Abstract of "Selecting Approximately-optimal Actions
in Complex Structured Domains" by Luis E. Ortiz, Ph.D., Brown University, May 2002.

We study the problem of action selection in structured domains. In general, the provided structural decomposition of the problem is not sufficient to allow tractable computation of the exact solution. Hence, we concentrate on obtaining near-optimal solutions with some guaranteed qualities.

In this work, the main intuition we exploit is that the problem of action selection is primarily a comparison rather than an estimation task. From this point of view, we consider sampling methods for action selection. We propose methods to reduce the number of samples required to obtain near-optimal actions. We present results on the number of samples needed to obtain highly probable, near-optimal actions. In addition, we present a comparison-based sampling method and a heuristic stopping rule that can potentially reduce the total number of samples.

Although estimation is not a primary task, better estimators lead to better action selection. We present update rules to adaptively improve the sampling distribution and hence the resulting estimators.

We present preliminary validation results on both made-up and real models. The results show the potential of the methods for action selection and improving estimation. Throughout the document, we present remaining open questions and suggest future work.

Selecting Approximately-optimal Actions in Complex Structured Domains

by

Luis E. Ortiz

B. S., Univerisity of Minnesota, 1995

Sc. M., Brown University, 1998

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island

May 2002

This dissertation by Luis E. Ortiz is accepted in its present form by
the Department of Computer Science as satisfying the dissertation requirement
for the degree of Doctor of Philosophy.

Date _____ _____

Leslie Pack Kaelbling, Director

Recommended to the Graduate Council

Date _____ _____

Thomas L. Dean, Reader

Date _____ _____

Stuart Geman, Reader
Division of Applied Mathematics

Date _____ _____

Peter Müller, Reader
Department of Biostatistics
University of Texas M. D. Anderson Cancer Center

Approved by the Graduate Council

Date _____ _____

Dean of the Graduate School and Research

Vita

Luis E. Ortiz was born in Ponce, Puerto Rico, on September 28, 1971. He spent most of his early years in Juana Díaz, Puerto Rico, where his family lives. His full name is Luis Enrique Ortiz Franceschi. He obtained a B.S. degree in Computer Science from the Institute of Technology at the University of Minnesota, and an Sc.M. in Computer Science from Brown University. He received an NSF Integrative Graduate Education and Research Training (IGERT) Fellowship (1999-2001), an NSF Minority Graduate Fellowship (1996-1999), a National Physical Science Consortium (NPSC) Ph.D. Fellowship (1995-1996), a NACME-IBM corporate Scholarship (1993-1995), a National Hispanic Scholarship (1994-1995), a Minnesota Hispanic Educational Scholarship (1994-1995), and a NACME Scholarship (1990-1991). His professional experiences include: (1) research assistant/independent researcher (1995-2001) and (2) lecturer assistant (2001) in the Department of Computer Science at Brown University under the advise of Thomas L. Dean and Leslie Pack Kaelbling, (3) mentor for undergraduate student in the Department of Cognitive Science at Brown University (work done as part of NSF IGERT Program; 2000), (4) member of research project in the Department of Computer Science at Brown University under the advise of Eli Upfal (1998-1999), (5) undergraduate teaching assistant at the Project Technology Power of the University of Minnesota—Twin Cities Campus (1994), and (6) programmer (Summer Pre-Professional Program) at IBM, Rochester, MN (1991-1993).

Acknowledgments

I realize that this thesis would have not been possible without the help and support of many people starting, from my early developments as a student in the public school system in Juana Díaz, Puerto Rico, to my graduate school years at Brown.

First of all, I would like to thank my advisor Leslie Kaelbling for her patience and support. Thanks for sharing your knowledge and wisdom on everything from technical subjects to writing and presenting material. For showing me how to say what I wanted to say. Also, thanks for keeping a bigger picture when I did not, for giving me the time to think and pursue my own interests, and for the timely encouragements during my early years as a graduate student.

I would like to thank my co-advisor Thomas Dean for all his support. Professor Dean has provided extremely valuable direction during my graduate student career. I will miss very much his insightfulness and critical thinking, and the many intellectual discussions. His methodological approach to research has had a great impact on my work as a researcher.

I would like to thank Stuart Geman for the innumerable discussions on mathematical, statistical, and probabilistic questions. Professor Geman has been a great teacher to me and I cannot thank him enough for sharing his knowledge with me and being a great source of inspiration. I will really miss our discussions.

I would like to thank Peter Müller for accepting to be a reader in my dissertation committee. Despite the limited interaction, his comments of my proposed work were very useful then, and I believe will continue to have an effect as I explore future directions of my work in this area.

Thanks also go to Eugene Charniak, who as my first-year advisor, allowed me to follow and develop my own interests.

I would also like to thank Peter Müller for suggesting the IctNeo ID and to Concha Bielza and Manuel Gómez Olmedo for providing it and allowing me to use it in my thesis. Also, I would like to thank Manuel Gómez Olmedo for all the time he spent helping me

with the model and patiently answering my questions.

I am very grateful I had the opportunity to be a graduate student at Brown. My graduate student experience was very rewarding. I will miss the academic, intellectual environment, and the interactions between groups in different areas and departments.

Also, my experience has been made even more rewarding, both from a personal and professional standpoint, by the interaction with other graduate students. Among the many of them, I would like to specially acknowledge Kee-Eung Kim, Hagit Shatkay, and the members of the AI group at Brown, for their friendship, which I will always cherish, and all the help they provided me during all my years at Brown.

Milos Hauskrecht, as a post-doctorate at Brown (1999-2000), was very influential in my development as a graduate student and an invaluable resource. I am very grateful to him for his patience and the long periods of time we spent in innumerable technical discussions.

Also, my officemates during all my years at Brown, Vasiliki Chatzi (Vaso) and Michael Benjamin (Mike) have been a source of support and inspiration in all aspects of my graduate experience. I cannot express in words how much I appreciate that they were *always* there for me, in both good and bad times, and their understanding. I will forever cherish their friendship.

To Dimitris Michailidis, whose knowledge and help is unlimited, as many people before me have already found out. I am really glad I met him.

To Stella Kakavouli, who has been a source of support since I met her and an incredible friend.

To the "Spanish-connection:" Luis J. Vega, Daniel Acevedo Feliz and their respective wives, Angela and María. They have been a source of support to both my wife and I during my last years at Brown. My wife and I cherish their friendship. Also, to Dan Keefe, who along with Luis and Daniel allowed me to play (i.e. "make some noise") with them. I will most definitely miss the jam sessions. To Luis and the Gapasutras, please keep writing and playing such interesting music and I wish you all the luck and success you deserve. Keep on jamming and having fun!

To my wife, Connie Arline Acosta López, for everything: her love, patience, sacrifice, understanding and support without bound.

To my mom, Nilsa Franceschi González, and my dad, Luis Antonio Ortiz Santiago, for supporting my decisions through all these years even when they were troubled by them. To my sisters, Nitza (Nilsita) and Ludian, for being a source of inspiration and support.

To my long-time hometown friends, Javier Antonio Montero Santiago, Antonio Radamés Alvarez Rodríguez, Angel Ramón Alvarez Rodríguez, Arturo Carlos Martínez, and Orlando

Enrique Zayas, and their respective families.

To my teachers, specially math teachers, from every educational period of my academic career, from elementary to graduate school.

To all the people who crossed path with me in this long and exciting journey!

Credits

Portions of the work presented in Chapter 2 and 3 have appeared in Ortiz and Kaelbling [2000b] and Ortiz and Kaelbling [2000a], respectively. Also, this document is an extension of my thesis proposal [Ortiz, 2000].

The dynamic weighting scheme and the $1/\sigma^2$ recommendation in Section 3.2.1 and the ϵ-boundary in Section 3.4.2 were independently developed by Jian Cheng and Marek Druzdzel. Both heuristics are reported in a manuscript that the author saw while he was working on the paper Ortiz and Kaelbling [2000a].

I would like to thank my advisor Leslie Kaelbling; Constantine Gatsonis for suggesting the MCB literature; Eli Upfal, Milos Hauskrecht, Kee-Eung Kim, Thomas Dean, Thomas Hofmann and Gopal Pandurangan for many useful discussions and suggestions. Also, the implementations of the methods used for the experimental results in this thesis use some of the functionality of the *Bayes Net Toolbox for Matlab* [Murphy, 1999], for which I thank Kevin Murphy.

During my years as a graduate student at Brown leading up to the work in this thesis, I was supported in part by a National Physical Science Consortium (NPSC) Ph.D. Fellowship, an NSF Graduate Fellowship and by NSF IGERT award SBR 9870676.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem studied in this thesis is the computational equivalent of a decision-maker or *agent* trying to make reasonable decisions or behave well in a large, complex, uncertain, but structured environment. The agent might be a robot, a doctor, a farmer, or a computer-repair person. The world or environment could be the floor of a building, a patient, a crop plantation, or a computer system.

Consider the case of a robot as an agent. The robot has available a *limited number of actions* it can take; for instance, the robot can move in different directions. It has *objectives* that affect its behavior; the robot needs to make deliveries from one office to another. The robot has some predefined notions of *optimality* and uses them to evaluate its behavior; it needs to accomplish its task efficiently, where the notion of efficiency is defined through characteristics of its behavior such as the time to completion of the task and its safety.

We can think of the coupling of the agent and its environment as a (sometimes dynamical) *system*. We can characterize the system through its *state* or the condition it is in. The robot can be in one corner of the floor facing north with an envelope to deliver. The door of the office it has to deliver to can be closed.

The sources of *uncertainty* are multiple. The robot's actions are *nondeterministic*; for example, depending on the condition of the floor, the action of moving forward is not always successful. The robot typically has *limited available information* about the true state of the environment and its own true state at the time of making its decision; detecting a corner is not enough to distinguish which corner it might be in. Also, the observations can be *noisy*; the robot's sensors sometimes fail to detect a closed door. Also, the compass reading is not certain; the robot might "observe" that it is facing north while it is really facing north-east.

The (global) state of a system is formed by the relevant (local) variables or *features* of the

system. The number of states grows exponentially with the number of features. Hence, large environments require a large number of features to be properly described. The complexity of the system results from the non-trivial (global) interactions among the features forming the state and the behavior of the system. Although those global interactions are complex, the systems we consider can be described through local interactions among their features. Exploiting the local decomposition allows for a more compact representation of the decision problem. However, it is not always possible to further exploit this local decomposition to provide efficient computational mechanisms for action selection.

In the following, we will present the decision-theoretic approach (von Neumann and Morgenstern [1944]; see also Pratt et al. [1995]), which is becoming popular in AI to deal with problems of decision-making under uncertainty [Boutilier et al., 1999]. Then, we will discuss how we can obtain exact solutions to the problem of action selection by exploiting the local decomposition. After describing why it is not always possible to provide efficient computational mechanisms to solve the action selection problem exactly, we state the motivating ideas behind this work. A more formal presentation of some of the models and techniques that we use in this work follows. Then we present a statement of our objectives in this work. Finally, we present an overview of the chapters to come, including a summary of our results and conclusions.

1.1 Decision-theoretic models (in general)

In this work, we concentrate on the problem of action selection in decision-theoretic models. Decision-theoretic models have become the standard framework for modeling problems of decision-making under uncertainty. Under this framework, we represent our uncertainty about the state of the system by means of a probability distribution over the states. We define how useful certain states and actions are by defining a utility function mapping states and actions into a utility value measuring the degree of usefulness. In this context, our goal is to select the action with the largest *expected utility*.

As mentioned earlier, the number of states grows exponentially with the number of features. Hence, the explicit representation of the probability distribution and the utility function becomes infeasible quickly. One way to get around this problem is to exploit any structural characteristics of the system to simplify the description of the decision problem. In particular, it is sometimes possible to express the probability distribution over the joint set of features using smaller distributions over combinations of the individual features. This is possible by exploiting conditional independencies associated with the features of

the particular problem. Also, we can represent the global utility over states and actions through smaller, more local utility descriptions. Both of these concepts will become clearer when we present the models used in this work more formally.

1.2 On exact solutions

Researchers have developed many computational methods to solve the problem of action selection exactly by exploiting the available compact representations of the decision problem. The basis for most of the methods proposed is to exploit the problem structure representation through *dynamic programming* [Bellman, 1957, Tatman and Shachter, 1990, Aji and McEliece, 2000]. Using dynamic programming, the methods decompose the problem into smaller problems whose solution provide solutions to the global problem of action selection. However, the local descriptions might not be sufficient for dynamic programming to allow efficient procedures in general.

1.3 Motivating ideas

Because it is highly intractable, in general, to compute the optimal solution, we are interested in solutions that are *good enough* as opposed to optimal. That is, we are interested in obtaining approximately optimal solutions within a certain confidence level. The approach we take is to exploit several intuitions about the decision-theoretic problem and model available, which we will define shortly.

Although the local descriptions and the decompositions of the decision problem may not allow efficient computation of exact solutions of optimal strategies, they do allow efficient generation of instances of the system state according to its probability distribution. The value of a particular action is related to how useful it is expected to be. In evaluating this value, the most important components are those states that are most likely and/or those that are very useful or very bad. Furthermore, since action selection is a process of comparing the values of different action choices, the exact value of each action choice by itself is not as important as its value with respect to that of the optimal action. By concentrating on comparing choices, the process of generating possible and useful states and actions should help us provide approximately optimal actions in those cases where the exact computation of the optimal action is intractable. Most of these ideas are studied in Chapter 2.

Several ideas are investigated to further improve on these intuitions. Among them is

to use information about the utility values to improve the generation of instances, an idea which we study further in the Chapter 3.

1.4 Notation

In this section, we establish notation used throughout this document. We denote one-dimensional random variables by capital letters and denote multi-dimensional random variables by bold capital letters. For instance, we denote a multi-dimensional random variable by $\boldsymbol{X}$ and denote all its components by $(X_1, \ldots, X_n)$ where X_i is the i^{th} one-dimensional random variable. We use small letters to denote assignments to random variables. For instance, $\boldsymbol{X} = \boldsymbol{x}$ means that for each component X_i of $\boldsymbol{X}$, $X_i = x_i$. We denote the *state space* or set of possible values that X_i can take by Ω_{X_i} and the state space set of $\boldsymbol{X}$ by $\Omega_{\boldsymbol{X}} = \prod_{i=1}^{n} \Omega_{X_i}$. We also denote by capital letters the nodes in a graph. The terms *node* and *variable* are often used interchangeably throughout this document. We denote by $Pa(Y)$ the parents of node Y in a directed graph.

We now introduce notation that will become useful during the description of the methods presented in this paper. We denote by the operator $\sum_{\boldsymbol{Z}}$ the sum over the possible values of the individual variables forming $\boldsymbol{Z}$, $\sum_{Z_1} \sum_{Z_2} \cdots \sum_{Z_{n_1}}$. For any function h with variables $\boldsymbol{Z}$ and $\boldsymbol{O}$, the expression $h(\boldsymbol{Z}, \boldsymbol{O})|_{\boldsymbol{O}=\boldsymbol{o}}$ stands for a function g over variables $\boldsymbol{Z}$ that results from setting the values of $\boldsymbol{O}$ in h with assignment $\boldsymbol{o}$ while letting the values for $\boldsymbol{Z}$ remain unassigned. In other words, $g(\boldsymbol{Z}) = h(\boldsymbol{Z}, \boldsymbol{O})|_{\boldsymbol{O}=\boldsymbol{o}} = h(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o})$. The notation $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{O})$ means that the variable $\boldsymbol{X}$ is formed by all the variables that form $\boldsymbol{Z}$ and $\boldsymbol{O}$. That is, $\boldsymbol{X} = (X_1, \ldots, X_n) = (Z_1, \ldots, Z_{n_1}, O_1, \ldots, O_{n_2}) = (\boldsymbol{Z}, \boldsymbol{O})$, where $n = n_1 + n_2$. Note that we are assuming that the set of variables forming $\boldsymbol{Z}$ and those forming $\boldsymbol{O}$ are disjoint. The notation $\boldsymbol{Z} \sim f$ means that the random variable $\boldsymbol{Z}$ is distributed according to probability distribution f.

1.5 Decision-theoretic models: Definitions

Before describing the decision-theoretic model used in this work, a description of the Bayesian network probabilistic model is given. This model is commonly used to exploit the structural characteristics of the system under study and provide a more compact representation of the distribution over its states.

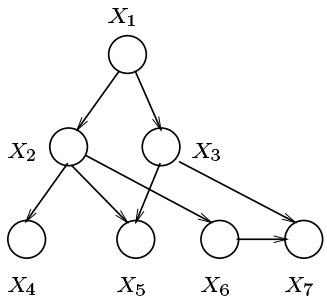


Figure 1.1: Example of a Bayesian network.

1.5.1 Bayesian networks

A *Bayesian network (BN)* is a graphical probabilistic model used to represent uncertainty in structured domains [Pearl, 1988, Jensen, 1996]. It compactly represents the joint probability distribution over the relevant variables of the system of interest. It uses a *directed acyclic graph (DAG)* to represent the relationship between the relevant variables. A node in the graph represents a variable. The model defines a local conditional distribution $P(X_i \mid \mathrm{Pa}(X_i))$ for each node or variable X_i given its parents $\mathrm{Pa}(X_i)$ in the graph. The joint distribution is then

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i)). \tag{1.1}$$

The assumption inherent in this model and represented in the graph is that a variable is (conditionally) independent of (any subset of) its non-descendants in the graph given (observations of) its parents. Applying the *chain rule* of probability to the joint distribution using a partial order of the nodes in the graph (where we condition on parents before children), and simplifying the conditional probability expressions by using the model assumptions represented by the graph, yields the decomposition of the joint probability distribution given in Equation 1.1. For instance, we can define a BN on the graph given in Figure 1.1. Using the structure of the graph, we write the joint distribution as

$$\begin{aligned} P(\boldsymbol{X}) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)P(X_4 \mid X_1, X_2, X_3) \times \\ &\quad P(X_5 \mid X_1, X_2, X_3, X_4)P(X_6 \mid X_1, X_2, X_3, X_4, X_5) \times \\ &\quad P(X_7 \mid X_1, X_2, X_3, X_4, X_5, X_6) \\ &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2) \times \\ &\quad P(X_5 \mid X_1, X_2)P(X_6 \mid X_2)P(X_7 \mid X_3, X_6). \end{aligned}$$

The inference problem in BNs is that of computing the posterior probability of an assignment to a subset of variables given evidence about (i.e., assignments to) another subset of variables in the system. Assume that the variables are discrete and their *sample spaces* are finite. Consider the inference problem of computing $P(X_7 = x_7 \mid X_4 = x_4, X_5 = x_5)$ in our example BN. By the definition of conditional probability, we can decompose this problem into computing the probabilities $P(X_4 = x_4, X_5 = x_5, X_7 = x_7)$ and $P(X_4 = x_4, X_5 = x_5)$. Using the decomposition of the joint probability distribution we can compute

$$\begin{aligned} P(X_4 = x_4, X_5 = x_5, X_7 = x_7) \quad = \quad & \sum_{X_1, X_2, X_3, X_6} P(X_1)P(X_2 \mid X_1) \times \\ & P(X_3 \mid X_1)P(X_6 \mid X_2) \times \\ & P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_1, X_2) \times \\ & P(X_7 = x_7 \mid X_3, X_6) \end{aligned}$$

Sometimes we can compute this quantity more efficiently by distributing the sums. However, this is not always feasible. In general, let $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{O})$ where $\boldsymbol{O}$ is the set of variables of interest, $\boldsymbol{o}$ is an assignment to it and $\boldsymbol{Z}$ are the remaining variables. For this problem we want to compute probabilities of the form

$$P(\boldsymbol{O} = \boldsymbol{o}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o}).$$

In the example above, $\boldsymbol{O} = (X_4, X_5, X_7)$, $\boldsymbol{o} = (x_4, x_5, x_7)$, $\boldsymbol{Z} = (X_1, X_2, X_3, X_6)$, and

$$\begin{aligned} P(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o}) \quad = \quad & P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_6 \mid X_2) \times \\ & P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_1, X_2)P(X_7 = x_7 \mid X_3, X_6). \end{aligned}$$

Often, the local decomposition of the joint distribution still leads to the evaluation of sums over a large number of variables. In general, this problem is computationally intractable [Cooper, 1990]. As a matter of fact, although approximation techniques have been developed, belief inference approximations (both deterministic and randomized) are also computationally intractable [Dagum and Luby, 1993].

1.5.2 Influence diagram

An *influence diagram (ID)* is a decision-theoretic model for decision-making under uncertainty [Howard and Matheson, 1981]. It consists of a directed acyclic graph along with a structural strategy model, a probabilistic model and a utility model. The graph represents

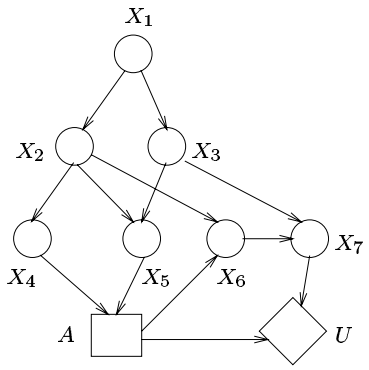


Figure 1.2: Example of an influence diagram.

the decomposition used to compactly define the different models. We can think of an ID as a BN with decision and utility nodes added. For instance, we can use our example BN to build an ID as shown in Figure 1.2. The vertices of the graph consist of three types of nodes: decision nodes, chance nodes and utility nodes. Decision nodes are square and represent the decisions or action choices in the decision problem. Chance nodes are circular and represent the variables of the system relevant to the decision problem. Utility nodes are diamonds and represent the utility associated with actions and *states*. A *state* is an assignment to the variables associated with the chance nodes of the ID.

Structural strategy model The structural strategy model defines locally the form of a decision rule for each decision node A_i. This rule is a function of the information available at the time of making that decision. The typical assumption is that the actions are ordered and that previous decisions and the information they are based on is not forgotten at the time of making a future decision (this is called the *no-forgetting* assumption). The values of the parent variables of a decision node will be available at the time of making that decision (although not all of them might actually be relevant—see Shachter [1998]). Hence, there is an implicit directed arc from previous decision nodes and their parents to the next decision node in order. Therefore, the local policy is a mapping from assignments of $\mathrm{Pa}(A_i)$ (including its implicit parents) to the set of actions available for A_i. For the most part, we will assume that the ID has a single decision node. We denote a strategy for our example model by π, the *state space* or set of possible assignments for the parents of the action node by $\Omega_{\mathrm{Pa}(A)}$ and the set of possible actions Ω_A. Then, $\pi : \Omega_{\mathrm{Pa}(A)} \rightarrow \Omega_A$.

