models that satisfy the principle by choosing the right structure, combining rule, or parameters.

2. Why Instance-level Marginals should equal Class-level Marginals

We provide different types of motivation as follows. (1) Intuitive plausibility and connections to other established principles/approaches. (2) If an inference algorithm is scored on its likelihood performance on marginals, the score is maximized if the algorithm assigns the class-level marginals as answers to marginal queries about ground atoms. (3) Examples of prominent probabilistic models that satisfy the equivalence, such as propositional models and latent factor models like those that arise from matrix factorization.

2.1. Intuitive Plausibility

By definition, a marginal probability is assessed without conditioning on any further information about the indvidual entity involved, except for its type/population. Therefore the individual is equivalent to all other members of the population and inference about its properties should reflect only properties of the population, such as the statistical frequency of a trait. A practical setting where this reasoning is applied is the cold-start problem for recommendation system: recommending a movie to a new user who has just entered the system. In this case the system cannot make use of specific information, like other movies the user has watched, and has to fall back on general statistics about the popularity of a movie and the demographics of the user (e.g., men like action movies).

Halpern proves that his instantiation principle is equivalent to Miller's principle (Halpern, 1990, Th.4.5).[1] Miller's principle applies to probability types of different orders. For marginals of ground atoms and frequencies, it states that conditional on

[1] This result depends on the assumption of rigid designators; for details please see Halpern's paper.

the value of a class-level frequency, the marginal probability of a ground atom equals that value. Miller's principle has been long accepted by probability theorists.

2.2. Score Maximization.

A commonly used score for a relational inference system, such as Markov Logic Networks, is based on the pseudo-likelihood measure: for each ground atom, apply inference to calculate its log-likelihood *given complete information about all other ground atoms.* Then define the conditional log-likelihood score overall to be the sum of the conditional log-likelihoods over all ground atoms in the test set. This score considers inferences about a ground atom given information about *all* others; at the other extreme are marginal inferences about a ground atom without *any* information about other atoms. If we apply the same method for scoring marginal inferences independently, we can define the *marginal log-likelihood score* to be the sum, over all ground atoms, of the marginal log-likelihood of the ground atom. Using standard likelihood maximization arguments, it is easy to see that *the marginal log-likelihood score is maximized if marginal probabilities equal population frequencies.* This result assumes exchangeability, meaning that predicted probabilities for individuals of the same type are the same.

Some formalization may clarify these concepts. Let P denote the probabilistic inferences provided by an SRL system that is to be scored. Let $X = x$ denote the event that ground node X has value x, and let $\overline{\mathbf{X}} = \overline{\mathbf{x}}$ denote that all ground nodes other than X are assigned the values specified by a list $\mathbf{x}$. For a given database $\mathcal{D}$ that specifies values for all ground nodes, let $x_{\mathcal{D}}$ be the value of X specified by $\mathcal{D}$, and similarly for $\overline{\mathbf{X}}$. Then the conditional log-likelihood score (CLL) of inferences P, as a function of a data $\mathcal{D}$, is given by

$$CLL(\mathcal{D}) \equiv \sum_{X} lnP(X = x_{\mathcal{D}} | \overline{\mathbf{X}} = \overline{\mathbf{x}}_{\mathcal{D}}).$$

The marginal log-likelihood score (MLL) is given by

$$MLL(\mathcal{D}) \equiv \sum_{X} lnP(X = x_{\mathcal{D}}).$$

2.3. Models that satisfy the Marginal Equivalence Principle

Our last motivation for the marginal-frequency equivalence principle is that prominent models satisfy it. This shows that it is possible to design accurate inference systems to do so. If standard approaches satisfy the principle, it is in a sense normal to have marginal instance-level probabilities track class-level probabilities.

2.3.1. Propositional Models

In a propositional model, direct inference from population frequencies to specific individuals is so natural it is usually not made explicit. For instance, suppose that we build a propositional model for properties of courses only. In terms of the database of Figure 1, this means that the only type of data table is the *Courses* table (no students or registration information). Suppose that the population statistics are such that half the courses are difficult. If we now query the model to predict the difficulty of course 250, an inference system that was learned from representative population samples from a single data table, would predict that the probability of course 250 being difficult was 1/2.