Probability model The probability model compactly defines the joint probability distribution of the relevant variables given the actions taken using a Bayesian network (BN). We now have potentially different joint distributions over the variables, for each action choice available. The joint distribution over the variables, given the action choices $\boldsymbol{a}$ assigned to the decision variable, is

$$P(X_1, \ldots, X_n \mid \boldsymbol{A} = \boldsymbol{a}) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}(X_i))\Big|_{\boldsymbol{A}=\boldsymbol{a}}.$$

Note that in our example ID there is only one decision node. In particular, for our example ID, the joint distribution is

$$\begin{aligned} P(\boldsymbol{X} \mid A = a) \;=\; & P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2) \;\times \\ & P(X_5 \mid X_2, X_3)P(X_6 \mid X_2, A = a)P(X_7 \mid X_3, X_6). \end{aligned}$$

Utility model Finally, the utility model defines the utility associated with actions resulting from the decisions made and states of the variables in the system. Although other possible definitions exist, in this work we assume that the total utility function U is the sum of local utility functions associated with each utility node (i.e., *additive utilities* if more than one) [Keeney and Raiffa, 1976]. For each utility node U_i, the utility function provides a utility value as a function of its parents $\mathrm{Pa}(U_i)$ in the graph. The total utility can be expressed as

$$U(\boldsymbol{X}, \boldsymbol{A}) = \sum_{i=1}^{m} U_i(\mathrm{Pa}(U_i)). \tag{1.2}$$

Note that we are also using the label of the utility node to denote the utility function associated with it. In our example, there is only one utility node and it is a function of the variable X_7 and the decision A; that is, the value of the utility is $U(X_7, A)$, where U is a function mapping values of X_7 and A to a real value.

In this work, we assume that the decisions are discrete and finite. The variables are discrete and finite, and the local utilities are bounded. Some of these assumptions can be relaxed.

Value of a strategy Assume that we have a finite number of discrete action choices. Then, one problem is to select the *best strategy* or function π^*, mapping each possible value of the parents of the decision node to an action choice. In our example, $A = \pi(X_4, X_5)$, where π is a function mapping each combination of values of X_4 and X_5 to an action choice.

The best strategy is the strategy with highest expected utility. Let $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{O})$ where the variables in $\boldsymbol{O}$ are parents of the decision node and $\boldsymbol{Z}$ are the remaining variables. The problem of obtaining an optimal strategy can be reduced to obtaining, for each assignment $\boldsymbol{O} = \boldsymbol{o}$, the action that maximizes the value associated with the action and the assignment.

The *value* V^π *of a strategy* π is the expected utility of the strategy:

$$\begin{aligned} V^\pi &= \sum_{\boldsymbol{X}} P(\boldsymbol{X} \mid A = \pi(\boldsymbol{O}))U(\boldsymbol{X}, A = \pi(\boldsymbol{O})) \\ &= \sum_{\boldsymbol{O}} \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}, \boldsymbol{O} \mid A = \pi(\boldsymbol{O}))U(\boldsymbol{Z}, \boldsymbol{O}, A = \pi(\boldsymbol{O})). \end{aligned}$$

The optimal strategy π^* is that which maximizes V^π over all π. We denote the value of the optimal strategy by V^*.

Note that we can decompose this maximization into maximizations over the set of actions for each observation. For each assignment to the observations $\boldsymbol{o}$, we define the *(unconditional) value of an action* a by

$$V_{\boldsymbol{o}}(a) \equiv V(\boldsymbol{o}, a) = \sum_{\boldsymbol{Z}} P(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o} \mid A = a)U(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o}, A = a). \tag{1.3}$$

(As we will see soon, we will be considering $V(\boldsymbol{o}, a)$ for each observation $\boldsymbol{o}$, so we use the notation $V_{\boldsymbol{o}}(a)$ often to simplify the expression and make it clear that we are considering a particular observation. The two notations are equivalent.) Hence, the value of a strategy is $V^\pi = \sum_{\boldsymbol{O}} V_{\boldsymbol{O}}(\pi(\boldsymbol{O}))$. Note that this is not the traditional definition of the value of an action. Typically, the value of an action is defined as the *conditional* expected utility of the action *given* an assignment of the observations. If we denote this value by $V(a \mid \boldsymbol{o})$, we can express the value of a policy as $V^\pi = \sum_{\boldsymbol{O}} P(\boldsymbol{O})V(\pi(\boldsymbol{O}) \mid \boldsymbol{O})$. We discuss later why we do not use the traditional definition.

If we denote by $a^* = \pi^*(\boldsymbol{o})$ the action that maximizes $V_{\boldsymbol{o}}(a)$ over all actions a, then the value of the optimal strategy is $V^* = \sum_{\boldsymbol{O}} V_{\boldsymbol{O}}(\pi^*(\boldsymbol{O})) = \sum_{\boldsymbol{O}} \max_a V_{\boldsymbol{O}}(a)$. Hence, the problem of strategy selection can be reduced in this way to that of action selection for each observation.

For instance, in our example ID, $\boldsymbol{O} = (X_4, X_5)$, $\boldsymbol{Z} = (X_1, X_2, X_3, X_6, X_7)$, $\boldsymbol{o} = (x_4, x_5)$,

$$\begin{aligned} P(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o} \mid A = a) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_6 \mid X_2, A = a) \times \\ &\quad P(X_7 \mid X_3, X_6)P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_2, X_3), \end{aligned}$$

and $U(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o}, A = a) = U(X_7, A = a)$. Note once again that computing $V_{\boldsymbol{o}}(a)$ requires the evaluation of a sum. For the same reasons as in the previous problem of belief inference in BNs, the exact computation of this value is intractable in general.

1.5.3 Markov decision process

Markov decision processes (MDPs) and *partially observable Markov decision processes (POMDPs)* are popular frameworks for modeling sequential decision-making under uncertainty. One can think of these models as IDs with multiple (and a possibly infinite number of) decision nodes and utility nodes (typically called *rewards* in their context) and restrictions on the probability model (both on the structure and parameters of the local probability models). We do not deal directly with these models in this work. However, we believe that some of the results from our work have potential extensions to problems of this kind.

1.6 Importance sampling

In this work we present methods based on Monte Carlo estimation (See Rubinstein [1981] and the references therein). In this section, we describe *importance sampling*, which is a general method for estimating integrals and sums in high dimensions (See Kahn and Marshall [1953] and the references therein). Importance sampling provides an alternative to the exact methods for evaluating sums in the form of an approximation. Let the quantity of interest be $G = \sum_{\boldsymbol{Z}} g(Z)$ for some real function g, which we call here the *target function*. We can turn the sum into an expectation by expressing

$$G = \sum_{\boldsymbol{Z}} f(\boldsymbol{Z}) \frac{g(\boldsymbol{Z})}{f(\boldsymbol{Z})},$$

where f is a probability distribution over $\boldsymbol{Z}$ satisfying, for all $\boldsymbol{Z}$, $g(\boldsymbol{Z}) \neq 0 \Rightarrow f(\boldsymbol{Z}) \neq 0$ (i.e., f is *absolutely continuous* with respect to g). We call f the *importance-sampling distribution*. We define the *weight function* $\omega(\boldsymbol{Z}) = g(\boldsymbol{Z})/f(\boldsymbol{Z})$ which allows us to express

$$G = \sum_{\boldsymbol{Z}} f(\boldsymbol{Z}) \omega(\boldsymbol{Z}).$$

Hence, we can obtain an unbiased estimate of G by obtaining N samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(N)}$ from $\boldsymbol{Z} \sim f$ and computing the estimate

$$\hat{G} = \frac{1}{N} \sum_{l=1}^{N} \omega(\boldsymbol{z}^{(l)}). \tag{1.4}$$

Note also that by the *central limit theorem (CLT)*, $\sqrt{N}(\hat{G} - G)$ is asymptotically normally distributed with mean 0 and variance $\mathrm{Var}_f[\omega(\boldsymbol{Z})] \equiv \sum_{\boldsymbol{Z}} f(\boldsymbol{Z}) \left(\omega(\boldsymbol{Z}) - G\right)^2$.

We can apply this technique to the problem of belief inference in BNs [Fung and Chang, 1989, Shachter and Peot, 1989]. Typically, we let

$$\begin{aligned} g(\boldsymbol{Z}) &= P(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o}) \\ &= \prod_{i=1}^{n_1} P(Z_i \mid \mathrm{Pa}(Z_i)) \prod_{j=1}^{n_2} P(O_j \mid \mathrm{Pa}(O_j)) \Big|_{\boldsymbol{O}=\boldsymbol{o}}, \\ f(\boldsymbol{Z}) &= \prod_{i=1}^{n_1} P(Z_i \mid \mathrm{Pa}(Z_i)) \Big|_{\boldsymbol{O}=\boldsymbol{o}}, \text{ which implies} \\ \omega(\boldsymbol{Z}) &= \prod_{j=1}^{n_2} P(O_j \mid \mathrm{Pa}(O_j)) \Big|_{\boldsymbol{O}=\boldsymbol{o}}. \end{aligned}$$

For instance, in our example, we can estimate $G = P(X_4 = x_4, X_5 = x_5, X_7 = x_7)$ by letting $\boldsymbol{Z} = (X_1, X_2, X_3, X_6)$, $\boldsymbol{O} = (X_4, X_5, X_7)$, $\boldsymbol{o} = (x_4, x_5, x_7)$,

$$\begin{aligned} g(\boldsymbol{Z}) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_6 \mid X_2) \times \\ &\quad P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_1, X_2)P(X_7 = x_7 \mid X_3, X_6), \\ f(\boldsymbol{Z}) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_6 \mid X_2), \\ \omega(\boldsymbol{Z}) &= P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_1, X_2)P(X_7 = x_7 \mid X_3, X_6). \end{aligned}$$

Note that, in the example above, we are defining the importance sampling distribution to be the prior distribution over the *hidden* variables $\boldsymbol{Z}$ of the BN. In general, the sampling distribution is that which results from the *do-operated* BN as defined by Pearl [2000] where the *do* operation or intervention is done on the variables that have assignments (in this case, $\boldsymbol{O}$). Graphically, we can represent this distribution by removing the arcs into nodes that have been assigned values (i.e., operated on). Consider the problem of computing $P(X_3 = x_3, X_4 = x_4, X_5 = x_5)$. The graph of the BN representing the do-operated distribution $P(\boldsymbol{X} \mid \mathrm{do}(X_3 = x_3, X_4 = x_4, X_5 = x_5))$ that we would use for this problem is given in Figure 1.3. We obtain samples from this distribution by sampling the variables in the (partial) order defined by the DAG of the do-operated BN. We obtain samples from each variable by traversing the nodes in the graph and sampling the variable corresponding to nodes that were not assigned by the do-operation, conditioned on the assignments to the parents of those variables. Note that by the way we are obtaining samples from the hidden variables, the parent assignments will always be available at the time a node is sampled. Therefore, the resulting samples will be assignments to those variables that are not in the evidence set (i.e., the hidden variables) according to the do-operated distribution of the BN.

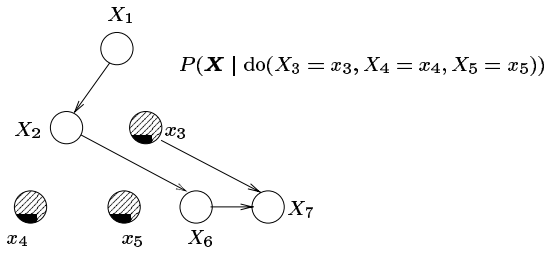


Figure 1.3: Example of *do-operated* BN for BN in Figure 1.1.

In cases when no variable in $\boldsymbol{O}$ has children in $\boldsymbol{Z}$ (as is the case of the running example from the previous paragraph), the do-operated distribution corresponds to the prior distribution given by the original BN.

Consider again the problem of computing $P(X_3 = x_3, X_4 = x_4, X_5 = x_5)$. We would

1. Sample $x_1 \sim P(X_1)$.
2. Sample $x_2 \sim P(X_2 \mid X_1 = x_1)$.
3. Sample $x_6 \sim P(X_6 \mid X_2 = x_2)$.
4. Sample $x_7 \sim P(X_7 \mid X_3 = x_3, X_6 = x_6)$.

We can extend this procedure beyond this example to perform more general computations in BNs.

We call the method resulting from this importance-sampling distribution the *traditional method*. In the context of belief inference, this method is called *likelihood-weighting (LW)* since the weight function is a "local likelihood" and thus each sample is weighted by a product of "local likelihoods."

We can similarly apply this technique in the context of action selection in IDs to evaluate $V_{\boldsymbol{o}}(a)$ (See Charnes and Shenoy [1999] for an example and previous references). Now the decision node is also do-operated ($A = a$). In general, we let

$$
\begin{aligned}
g(\boldsymbol{Z}) &= P(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o} \mid A = a) U(\boldsymbol{Z}, \boldsymbol{O} = \boldsymbol{o}, A = a), \\
f(\boldsymbol{Z}) &= \prod_{i=1}^{n_1} P(Z_i \mid \mathrm{Pa}(Z_i)) \Big|_{\boldsymbol{O}=\boldsymbol{o}, A=a}, \\
\omega(\boldsymbol{Z}) &= \prod_{j=1}^{n_2} P(O_j \mid \mathrm{Pa}(O_j)) U(\boldsymbol{Z}, \boldsymbol{O}, A) \Big|_{\boldsymbol{O}=\boldsymbol{o}, A=a}.
\end{aligned}
$$

In particular, for our example,

$$\begin{aligned} g(\boldsymbol{Z}) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_6 \mid X_2, A=a)P(X_7 \mid X_3, X_6) \times \\ &\quad P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_2, X_3)U(X_7, A=a), \\ f(\boldsymbol{Z}) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_6 \mid X_2, A=a)P(X_7 \mid X_3, X_6), \\ \omega(\boldsymbol{Z}) &= P(X_4 = x_4 \mid X_2)P(X_5 = x_5 \mid X_2, X_3)U(X_7, A=a). \end{aligned}$$

1.7 General objectives

Our primary objective is to assess the properties of the simple importance sampling or traditional method for the problem of solving IDs, and more specifically, selecting "good" actions. We look at how, by using simple stopping rules and allocation schedules based on the idea of multiple comparisons, we can improve over the naive use of this sampling method based on direct estimation.

There are many different exact methods for solving IDs directly. Also, if we could normalize the utilities (bring the utilities to a 0-1 scale through linear transformations), we can reduce the problem of solving IDs to the problem of inference in BNs (See Shachter and Peot [1992] and Zhang [1998], who also refer to Cooper [1988]). Hence, in principle, we can apply any method (either exact or approximate) for the belief inference in BNs to solving IDs. There are many methods for exact computation in BNs. Given the general intractability of exact inference, I believe they will all have problems with some class of models where one might have to consider a very large number of possible joint outcomes for a subset of the domain variables. It is this class for which we believe the methods developed in this thesis can be most effective. In particular, the effectiveness of sampling methods is primarily tied to the *numerical* properties of the model, instead of the global *structural* properties conveyed in the graph. Therefore, I believe, in general, the properties that make exact methods efficient are different from those that make sampling methods effective. Whether exact methods are efficient or not can be determined primarily from the properties of the graph. (See Appendix B for an example of a model that is computationally problematic for exact methods.) Thus, which method would be more applicable is mostly problem-domain dependent. Hence, further analytical and empirical study than that provided in this thesis is still necessary to compare exact methods to the approximation methods proposed in this thesis.

Also, other kinds of approximation methods, including deterministic methods, that exploit or force different particular *structural* properties of the problem for computing optimal

strategies have been suggested in the context of IDs [Nilsson and Lauritzen, 2000]. In addition, approximation techniques have been suggested in the related context of MDPs and POMDPs[Dearden and Boutilier, 1997, Boutilier et al., 2000, Koller and Parr, 1999, 2000, Kim and Dean, 2001]. In principle, all of these techniques can be applied to the problem we consider in this thesis, and in principle, some can be combined with the methods developed in this work. I believe such approximations, by their nature, will primarily be effective under different conditions, as they are mostly based on exploiting further structural properties of the model. I also believe that they can be combined with the methods presented here. However, further analysis and study comparing the methods presented here with other type of approximation methods is still required.

It has been suggested that the simple importance sampling or traditional importance sampling method is used most often in practice because of both its simplicity and effectiveness as compared to other methods (at least for problems in BNs). Hence, I use *adaptive importance sampling (AIS)* as a method to update our sampling distribution while remaining in the simple importance-sampling class of methods. To this degree we study the general properties of methods of this kind, and how they compare with the simple importance sampler for the estimation problem. A careful, general comparison study of the methods presented here with all other sampling methods developed was not performed in this thesis but will certainly be required in the future.

Finally, our empirical study of the AIS method involves a very special type of model. A deterministic approximation method based on variational methods in statistical physics [Jordan et al., 1997] has been shown quite successful for this model [Jaakkola and Jordan, 1999]. We do not empirically compare this technique to ours. However, we comment on the connection of variational methods to ours from a theoretical perspective.

1.8 Overview

We will now provide an overview of the chapters forming this document.

- Chapter 2 deals with the problem of action selection in IDs with a single decision. For the most part, we will concentrate on studying methods for which we can prove theoretically sound approximations, as opposed to methods based on *asymptotic* approximations and/or other heuristics. We will present theoretical results establishing bounds on the number of samples required to select actions that are near optimal with high probability. The final objective behind the approximation is the computation of a

full strategy that is near optimal with high probability. We view the problem of action selection as a problem of multiple-comparisons and exploit results from the statistical literature in *multiple-comparisons with the best (MCB)* [Hsu, 1996] throughout.

We also developed a general class of comparison-based or multi-stage sequential methods similar to *group sequential methods* used in the statistical literature on experimental design and clinical trials. For the most part, comparison-based methods achieve the theoretical guarantees of other estimation-based methods presented in the chapter. Although we provide stopping rules for which the comparison-based methods are theoretically valid, we do not present rigorous bounds on the number of samples. Hence, we will empirically show that comparison-based methods can perform theoretically-guaranteed, near-optimal action selections with significantly fewer samples than those needed by estimation-based methods. We also suggest a heuristic version of the comparison-based method in which we allow adaptive sample-allocation schedules. The motivation is to reduce the total number of samples for action selection by reallocating the samples given information from the sample outcomes. We present a very simple version of this idea. Even in simple formulations, adaptive reallocation of samples presents problems for theoretical analysis. However, we believe this heuristic method can be very effective in reducing the number of samples needed for near-optimal action selection. At the end of the chapter we discuss issues of optimality associated with the adaptive reallocation heuristic method, as well as some potentially effective (though not necessarily theoretically grounded) practical extensions of our methods.

Section 2.9 presents an empirical study of the methods in this chapter in a real ID known as the IctNeo ID [Bielza et al., 2000, Gómez et al., 2000]. This ID was developed to treat jaundice in newborns and is still under development. At this time, the model is not amenable to exact methods. Since the ID is technically a multi-stage ID, we considered several modified versions of it in order to make a single decision model. Because the ID models much of the background information associated with the patient, the number of observations potentially available at the time of decision-making is quite large for this ID. However, the exact methods we tried could not solve one modified version of this ID used in the experiments, even when considering just a single observation scenario at a time. Not only were the sampling methods developed and presented in this chapter very effective in obtaining optimal actions for each observation (from the set of randomly generated observations that we tried), but also

they theoretically guarantee very close-to-optimal behavior. Also, an instantiation of the theoretically-grounded comparison-based methods effectively reduced the number of samples needed to achieve near-optimal action selection. We evaluated the adaptive-reallocation heuristic method in this problem and found it to be very effective too. We also evaluated the methods in another modified version of this ID for which exact methods were efficient. This was done primarily to provide another version of the problem in which to compare the importance sampling methods among themselves.

- Chapter 3 presents a class of *adaptive importance sampling (AIS)* methods for estimation in graphical models. Although we believe the methods are more general, we will concentrate on problems in BNs and IDs. The main motivation for these methods is to improve the quality of the estimators of the traditional importance sampling method. We view the problem as a learning or optimization problem. The objective is to find the best probability distribution to use for sampling over a parameterized class. We argue that an immediate and useful class for this problem is that of BNs. This is because BNs allow efficient simulation. We consider particularly relevant error measures and developed update rules for *learning* the importance sampling BN from the samples themselves. We suggest a class of estimators that make more efficient use of the samples generated. We theoretically study those estimators and show that some instantiations of the class of estimators have interesting properties, the ability to be made unbiased, to converge in probability to the true value, and to allow computation of confidence bounds. We also theoretically study the *learning* process. Our main objective in doing this is to show that by approaching the update process as a stochastic gradient problem we can borrow convergence (and other) results from the theory of stochastic approximation to analyze theoretically the behavior of the AIS methods presented in this thesis.

  We performed a preliminary empirical study on a simple ID problem. We also performed a simple empirical study that involved computing the likelihood of a random sample on a synthetic BN. The synthetic BN was randomly generated from a class of BNs known as QMR-DT-type BNs, for their similarity with a large real BN developed for medical diagnosis. This model has been studied in the community and is still the subject of current interest. In general, exact computations in this special model are intractable. Inspired by the empirical results, we theoretically study some properties of the AIS methods for this class of problems and in doing so we start to establish the connection between AIS and variational method. We also connect this work to

additional work on adaptive importance sampling in our field and others.

1.9 Main theme

The main theme behind the arguments in this dissertation is that sampling methods can be an effective tool for problems in Bayesian networks and influence diagrams when they exploit particular properties of the model under consideration. In particular, sampling methods are very effective for selecting approximately optimal actions in influence diagrams. For instance, in Chapter 2, I exploit the fact that optimal action selection is primarily a comparison problem to provide more effective sampling methods than those resulting from more naive application of the sampling process. In Chapter 3, I exploit the fact that one can simulate Bayesian networks efficiently and use sample information to provide methods for adapting the sampling distribution that resulted in significantly more accurate estimates than those produced by naive applications of importance sampling. In conclusion, I believe this theme applies to models beyond those explicitly considered in this thesis.

Chapter 2

Action Selection

In this chapter, the problem under study is selecting an optimal strategy in an influence diagram, concentrating on the case in which there is only one decision to be made. Our motivation for this focus is that we can decompose the problem of multiple decisions into many sub-problems involving single decisions (i.e., by using the technique presented by Charnes and Shenoy [1999]). We believe we can extend methods developed to solve IDs of this kind to obtain methods to solve finite-horizon Markov decision processes (MDPs) and partially observable Markov decision processes (POMDPs) expressed as dynamic Bayesian networks (DBNs) (i.e., by modifying the technique presented by Kearns et al. [1999b]). Figure 2.1 shows the general structure of the example ID we consider.

The problem of strategy selection involves the sub-problem of selecting an *optimal action*, from the set of action choices available for that decision, *for each possible observation* available at the time of making the decision. Therefore, we want to select the action that maximizes the expected utility for each observation. One way to do action selection is to compute, exactly or approximately, the probabilities of the sub-states of the system directly relevant to our utility in order to evaluate the expected utility or *value* of each action. A sub-state is formed from the state of a subset of variables in the system. This approach fails to take advantage of an important intuition: the expected utilities of the actions are unimportant—it only matters which action is best. Therefore, the problem of action selection is primarily one of comparing the values of the actions. Hence, we can combine this with the intuition that actions that are close to optimal are also good. In this chapter, methods for action selection in IDs are presented that take advantage of these intuitions to make major gains in efficiency.

Exact methods exist for computing the optimal strategy in an ID (see Charnes and

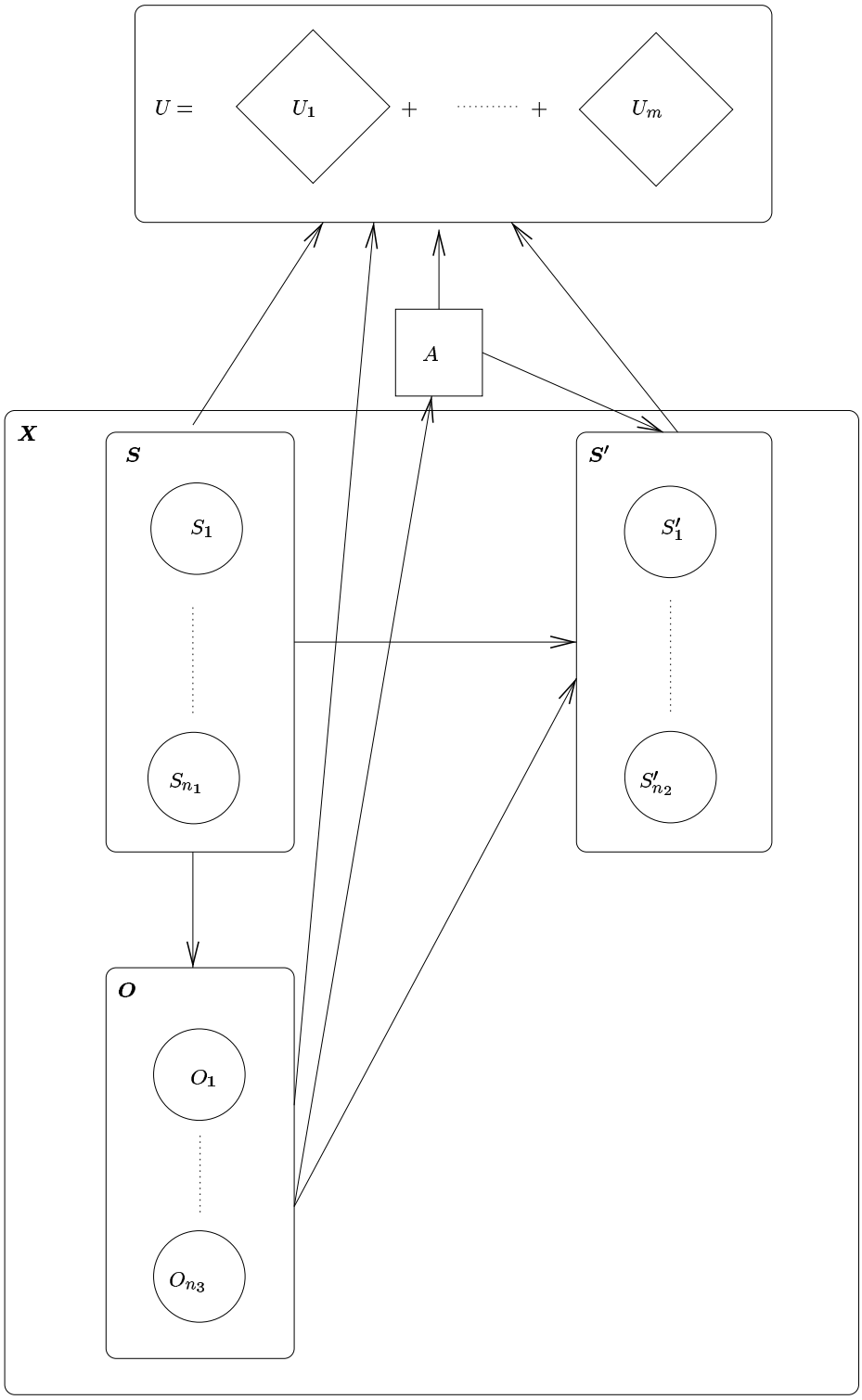


Figure 2.1: General structure of ID we consider in this chapter.

Shenoy [1999] and Jensen [1996] for short descriptions and a list of references). However, this problem is hard in general. In this chapter, we concentrate on obtaining approximations to the optimal strategy with certain guarantees. Our objective is to find policies that are close to optimal with high probability. That is, for a given accuracy parameter ϵ^* and confidence parameter δ^*, we want to obtain a strategy $\hat{\pi}$ such that $V^* - V^{\hat{\pi}} < \epsilon^*$ with probability at least $1 - \delta^*$, where V^* is the (true) value of the optimal strategy and $V^{\hat{\pi}}$ is the (true) value of the strategy $\hat{\pi}$. From the decomposition of the value of a strategy described in the introduction, if we obtain actions for each observation such that their values are *sufficiently* close to optimal with *sufficiently* high probability, then we obtain a near-optimal strategy with high probability. One simple, albeit naive, way to do this is as follows. Let l be the number of possible assignments to the observations. If for each observation $\boldsymbol{o}$ we select action $\hat{a}$ such that $V_{\boldsymbol{o}}(a^*) - V_{\boldsymbol{o}}(\hat{a}) < 2\epsilon$ with probability at least $1 - \delta$, where $\epsilon = \epsilon^*/(2l)$ and $\delta = \delta^*/l$, then we obtain a strategy that is within ϵ^* from the optimal with probability at least $1 - \delta^*$. This approach is naive because it allocates error and precision parameters equally among observations. A smarter way to do this allocation is given later in this chapter. Based on the argument just presented, we concentrate on finding a *good* action for each observation.

Typically the value of an action is defined as the *conditional* expected utility of the action *given* an assignment of the observations. If we denote this value by $V(a \mid \boldsymbol{o})$, we can express the value of a policy as $V^\pi = \sum_{\boldsymbol{O}} P(\boldsymbol{O}) V(\pi(\boldsymbol{O}) \mid \boldsymbol{O})$. This definition is not used because it is harder to obtain estimates for $V(a \mid \boldsymbol{o})$ with guaranteed confidence bounds than it is to obtain estimates for $V_{\boldsymbol{o}}(a)$. Let us argue briefly why it is harder to deal with $V(a \mid \boldsymbol{o})$. The main reason is that, in general, obtaining an approximation for $V(a \mid \boldsymbol{o})$ requires obtaining the approximation of a ratio (i.e., $V_{\boldsymbol{o}}(a)/P(\boldsymbol{O} = \boldsymbol{o})$). In general, to obtain any kind of approximation (either absolute or relative) of a ratio requires relative approximations for the numerator and denominator. Relative approximations are in general intractable computationally, potentially requiring an unreasonable number of samples. We also argue that we do not need to estimate $V(a \mid \boldsymbol{o})$ to get absolute approximations as described above. In general, computing the conditional value of an action given an observation $\boldsymbol{o}$ requires that we compute the normalizing factor probability $P(\boldsymbol{O} = \boldsymbol{o})$, as well as the unconditional value $V_{\boldsymbol{o}}(a)$. However, the normalizing factor is constant for all possible action choices available given the observation. Hence, to determine the optimal action we do not need to compute this probability. Also, with respect to the *global* value of the strategy, this term will be canceled by the expectation. In addition, if the unconditional value is very small (in the case that the total utility has range $[0, 1]$, the unconditional value will be smaller

than the probability of the observation: $V_{\boldsymbol{o}}(a) < P(\boldsymbol{O} = \boldsymbol{o})$), and if its variance is not small enough, obtaining relative bounds for it might require an extremely large number of samples (inversely proportional to how small that value is). To summarize, the type of approximations considered here are absolute approximations. Relative approximations can be useful in some cases and we will consider them in Section 2.4.

2.1 Useful mathematical results

In this section, we present some mathematical results from large deviation theory and the statistical work on multiple comparisons with the best (MCB) that will be useful during the analysis of the methods presented in this thesis. The results are presented for completeness and in order to refer back to them during the analysis and proofs.

2.1.1 Large deviation results

We first introduce the following notation and definitions which will be used in the statements of the results below. For fixed n, let

1. $X_1, X_2, \ldots, X_n$ be random variables,
2. $S = X_1 + X_2 + \cdots + X_n$,
3. $\overline{X} = S/n$,
4. $-\infty < \mu = \mathrm{E}[\overline{X}] = \mathrm{E}[S]/n < \infty$ (finite) and
5. $\sigma^2 = n\,\mathrm{Var}[\overline{X}] = \mathrm{Var}[S]/n < \infty$ (finite).

Theorem 1 *1.* (Hoeffding's traditional bound [Hoeffding, 1963]) *If*

(a) $X_1, X_2, \ldots, X_n$ *are independent, and*

(b) $a_i \le X_i \le b_i$, *and* $a_i \le b_i$ *constants, for* $i = 1, 2, \ldots, n$,

then for $t > 0$,

$$\Pr\left\{\overline{X} - \mu \ge t\right\} \le \exp\left(-2n^2t^2 / \sum_{i=1}^{n}(b_i - a_i)^2\right). \tag{2.1}$$

2. (Hoeffding's strengthened bound—See Hoeffding [1963], page 17–18, for a discussion) *The bound given in (2.1) above still holds if condition 1a above is replaced by the weaker condition*

(a) the sequence $S'_m = S_m - \mathrm{E}[S_m], m = 1, 2, \ldots, n$, *is a martingale; that is*

$$\mathrm{E}[S'_m \mid S'_1, \ldots, S'_j] = S'_j, 1 \leq j \leq m \leq n, \tag{2.2}$$

with probability one,

Theorem 2 (Bernstein's inequality—as presented in Devroye et al. [1996] and attributed to Bernstein [1946]) *If*

1. $X_1, X_2, \ldots, X_n$ *are independent, and*
2. $\mathrm{E}[X_i] = 0$, *and* $X_i \leq c$, *for* $i = 1, 2, \ldots, n$, *and* c *constant,*

then for $0 < t < c$,

$$\Pr\left\{\overline{X} \geq t\right\} \leq \exp\left(-nt^2/(2\sigma^2 + 2ct/3)\right). \tag{2.3}$$

In applying Bernstein's inequality, typically we have $a \leq X_i \leq b$ for $i = 1, 2, \ldots, n$, and constants a, b, such that $a \leq b$. Hence, we use the random variables $X'_i = X_i - \mu$, with $\mu = \mathrm{E}[X_i]$, such that $\mathrm{E}[X'_i] = 0$, for $i = 1, 2, \ldots, n$. This leads to $\overline{X}' = \overline{X} - \mu/n$, and $c = b - a$. (Note that σ^2 is as defined above, since $\mathrm{Var}[\overline{X}'] = \mathrm{Var}[\overline{X}]$.) The result is a statement about the deviation of the sample mean from the true mean (in the case that the common mean is not necessarily zero, just as in Hoeffding's bound). Bernstein's bound is tighter than Hoeffding's when $\sigma^2 \ll (b-a)^2$ and t is sufficiently small.

Theorem 3 *(*Bonferroni's inequality or Union bound*) If* $A_1, A_2, \ldots$ *be a (countably infinite) sequence of events (or sets), then* $\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \Pr(A_i)$.

2.1.2 Multiple comparisons with the best (MCB)

There are two important results from the field of *multiple comparisons* and in particular from the field of *multiple comparisons with the best* that are used in this work. These results are based on the work of Hsu [1981] (see Hsu [1996] for more information). Before presenting the results, let us introduce the following notation: denote $x^+ = \max(x, 0)$ and $-x^- = \min(0, x)$. The first result is known as *Hsu's single-bound lemma*, which is presented as Lemma 1 by Matejcik and Nelson [1995].

Lemma 1 *Let* $\mu_{(1)} \leq \mu_{(2)} \leq \cdots \leq \mu_{(k)}$ *be the (unknown) ordered performance parameters of* k *systems, and let* $\hat{\mu}_{(1)}, \hat{\mu}_{(2)}, \ldots, \hat{\mu}_{(k)}$ *be any estimators of the parameters. If*

$$\Pr\{\hat{\mu}_{(k)} - \hat{\mu}_{(i)} - (\mu_{(k)} - \mu_{(i)}) > -w, i = 1, \ldots, k-1\} = 1 - \alpha, \tag{2.4}$$

then

$$\Pr\{\mu_i - \max_{j\neq i}\mu_j \in [-(\hat{\mu}_i - \max_{j\neq i}\hat{\mu}_j - w)^-, (\hat{\mu}_i - \max_{j\neq i}\hat{\mu}_j + w)^+], \text{ for all } i\} \geq 1-\alpha. \quad (2.5)$$

If we replace the $=$ *in* (2.4) *with* $\geq$, *then* (2.5) *still holds.*

In the context of the work presented in this chapter, let for each action a, the true value $\mu_a \equiv V_{\boldsymbol{o}}(a)$ and the estimate $\hat{\mu}_a \equiv \hat{V}_{\boldsymbol{o}}(a)$. Also, the i^{th} smallest true value corresponds to $\mu_{(i)}$. That is, if $V_{\boldsymbol{o}}(a_1) \leq V_{\boldsymbol{o}}(a_2) \leq \cdots \leq V_{\boldsymbol{o}}(a_k)$, then for all i, $\mu_{(i)} \equiv V_{\boldsymbol{o}}(a_i)$. Note that in practice, we do not know which action has the largest value. In order to apply Hsu's single-bound lemma, we obtain the bound $\Pr\{\hat{\mu}_j - \hat{\mu}_i - (\mu_j - \mu_i) > -w, \text{ for all } i \neq j\} \geq 1-\alpha$, for each action j, individually. This implies that $\Pr\{\hat{\mu}_{(k)} - \hat{\mu}_{(i)} - (\mu_{(k)} - \mu_{(i)}) > -w, i = 1, \ldots, k-1\} \geq 1-\alpha$, which allow us to apply the lemma. Figure 2.2 graphically describes this practical interpretation of the lemma. For each action i, individually, the upper bounds on the true differences, drawn on the left-hand side, $V_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(j) < \hat{V}_{\boldsymbol{o}}(i) - \hat{V}_{\boldsymbol{o}}(j) + w$, for each $j \neq i$, hold simultaneously with probability at least $1-\alpha$. Note that the "lower bounds" on the left-hand side are $-\infty$. The confidence intervals, drawn on the right-hand side, $V_{\boldsymbol{o}}(i) - \max_{j\neq i} V_{\boldsymbol{o}}(j) \in [-(\hat{V}_{\boldsymbol{o}}(i) - \max_{j\neq i}\hat{V}_{\boldsymbol{o}}(j) - w)^-, (\hat{V}_{\boldsymbol{o}}(i) - \max_{j\neq i}\hat{V}_{\boldsymbol{o}}(j) + w)^+]$, for each action i, hold simultaneously with probability at least $1-\alpha$.

The second result allows us to assess joint confidence intervals on the difference between the value of each action from the value of the best action when we have estimates of the differences between values of each pair of actions with different degrees of accuracy. The result is known as *Hsu's multiple-bound lemma.* It is presented as Lemma 2 by Matejcik and Nelson [1995], and credited to Chang and Hsu [1992].

Lemma 2 *Let* $\mu_{(1)} \leq \mu_{(2)} \leq \cdots \leq \mu_{(k)}$ *be the (unknown) ordered performance parameters of* k *systems. Let* T_{ij} *be a point estimator of the parameter* $\mu_i - \mu_j$. *If for each* i *individually*

$$\Pr\{T_{ij} - (\mu_i - \mu_j) > -w_{ij}, \text{for all } j \neq i\} = 1-\alpha, \quad (2.6)$$

then we can make the joint probability statement

$$\Pr\{\mu_i - \max_{j\neq i}\mu_j \in [D_i^-, D_i^+], \text{for all } i\} \geq 1-\alpha, \quad (2.7)$$

where $D_i^+ = (\min_{j\neq i}[T_{ij} + w_{ij}])^+$, $\mathcal{G} = \{l : D_l^+ > 0\}$, *and*

$$D_i^- = \begin{cases} 0 & \text{if } \mathcal{G} = \{i\} \\ -(\min_{j\in\mathcal{G}, j\neq i}[-T_{ji} - w_{ji}])^- & \text{otherwise.} \end{cases}$$

If we replace the $=$ *in* (2.6) *with* $\geq$, *then* (2.7) *still holds.*

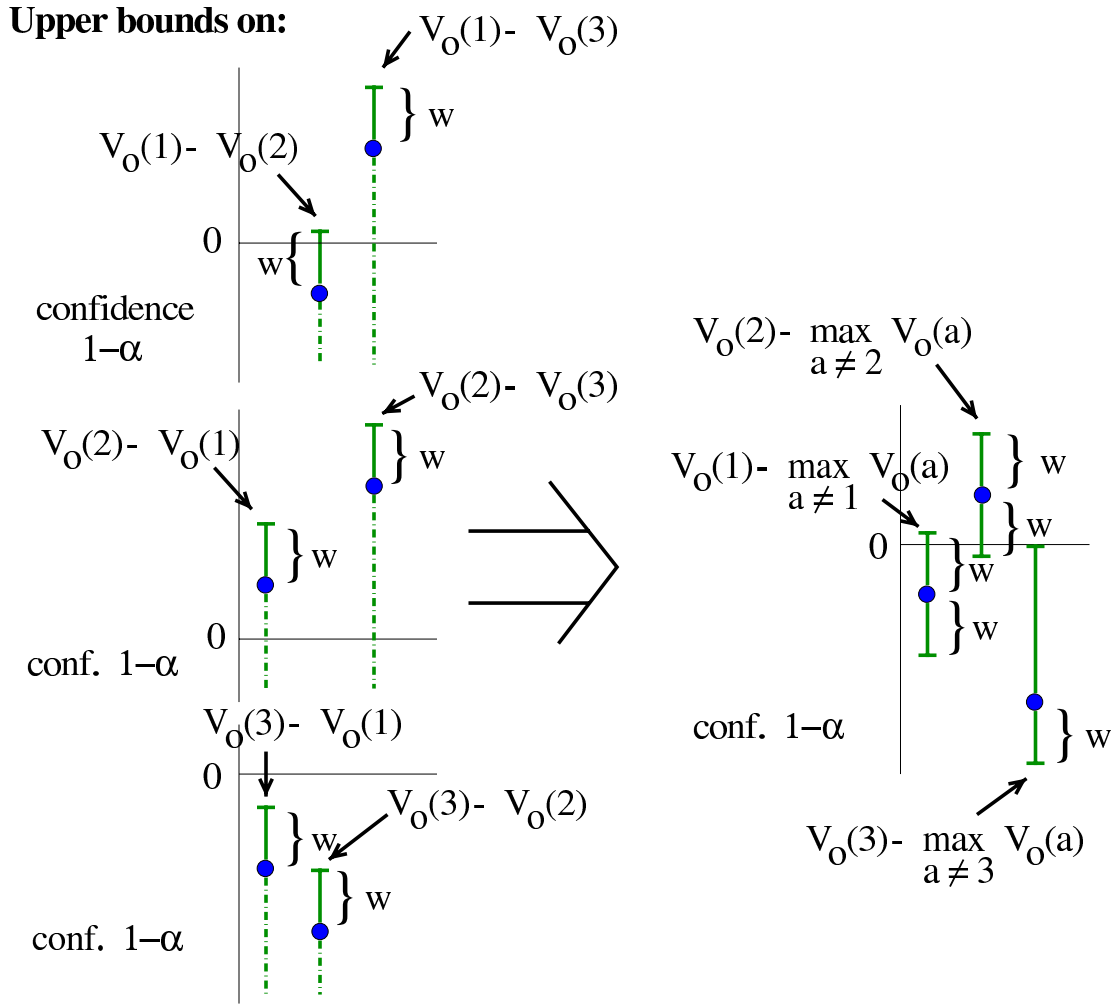


Figure 2.2: Graphical description for practical application of Hsu's single-bound lemma.

Figure 2.3 presents a graphical description of this lemma. Let, for all actions i, D_i^- and D_i^+, be as defined in Hsu's multiple-bound lemma, with $\mu_i = V_{\boldsymbol{o}}(i)$ and for all $j \neq i$, $T_{ij} = \hat{V}_{\boldsymbol{o}}(i) - \hat{V}_{\boldsymbol{o}}(j)$. For each action i, individually, the upper bounds on the true differences, drawn on the left-hand side, $V_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(j) < T_{ij} + w_{ij}$, for each $j \neq i$, hold simultaneously with probability at least $1-\alpha$. Note again that the "lower bounds" on the left-hand side are $-\infty$. The confidence intervals, drawn on the right-hand side, $V_{\boldsymbol{o}}(i) - \max_{j \neq i} V_{\boldsymbol{o}}(j) \in [D_i^-, D_i^+]$, for each action i, hold simultaneously with probability at least $1-\alpha$. Also, in this example, $\mathcal{G} = \{1, 2\}$. In the context of the work presented in this proposal, $\mathcal{G}$ is the set of all the actions that could potentially be the best with probability at least $1 - \alpha$. That is, for each action a in $\mathcal{G}$, the upper bound D_a^+ on the difference of the true value of action a and the best of *all* the other actions, including those in $\mathcal{G}$, is positive.

2.2 Estimation-based methods

The most straightforward approach to selecting the best action is to obtain estimates of $V_{\boldsymbol{o}}(a)$ for each a by sampling, using the probability model of the ID conditioned on a, then select the action with the largest estimated value.