What is interesting about comparing propositional and relational models is that, while a relational system can *potentially* utilize information about linked entities, marginal probabilities do not *actually* utilize this information. If one accepts the marginal equivalence principle for propositional data, it seems to require an explanation why the mere fact that relational information could be relevant, but is not actually used, should change the marginal probabilities. Figure 2 aims to visualize this point. On the left is a trivial model of the difficulty of courses where the probability of a specific course being difficult would be based on class frequencies, or estimates thereof from a single data table, such as the *Courses* table shown in Figure 1. On the right is a more complex model where the intelligence of students who are taking the course is relevant. Suppose we ask the marginal query $P(\mathit{diff}(250)) =?$ in both models. This query does not involve students registered in course 250 at all, so one may expect both models to give the same answer. Thus if marginal inferences in one model track the class-level probabilities, so should inferences in the other.

2.3.2. Matrix Factorization Models

Latent factor models are among the most predictively powerful relational models (Chiang & Poole, 2012). Figure 3 gives a latent factor model for the student example. A latent factor $U(c)$ is assigned to each course c, and another $U(s)$ to each student s. Figure 4 shows a database with imputed factors. There is a large literature on how to learn the latent factors, but most methods are based on maximizing the

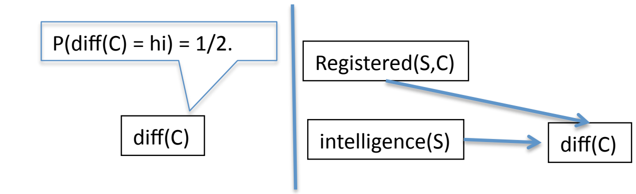

Figure 2. Left: A one-node propositional Bayes net, which satisfies Halpern's instantiation principe. Right: A Bayes net expanded with relational information. Should the relational information be relevant for predicting the *marginal* probability of the difficulty of a course? For instance, should $P(\mathit{diff}(C) = hi)$ be the same in both models?

likelihood of the observed data. In a Bayes net model, this means that the conditional probability parameters will be (smoothed) versions of the empirical frequencies with the imputed factors. Now the imputed latent factors do not affect the marginal frequencies over observed variables. Thus inference regarding marginals of observed variables will be based on the population frequency over observed variables. For instance, the prior probability parameter $\mathit{diff}(C) = hi$ will be 1/2 given the toy database shown in Figure 4 (perhaps with some smoothing).

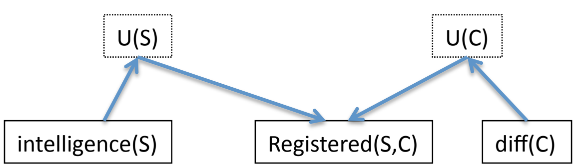

Figure 3. A graphical model for factorizing the Registration relationship, based on two types of latent factors, one for students and one for courses. In such models, instance-level marginal probabilities typically equal class-level marginal probabilities.

Students

Name	intelligence	U
Anna	lo	30
Bob	hi	20

Courses

ID	difficulty	U
100	lo	10
200	hi	20
300	hi	15
400	lo	12

Registered

Name	ID
Anna	100
Anna	300
Bob	200
Bob	300

Figure 4. A small database instance for the model of Figure 3 with imputed latent factors. The population distribution over *observed* attributes is the same as in the database of Figure 1 that has no latent factors.

We believe that the considerations we have given motivate the marginal equivalence principle sufficiently to consider how it may be realized in a statistical-relational model. This is the technical part of the paper.

3. Bayes Net Examples

We consider Parametrized Bayes net (PBN) models without latent factors. We briefly review the relevant definitions. Parametrized Bayes nets are a basic graphical model for relational data due to Poole (Poole, 2003). The syntax of PBNs is as follows. A **functor** is a function symbol or a predicate symbol. Each functor has a set of values (constants) called the **range** of the functor. To conform to statistical terminology, Poole refers to 1st-order variables as population variables. A **population variable** X is associated with a population, a set of individuals, corresponding to a type, domain, or class in logic. A **Parametrized random variable** is of the form $f(X_1, \ldots, X_k)$. A **Parametrized Bayes Net** is a Bayes net whose nodes are Parametrized random variables. Poole specifies a grounding semantics where the 1st-order Bayes net is instantiated with all possible groundings of its 1st-order variables to obtain a directed graph whose nodes are functors with constants. Figure 5 shows an example with the ground nodes obtained by instantiating $S := \mathit{Anna}, C := \mathit{250}$.