As stated in the introduction, we can apply the idea of *importance sampling* to this estimation problem by using the probability distribution defined by the ID as *the importance function* or *sampling distribution*. We now quickly review this method in the context of the example ID in Figure 2.1.

First, let us present definitions that will allow us to rewrite $V_{\boldsymbol{o}}(a)$ more clearly. First, let $\boldsymbol{Z} = (\boldsymbol{S}, \boldsymbol{S}')$ and define the *target function* (in our case, the *weighted utilities*)

$$\begin{aligned} g_{a,\boldsymbol{o}}(\boldsymbol{Z}) &= g_{a,\boldsymbol{o}}(\boldsymbol{S}, \boldsymbol{S}') \\ &= P(\boldsymbol{S})P(\boldsymbol{S}' \mid \boldsymbol{S}, \boldsymbol{O} = \boldsymbol{o}, A = a) \times \\ &\qquad P(\boldsymbol{O} = \boldsymbol{o} \mid \boldsymbol{S})U(\boldsymbol{S}, \boldsymbol{S}', \boldsymbol{O} = \boldsymbol{o}, A = a). \end{aligned}$$

Note that $V_{\boldsymbol{o}}(a) = \sum_{\boldsymbol{Z}} g_{a,\boldsymbol{o}}(\boldsymbol{Z})$. Now, we can define the *importance function* as

$$f_{a,\boldsymbol{o}}(\boldsymbol{Z}) = P(\boldsymbol{S})P(\boldsymbol{S}' \mid \boldsymbol{S}, \boldsymbol{O} = \boldsymbol{o}, A = a), \tag{2.8}$$

which lets us define the *weight function* $\omega_{a,\boldsymbol{o}}(\boldsymbol{Z}) = g_{a,\boldsymbol{o}}(\boldsymbol{Z}) / f_{a,\boldsymbol{o}}(\boldsymbol{Z})$. In this case,

$$\omega_{a,\boldsymbol{o}}(\boldsymbol{Z}) = P(\boldsymbol{O} = \boldsymbol{o} \mid \boldsymbol{S})U(\boldsymbol{S}, \boldsymbol{S}', \boldsymbol{O} = \boldsymbol{o}, A = a). \tag{2.9}$$

Finally, we can express $V_{\boldsymbol{o}}(a) = \sum_{\boldsymbol{Z}} f_{a,\boldsymbol{o}}(\boldsymbol{Z})(g_{a,\boldsymbol{o}}(\boldsymbol{Z}) / f_{a,\boldsymbol{o}}(\boldsymbol{Z}))$. The idea of the sampling methods described in this section is to obtain independent samples according to $f_{a,\boldsymbol{o}}$, use

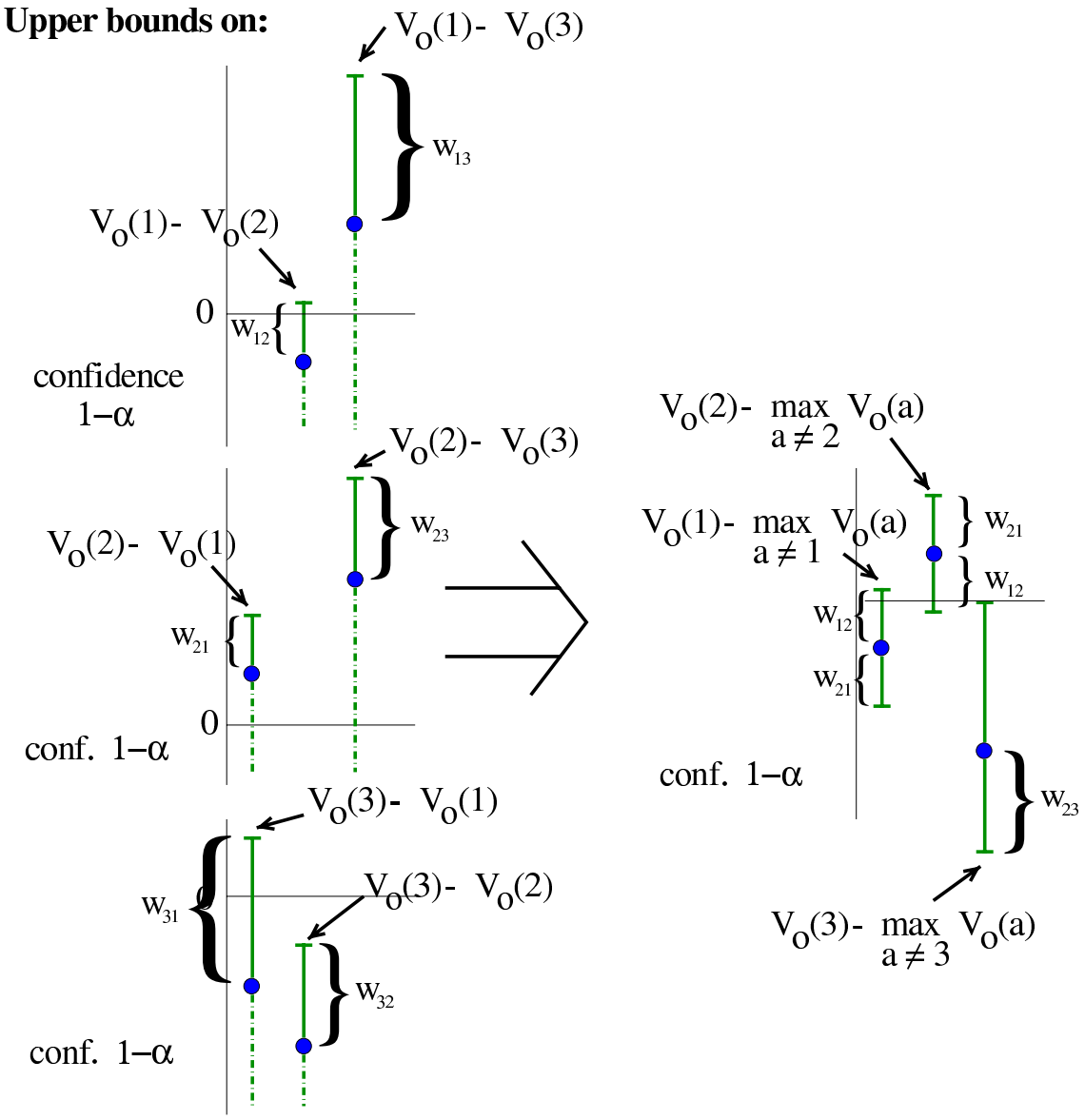


Figure 2.3: Graphical description of Hsu's multiple-bound lemma.

those samples to estimate the value of the actions, and finally select an approximately optimal action by taking the action with largest value estimate. Denote the *weight of a sample* $\boldsymbol{z}^{(i)}$ from $\boldsymbol{Z} \sim f_{a,\boldsymbol{o}}$ as $\omega_{a,\boldsymbol{o}}^{(i)} \equiv \omega_{a,\boldsymbol{o}}(\boldsymbol{z}^{(i)})$. Then an unbiased estimate of $V_{\boldsymbol{o}}(a)$ is $\hat{V}_{\boldsymbol{o}}(a) = \frac{1}{N_{a,\boldsymbol{o}}} \sum_{i=1}^{N_{a,\boldsymbol{o}}} \omega_{a,\boldsymbol{o}}^{(i)}$.

2.2.1 Traditional Method

We can obtain an estimate of $V_{\boldsymbol{o}}(a)$ using the straightforward method presented in Algorithm 1; it requires parameters $N_{a,\boldsymbol{o}}$ that will be defined in Theorem 4.

Algorithm 1 Traditional Method

1. Obtain independent samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(N_{a,o})}$ from $\boldsymbol{Z} \sim f_{a,\boldsymbol{o}}$.
2. Compute the weights $\omega_{a,\boldsymbol{o}}^{(1)}, \ldots, \omega_{a,\boldsymbol{o}}^{(N_{a,o})}$.
3. Output $\hat{V}_{\boldsymbol{o}}(a)$ = average of the weights.

This is the traditional sampling-based method used for action selection. However, the author is unaware of any previous result regarding the number of samples needed to obtain a near-optimal strategy with high probability using this method in this context.

The following small lemma is useful for the proof of the next theorem.

Lemma 3 *Let A and B be two events. If $A \Rightarrow B$, then* $\Pr\{B\} \geq \Pr\{A\}$.

Proof: Let $\overline{A}$ be the complement of event A. From, $A \Rightarrow B$, $\Pr\{B \mid A\} = 1$. Also,

$$\begin{aligned} \Pr\{B\} &= \Pr\{B \mid A\} \Pr\{A\} + \Pr\{B \mid \overline{A}\} \Pr\{\overline{A}\} \\ &\geq \Pr\{B \mid A\} \Pr\{A\} \\ &= \Pr\{A\}. \end{aligned}$$

□

Theorem 4 *If for each possible action $i = 1, \ldots, k$, we estimate $V_{\boldsymbol{o}}(i)$ using the traditional method, the weight function satisfies $l_{i,\boldsymbol{o}} \leq \omega_{i,\boldsymbol{o}}(\boldsymbol{Z}) \leq u_{i,\boldsymbol{o}}$, and the estimate uses*

$$N_{i,\boldsymbol{o}} = \left\lceil \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^2}{2\epsilon^2} \ln \frac{k}{\delta} \right\rceil$$

samples, then the action with the largest value estimate has a true value that is within

$$- \left[- \left(\hat{V}_{\boldsymbol{o}}(\hat{a}) - \max_{a \neq \hat{a}} \hat{V}_{\boldsymbol{o}}(a) - 2\epsilon \right)^{-} \right] \leq 2\epsilon$$

of the optimal with probability at least $1 - \delta$.

Proof sketch. The proof goes in three basic steps. First, we apply *Hoeffding's bounds* [Hoeffding, 1963] to obtain a bound on the probability that each estimate deviates from its true mean by some amount ϵ. Then, we apply the *Bonferroni's inequality (Union bound)* to obtain joint bounds on the probability that the difference of each estimate from all the others deviates from the true difference by 2ϵ. Finally, we apply Hsu's single-bound lemma to obtain the result.

Proof: By Hoeffding's bounds, for each i, if we use $N_{i,\boldsymbol{o}}$ as defined above, individually,

$$\Pr\{\hat{V}_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(i) \geq \epsilon\} \leq \frac{\delta}{k}$$

and

$$\Pr\{\hat{V}_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(i) \leq -\epsilon\} \leq \frac{\delta}{k}.$$

Using Bonferroni's inequality (Union bound), we can state for each i individually,

$$\Pr\left(\{\hat{V}_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(i) \geq \epsilon\} \bigcup_{j=1, j\neq i}^{k} \{\hat{V}_{\boldsymbol{o}}(j) - V_{\boldsymbol{o}}(j) \leq -\epsilon\}\right) \leq \delta.$$

This implies, for each i, individually,

$$\Pr\left(\{\hat{V}_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(i) < \epsilon\} \bigcap_{j=1, j\neq i}^{k} \{\hat{V}_{\boldsymbol{o}}(j) - V_{\boldsymbol{o}}(j) > -\epsilon\}\right) > 1 - \delta.$$

By Lemma 3, for each i, individually,

$$\Pr\{V_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(j) - (\hat{V}_{\boldsymbol{o}}(i) - \hat{V}_{\boldsymbol{o}}(j)) > -2\epsilon, \text{ for all } j = 1, \ldots, k, j \neq i\} > 1 - \delta.$$

Using Hsu's single-bound lemma with $\mu_i = V_{\boldsymbol{o}}(i)$, $\hat{\mu}_i = \hat{V}_{\boldsymbol{o}}(i)$ for all i, we obtain,

$$\Pr\{V_{\boldsymbol{o}}(i) - \max_{j\neq i} V_{\boldsymbol{o}}(j) \in [-(\hat{V}_{\boldsymbol{o}}(i) - \max_{j\neq i} \hat{V}_{\boldsymbol{o}}(j) - 2\epsilon)^{-}, (\hat{V}_{\boldsymbol{o}}(i) - \max_{j\neq i} \hat{V}_{\boldsymbol{o}}(j) + 2\epsilon)^{+}], \text{ for all } i = 1, \ldots, k\} \geq 1 - \delta.$$

Let $a^* = \text{argmax}_a V_{\boldsymbol{o}}(a)$, $\hat{a} = \text{argmax}_a \hat{V}_{\boldsymbol{o}}(a)$. From the last statement, $V_{\boldsymbol{o}}(a^*) - V_{\boldsymbol{o}}(\hat{a}) \leq -\left[-(\hat{V}_{\boldsymbol{o}}(\hat{a}) - \max_{j\neq i} \hat{V}_{\boldsymbol{o}}(j) - 2\epsilon)^{-}\right] \leq 2\epsilon$ with probability at least $1 - \delta$. □

A simple way to compute $l_{i,\boldsymbol{o}}$ and $u_{i,\boldsymbol{o}}$ is immediately from information local to each node in the graph. Assuming that we have non-negative utilities, we can let

$$u_{i,\boldsymbol{o}} = \left[\prod_{j=1}^{n_3} \max_{\text{Pa}(O_j)} P(O_j \mid \text{Pa}(O_j))|_{\boldsymbol{O}=\boldsymbol{o}}\right]\left[\sum_{j=1}^{m} \max_{\text{Pa}(U_j)} U_j(\text{Pa}(U_j))|_{\boldsymbol{O}=\boldsymbol{o}, A=i}\right], \quad (2.10)$$

$$l_{i,\boldsymbol{o}} = \left[\prod_{j=1}^{n_3} \min_{\text{Pa}(O_j)} P(O_j \mid \text{Pa}(O_j))|_{\boldsymbol{O}=\boldsymbol{o}}\right]\left[\sum_{j=1}^{m} \min_{\text{Pa}(U_j)} U_j(\text{Pa}(U_j))|_{\boldsymbol{O}=\boldsymbol{o}, A=i}\right]. \quad (2.11)$$

However, these bounds can be very loose. One can try to improve on those bounds by applying a two-stage estimation method, similar to that which we will present next, where we would first estimate the bounds on the weights. This idea was heuristically implemented by Cheng and Druzdzel [2001] in the context of approximate belief inference in BNs.

It is also important to note that as a result of using Bonferroni's inequality, the results allow us to share the random numbers used to generate samples among actions. Sharing of the random numbers among observations is also possible as well if we use Bonferroni's inequality to combine the statements for each observation to form a global bound or statement about the quality of the approximate strategy. Sharing of random numbers can be very effective for the problem of action selection since it creates (positive) correlations among the action value estimates yielding better estimates for the pairwise value differences between actions; that is, the difference estimates will have smaller variance than if completely independent samples had been used for each action (See Kahn and Marshall [1953] and Rubinstein [1981] and the references therein). All of this is allowed by the results of Theorem 4. As a matter of fact, if we obtained strictly independent samples for each action (and observation), the bounds above can be strictly improved [Hsu, 1996], but we believe the amount of improvement would be negligable compared to the variance-reduction effect of the positive correlations.

2.2.2 Two-stage Sequential Method

The sequential method tries to reduce the number of samples needed by the traditional method, using ideas from sequential analysis. The idea is to first obtain an estimate of the variance and then use it to compute the number of samples needed to estimate the mean. The method, presented in Algorithm 2, requires the parameters $N'_{a,\boldsymbol{o}}$ and the function $N''_{a,\boldsymbol{o}}(s)$, for $s > 0$, which will be defined in Theorem 6.

Algorithm 2 Sequential Method

1. Obtain independent samples $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(2N'_{a,\boldsymbol{o}})}$ from $\boldsymbol{Z} \sim f_{a,\boldsymbol{o}}$.
2. Compute the weights $\omega_{a,\boldsymbol{o}}^{(1)}, \ldots, \omega_{a,\boldsymbol{o}}^{(2N'_{a,\boldsymbol{o}})}$.
3. For $j = 1, \ldots, N'_{a,\boldsymbol{o}}$, let $y_j = (\omega_{a,\boldsymbol{o}}^{(2j-1)} - \omega_{a,\boldsymbol{o}}^{(2j)})^2/2$.
4. Compute $\hat{\sigma}^2_{a,\boldsymbol{o}}$ = average of y_j's.
5. Let $N_{a,\boldsymbol{o}}(\hat{\sigma}^2_{a,\boldsymbol{o}}) = 2N'_{a,\boldsymbol{o}} + N''_{a,\boldsymbol{o}}(\hat{\sigma}^2_{a,\boldsymbol{o}})$.
6. Obtain $N''_{a,\boldsymbol{o}}(\hat{\sigma}^2_{a,\boldsymbol{o}})$ new independent samples $\boldsymbol{z}^{(2N'_{a,o}+1)}, \ldots, \boldsymbol{z}^{(N_{a,o}(\hat{\sigma}^2_{a,o}))}$ from $\boldsymbol{Z} \sim f_{a,\boldsymbol{o}}$.
7. Compute the new weights $\omega_{a,\boldsymbol{o}}^{(2N'_{a,o}+1)}, \ldots, \omega_{a,\boldsymbol{o}}^{(N_{a,o}(\hat{\sigma}^2_{a,o}))}$.
8. Output $\hat{V}_{\boldsymbol{o}}(a)$ = average of the new weights.

Note that given the sequential nature of the method, the total number of samples is now a random variable (a function of the variance estimate). While two-stage sequential procedures of this kind are commonly used in the statistical literature, they all seem to be based on restricting assumptions on the distribution of the random variables (i.e., parametric families like normal and binomial distributions) [Bechhofer et al., 1995, Jennison and Turnbull, 2000].

The next theorem is the basis of the result on the number of samples needed by the sequential method. It provides a result on using the two-stage procedure described in Algorithm 2 for non-parametric estimation of the mean of bounded random variables. It provides distribution-free, absolute-error bounds on the estimates obtained using the sequential method. The theorem was originally presented by Ortiz and Kaelbling [2000b] and the proof by Ortiz [2000]. The statement of the theorem and the proof have been modified here for clarity.

Theorem 5 *Let $X \sim f$ be a random variable such that $X \in [l, u]$, $\mu = \mathrm{E}[X]$, $\sigma^2 = \mathrm{Var}[X]$. If for $0 < \epsilon < u - l, 0 < \delta < 1$*

1. *N' is defined as*

$$N' = \left\lceil (((u-l)/\epsilon)^{4/3}/2^{5/3}) \ln(2/\delta) \right\rceil,$$

2. *for $s > 0$,*

$$N''(s) = \left\lceil \left(2\left(s/\epsilon^2\right) + 2((u-l)/\epsilon)/3 + 2^{1/3}((u-l)/\epsilon)^{4/3}\right) \ln(2/\delta) \right\rceil,$$

3. *for $s > 0$, $N(s) = 2N' + N''(s)$,*

4. *for $i = 1, 2, \ldots, 2N'$, $X^{(i)} \sim f$, i.i.d. (*independent identically distributed*),*

$$\hat{\sigma}^2 = (1/N') \sum_{i=1}^{N'} \left(X^{(2i-1)} - X^{(2i)}\right)^2 /2,$$

5. *for $i = 2N' + 1, \ldots, N(\hat{\sigma}^2)$, $X^{(i)} \sim f$, i.i.d.,*

$$\hat{\mu} = (1/N''(\hat{\sigma}^2)) \sum_{i=2N'+1}^{N(\hat{\sigma}^2)} X^{(i)}$$

then the following statements hold:

1. $\Pr\{\hat{\mu} - \mu > \epsilon\} < \delta$,

2. $\Pr\{\hat{\mu} - \mu < -\epsilon\} < \delta$,

3. *with probability at least* $1 - \delta/2$,

$$\begin{aligned} N(\hat{\sigma}^2) &< \left(2(\sigma/\epsilon)^2 + 2((u-l)/\epsilon)/3 + 5((u-l)/\epsilon)^{4/3}/2^{2/3}\right)\ln(2/\delta) + 3 \\ &= O\left(\max\left(\frac{\sigma^2}{\epsilon^2}, \left(\frac{u-l}{\epsilon}\right)^{4/3}\right)\ln\frac{1}{\delta}\right), \end{aligned}$$

4. *the expected total number of samples is*

$$\begin{aligned} \mathrm{E}[N(\hat{\sigma}^2)] &< \left(2(\sigma/\epsilon)^2 + 2((u-l)/\epsilon)/3 + 3((u-l)/\epsilon)^{4/3}/2^{2/3}\right)\ln(2/\delta) + 3 \\ &= O\left(\max\left(\frac{\sigma^2}{\epsilon^2}, \left(\frac{u-l}{\epsilon}\right)^{4/3}\right)\ln\frac{1}{\delta}\right). \end{aligned}$$

Proof sketch. Instead of using Hoeffding's bounds to bound the probability that each estimate deviates from its true mean, we use a combination of *Bernstein's inequality* (as presented by Devroye et al. [1996] and credited to Bernstein [1946]) and Hoeffding's bounds as follows. We first use Hoeffding's bound to bound the probability that the estimate of the variance after taking $2N'$ samples deviates from the true variance by some amount ϵ'. We then use Bernstein's inequality to bound the probability that the estimate we obtain after taking additional $N''(\hat{\sigma}^2)$ samples deviates from its true mean by ϵ *given* that the true variance is no larger than our estimate of the variance plus ϵ'. We then find the value of ϵ' (in terms of ϵ) that minimizes the total number of samples $N = N''(\hat{\sigma}^2) + 2N'$. The results on the number of samples follow by substituting the minimizing ϵ' back into the expressions for $N''(\hat{\sigma}^2)$ and N'.

Proof: In order to simplify proof of the theorem, let us first introduce the following notation. The parameter ϵ_1 used throughout is related to the accuracy with which we estimate the variance. Let

1. for $0 < \epsilon_1 < (u-l)^2/2$,

$$N_1^l(\epsilon_1) = \frac{(u-l)^4}{8\epsilon_1^2}\ln\frac{2}{\delta}$$

be a lower bound on the number of *pairs* of samples we might use to estimate the variance as a function of the error parameter ϵ_1, $N_1(\epsilon_1) = \lceil N_1^l(\epsilon_1)\rceil$ be the *actual* number of samples we might use (after taking into account that it should be an integer) as a function of ϵ_1, and $N_1^u(\epsilon_1) = N_1^l(\epsilon_1) + 1 > N_1(\epsilon_1)$ be an upper bound on

the number of samples (We need the function N_1^l to define N_1^u, which in turn, we need to remove the discreteness on the "number of samples" so that we have a continuously differentiable function of ϵ_1 we can easily minimize);

2. for $s > 0$, $0 < \epsilon_1 < (u-l)^2/2$,

$$N_2^l(s, \epsilon_1) = \frac{2s + 2\epsilon_1 + 2\epsilon(u-l)/3}{\epsilon^2} \ln \frac{2}{\delta}$$

be the number of samples we might use to estimate the mean (as a function of ϵ_1), $N_2(s, \epsilon_1) = \lceil N_2^l(s, \epsilon_1) \rceil$ be the (*actual* number of samples, and $N_2^u(s, \epsilon_1) = N_2^l(s, \epsilon_1) + 1 > N_2(s, \epsilon_1)$ be an upper bound on the number of samples (We need the functions N_2^l and N_2^u for the same reasons as above);

3. for $s > 0$, $0 < \epsilon_1 < (u-l)^2/2$, $N(s, \epsilon_1) = 2N_1(\epsilon_1) + N_2(s, \epsilon_1)$ be the total number of samples used for the variance, and $N^u(s, \epsilon_1) = 2N_1^u(\epsilon_1) + N_2^u(s, \epsilon_1) > N(s, \epsilon_1)$ be an upper bound on the number.

4. for $i = 1, \ldots, 2N_1(\epsilon_1)$, $X^{(i)} \sim f$, independent, be the samples used to estimate the variance,

5. $Y^{(i)} = \left(X^{(2i-1)} - X^{(2i)}\right)^2/2$, for $i = 1, \ldots, N_1(\epsilon_1)$, be a function on the difference of the individual sample pairs (i.e., a random variable whose expectation is σ^2)

6. $\hat{\sigma}^2 = (1/N_1(\epsilon_1)) \sum_{i=1}^{N_1(\epsilon_1)} Y^{(i)}$ be the estimator of the variance,

7. for $i = 2N_1(\epsilon_1)+1, \ldots, N(\hat{\sigma}^2, \epsilon_1)$, $X^{(i)} \sim f$, independent, be the samples used for the mean,

8. $\hat{\mu} = (1/N_2(\hat{\sigma}^2, \epsilon_1)) \sum_{i=2N_1(\epsilon_1)+1}^{N(\hat{\sigma}^2, \epsilon_1)} X^{(i)}$ be the estimator of the mean.

Note that $0 < Y^{(i)} < (u-l)^2/2$ for $i = 1, \ldots, N_1(\epsilon_1)$. Also, since the $Y^{(i)}$, for $i = 1, \ldots, N_1(\epsilon_1)$ are independent, and $E[Y^{(i)}] = \sigma^2$, $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$.

By Hoeffding's bound,

$$\Pr\{\hat{\sigma}^2 - \sigma^2 \geq \epsilon_1\} \leq \frac{\delta}{2},$$

and

$$\Pr\{\hat{\sigma}^2 - \sigma^2 \leq -\epsilon_1\} \leq \frac{\delta}{2}.$$

By Bernstein's inequality,

$$\Pr\{\hat{\mu} - \mu \geq \epsilon \mid \hat{\sigma}^2 - \sigma^2 > -\epsilon_1\} \leq \frac{\delta}{2},$$

and

$$\Pr\{\hat{\mu} - \mu \leq -\epsilon \mid \hat{\sigma}^2 - \sigma^2 > -\epsilon_1\} \leq \frac{\delta}{2}.$$

Now note that, for any $\epsilon_1 > 0$,

$$\begin{aligned} \Pr\{\hat{\mu} - \mu \geq \epsilon\} &\leq \Pr\{\hat{\mu} - \mu \geq \epsilon \mid \hat{\sigma}^2 - \sigma^2 > -\epsilon_1\} + \Pr\{\hat{\sigma} - \sigma^2 \leq -\epsilon_1\} \\ &< \delta, \end{aligned}$$

and

$$\begin{aligned} \Pr\{\hat{\mu} - \mu \leq -\epsilon\} &\leq \Pr\{\hat{\mu} - \mu \leq -\epsilon \mid \hat{\sigma}^2 - \sigma^2 > -\epsilon_1\} + \Pr\{\hat{\sigma} - \sigma^2 \leq -\epsilon_1\} \\ &< \delta. \end{aligned}$$

Let us select ϵ_1 which minimizes (an upper bound on) the sum of the "total" number of samples. That is, let $\epsilon_1^* = \text{argmin}_{\epsilon_1} N^u(s, \epsilon_1)$. Then, after taking the derivative of $N^u(s, \epsilon_1)$ with respect to ϵ_1 and setting to 0, we note that ϵ_1^* satisfies the equation

$$\frac{2}{\epsilon^2} + 2\frac{(u-l)^4}{8}\left(\frac{-2}{\epsilon_1^{*3}}\right) = 0.$$

Solving for ϵ_1^* we obtain that

$$\epsilon_1^* = \left(\frac{1}{4}(u-l)^4\epsilon^2\right)^{\frac{1}{3}}.$$

Since $\partial^2 N^u(s, \epsilon_1)/(\partial\epsilon_1)^2 > 0$, ϵ_1^* is indeed a minimum of $N^u(s, \epsilon)$ with respect to ϵ_1. Letting $N' = N_1(\epsilon_1^*)$, and $N''(\hat{\sigma}^2) = N_2(\hat{\sigma}^2, \epsilon_1^*)$ we obtain the first two conclusions of the theorem.

For the last two conclusions of the theorem, note that since

$$\Pr\{\hat{\sigma}^2 - \sigma^2 < \epsilon_1\} > 1 - \frac{\delta}{2},$$

from Lemma 3,

$$\Pr\left\{N_2^u(\hat{\sigma}^2, \epsilon_1) < \left(\frac{2\sigma^2 + 4\epsilon_1 + 2\epsilon(u-l)/3}{\epsilon^2}\right)\ln\frac{2}{\delta} + 1\right\} > 1 - \frac{\delta}{2}.$$

Also, since $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$, and because N_2^u is linear in its first parameter,

$$\mathrm{E}[N_2^u(\hat{\sigma}^2, \epsilon_1)] = \left(\frac{2\sigma^2 + 2\epsilon_1 + 2\epsilon(u-l)/3}{\epsilon^2}\right)\ln\frac{2}{\delta} + 1.$$

Noting that $N(\hat{\sigma}^2) = N(\hat{\sigma}^2, \epsilon_1^*) < N^u(\hat{\sigma}^2, \epsilon_1^*)$ yields the last two conclusions of the theorem.

□

Note that in the process described in the previous theorem, we are not re-using the samples from $\hat{\sigma}^2$ in the estimate for $\hat{\mu}$. This is because conditioned on knowing $\hat{\sigma}^2$, the first set of samples used for $\hat{\sigma}^2$ become dependent. Hence, Bernstein's inequality does not hold anymore, and we would have to use another large-deviation bound to apply for $\hat{mu}$. At this point, I do not know of any bound that would apply in such a case.

We now use the last theorem, combined with Hsu's single-bound lemma, to show in the next theorem, that the two-stage sequential method can reduce the total number of samples taken by the traditional method (both with high probability and in expectation).

Theorem 6 *If, for each possible action $i = 1, \ldots, k$, we estimate $V_{\boldsymbol{o}}(i)$ using the sequential method, the weight function satisfies $l_{i,\boldsymbol{o}} \leq \omega_{i,\boldsymbol{o}}(\boldsymbol{Z}) \leq u_{i,\boldsymbol{o}}$, $\sigma^2_{i,\boldsymbol{o}} = Var[\omega_{i,\boldsymbol{o}}(\boldsymbol{Z})]$,*

$$N'_{i,\boldsymbol{o}} = \left\lceil \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^{4/3}}{2\ 2^{2/3}\epsilon^{4/3}} \ln \frac{2k}{\delta} \right\rceil,$$

and

$$N''_{i,\boldsymbol{o}}(\hat{\sigma}^2_{i,\boldsymbol{o}}) = \left\lceil \left(\frac{2\hat{\sigma}^2_{i,\boldsymbol{o}} + 2(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})\epsilon/3}{\epsilon^2} + 2^{1/3} \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^{4/3}}{\epsilon^{4/3}} \right) \ln \frac{2k}{\delta} \right\rceil,$$

then the action $\hat{a}$ with the largest value estimate has a true value that is within

$$-\left[-\left(\hat{V}_{\boldsymbol{o}}(\hat{a}) - \max_{a \neq \hat{a}} \hat{V}_{\boldsymbol{o}}(a) - 2\epsilon \right)^{-} \right] \leq 2\epsilon$$

of the optimal with probability at least $1 - \delta$. Also,

$$\begin{aligned} N_{i,\boldsymbol{o}}(\hat{\sigma}^2_{a,\boldsymbol{o}}) &< \left(\frac{2\sigma^2_{i,\boldsymbol{o}} + 2(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})\epsilon/3}{\epsilon^2} + \frac{5}{2^{2/3}} \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^{4/3}}{\epsilon^{4/3}} \right) \ln \frac{2k}{\delta} + 3 \\ &= O\left(\max\left(\frac{\sigma^2_{i,\boldsymbol{o}}}{\epsilon^2}, \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^{4/3}}{\epsilon^{4/3}} \right) \ln \frac{k}{\delta} \right), \end{aligned}$$

with probability at least $1 - \delta/(2k)$, and

$$\begin{aligned} E[N_{i,\boldsymbol{o}}(\hat{\sigma}^2_{a,\boldsymbol{o}})] &< \left(\frac{2\sigma^2_{i,\boldsymbol{o}} + 2(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})\epsilon/3}{\epsilon^2} + \frac{3}{2^{2/3}} \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^{4/3}}{\epsilon^{4/3}} \right) \ln \frac{2k}{\delta} + 3 \\ &= O\left(\max\left(\frac{\sigma^2_{i,\boldsymbol{o}}}{\epsilon^2}, \frac{(u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^{4/3}}{\epsilon^{4/3}} \right) \ln \frac{k}{\delta} \right). \end{aligned}$$

Proof sketch. The only difference from the proof of Theorem 1 is the first step. We use the result from Theorem 5 for estimating the values of $V_{\boldsymbol{o}}(i)$ for all i. Steps 2 and 3 are as in Theorem 4.

Proof: Using Theorem 5, for each i, if we use in the sequential method $N'_{i,\boldsymbol{o}}$ and $N''_{i,\boldsymbol{o}}(\hat{\sigma}^2_{i,\boldsymbol{o}})$ as defined above, individually,

$$\Pr\{\hat{V}_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(i) \geq \epsilon\} \leq \frac{\delta}{k}$$

and

$$\Pr\{\hat{V}_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(i) \leq -\epsilon\} \leq \frac{\delta}{k}.$$

The rest of the proof is as in Theorem 4. □

The sequential method is particularly more effective than the traditional method when $\sigma^2_{i,\boldsymbol{o}} \ll (u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^2$ and ϵ is small enough. We will observe this improvement in the empirical studies.

Note that we can derive similar results bounding the number of samples for methods that concentrate on estimating pairwise differences $V_{\boldsymbol{o}}(i) - V_{\boldsymbol{o}}(j)$, for all $i, j, i \neq j$ instead of individual values $V_{\boldsymbol{o}}(i)$ for all i. These bounds can be better if the upper and lower bounds on the difference in weights is smaller than those on the individual weights. Note that we have to compute $O(k^2)$ difference estimates but we only need to obtain a given number of samples (batch) for each action since we can reuse those samples to obtain our estimates. For non-sequential methods like the estimation-based traditional method above, we can determine before-hand whether a method based on differences requires fewer samples than one based on individual estimations.

We also note the large-deviation bounds used here are the simplest to work with. For instance, Hoeffding [1963] presents other tighter but more involved bounds, which typically require unknown quantities like the mean or the variance. One could try to use an idea similar to that used for the two-stage sequential method to estimate the unknown parameters and use the estimates in the tighter bound but this is left for future study.

2.3 Comparison-based Method

Using the results from MCB, we can compute simultaneous or joint confidence intervals on the difference between the value of $V_{\boldsymbol{o}}(a)$ and the best of all the others, for all actions a. Therefore, MCB allows us to select the best action choice or an action with value close to it, within a confidence level.

The methods presented in the previous section required that we have estimates with the same precision in order to select a good action. Hsu's multiple-bound lemma applies when

we do not have estimates of $V_{\boldsymbol{o}}(a)$ for each a with the same precision. Based on this result, we propose the method presented in Algorithm 3 for action selection.

Algorithm 3 Comparison-based Method (General description)

1. Obtain an *initial number of samples* for each action a.
2. Compute *MCB confidence intervals* on the difference in value of each action from the best of the other actions using those samples.

while *not able to select a good action with high certainty* **do**

3(a). Obtain *additional samples.*

3(b). Recompute MCB confidence intervals using total samples so far.

end while

Let us briefly expand on the stopping condition used for the while statement in the algorithm above. Let us assume that we have constructed the MCB confidence intervals correctly with confidence at least $1 - \delta$. If there is an action choice for which the lower bound of its corresponding MCB confidence interval is greater or equal to -2ϵ, then we can say that should we select that action, its true value will be no less than 2ϵ from the optimal with probability at least $1 - \delta$. Hence, the MCB confidence intervals provides us with the right information on when to stop to obtain a given approximation guarantee. Following, we will be more specific about how we can build the MCB confidence intervals, and hence, turn the description above into an actual method with the same theoretical guarantees on the quality of the solution as the previous two estimation-based methods presented.

2.3.1 Formalization and analysis

The method above is really a multi-stage group sequential method [Jennison and Turnbull, 2000]. Hence, the analysis below is similar to that presented in the statistical literature on group sequential methods, now also combined with work on MCB. In what follows, we index *stages* by t ($t = 1, 2, \dots$). The discussion is made somewhat general with regard to the allocation of the confidence parameters $0 < \delta_{ij}^{(t)} < 1$ and the number of samples per stage $N_i^{(t)}$, where i, j stands for the pair of actions i and j, $i \neq j$, and the superscript denotes the stage (or iteration) t. (For the purpose of this discussion, it might help to think of $\delta_{ij}^{(t)}$ as the maximum allowed probability that the estimate of the difference in value of action i to that of j at stage t deviates a large amount from the true value.) We also assume that the allocation of the confidence parameters is *sensible*; that is, the resulting confidence statements have probabilistic meaning—the confidence probabilities all lie between 0 and 1 and preferably are at least larger then 0.5.

To simplify notation, assume we are considering a particular (fixed) observation $\boldsymbol{o}$, and

let

1. k be the total number of action choices,
2. $\mu_i \equiv V(\boldsymbol{o}, i)$ be the true (unnormalized) value of action i,
3. $\overline{X}_i \equiv \hat{V}(\boldsymbol{o}, i)$ be our (importance-sampling) estimate of μ_i (the average of the sample weights for action i),
4. $\mu^* \equiv \max_{i=1,\ldots,k} \mu_i$ be the value of of the best action,
5. $N_i^{(1:t)} = \sum_{l=1}^{t} N_i^{(l)}$ be the *cumulative* number of samples at the end of sampling stage t for action i (and observation $\boldsymbol{o}$),
6. $\overline{X}_i^{(t)} = (1/N_i^{(1:t)}) \sum_{l=1}^{N_i^{(1:t)}} \omega_{\boldsymbol{o},\boldsymbol{i}}^{(l)}$ be the value estimate of action i using all the samples taken until stage t,
7. $\delta_i^{(t)} = \sum_{j=1, j\neq i}^{k} \delta_{ij}^{(t)}$ (for the purpose of this discussion, it might help to think of $\delta_i^{(t)}$ as the maximum allowed probability that the estimate of the difference in value of action i to that of j at stage t deviates a large amount from their true value, for any $j \neq i$),
8. $\delta^{(t)} = \max_{i=1,\ldots,k} \delta_i^{(t)}$ (i.e., $1 - \delta^{(t)}$ would be the confidence we have on the MCB intervals *individually* for each stage t), and
9. $\delta^{(1:t)} = \sum_{l=1}^{t} \delta^{(t)} < 1$ ($1-\delta^{(1:t)}$ would be the confidence we have on the MCB intervals, *jointly*, for any stage form 1 to t).

Assume that the sequence $N_i^{(t)}$ is fixed in advance. (We will later discuss the most general case in which the allocation of the number of samples changes as a function of the outcomes so far – this is known in the clinical trials community as *data dependent timing of analysis*.)

For each pair i, j. $i \neq j$, given $\delta_{ij}^{(1)}$, $N_i^{(1)}$ and $N_j^{(1)}$, by Hoeffding's traditional bound and Bonferroni's inequality, there exists *error widths* $w_{ij}^{(1)} > 0$ such that individually for each i,

$$\Pr\left\{\overline{X}_i^{(1)} - \overline{X}_j^{(1)} - (\mu_i - \mu_j) > -w_{ij}^{(1)}, \forall j \neq i\right\} \geq 1 - \delta_i^{(1)} \tag{2.12}$$

holds. Although not directly denoted for the sake of simple notation, the $w_{ij}^{(1)}$ are functions of $N_i^{(1)}$, $N_j^{(1)}$, and $\delta_{ij}^{(1)}$ (as well as other quantities based on $\omega_{\boldsymbol{o},i}$, but let us ignore that for now). From the last statement, by Hsu's multiple-bound lemma, there exists confidence bounds $D_i^{-(1)}$ such that

$$\Pr\left\{\mu_i - \max_{j\neq i} \mu_j \in [D_i^{-(1)}, D_i^{+(1)}], \forall i\right\} \geq 1 - \delta^{(1)}. \tag{2.13}$$

Although not explicitly stated, the confidence intervals are a function of the estimators $\overline{X}_i^{(1)}$, and the precisions $w_{ij}^{(1)}$. (Recall that the precisions are in turn functions of the number of samples $N_i^{(1)}$ and the confidence parameters $\delta_{ij}^{(1)}$.) Note that the last confidence interval expression is for *simultaneous* or *joint* confidence intervals [Hsu, 1996].

Therefore, if we stop at the end of stage 1 and select action $\hat{a}^{(1)}$, then the *true* value of this action $\mu_{\hat{a}}^{(1)}$ is guaranteed to satisfy the inequality $\mu_{\hat{a}^{(1)}} - \mu^* > D_i^{-(1)}$ (i.e., have value no more than $-D_i^{-(1)}$ less than the best value μ^*), with probability at least $1 - \delta^{(1)}$. If we had to select at the end of this stage and we want to cover for the worst case lost, we would select the action for which $D_i^{-(1)}$ is largest (over i).

If we are not satisfied with the precisions obtained at the end of the first stage, we take additional samples during a second stage. Let us ignore for now the fact that we have looked at the data and realized that it does not guarantee us enough precision to stop after a single stage. In other words, think that we actually did not have a first stage at all. In that case, given $\delta_{ij}^{(2)}$, $N_i^{(1:2)}$, and $N_j^{(1:2)}$, by Hoeffding's (traditional) bound (*because all the samples are independent*) and Bonferroni's inequality, there exists $w_{ij}^{(2)}$ such that, at the end of the *second* stage, individually for each i,

$$\Pr(\overline{X}_i^{(2)} - \overline{X}_j^{(2)} - (\mu_i - \mu_j) > -w_{ij}^{(2)}, \forall j \neq i) \geq 1 - \delta_{ij}^{(2)} \tag{2.14}$$

and by Hsu's multiple-bound lemma

$$\Pr(\mu_i - \max_{j \neq i} \mu_j \in [D_i^{-(2)}, D_i^{+(2)}], \forall i) \geq 1 - \max_{i=1,\ldots,k} \delta_i^{(2)} = 1 - \delta^{(2)}. \tag{2.15}$$

(Again $D_i^{-(2)}$ and $D_i^{+(2)}$ are random functions since they will depend on $\overline{X}_i^{(2)}$). From the last statement, we would like to infer that if we select action $\hat{a}^{(2)}$ at the end of stage 2, if $D_{\hat{a}}^{-(2)}$ is sufficiently large, then this action will be *good* with high probability (i.e., $\mu_{\hat{a}^{(2)}} \geq \mu^* + D_{\hat{a}^{(2)}}^{-(2)}$ with probability at least $1 - \delta^{(2)}$). However this would only be true if we had decided to sample until the end of second stage right from the start. To make our statements on the selected action of the multi-stage process independent of when we stop, we need to correct for the fact that we *looked* at the data at the end of the first stage to determine whether we needed to continue to a second stage. This is known in statistics as the *multiple-looks* problem (See for instance Anscombe [1954], Armitage et al. [1969], Jennison and Turnbull [2000] and the references therein) [1]. We can fix this by making *simultaneous* confidence intervals for all stages. We can do that by applying Bonferroni's

[1] Actually, Anscombe [1954] informally suggests general conditions under which this problem is and is not significant (Pages 99-100).

inequality to the statements for each stage. In the case of 2 stages, we have

$$\Pr\left\{\mu_i - \mu_j < \overline{X}_i^{(t)} - \overline{X}_j^{(t)} + w_{ij}^{(t)}, \forall j \neq i, t = 1, 2\right\} = \Pr\left\{\mu_i - \mu_j < \min_{t=1,2} \overline{X}_i^{(t)} - \overline{X}_j^{(t)} + w_{ij}^{(t)}, \forall j \neq i\right\} \geq 1 - \delta^{(1:t)}.$$

This implies, by Hsu's multiple-bound lemma, that there exist $D_i^{-(1:2)}$ and $D_i^{+(1:2)}$, for $i = 1, \ldots, k$ and $t = 1, 2$ (maybe different from the ones used previously), such that

$$\Pr\left\{\mu_i - \max_{j \neq i} \mu_j \in [D_i^{-(1:2)}, D_i^{+(1:2)}], \forall i\right\} \geq 1 - \delta^{(1:2)}.$$

We can apply this analysis to any arbitrary final stage t or even to an infinite stage setting, for appropriate allocation of $\delta_{ij}^{(t)}$, as we will soon see. Let us now describe how we *really* compute $D_i^{-(1:t)}$ and $D_i^{+(1:t)}$ given $\overline{X}_i^{(l)}$, and $w_{ij}^{(l)}$, for $l = 1, \ldots, t$, to take as much advantage of the theory as we can. (We will see soon how to set $w_{ij}^{(t)}$.) Let

$$Y_{ij}^{(t)} = \overline{X}_i^{(t)} - \overline{X}_j^{(t)} + w_{ij}^{(t)} \tag{2.16}$$

be the (confidence) upper bound obtained *individually* at stage t on the true difference between the value of action i and j,

$$Y_{ij}^{(1:t)} = \min_{l=1,\ldots,t} Y_{ij}^{(l)} \tag{2.17}$$

be the *best* (confidence) upper bound obtained up to stage at stage t on the true difference between the value of action i and j [2],

$$D_i^{+(1:t)} = \left(\min_{j \neq i} Y_{ij}^{(1:t)}\right)^+ \tag{2.18}$$

be the MCB-confidence-interval upper bound for action i up to stage t,

$$\mathcal{G}^{(t)} = \left\{i : D_i^{+(1:t)} > 0\right\} \tag{2.19}$$

be the set of possible best actions at time t (their corresponding MCB-confidence-interval upper bound is positive), and

$$D_i^{-(1:t)} = \begin{cases} 0 & \text{if} \mathcal{G}^{(t)} = \{i\} \\ -\left(\min_{j \neq i} -Y_{ij}^{(1:t)}\right)^- & \text{otherwise} \end{cases} \tag{2.20}$$

[2] By *best* here we mean the smallest upper bound of all those obtained up to stage t.

be the MCB-confidence-interval lower bound up to stage t. Typically, $Y_{ij}^{(t)} = -Y_{ij}^{(t)}$, but it need not be in general. We note that the upper bounds on the value differences $(Y_{ij}^{(1:t)})$ form a monotonically non-increasing sequence of t. From this we can show that the MCB confidence intervals form a sequence of sets (in this case, the sets are actually intervals indexed by t) $I_t = [D_i^{-(1:t)}, D_i^{+(1:t)}]$ that *decreases* to $\bigcap_t I_t$ (denoted $I_t \nearrow \bigcap_t I_t$; see Wheeden and Zygmund [1977]). We believe this is interesting because it helps us monitor the *behavior* of the MCB confidence intervals in practice. Given the above definitions, if there exists $w_{ij}^{(t)}$ (which in this case we can find by Hoeffding's traditional bound and Bonferrni's inequality) such that at each stage t we can guarantee

$$\Pr\left\{\mu_i - \mu_j < \overline{X}_i^{(l)} - \overline{X}_j^{(l)} + w_{ij}^{(l)}, \forall j \neq i, l = 1, \ldots, t\right\} = \Pr\left\{\mu_i - \mu_j < Y_{ij}^{(1:t)}, \forall j \neq i\right\} \geq 1 - \delta^{(1:t)}, \quad (2.21)$$

then, by Hsu's multiple-bound lemma, at each stage t we can guarantee

$$\Pr\left\{\mu_i - \max_{j \neq i} \mu_j \in [D_i^{-1:t}, D_i^{+1:t}], \forall i\right\} \geq 1 - \delta^{(1:t)}. \quad (2.22)$$

From this we derive a version of the comparison-based method: continue sampling and computing the MCB confidence intervals until one of the MCB confidence lower-bounds jumps above some prescribed precision -2ϵ for $\epsilon > 0$. Algorithm 4 describes the general framework for the comparison-based method. The algorithm requires the *error width function* W which we will shortly describe, since it will depend on the particular instantiation of the general framework.