In what follows *we assume that the marginal class-level probabilities can be computed from the Bayes net model.* This will generally be the case if parameter estimates are based on event counts, and enough data is available. To illustrate, the conditional probabilities of Figure 5 were derived from the database in Figure 1. if we apply standard Bayes net calculations to the top Bayes net, the marginal probabilities are uniformly 1/2, which is the frequency in the database. Thus $1/2 = P(\mathit{Registered}(S, C) = T) = P(\mathit{diff}(C) = hi) = P(\mathit{intelligence}(S) = hi)$. Assuming that the PBN model represents the class-level probabilities means that whether predicted marginal probabilities match class statistics is only a question of inference, not of learning.

3.1. Unique Parents

In the structure of Figure 5, each ground node has a unique set of parents. A simple observation is that *if each ground node has a unique set of parents, then the grounding semantics satisfies the marginal equivalence principle.* The general argument goes like this. Each ground source node with indegree 0 has a marginal distribution that matches the class-level distribution. Different source nodes are inde-

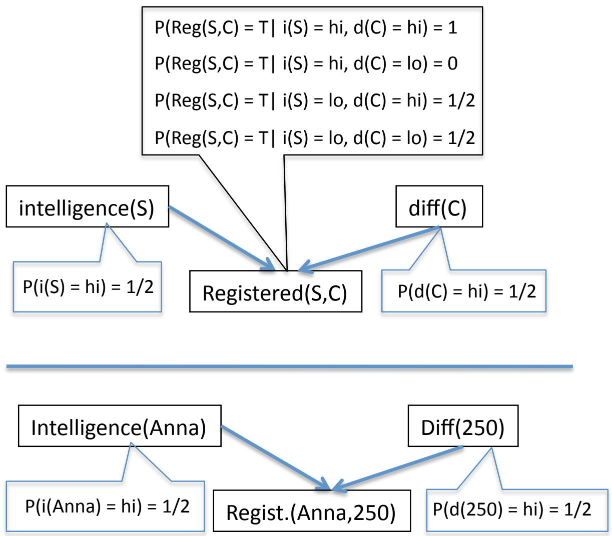

Figure 5. To illustrate the grounding semantics of Parametrized Bayes Nets. Top: A Parametrized Bayes Net with conditional probability parameters specified following the database frequencies in Figure 1. Bottom: A fragment of the ground graph for the child node *Registered*(*Anna*, *250*). Conditional probability parameters are inherited as shown (we omit the parameters for *Registered*(*Anna*, *250*) for brevity).

pendent, so for non-source nodes, the joint distribution over the parent-instantiation matches the class-level model as well. So the marginal distribution of a non-source node matches the class-level distribution as well. To illustrate, Table 1 goes through the computational steps for computing the marginal probability $P(Registered(Anna, 250) = T)$.

Table 1. If each ground node has a unique parent set, the grounding semantics satisfies the marginal equivalence principle. For instance $P(Registered(Anna, 250) = T) = 1/2 = P(Registered(S, C) = T)$. The table shows the computation of the marginal probability $P(Registered(Anna, 250) = T)$.

int(Anna)	*diff(250)*	$P(Reg(Anna, 250) = T \mid parents)$	P(parents)	product
hi	hi	1	1/4	1/4
hi	lo	0	1/4	0
lo	hi	1/2	1/4	1/8
lo	lo	1/2	1/4	1/8
			Sum	1/2

Let us review the assumptions in the argument. (1) The structure is as shown in Figure 5. (2) The BN parameters agree with the true population conditional probabilities.

3.2. Average Combining Rule

A ground node may have several instantiations of its class-level parents as illustrated in Figure 6. In this case Poole suggests using a combining rule for combining the conditional probabilities defined by each parent instance, as in Bayes Logic Progams (Kersting & de Raedt, 2007). Prominent examples of combining rules include average and noisy-or. We begin by considering the average combining rule.

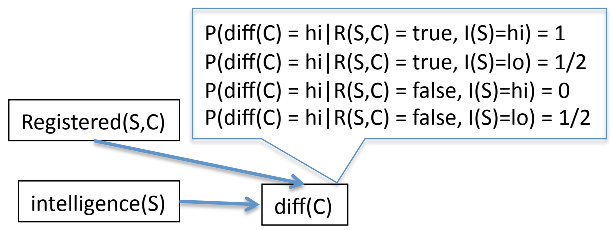

Figure 6. In this Bayes net, a grounding the node *diff*(C) has multiple parent instantiations, one for each student. Conditional probability parameters are set to match the database frequencies in Figure 1. The priors over the nodes *Registered*(S, C) and *intelligence*(S) are uniform.