Now, we can make the following statement about the comparison-based method presented above.

Theorem 7 *Given the respective definitions above, if the* $w_{ij}^{(t)}$ *are* proper *(i.e. satisfy condition 2.21 above) and the comparison-based method (as presented in Algorithm 4) stops at some stage (stopping time)* $\tau > 1$*, its output action* $\hat{a}$ *is such that*

$$V(a^*, \boldsymbol{o}) - V(\hat{a}, \boldsymbol{o}) \leq -D_{\hat{a}}^{(\tau)} \leq 2\epsilon$$

with probability at least $1 - \delta^{(1:\tau)}$.

Proof: Refer to the discussion above. □

We are typically interested in the case that $\delta^{(1:\tau)} < \delta$ for some prescribed $0 < \delta < 1/2$, sufficiently small (say $\delta < 0.1$). We will now see how to achieve this.

We now consider some example instantiations of the general framework. The instantiations presented here are by no mean exhaustive.

Algorithm 4 Comparison-based Method (General algorithmic framework)

Input:

1. for $t = 1, 2, \ldots$, for all $i, j \in \{1, \ldots, k\}$, $i \neq j$, $0 < \delta_{ij}^{(t)} < 1$ (fixed),
2. for $t = 1, 2, \ldots$, for $i = 1, \ldots, k$, $N_i^{(t)} \geq 0$ (fixed),
3. $\omega_{i,\boldsymbol{o}}$ (weight function),
4. W (error width function), and
5. $\epsilon > 0$.

$t \leftarrow 0$
$\forall i, S_i^{(1:t)} \leftarrow 0, N_i^{(t)} \leftarrow 0, D_i^{-(1:t)} \leftarrow -\infty, D_i^{-(1:t)} \leftarrow +\infty$
$\forall i, j, i \neq j, B_{ij}^{(1:t)} \leftarrow +\infty$
$\forall i, j, i \neq j,$

$$w_{ij}^{(t)} \leftarrow W(N_i^{(1:t)}, N_j^{(1:t)}, \delta_{ij}^{(t)}, \omega_{i,\boldsymbol{o}}, \omega_{j,\boldsymbol{o}})$$

We are precomputing the error widths to emphasize that they are fixed by the fixed number of samples per stage and precision allocations. There is no need to do this in practice.
$\hat{a} \leftarrow 1$
while $D_{\hat{a}}^{-(1:t)} < -2\epsilon$ **do**
$t \leftarrow t + 1$
$\forall i$, obtain $N_i^{(t)}$ new *i.i.d.* samples for each action i
and compute their weights $\omega_{i,\boldsymbol{o}}^{(l)}$ for $l = (N_i^{(1:(t-1))} + 1), \ldots, N_i^{(1:t)}$.
$\forall i,$

$$S_i^{(t)} \leftarrow \sum_{l=N_i^{(1:(t-1))}+1}^{N_i^{(1:t)}} \omega_{i,\boldsymbol{o}}^{(l)},$$

$$S_i^{(1:t)} \leftarrow S_i^{(1:(t-1))} + S_i^{(t)},$$

$$\overline{X}_i^{(t)} \leftarrow S_i^{(1:t)} / N_i^{(1:t)}$$

$\forall i, j, i \neq j,$

$$Y_{ij}^{(t)} \leftarrow \overline{X}_i^{(t)} - \overline{X}_j^{(t)} + w_{ij}^{(t)},$$

$$B_{ij}^{(t)} \leftarrow \min(Y_{ij}^{(t)}, B_{ij}^{(t-1)})$$

$\forall i, D_i^{+(1:t)} \leftarrow \min((\min_{j \neq i} B_{ij}^{(t)})^+, D_i^{+(1:(t-1))})$
$\mathcal{G}^{(t)} \leftarrow \{i : D_i^{+(1:t)} > 0\}$
$\forall i,$

$$D_i^{-(1:t)} \leftarrow \begin{cases} 0 & \text{if } \mathcal{G}^{(t)} = \{i\} \\ \max\left(-\left(\min_{j \in \mathcal{G}^{(t)}, j \neq i} -B_{ji}^{(t)}\right)^-, D_i^{-(1:(t-1))}\right) & \text{otherwise.} \end{cases}$$

$\hat{a} \rightarrow \text{argmax}_i D_i^{(1:t)}$
end while
Output: $\hat{a}$, $-D_{\hat{a}}^{-(1:t)}$, $1 - \delta^{(1:t)}$.

Algorithm 5 Comparison-based Method with Adaptive Allocations (General algorithmic framework): The differences from the comparison-based method in Algorithm 4 are emphasized.

Input:

1. for $t = 1, 2, \ldots$, for all $i, j \in \{1, \ldots, k\}$, $i \neq j$, $0 < \delta_{ij}^{(t)} < 1$ (fixed), *Note that now only* $N_i^{(1)}$ *is fixed.*
2. *for* $i = 1, \ldots, k$, $N_i^{(1)} \geq 0$ *(fixed),*
3. $\omega_{i,\boldsymbol{o}}$ (weight function),
4. W (error width function),
5. M *(number-of-samples allocation function)*, and
6. $\epsilon > 0$.

$t \leftarrow 0$
$\forall i, S_i^{(1:t)} \leftarrow 0, N_i^{(t)} \leftarrow 0, D_i^{-(1:t)} \leftarrow -\infty, D_i^{-(1:t)} \leftarrow +\infty$
$\forall i, j, i \neq j, B_{ij}^{(1:t)} \leftarrow +\infty$
$\hat{a} \leftarrow 1$
while $D_{\hat{a}}^{-(1:t)} < -2\epsilon$ **do**
$t \leftarrow t + 1$
$\forall i$, obtain $N_i^{(t)}$ new *i.i.d.* samples for each action i and compute their weights $\omega_{i,\boldsymbol{o}}^{(l)}$ for $l = N_i^{(1:(t-1))} + 1, \ldots, N_i^{(1:t)}$.
$\forall i$,

$$S_i^{(t)} \leftarrow \sum_{l=N_i^{(1:(t-1))}+1}^{N_i^{(1:t)}} \omega_{i,\boldsymbol{o}}^{(l)},$$

$$S_i^{(1:t)} \leftarrow S_i^{(1:(t-1))} + S_i^{(t)},$$

$$\overline{X}_i^{(t)} \leftarrow S_i^{(1:t)} / N_i^{(1:t)}$$

$\forall i, j, i \neq j$,

$$w_{ij}^{(t)} \leftarrow W(N_i^{(1:t)}, N_j^{(1:t)}, \delta_{ij}^{(t)}, \omega_{i,\boldsymbol{o}}, \omega_{j,\boldsymbol{o}})$$

Note that now we cannot precompute the error widths, since they depend on the adaptive allocation of the number of samples.
$\forall i, j, i \neq j$,

$$Y_{ij}^{(t)} \leftarrow \overline{X}_i^{(t)} - \overline{X}_j^{(t)} + w_{ij}^{(t)},$$

$$B_{ij}^{(t)} \leftarrow \min(Y_{ij}^{(t)}, B_{ij}^{(t-1)})$$

$\forall i, D_i^{+(1:t)} \leftarrow \left(\min_{j \neq i} B_{ij}^{(t)}\right)^+$
$\mathcal{G}^{(t)} \leftarrow \{i : D_i^{+(1:t)} > 0\}$
$\forall i$,

$$D_i^{-(1:t)} \leftarrow \begin{cases} 0 & \text{if } \mathcal{G}^{(t)} = \{i\} \\ -\left(\min_{j \in \mathcal{G}^{(t)}, j \neq i} -B_{ji}^{(t)}\right)^- & \text{otherwise.} \end{cases}$$

$\hat{a} \leftarrow \operatorname{argmax}_i D_i^{-(1:t)}$
$(N_1^{(t)}, \ldots, N_k^{(t)}) \leftarrow M(D_1^{-(1:t)}, D_1^{+(1:t)}, \ldots, D_k^{-(1:t)}, D_k^{+(1:t)})$.
end while
Output: $\hat{a}$, $-D_{\hat{a}}^{-(1:t)}$, $1 - \delta^{(1:t)}$.

Finite number of stages and uniform allocation of confidence parameters

Let us say we want to stop on or before considering a maximum of $T^{\max}$ stages. One (of many) alternatives for setting the number of samples per stage for each action while keeping approximation guarantees results from using the traditional version of Hoeffding's bounds to compute $w_{ij}(t)$. Let

$$\delta_{ij}^{(t)} = \delta/((k-1)T^{\max}) \tag{2.23}$$

$$l_i \leq \omega_{\boldsymbol{o},i}(\boldsymbol{Z}) \leq u_i,$$

$$N_i^{(t)} = \left\lceil \frac{(u_i - l_i)^2}{2\epsilon^2 T^{\max}} \ln \frac{kT^{\max}}{\delta} \right\rceil, \tag{2.24}$$

$$\begin{aligned} W(N_i, N_j, \delta_{ij}, \omega_{i,\boldsymbol{o}}, \omega_{j,\boldsymbol{o}}) &= \left((u_i - l_i)\sqrt{\frac{1}{N_i}} + (u_j - l_j)\sqrt{\frac{1}{N_j}} \right) \sqrt{\frac{1}{2} \ln \frac{k}{\delta_{ij}(k-1)}} \\ &= \left((u_i - l_i)\sqrt{\frac{1}{N_i}} + (u_j - l_j)\sqrt{\frac{1}{N_j}} \right) \sqrt{\frac{1}{2} \ln \frac{kT^{\max}}{\delta}}. \end{aligned} \tag{2.25}$$

Note that by this definition of the error width function the error widths monotonically decrease with t as $O(1/\sqrt{t})$.

Theorem 8 *Consider the instantiation of the comparison-based method resulting from executing Algorithm 4 with assignments for input $\delta_{ij}^{(t)}, N_i^{(t)}$, and W given by 2.23, 2.24, and 2.25, respectively. This instantiation will stop at some time $\tau \in \{1, \ldots, T^{\max}\}$. Furthermore, its output is such that the action $\hat{a}$ satisfies*

$$V(a^*, \boldsymbol{o}) - V(\hat{a}, \boldsymbol{o}) \leq -D_{\hat{a}}^{-(\tau)} \leq 2\epsilon$$

with probability at least $1 - \delta^{(1:\tau)} = 1 - \tau\delta/T^{\max} \geq 1 - \delta$.

Proof: Let

$$w_i^{(t)} = (u_i - l_i)\sqrt{\frac{1}{2N_i^{(1:t)}} \ln \frac{kT^{\max}}{\delta}}$$

and

$$w_{ij}^{(t)} = w_i^{(t)} + w_j^{(t)}.$$

By Hoeffding's traditional bound,

$$\Pr\left\{\overline{X}_i^{(t)} - \mu_i \geq w_i^{(t)}\right\} \leq \delta/(kT^{\max})$$

and

$$\Pr\left\{\overline{X}_i^{(t)} - \mu_i \leq -w_i^{(t)}\right\} \leq \delta/(kT^{\max}).$$

Now, after some probabilistic manipulations similar to those of the proof in Theorem 4, we get

$$\Pr\left\{\overline{X}_i^{(t)} - \overline{X}_j^{(t)} - (\mu_i - \mu_j) \geq -w_{ij}^{(t)}, \forall j \neq i\right\} \geq 1 - \delta/T^{\max}.$$

Therefore, the computed $w_{ij}^{(t)}$ satisfy, for each i,

$$\Pr\left\{\overline{X}_i^{(t)} - \overline{X}_j^{(t)} - (\mu_i - \mu_j) \geq -w_{ij}^{(t)}, \forall j \neq i\right\} \geq 1 - \sum_{j=1, j\neq i}^{k} \delta_{ij}^{(t)}.$$

This, along with Bonferroni's inequality, implies the condition on the $w_{ij}^{(t)}$ given in Equation 2.21. Note that by definition $w_{ij}^{(T^{\max})} \leq 2\epsilon$. At every time t, there exists a pair $i, j, i \neq j$, such that $\overline{X}_i^{(t)} - \overline{X}_j^{(t)} \geq 0$. Therefore, for some time $\tau \in \{1, \ldots, T^{\max}\}$, there exists a pair $i, j, i \neq j$, such that $\overline{X}_i^{(\tau)} - \overline{X}_j^{(\tau)} - w_{ij}^{(\tau)} \geq -2\epsilon$, and therefore there exists an action $\hat{a}$ such that $D_{\hat{a}}^{-(1:\tau)} \geq -2\epsilon$. Therefore, this instantiation of the comparison method will stop. By Theorem 7, and noting that by the allocation of the confidence parameters used, $\delta^{(1:\tau)} = \tau\delta/T^{\max} \leq \delta$, the theorem follows. □

This alternative is interesting because the samples can be dependent between actions. Hence, we can share random numbers between actions to generate samples, which will lead to better estimates of the value differences.

We can also get exactly the same approximation guarantees as the instantiation presented in the previous paragraph if we use another instantiation based on independent samples for each action. Let

$$\delta_{ij}^{(t)} = \delta/((k-1)T^{\max}) \tag{2.26}$$

$$l_{ij} \leq \omega_{\boldsymbol{o},\boldsymbol{i}}(\boldsymbol{Z}) \leq u_{ij}, l_{ij} \leq \omega_{\boldsymbol{o},\boldsymbol{j}}(\boldsymbol{Z}) \leq u_{ij}$$

$$N_i^{(t)} = \left\lceil \frac{\max_{ij}(u_{ij} - l_{ij})^2}{2\epsilon^2 T^{\max}} \ln \frac{kT^{\max}}{\delta} \right\rceil, \tag{2.27}$$

$$W(N_i, N_j, \delta_{ij}, \omega_{i,\mathbf{o}}, \omega_{j,\mathbf{o}}) = (u_{ij} - l_{ij})\sqrt{\frac{1}{2}\left(\frac{1}{N_i} + \frac{1}{N_j}\right)\ln\frac{1}{\delta_{ij}}}$$
$$= (u_{ij} - l_{ij})\sqrt{\frac{1}{2}\left(\frac{1}{N_i} + \frac{1}{N_j}\right)\ln\frac{(k-1)T^{\max}}{\delta}} \quad (2.28)$$

Although this can lead to smaller error widths in general, which way is better depends on the bounds on the weight functions and how much difference in (unnormalized) value there is between the first and second best action, as well as how much we lose in variance by not correlating the samples. Other versions can be derived if we also use extra samples for variance estimation, and/or consider confidence intervals on the difference in value instead of the values themselves.

Informal analysis on the number of samples A simple analysis shows that, in the worst case, the number of samples for each action has just increased by a $O(\ln T^{\max})$ as compared to the estimation-based bounds. Therefore, in the worst case, asymptotically, by proceeding in multiple stages versus proceeding in one or two stages as in the estimation based methods, we will not lose much and have potentially a lot to gain.

We believe, although we do not have proof, that the total number of samples required by this instantiation of the comparison-based method is smaller than those required by the traditional method both with high probability and in expectation. Also, the bounds on the number of samples will depend heavily on the amount of separation between the best and second best (unnormalized) action values. We will present evidence of this in the empirical study. We now present an informal analysis leading to this connection.

Assume that the bound on the range of the weight functions is the same for all the actions. Denote $b = u_i$, $a = l_i$, for all $i = 1, \ldots, k$, and $w_t = w_{ij}^{(t)}$, for all i, j (note that equal bounds on the ranges gives the same error widths for all the actions). [3] Denote by Δ the difference in value between the best action and the second best (i.e., $\Delta = \mu^* - \max_{a \neq a^*} \mu_a$). A good approximation to our stopping condition is the condition: we stop at the first stage t such that $\Delta - w_t \geq -2\epsilon$. This assumes that our estimate of the difference is indeed the difference (this is true in expectation). Note that,

$$w_t \approx 2\epsilon\sqrt{T^{\max}/t}.$$

[3] This implies that at each stage we have *almost* equally accurate estimates for the action values, something we wanted to avoid in general. However, the accuracy obtained during the inital stages will be lower than that implicitly used by the traditional method.

Hence, a little algebra shows that we "expect" to stop at about

$$t^* \approx T^{\max}\left(\frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\right).$$

Taking into account that t^* is discrete, we get

$$t^* \approx \left\lceil T^{\max}\left(\frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\right)\right\rceil. \tag{2.29}$$

Let

$$N_s = \left\lceil \frac{(b-a)^2}{2T^{\max}\epsilon^2}\ln(kT^{\max}/\delta)\right\rceil$$

be the number of samples per stage per action (the same for all the actions because of the equal range bounds). Thus, the "expected" number of samples per action is

$$\begin{aligned} N_s t^* &\approx \frac{2(b-a)^2}{(\Delta+2\epsilon)^2}\ln(kT^{\max}/\delta) \\ &\approx \left\lceil \frac{(b-a)^2}{2T^{\max}\epsilon^2}\ln(kT^{\max}/\delta)\right\rceil \times \left\lceil T^{\max}\left(\frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\right)\right\rceil, \end{aligned} \tag{2.30}$$

where the last expression is to take into account the discreteness of the number of samples and the stages. Hence, the larger Δ, the smaller the "expected" number of samples. A formalization of this analysis is left for future work.

From this we can get an expression for the ratio of the "expected" number of samples of this method to that of the traditional method; let N be the number of samples needed by the traditional method per action (the same for all the actions because of the equal range bounds),

$$\begin{aligned} \frac{kN_s t^*}{kN} &= \frac{N_s t^*}{N} \\ &\approx \frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\left(1+\ln(T^{\max})/\ln(k/\delta)\right) \\ &\approx \left\lceil \frac{1}{T^{\max}(2\epsilon)^2}\ln(kT^{\max}/\delta)\right\rceil \times \left\lceil T^{\max}\left(\frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\right)\right\rceil / \left\lceil \frac{1}{(2\epsilon)^2}\ln(k/\delta)\right\rceil \end{aligned} \tag{2.31}$$

where the last expression is to take into account the discreteness of the number of samples and the stages. The comparison-based method with uniform allocations presented in this Sub-section is more effective than the traditional method if this ratio is smaller than one. Using this information, we can get an expression involving $\epsilon, \delta, k, T^{\max}$, and Δ for determining when this method will be more effective than the traditional method.

The expected achieved confidence parameter is

$$\begin{aligned}\delta t^*/T^{\max} &\approx \delta\left(\frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\right)\\ &\approx (\delta/T^{\max})\left\lceil T^{\max}\left(\frac{(2\epsilon)^2}{(\Delta+2\epsilon)^2}\right)\right\rceil.\end{aligned}$$

The expressions above are useful in analyzing the empirical results for this method which we present later.

Unbounded number of stages and constant number of samples per stage (per action)

If we do not impose a bound on the number of stages but would like to guarantee that when the method stops, it outputs an action with the right requirements, we need to define a sequence of confidence parameters that (1) converges to zero as the number of stages and (2) decreases at the right rate. For instance, we can let

$$\delta_{ij}^{(t)} = \delta/((k-1)t(t+1)) \tag{2.32}$$

$$N_i^{(t)} = N^{\text{stage}} \tag{2.33}$$

where N^{stage} is a constant number of samples that we take for each action at each stage, and

$$\begin{aligned}W(N_i, N_j, \delta_{ij}, \omega_{i,\boldsymbol{o}}, \omega_{j,\boldsymbol{o}}) &= \left((u_i-l_i)\sqrt{\frac{1}{N_i}}+(u_j-l_j)\sqrt{\frac{1}{N_j}}\right)\sqrt{\frac{1}{2}\ln\frac{k}{\delta_{ij}(k-1)}}\\ &= ((u_i-l_i)+(u_j-l_j))\sqrt{\frac{1}{2tN^{\text{stage}}}\ln\frac{kt(t+1)}{\delta}}.\end{aligned} \tag{2.34}$$

From this we can see that the process will not go on forever (i.e., it will stop at some time and select and action), since $w_{ij}^{(t)} \to 0$ as $t \to +\infty$, and does this as $O(\sqrt{\ln t/t})$. The shrinking rate is smaller than in the instantiation with bounded stages ($O(\sqrt{1/t})$), but if there is enough separation between the value of the best and second best action to compensate for the range in weight function values, there is a chance that we can stop after taking a smaller total number of samples than if we had used a fixed total number of samples or a fixed maximum number of stages.