For the Bayes net structure shown in Figure 6, the average combining rule satisfies the marginal equivalence principle. We provide a general argument to show that this does not depend on our specific database example. Consider the ground graph of Figure 5. Let us write π_1 for an assignment of values to the first parent set with $S := Anna$ and π_2 for the second parent set with $S := Bob$. Consider predicting the marginal probability for a new course with number 250, and write

$$\theta_1(\pi_1) \equiv P(diff(250) = hi|\pi_1)$$

for the conditional probability of high course difficulty given information about Anna, and similarly

$$\theta_2(\pi_2) \equiv P(diff(250) = hi|\pi_2)$$

for the conditional probability of high course difficulty given information about Bob. Notice that the distribution of θ_1 and θ_2 is identical. Then the average combining rule sets the conditional probability of the course difficulty given information about both Bob and Anna to be the average of the two conditional probabilities θ_1 and θ_2. So the marginal probability that course 250 is highly difficult is the expected value of this average:

$$P(diff(250) = hi) = E[(\theta_1(\pi_1) + \theta_2(\pi_2))/2].$$

The expectation of the sum of two random variables is the sum of their expectations, regardless of whether the two random variables are independent. Therefore we have

$$P(\mathit{diff}(250) = hi) = E[(\theta_1(\pi_1)]$$

where we could also have used θ_2 because both θ_1, θ_2 are identically distributed. This results means that, *with the average combining rule*, the instance-level marginal can be computed by just considering a single parent. Table 2 shows the computation. Since we assumed that the Bayes net model agrees with the class-level distribution, the marginal instance-level probability equals the marginal class-level probability.

Let us review the assumptions in the argument. (1) The structure is as shown in Figure 2. (2) The BN parameters agree with the true population conditional probabilities. (3) The average combining rule is used.

Table 2. Computing the instance-level marginal probability $P(\mathit{diff}(250) = hi)$ using the average combining rule and the Bayes net of Figure 2. In this structure, the average combining rule ensures the equality of the instance-level probability and the class-level probability.

Reg(Anna,250)	Int(Anna)	$P(\mathit{diff}(250) = hi \mid parents)$	P(parents)	product
T	hi	1	1/4	1/4
T	lo	1/2	1/4	1/8
F	hi	0	1/4	0
F	lo	1/2	1/4	1/8
			Sum	1/2

3.3. Disjunctive Combining Rule

A deterministic disjunctive model may specify that a course is highly difficult if and only if it is taken by at least one highly intelligent student. Depending on which students take the course, we specify the conditional probabilities for course difficulty as follows.

1. If no student takes the course, we use a default prior set to 1/4: $P(\mathit{diff}(250) = hi|\#Reg = 0) = 1/4$.
2. If exactly one student takes the course, the course is difficult iff that student is highly intelligent, so $P(\mathit{diff}(250) = hi|\#Reg = 1) = 1/2$.
3. If both Bob and Anna take course 250, the course is difficult iff at least one of the is highly intelligent, so

$$\begin{aligned} P(\mathit{diff}(250) = hi|\#Reg = 2) = \\ = 1 - P(intelligence(Bob) = lo) \cdot P(intelligence(Anna) = lo) \\ = 3/4 \end{aligned}$$

Each number of student participants is equally likely, so the overall marginal probability of course 250 being difficulty is the average of these numbers, which comes out to the class-level probability 1/2. The computation is shown in Table 3. The default value 1/4 for no students registered has to be set exactly to balance out the 3/4 probability from the disjunctive rule when there is more than one student, which is higher than the class-level probability.

Table 3. Computing the instance-level marginal probability $P(\mathit{diff}(250) = hi)$ using a deterministic or-gate (disjunctive rule) and the Bayes net of Figure 2. With the right parameter settings, the instance-level probability equals the class-level probability.

$\#Reg = r$	$P(\mathit{diff}(250) = hi|r)$	$P(r)$	product
0	1/4	1/4	1/16
1	1/2	1/2	1/4
2	3/4	1/4	3/16
		Sum	1/2

Let us review the assumptions in the argument. (1) The structure is as shown in Figure 6. (2) The source/parent nodes have prior probabilities set to match population probabilities. (3) A deterministic-or combining rule is used. (4) The parameters of the combining rule are set exactly so that the class-instance marginal equivalence holds.

3.4. Discussion.

The examples illustrate how the equivalence principle motivates some choices of structures, combining rules (the average rule), and parameter settings (for the disjunction rule). Open questions for future research include: does the average combining rule ensure the equality of instance and class level marginals for every structure, regardless of the exact parameters? If not, are there constraints on structures that work with the average rule? Are there always parameter settings for the noisy-or rule that can ensure the equality, and for what type of structures?