Theorem 9 *Consider the instantiation of the comparison-based method resulting from executing Algorithm 4 with assignments for input $\delta_{ij}^{(t)}, N_i^{(t)}$, and W given by 2.32, 2.33, and*

2.34, respectively. This instantiation of the comparison-based method will stop at some time $\tau < \infty$. *Furthermore, its output is such that the action* $\hat{a}$ *satisfies*

$$V(a^*, \boldsymbol{o}) - V(\hat{a}, \boldsymbol{o}) \leq -D_{\hat{a}}^{-(\tau)} \leq 2\epsilon$$

with probability at least $1 - \delta^{(1:\tau)} = 1 - (1 - (1/(\tau+1)))\delta \geq 1 - \delta$.

Proof sketch: The idea is to realize that the computed error widths will monotonically decrease at every stage and we have allocated enough precision to allow for every possible probabilistic event to hold simultaneously with precision no larger than δ.

Proof: The proof follows closely that of Theorem 8 (combined with the discussion just presented in the last paragraph). Note that now, by Hoeffding's traditional bound,

$$\Pr\left\{\overline{X}_i^{(t)} - \mu_i > w_i^{(t)}\right\} < \delta/(kt(t+1))$$

and

$$\Pr\left\{\overline{X}_i^{(t)} - \mu_i < -w_i^{(t)}\right\} < \delta/(kt(t+1)).$$

After some probabilistic manipulations similar to those of the proof in Theorem 4, we get

$$\Pr\left\{\overline{X}_i^{(t)} - \overline{X}_j^{(t)} - (\mu_i - \mu_j) \geq -w_{ij}^{(t)}, \forall j \neq i\right\} \geq 1 - \delta/(t(t+1)).$$

Note also that in this case $\delta^{(1:\tau)} = \sum_{t=1}^{\tau} \delta/(t(t+1)) = (1 - 1/(\tau+1))\delta \leq \delta$. □

Adaptive sample allocation

Let us revisit the assumption that the number of samples per stage for each action be scheduled and fixed in advance. Let us now consider removing this constraint to allow arbitrary adaptive (dynamic) allocation of samples, and making this allocation of samples at each stage a function of the MCB lower-bounds (and hence of the previous samples). One problem we have to deal with is that the total number of samples we would have taken for any action at the the end of any stage but the first is a random variable. Another problem is that we have introduced dependencies between the previous samples conditioned on knowledge about the total number of samples taken for an action choice. For instance, let us say we take 40 (independent) samples for each action during the first stage. Let us also say at the end of the first stage we found that action a has a larger lower bound on the MCB confidence interval (D_a^-) than all the others and decide to take 40 more samples from this action and 10 from the others. Consider the sample average computed for an arbitrary

action i. This average uses 40 independent samples from the first stage and 40 (if $i = a$) or 10 (otherwise) independent samples from the second stage. Although these two sets of samples are independent *among* themselves, they are not independent *between* them. The reason is that the number of samples we decided to use for a particular action at the second stage certainly seems to depend on the outcome from the first stage. We believe it is in general unreasonable to expect that knowing that number of samples used for the second stage provides no information about the outcomes of the samples from the first stage (i.e., it is unreasonable to expect that they are independent). However, Hoeffding's traditional bounds, the tool we have been using to get our confidence intervals on the differences in value between action, from which we then obtained the MCB confidence intervals, requires the random variables involved in the sum (or average) to be independent. Hence, we can't apply that technique.

Unfortunately, the author is not aware of another way of computing distribution-free confidence intervals that can be applied in this context. Hoeffding's strengthened bound allows us to remove the assumptiuon of independence for a weaker assumption of the sum of the samples weights for each action forming a martingale, *for a fixed total number of samples*. We would have to show that, conditioned on a fixed total number of samples, the sum of the weights form a martingale, which is not obviously true. Even if it is, we still have to deal with the fact that the total number of samples is not fixed, but a random variable. (I believe this is formally called a *stopping time*.) Under the assumption that the partial sums form a martingale, simple analysis does not lead to sufficiently interesting bounds. Hence, obtaining interesting bounds seems to require a more sophisticated analysis.

The discussion of Jennison and Turnbull [2000] in Chapter 17 also suggests what is theoretically wrong with the adaptive allocation rule that uses the smallest lower bound on the MCB confidence intervals: the *sample averages* are not independent of the *information levels*. In our context, we can interpret the information levels to mean the information that allows the computation of the precisions achieved at each stage. If we fix the information levels in advance as we are doing when we predefine the allocation schedule, then the sample averages and the information levels are indeed independent and the confidence intervals computed hold. This is not true if, for instance, we allocate more samples to one action than others because its corresponding MCB lower-bound was smallest at the previous stage! This seems to be a hard issue, even when dealing when very specific distributions like normal and binomial, which are to some degree "nicer" to handle theoretically. We do not have an answer to the distribution-free, finite outcomes, bounded-random-variable needed for the general ID model considered here, and we are not aware of any results of a similar this kind

in the probability or statistics community.

In conclusion, the instantiation of the comparison-based method with adaptive allocation as presented in Algorithm 5 is a heuristic based on the assumptions that (1) the way the reallocation is done does not introduce dependencies between the samples used to compute the total averages at the end of each stage and that (2) the total number of samples at the end of each stage is not a random variable. The distribution-free bounds are so loose in general that we believe to be unlikely to find a ID model in practice for which using this assumption will not work well. In the algorithm, the function M is an allocation function and maps the MCB confidence intervals into the number of samples we will take for each action at the next stage. We now present a particular version of the comparison-based method with adaptive allocation.

A heuristic comparison-based method with adaptive allocation

In this section, we present an instantiation of the comparison-based method with adaptive allocations. In this instantiation, we compute the MCB confidence intervals heuristically. To do this, we approximate the precisions w_{ij} that satisfy the conditions required by Hsu's multiple-bound lemma (Equation 2.6) using Hoeffding's bounds. Using this approach, for each pair of actions i and j, and values $l_{ij,\boldsymbol{o}}$ and $u_{ij,\boldsymbol{o}}$ such that $l_{ij,\boldsymbol{o}} \leq \omega_{i,\boldsymbol{o}}(\boldsymbol{Z}) \leq u_{ij,\boldsymbol{o}}$ and $l_{ij,\boldsymbol{o}} \leq \omega_{j,\boldsymbol{o}}(\boldsymbol{Z}) \leq u_{ij,\boldsymbol{o}}$, we approximate w_{ij} as

$$w_{ij} = (u_{ij,\boldsymbol{o}} - l_{ij,\boldsymbol{o}})\sqrt{\frac{1}{2}\left(\frac{1}{N_{i,\boldsymbol{o}}} + \frac{1}{N_{j,\boldsymbol{o}}}\right)\ln\frac{k-1}{\delta}}, \tag{2.35}$$

where $N_{i,\boldsymbol{o}}$ is the number of samples taken for action i thus far. Note that in computing the approximate precisions we have ignored the multiple looks issue (We do not allocate confidence parameters so as to adjust for previous evaluations of the MCB confidence intervals). Note that, because of the way in which we compute w_{ij}, in theory, we cannot share the random numbers computing the estimates of the different actions. If we want to share the random numbers we should use

$$w_{ij} = \left((u_i - l_i)\sqrt{\frac{1}{N_i}} + (u_j - l_j)\sqrt{\frac{1}{N_j}}\right)\sqrt{\frac{1}{2}\ln\frac{k}{\delta}}. \tag{2.36}$$

But then again, in the global scheme, they are both heuristics anyway. We then use these approximate precisions and the value-difference estimates to compute the MCB confidence intervals (as specified by Equation 2.7). There are alternative ways of heuristically approximating the precisions w_{ij} but, in this document, we use the ones above for simplicity.

The value of *initial number of samples* ($N_i^{(1)}$ in Algorithm 5) used in the experiments is 40. When taking additional samples, a sampling schedule (the allocation function M in Algorithm 5) is used that is somewhat selective in that it takes more samples from more promising actions as suggested by the MCB confidence intervals. After finding the action whose corresponding MCB confidence interval has an upper bound greater than 0 (i.e., from the set $\mathcal{G}$ as defined in Hsu's multiple-bound lemma) and whose lower bound is the largest, the method proceeds by taking 40 additional samples from this action and 10 from all the others. It is understood that these sample sizes are very arbitrary. Potentially, other setting of these sample sizes could be more effective but they were not optimized for the experiments. Algorithm 6 presents a detailed description of the particular instantiation of the general comparison-based method with adaptive allocations presented in Algorithm 5. Algorithm 6 is the version used in the experiments.

Although this method may seem well-grounded, we know from the previous discussion on adaptive sample allocation that the bounds might not hold rigorously.

Before we present the related work, in the following three sections, we make an aside to present some brief notes on relative approximations, allocation of approximation parameters and other practical considerations. The first deals with an alternative form of approximation than the one used thus far in this chapter while the other sections deal primarily with general ways to improve the effectiveness of the methods presented here in practice.

2.4 A note on relative approximations

Thus far, we have only considered absolute approximations, where the *accuracy* or approximation error does not depend on the true (unknown) value of the optimal strategy. Relative approximations offer an alternative form of approximation, where the approximation error depends on the true value of the optimal strategy. In this section, we consider how we can extend the estimation-based traditional method presented previously to obtain relative approximations. We also discuss some of the problems keeping us from similarly extending the other methods.

First, relative approximations are appealing for several reasons. They do not depend on the range of the utility functions. The required accuracy is expressed as a percent from optimal and as such it is a relative factor. The allocation of error for each observation does not depend on the number of observations, since relative approximations are multiplicative in nature. Hence to obtain a relative approximation of the optimal (global) strategy, we can require the same relative approximation on the value of the action selected *for each*

Algorithm 6 Algorithmic description of the instance of the comparison-based method used in the experiments.

for each observation $\boldsymbol{o}$ **do**
  $l \leftarrow 1$
  for each action $i = 1, \ldots, k$ **do**
    Compute $u_{i,\boldsymbol{o}}$ and $l_{i,\boldsymbol{o}}$ using equations 2.10 and 2.11, respectively.
    $D_i^- \leftarrow -\infty$; $N_{i,\boldsymbol{o}}^{(l)} \leftarrow 40$; $N_{i,\boldsymbol{o}} \leftarrow 0$; $\hat{V}_{\boldsymbol{o}}(i) \leftarrow 0$.
  end for
  for each pair of actions (i, j), $i \neq j$ **do**
    $u_{ij,\boldsymbol{o}} \leftarrow \max(u_{i,\boldsymbol{o}}, u_{j,\boldsymbol{o}})$; $l_{ij,\boldsymbol{o}} \leftarrow \max(l_{i,\boldsymbol{o}}, l_{j,\boldsymbol{o}})$.
  end for
  while there is no action i such that $D_i^- > -2\epsilon$ **do**
    for each action i **do**
      Obtain $N_{i,\boldsymbol{o}}^{(l)}$ samples $\boldsymbol{z}^{(N_{i,\boldsymbol{o}}+1)}, \ldots, \boldsymbol{z}^{(N_{i,\boldsymbol{o}}+N_{i,\boldsymbol{o}}^{(l)})}$ from $\boldsymbol{Z} \sim f_{i,\boldsymbol{o}}$, as in equation 2.8.
      Compute weights $\omega_{i,\boldsymbol{o}}^{(N_{i,\boldsymbol{o}}+1)}, \ldots, \omega_{i,\boldsymbol{o}}^{(N_{i,\boldsymbol{o}}+N_{i,\boldsymbol{o}}^{(l)})}$.
      $\hat{V}_{\boldsymbol{o}}(i) \leftarrow (N_{i,\boldsymbol{o}} \hat{V}_{\boldsymbol{o}}(i) + \sum_{j=1}^{N_{i,\boldsymbol{o}}^{(l)}} \omega_{i,\boldsymbol{o}}^{(N_{i,\boldsymbol{o}}+j)}) / (N_{i,\boldsymbol{o}} + N_{i,\boldsymbol{o}}^{(l)})$.
      $N_{i,\boldsymbol{o}} \leftarrow N_{i,\boldsymbol{o}} + N_{i,\boldsymbol{o}}^{(l)}$.
    end for
    for each pair of actions (i, j), $i \neq j$ **do**
      $T_{ij} \leftarrow \hat{V}_{\boldsymbol{o}}(i) - \hat{V}_{\boldsymbol{o}}(j)$; $T_{ji} \leftarrow -T_{ij}$.
      Compute w_{ij} using equation 2.35; $w_{ji} \leftarrow w_{ij}$.
    end for
    for each action i **do**
      Compute D_i^+, $\mathcal{G}$, and D_i^- using Hsu's multiple-bound lemma.
    end for
    for each action i **do**
      if $D_i^- == \max_{j \in \mathcal{G}} D_j^-$ **then** $N_{i,\boldsymbol{o}}^{(l+1)} \leftarrow 40$
      else $N_{i,\boldsymbol{o}}^{(l+1)} \leftarrow 10$.
    end for
    $l \leftarrow l + 1$.
  end while
  $\hat{\pi}(\boldsymbol{o}) \leftarrow \text{argmax}_i D_i^-$.
end for

observation.

The problems with relative approximations in this context is that they might not be very useful if the value is large and we are interested being close to the optimal value in actual unit terms, not percent terms. Because we do not know the value that we are estimating and the error terms depend on that value, the number of samples needed to achieve certain quality also involves this unknown quantity. In particular, the smaller the value we are trying to estimate, the larger the number of samples required. This seems somewhat counterintuitive because (1) we are putting as much weight to act well under observations whose value does not contribute significantly to the (global) value of the strategy as to those observations whose value does contribute significantly (so we will be doing "equally well" in all cases); and (2) we will be spending more samples in those cases where the value associated with actions and observations is not significant than in those where the value is. If the particular problem requires relative approximations, one way to get around not knowing the exact value when determining *a priori* the number of samples to get a certain accuracy is to lower-bound the value of the weight of the samples used. However, the lower bounds that can be easily obtained are typically very loose, leading to an unreasonably large number of samples.

Another way is through the use of sequential estimation. The idea is to derive stopping rules that allow us to stop when we have reached the accuracy we want based on information from the sample weights. This is similar to the sequential estimation method presented previously to guarantee a given absolute approximation. The idea behind an "optimal" method of computing relative approximations called the $\mathcal{AA}$ algorithm [Dagum et al., 2000] is precisely that (See also Pradhan and Dagum [1996] for an application of this algorithm to belief inference in BNs). This method is optimal in that no other method that uses independent identically distributed samples would be able to obtain a relative approximation using an expected number of samples that is more than a constant amount smaller than the expected number of samples that the $\mathcal{AA}$ method takes before it stops. In other words, it can take a long time if the value we are estimating is very small (and the variance in our estimates is large), but any other method cannot be much faster.

The basis for the theoretical analysis of the $\mathcal{AA}$ algorithm is the generalized zero-one estimator theorem [Dagum et al., 2000]. Hence, to apply this algorithm to our context, as described, we need to bring the value of the weights to be between zero and one. We can do this by finding upper and lower bounds on the utility functions of the utility nodes and then performing a simple linear transformation on the weights. Note that the error parameter need not be transformed as long as the utilities are nonnegative (Recal that the

error parameter for a relative approximation is given as a percent from optimal). However, if the lower bound is very large, the actual percent error required from the $\mathcal{AA}$ algorithm would be smaller than necessary. To see this, consider that we want to estimate $\mu_X = \mathrm{E}[X]$ for some random variable X which has all outcomes in $[l, u]$ for some constants u and l, such that $0 \leq l < u$. Define $Z = X - l/(u - l)$. Then Z has all outcomes in $[0, 1]$. If we obtain an estimate $\hat{\mu}_Z$ of $\mu_Z = \mathrm{E}[Z]$, such that $\hat{\mu}_Z \geq (1-\epsilon)\mu_Z$, then $\hat{\mu}_X \geq (1-\epsilon)\mu_X + \epsilon l \geq (1-\epsilon)\mu_X$.

We now present how we can use the $\mathcal{AA}$ algorithm to obtain *relatively* near-optimal strategies with high probability. Let $M \equiv |\Omega_{\boldsymbol{O}}|$ be the total number of observations. To obtain an (ϵ^*, δ^*)-relative approximation to the global strategy, where ϵ^* is the (percent) error parameter and δ^* the confidence parameter, all we need is to obtain an $(\epsilon = \epsilon^*/2, \delta = \delta^*/M)$-relative approximation for each observation. The reason for this is as follows. Let $\hat{pi}(\boldsymbol{o})$ be the action selected by the approximate strategy for observation $\boldsymbol{o}$, $\hat{pi}^*(\boldsymbol{o})$ be the optimal action for the same observation, and $V^{\hat{\pi}}$ and V^* be the trues values of the approximate strategy and the optimal, respectively. If, for each observation $\boldsymbol{o}$,

$$V(\hat{pi}(\boldsymbol{o}), \boldsymbol{o}) \geq (1 - 2\epsilon) V(\pi^*(\boldsymbol{o}), \boldsymbol{o})$$

with probability at least $1 - \delta$, then

$$V^{\hat{\pi}} = \sum_{\boldsymbol{O}} V(\hat{pi}(\boldsymbol{O}), \boldsymbol{O}) \geq (1 - \epsilon^*) V^*$$

with probability at least $1 - \delta^*$.

To simplify notation for the discussion that follows, let $\mu_i \equiv V(i, \boldsymbol{o})$, $\hat{\mu}_i \equiv V(i, \boldsymbol{o})$ and $\mu^* \equiv \max_i \mu_i$. Using the $\mathcal{AA}$ algorithm, for each observation $\boldsymbol{o}$, and action a, we can obtain estimates $\hat{\mu}_a$ for μ_a such that for a given ϵ and δ, $0 < \epsilon < 1$, and $0 < \delta < 1$,

$$\Pr\{\hat{\mu}_a - \mu_a \leq -\epsilon\mu_a\} \leq \frac{\delta}{k},$$

and

$$\Pr\{\hat{\mu}_a - \mu_a \geq \epsilon\mu_a\} \leq \frac{\delta}{k}.$$

Hence, we can obtain, for each i individually,

$$\Pr\{\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j) > -\epsilon(\mu_i + \mu_j), \forall j \neq i\} \geq 1 - \delta.$$

Since $\mu_i + \mu_j \leq 2\mu^*$, we get

$$\begin{aligned} \Pr\{\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j) > -2\epsilon\mu^*, \forall j \neq i\} &\geq \Pr\{\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j) > -\epsilon(\mu_i + \mu_j), \\ &\qquad \forall j \neq i\} \\ &\geq 1 - \delta. \end{aligned}$$

Using Hsu's single-bound lemma, we get

$$\Pr\left\{\mu_i - \max_{j\neq i}\mu_j \in [-(\hat{\mu}_i - \max_{j\neq i}\hat{\mu}_j - 2\epsilon\mu^*)^-, (\hat{\mu}_i - \max_{j\neq i}\hat{\mu}_j + 2\epsilon\mu^*)^+], \forall i\right\} \geq 1-\delta.$$

Hence, with probability at least $1-\delta$, for the selected action $\hat{a} = \text{argmax}_a\, \hat{\mu}_a$, $\mu_{\hat{a}} \geq (1-2\epsilon)\mu^*$. Therefore, by using the $\mathcal{AA}$ algorithm, we have essentially extended the estimation-based traditional method (presented previously for absolute approximation) to obtain relative approximations.

Unfortunately, it is not immediately evident how to extend this approach to the estimation-based sequential method and/or the comparison-based method. The problem is that the way the $\mathcal{AA}$ algorithm works, it does not provide "intermediate" confidence intervals, but only outputs an estimate with the right accuracy and confidence when it stops. One can apply the idea of using a lower bound to the possible actions values $V(\boldsymbol{o}, a)$ (for each observation) to obtain such confidence intervals. However, the reader should be aware that such lower bounds are typically very loose, leading to unreasonably large confidence intervals for reasonable sample sizes.

2.5 Allocating precision and confidence parameters for each observation

In this section, we consider a possible way to handle IDs with large number of observations by using a smarter allocation of the approximation parameters.

In many cases, we have to deal with models that have a large number of observations. If we use the estimation-based traditional method to solve our problem, we can try to reduce the total number of samples taken by allocating the precision and confidence parameters differently. In this subsection, we describe one way we can do this.

Let (ϵ^*, δ^*) be the parameters defining an approximation with error ϵ^* (the *accuracy* or *error* parameter) with probability at least $1-\delta^*$ (hence, δ^* is the confidence parameter). Also, let, for observation each $\boldsymbol{o}$,

1. $r_{\boldsymbol{o}}$ be the *percent* used to allocate the error for that observation from the total maximum error, such that $\sum_{\boldsymbol{O}} r_{\boldsymbol{O}} = 1$, and $r_{\boldsymbol{o}} > 0$,
2. $\epsilon_{\boldsymbol{o}} \equiv \epsilon^* r_{\boldsymbol{o}}$ be the amount of error allocated for that observation,
3. $R^2_{i,\boldsymbol{o}} \equiv (u_{i,\boldsymbol{o}} - l_{i,\boldsymbol{o}})^2$ be the range size of the weight functions for action i and that observation, squared, and $R^2_{\boldsymbol{o}} \equiv \sum_{i=1}^{k} R^2_{i,\boldsymbol{o}}$ be the total sum over all actions of those squared ranges,

4. $s_{i,\boldsymbol{o}}$ be the *percent* used to allocate the confidence parameter for action i and that observation, such that $\sum_{i,\boldsymbol{O}} s_{i,\boldsymbol{O}} = 1$, and $s_{i,\boldsymbol{o}} \geq 0$,

5. $\delta_{i,\boldsymbol{o}} \equiv \delta^* s_{i,\boldsymbol{o}}$ be the amount from the confidence parameter allocated for action i and that observation (it might help to think about this value as follows: in some sense, statements associated with action i and that observation will be made with confidence $1 - \delta_{i,\boldsymbol{o}}$).