The equivalence of class and instance marginals is a fruitful question for other types of SRL models as well.

1. Probabilistic Relational Models use aggregate functions rather than combining rules. Our discussion of latent factor models (Section 2.3.2) suggests that introducing new nodes that represent aggregate functions (Kersting & de Raedt, 2007) is a promising option for ensuring the equality of class-level and instance-level marginals.

2. Do log-linear models, like Markov Logic Networks, satisfy the marginal equivalence principle?

3. For dependency networks with Gibbs sampling, good empirical performance on marginals has been noted (Neville & Jensen, 2007). Do sampling methods satisfy the equivalence principle?

To provide concrete examples, we made the simplifying assumption that the model parameters represent class-level probabilities. In practice, learning is based on a limited sample, and some learning methods may be more suited to class-instance marginal equivalence than others. A mathematical analysis can aim for a statistical consistency result, showing that as more and more data are observed, inferred marginal instance-level probabilities agree with marginal class-level probabilities in the population. A concrete research question would be whether maximum likelihood methods for learning combining rule parameters (Natarajan et al., 2008) satisfy this marginal consistency property.

4. Conclusion

In classic AI research, Halpern proposed the principle that instance-level marginal probabilities should match class-level marginal probabilities, where the latter are based on population/class frequencies. For instance, if the only thing we know about Tweety is that Tweety is a bird, then the probability that Tweety flies should be the frequency of flyers in the class of birds. We proposed several motivations for Halpern's principle, including intuitive plausibility and predictive performance on marginal queries. Our challenge to the SRL community is to design inference models that ensure the equivalence of marginal and class-level frequencies. We presented several Parametrized Bayes net examples to show that the equivalence principle leads to nontrivial constraints on model structures and parameters. Constraints that have a strong theoretical motivation can be exploited to achieve faster and more accurate learning and inference.

Acknowledgements

This research was supported by a Discovery Grant from the National Sciences and Engineering Research Council of Canada. Anonymous reviewers for the SRL Workshop provided helpful comments, especially concerning diagrams and definitions. I am indebted to Hector Levesque for extensive discussions of the marginal equivalence principle.

References

Bacchus, Fahiem. *Representing and reasoning with probabilistic knowledge: a logical approach to probabilities*. MIT Press, Cambridge, MA, USA, 1990.

Bacchus, Fahiem, Grove, Adam J., Koller, Daphne, and Halpern, Joseph Y. From statistics to beliefs. In *AAAI*, pp. 602–608, 1992.

Chiang, Michael and Poole, David. Reference classes and relational learning. *Int. J. Approx. Reasoning*, 53(3):326–346, 2012.

Domingos, Pedro and Richardson, Matthew. Markov logic: A unifying framework for statistical relational learning. In *Introduction to Statistical Relational Learning* Getoor & Tasker (2007).

Getoor, Lise and Tasker, Ben. *Introduction to statistical relational learning*. MIT Press, 2007.

Getoor, Lise, Taskar, Benjamin, and Koller, Daphne. Selectivity estimation using probabilistic models. *ACM SIGMOD Record*, 30(2):461–472, 2001.

Getoor, Lise, Friedman, Nir, Koller, Daphne, Pfeffer, Avi, and Taskar, Benjamin. Probabilistic relational models. In *Introduction to Statistical Relational Learning* Getoor & Tasker (2007), chapter 5, pp. 129–173.

Halpern, Joseph Y. An analysis of first-order logics of probability. *Artificial Intelligence*, 46(3):311–350, 1990.

Halpern, Joseph Y. From statistical knowledge bases to degrees of belief: an overview. In *PODS*, pp. 110–113. ACM, 2006.

Kersting, Kristian and de Raedt, Luc. Bayesian logic programming: Theory and tool. In *Introduction to Statistical Relational Learning* Getoor & Tasker (2007), chapter 10, pp. 291–318.

Natarajan, Sriraam, Tadepalli, Prasad, Dietterich, Thomas G., and Fern, Alan. Learning first-order probabilistic models with combining rules. *Annals of Mathematics and Artifical Intelligence*, 54(1-3): 223–256, 2008.

Neville, Jennifer and Jensen, David. Relational dependency networks. In *An Introduction to Statistical Relational Learning* Getoor & Tasker (2007), chapter 8.

Poole, David. First-order probabilistic inference. In *IJCAI*, pp. 985–991, 2003.