Note that that for each observation, the number of samples needed by the estimation-based traditional method for each observation $\boldsymbol{o}$ and action a can be written as $\left\lceil (R^2_{a,\boldsymbol{o}}/(2\epsilon^2_{\boldsymbol{o}})) \ln (k/\delta_{a,\boldsymbol{o}}) \right\rceil$. Hence, to reduce the total number of samples, we want to solve the following optimization problem:

$$(\boldsymbol{r}^*, \boldsymbol{s}^*) = \operatorname*{argmin}_{\boldsymbol{r},\boldsymbol{s}} \sum_{\boldsymbol{O}} \sum_{i=1}^{k} \frac{R^2_{i,\boldsymbol{O}}}{2\epsilon^2_{\boldsymbol{O}}} \ln \frac{k}{\delta_{i,\boldsymbol{O}}}$$

subject to the constraints on $\boldsymbol{r}$ and $\boldsymbol{s}$. Using Lagrange multipliers, we will find that the solution satisfies the following fixed-point equations

$$\begin{aligned} r^*_{\boldsymbol{o}} &= \frac{(\sum_i R^2_{\boldsymbol{o}} \ln \frac{k}{\delta s^*_{i,\boldsymbol{o}}})^{\frac{1}{3}}}{\sum_{\boldsymbol{O}} (\sum_i R^2_{i,\boldsymbol{O}} \ln \frac{k}{\delta s^*_{i,\boldsymbol{O}}})^{\frac{1}{3}}} \\ s^*_{i,\boldsymbol{o}} &= \frac{\frac{R^2_{i,\boldsymbol{o}}}{r^2_{\boldsymbol{o}}}}{\sum_i \sum_{\boldsymbol{O}} \frac{R^2_{i,\boldsymbol{O}}}{r^2_{\boldsymbol{O}}}}. \end{aligned}$$

We can use several methods to try to solve them. For instance, we can just turn them into an iterative method starting from some assignment for $\boldsymbol{s}$, get an assignment for $\boldsymbol{r}$ using the first fix-point equation, then get a new value for $\boldsymbol{s}$ using the second fix-point equation, and so on and so forth. This process will converge to the global minimum of the total-number-of-samples function, since this function is convex in $\boldsymbol{r}$ and $\boldsymbol{s}$.

Note that in the case that we use a uniform allocation of the confidence parameters $\delta_{i,\boldsymbol{o}}$,

$$r^*_{\boldsymbol{o}} = \frac{(R^2_{\boldsymbol{o}})^{1/3}}{\sum_{\boldsymbol{O}} (R^2_{\boldsymbol{O}})^{1/3}}.$$

Thus, as intuitively expected,the larger the range size of the weights of the actions associated with an observation, the larger the precision $\epsilon_{\boldsymbol{o}}$ we should use. Also, for uniform allocation of the precision parameters $\epsilon_{\boldsymbol{o}}$,

$$s^*_{i,\boldsymbol{o}} = \frac{R^2_{i,\boldsymbol{o}}}{\sum_i \sum_{\boldsymbol{O}} R^2_{i,\boldsymbol{O}}}$$

Hence, the larger the range size of the weight of the action and observation, the larger the confidence parameter δ_o we should allocate to it (i.e., the less confidence we should require for the observation). Unfortunately, note that in the worst case that all the ranges are the same, this expression reduces to the uniform allocation suggested at the beginning of this chapter. Hence, another approach needs to be used to deal with models with a large number of observations. This is something we have partially studied but leave primarily for future work.

2.6 Other practical considerations

In this section, we briefly state some additional considerations leading to improvements of the efficiency and effectiveness of the sampling methods presented here in practice.

The theoretical results presented above for the most part allow us to share the random numbers used to generate the samples used to estimate the value of all the actions and observations (except in some instantiations of the comparison-based method that specifically use bounds based on independence of action value estimates). If we allow independence, we can obtain immediately tighter theoretical results. However, we believe for the problem of action selection, the improvements on the difference estimates based on the variance-reduction resulting from the positive correlations created by the random-number sharing will be more significant that any theoretical improvement based on independent estimates. This is because, in the particular problem of action selection, we are ultimately most interested in having good estimates of the value differences rather than the individual value estimates. Also note that sharing random numbers for sampling in IDs will have the following results: (1) those nodes that are not descendants of the decision node will have the same instantiations for all the samples used to estimate any of the action choices; and (2) the instantiations of all non-descendants of observation nodes will be the same between samples used to process each observation. Now, sharing random numbers might require us to keep an extra storage. This is not required by the instantiation of the method based on completely independent samples between actions and observations, where we get new samples of all the nodes for every action. When using completely independent samples, we just process the weight value and do not store the sample and weight. However, we do not really need to store the random numbers themselves, since we can just save the random seed and regenerate the random numbers.

In some cases, if the number of evidence or observation scenarios is not too large, we can save some time, by considering all the scenarios at once. (This is also true for exact

methods.) In general, we might need to consider each scenario individually, however, as the number of observation scenarios can be prohibitive for some IDs.

Finally, although we do not deal with IDs with multiple decision nodes, I would like to point out some properties we can exploit in this models, which can be useful in extending our sampling methods to multiple decisions. This problems are typically solved by dynamic programming. This approach leads to a decomposition of the original ID problem into a sequence of smaller ID problems (one for each decision variable), which we can construct from the original ID (See, for instance, Zhang [1998], and Charnes and Shenoy [1999]). When we are allocating precisions for the different sub-problems, we only need to consider scenarios constructed from the joint outcomes of observations available at the time of the decision that have chance nodes as parents in the graph of the sub-problem ID for that decision. The factor terms for observations that do not have chance nodes in that sub-problem ID can be moved outside the summation and hence do not play a role in the estimation (i.e., won't affect precision).

Before we present the empirical results for the methods, we present related work on the problem of action selection in IDs, and discuss further connections to the statistical literature of the methods presented in this chapter.

2.7 Related Work

Charnes and Shenoy [1999] present a Monte Carlo method similar to the "traditional method" presented here. They estimate the *conditional* values of each action, instead of the unconditional value as done here. Also, they use a heuristic stopping rule based on a normal approximation (i.e., the estimates have an *asymptotically* normal distribution). Their method takes samples until all the estimates achieve a required standard error to provide the correct confidence interval on each conditional value under the assumption that the estimates are normally distributed and the estimate of the variance is equal to the true variance. These assumptions are only asymptotically valid in general and are hard to verify. They do not give bounds on the number of samples needed to obtain a near-optimal action with the required confidence. We refer the reader to Charnes and Shenoy [1999] for a short description and references to other similar Monte Carlo methods for IDs.

Bielza et al. [1999] present a method based on Markov chain Monte Carlo (MCMC) for solving IDs. Although their primary motivation is to handle continuous action spaces, their method also applies to discrete action spaces. Because of the typical complications in analyzing MCMC methods, they do not provide bounds on the number of samples needed.

Instead, they use a heuristic stopping rule which does not guarantee the selection of a near-optimal action. Other MCMC-based methods have been proposed (See Bielza et al. [1999] for more information).

The general notion of "repeated confidence intervals" in the clinical trials (group sequential) literature is the same as that forming the base of the comparison-based method and its analysis, except that in the work on clinical trials the typical underlying assumption behind the computation of the intervals and proof of correctness is that the sampled values (actually, in our case, the weights of the samples) are normally distributed [Jennison and Turnbull, 2000]. It is certainly possible that there exist *distributions-free* extensions of that work of which the author is not aware.

Certainly, if we ignored the fact that the normality assumption does not hold in a general ID model, we could apply these methods to the problem of action selection considered in this thesis. It is possible that the direct application of those techniques could be useful in this context, even if no theoretical guarantees hold for general ID models. Under the normality assumptions, the confidence bounds will certainly be tighter than presented here. We do not pursue this approach in this thesis, but for practitioners, it would be interesting to evaluate them in the future to verify their potentially practical effectiveness.

Also, we can follow a Bayesian approach to compute the confidence intervals necessary to determine the stopping rules (MCB intervals) for the comparison-based methods.[4] The Bayesian approach is not pursued in this thesis, but it is certainly of future interest.

The ideas behind the adaptive allocation schedule used in the heuristic MCB-based method presented above is in the same spirit (and faces similar issues) as the IE (interval estimation) method of Kaelbling [1993], developed in the context of reinforcement learning. In the IE method, actions with higher value upper-bounds are given preference. This heuristic can also be applied to get another similar MCB-based heuristic method, similar to the one presented here. That is, instead of taking more samples from actions with largest MCB confidence lower bound, we can take more samples from actions with largest MCB confidence *upper* bounds. Since having a large MCB upper bound means that the difference of that action from the rest can be at most that large, there is a possibility of that action being that much better than the rest. Needless to say, there are many other heuristics that we can be used, depending on how optimistic or pessimistic we want to be.

Also, the idea and analysis of the method of Maron and Moore [1994] called "Hoeffding

[4]The idea of using a Bayesian approach was suggested to me by Peter Müller.

races" for model selection in a machine learning setting is very similar to the comparison-based methods presented here. In their setting they have a finite set of samples (the dataset) and the action choices are the finite set of models available. They evaluate the quality of each model on each sample, one sample at a time. After a model is determined inferior to any other (i.e., the confidence upper-bound on its true value given by Hoeffding's bounds is smaller than the confidence lower-bound of another), they do not evaluate that model anymore. Hence, their "allocation schedule" is "adaptive." Therefore, I believe that the arguments given in the analysis there are only valid if *all* the models are evaluated on *all* the samples, for the reasons given in the discussion regarding the adaptive-allocation schedule here.

2.8 Preliminary empirical results

In this section, we present the results for running some of the methods described in this chapter on the computer-mouse problem described in Appendix A. Unless specifically noted otherwise, in the rest of this section, we refer to the heuristic MCB-based method (with adaptive allocation and ignoring multiple-looks) presented above as the comparison-based method.

Table 2.1 presents the results on the effectiveness of the sampling methods for this problem. We set the final desired accuracy for the output strategy to $\epsilon^* = 5$ and confidence level $\delta^* = 0.05$. This leads to the individual accuracy $2\epsilon = 2.5$ and confidence level $\delta = 0.025$ for each subproblem. The sequential method and the comparison-based method were executed 100 times. The comparison-based method produces major reductions in the number of samples. When we observe the mouse pointer not working, the comparison-based method always selects the optimal action of buying a new mouse. When we observe the mouse pointer working, the comparison-based method failed to select the optimal action of *taking no action* 4 times out of the 100. In those cases, it selected the next-to-optimal action of upgrading the operating system ($A = 2$). This action is within the accuracy requirements since the difference in value with respect to the optimal action is 0.91.

The comparison-based method is highly effective in cases where there is a clear optimal action to take. For instance, in the computer mouse problem, buying a new mouse when we observe the mouse not working is clearly the best option. The differences in value between the optimal action and the rest are not as large as when we observe the mouse working.

In this problem, the results for the sequential method should not fully discourage us from its use, because the variances are still relatively large. Major reductions have been

A	MP_t	Method		
		Traditional	Sequential	Comp-based
1	0	2403	3802 (188)	335 (151)
2	0	3007	2266 (142)	115 (37)
3	0	3679	2426 (129)	118 (39)
1	1	2213	2508 (178)	521 (216)
2	1	2794	2969 (201)	695 (421)
3	1	3443	3468 (202)	1361 (560)
Total		17539	17438 (434)	**3145** (809)

Table 2.1: Number of samples taken by the different methods for each action and observation. For the sequential and the comparison-based methods, the table displays the average number of samples over 100 runs. The values in parentheses are the sample standard deviations.

seen in problems where the variance is significantly smaller than the square of the range of the variable whose mean we are estimating. We will see an example in the next section where we deal with a real ID.

2.9 Emprical results on IctNeo ID

In this section, we present empirical results on an ID from a medical domain [Bielza et al., 2000, Gómez et al., 2000]. The ID is being developed for the treatment of jaundice in newborns, and is still under development. Figure 2.4 shows a graphical plot of the IctNeo ID (all arcs point down). This ID has a total of 72 nodes (including 5 decision nodes and 1 utility node). It is important to point out that most of the characteristics of this problem do not really make it amenable for the type of approximation and model assumptions considered here. First it is a multi-stage problem: it has five decision nodes. Actually, only four are really interesting since, due to constraints on the possible set of actions, the optimal action for the last decision is really a deterministic function of the fourth action. Hence, we converted this node into a deterministic node in the graph (a node whose conditional probability defines a deterministic function). Because a large number of variables is necessary to represent historical and/or background information relevant to the patient's treatment, there is an extremely large number of possible observations available at the time the first decision is made. Similarly, due to the no-forgetting assumption, the number of observations for the fourth decision variable is also extremely large. Therefore, we believe the main roadblock in solving this problem is not the evaluation of the values for a given action and observation, since this is easy to compute exactly, but the extremely large number of

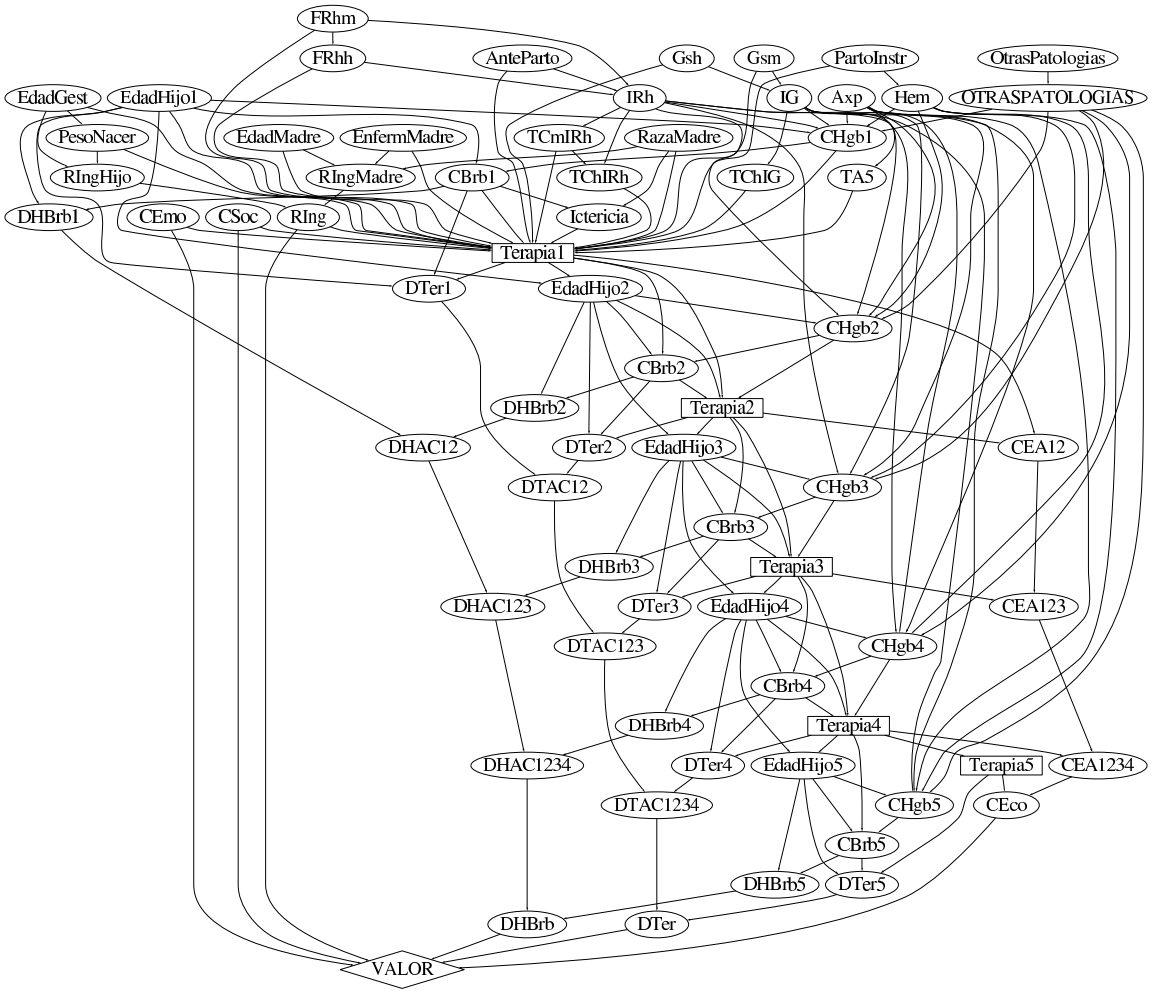


Figure 2.4: The IctNeo ID.

observations that are relevant and are available at the time of the decisions. Thus, for the most part, exact methods are tractable in this model for solving a particular observation scenario (conditioned on other decisions being assigned). Approximation methods that try to provide compact representations of policies have been developed for IDs as well as MDPs and POMDPs that could be more applicable to this model. However, my objective in trying the methods in this model is that this is a model for a real domain and has been sufficiently carefully developed. Therefore, it provides more realistic probabilistic and utility models that can be produced by generating random models or creating unrealistic models.

We performed our experiments in several modified versions of the IctNeo ID.[5]. The discussion of the experiments follows

2.9.1 Experiment 1: (Partially) solving the fourth decision stage

In this experiment, we consider solving the fourth (i.e., last interesting) decision stage. Because the number of evidence scenarios available for this decision (i.e., relevant observations and previous actions) is still very large, we only generated a subset of all the evidence scenarios at random using the following scheme. We first determined all possible three-action sequences prior to the fourth decision. Taking into account the constraints on those action sequences, there were 415 action sequences possible. For each action sequence, we instantiated the three initial decision nodes with the action sequence and simulated the ID to generate a single observation resulting from that action sequence. These observations are a sample from the model distribution under a fixed, blind strategy for the first three decisions (i.e., a strategy on which the action to execute for each decision is determined in advance and does not depend on the available observation at the time of the decision). The idea is to make a good decision after having executed those actions and have the available observations for that decision.

For each action sequence and the corresponding generated observation, we ran the estimation-based traditional and sequential methods, an instantiation of the comparison-based method with fixed, uniform allocation presented in Section 2.3.1 which uses a maximum number of stages equal to 10 and the heuristic instantiation of the comparison-based method with adaptive allocation. The instantiation of the comparison-based method with fixed, uniform allocation is given by Equations 2.23, 2.24, and 2.25 or Equations 2.26, 2.27 and 2.28 depending on whether we share random numbers (see Section 2.3.1 for details).

[5]The modified IDs were further simplified through the typical relevance-based reduction operations in IDs (see for example Shachter [1998]).

For a given approximation requirement (ϵ^*, δ^*), the first three methods guarantee to select an action with the given requirements. We should point out that the number of samples required for any reasonable guaranteed approximation in this model is huge. (A rough, yet I believe still informative, analysis based on rough estimates of some of the numerical quantities in the theoretical bounds suggests that in order to compute a full strategy for this problem with guarantees that it will be no less than 0.01 from the optimal with probability at least 0.95 would require approximately $8e24$ samples in total for the traditional method, $4e23$ for the two-stage sequential method, and $3e23$ for the comparison-based method with uniform allocations; tightening the error requirement to 0.001 would require approximately $8e26$, $3e26$ and $3e23$ for the same methods, respectively.) Nevertheless, we can still test, for any given value of ϵ and δ, how many samples each method took to select an action with those requirements. In other words, the objective is to study how the methods proposed that have theoretical guarantees compare to each other in terms of the total number of samples needed, and the effectiveness of the heuristic method relative to the more theoretically grounded methods. To both illustrate the looseness of the theoretical bounds behind the methods and evaluate the quality of the selected actions, we compute the average of the difference in conditional value between the action selected by each of the sampling methods and the optimal action computed using an exact method. To establish a baseline for comparison, we also evaluated a method that just selects actions uniformly at random from the set of possible action choices. Note also, that if for a given observation and previous action sequence scenario, the constraint in the action choices of the fourth decision is such that only one action is possible, we ignore that scenario in our evaluations. Hence, only 348 of the 415 scenarios generated were considered for the evaluations. Hence, the results are based on just those 384 scenarios.

The results for the case in which we obtain independent samples for each action (and observation) are summarized in Table 2.2. The column labeled *total # samples* has the total number of samples over all the 384 evidence scenarios taken by the methods before they stopped. The column labeled *mean* has the sample average of the *conditional* error based on the difference between the *conditional* value of the optimal action and the action selected by the methods. This quantity is an unbiased estimate of the *regret* of the *randomized* policy resulting from using the different methods. The column labeled *std. dev.* has the sampled standard-deviation (the square of this quantity is an unbiased estimate of the variance of the error of the respective *randomized* policy associated with each method; note that were we to compute confidence interval based on the typical normal approximation, we would take this value, divide it by $\sqrt{384}$—the number of evidence cases—multiply it by plus

or minus the appropriate critial value constant—1.96 for the case of approximately 95% confidence intervals—and add the emprical mean—the average). The column labeled *CI* has 95% bootstrap confidence intervals on the *regret*, where in the bootstrap method we use 1000 bootstrap-samples. (We computed the bootstrap confidence intervals as follows. First, we resampled *with replacement* from the set of error outcomes to generate 1000 resampled datasets of size 384—the same size as the original set of error outcomes. Then we computed the average from each new resampled set and order the averages in increasing value—computed order statistics. Finally, we selected the 50^{th} and 950^{th} element of the ordered set of averages as the lower and upper bounds, respectively, of the confidence interval. For a simple introduction to the boostrap method see Cohen [1995]. I should point out that in this case, should we have used the traditional approach to compute the confidence intervals based on Normal approximation, there would not have been much difference in the computed intervals.) The columns labeled *max* and *median* have the empirical maximum and median of *conditional* error. (The empirical minimum was always zero.) The row labeled *Random* has the results for a method where we select uniformly at random from the set of possible action choices for each scenario. The rows labeled *Traditional, Sequential, MCB, and MCB heuristic* have the results for the estimation-based traditional method, the estimation-based sequential method, the comparison-based method with uniform allocation and fixed maximum number of stages (which we set to $T^{\max} = 10$ in this case) and the heuristic comparison-based method with adaptive allocation, respectively. The comparison-based methods compute the error widths using expressions that assume no random number sharing.

Method	Total # samples	avg. regret	std. dev.	CI	max	median
Random		0.0133	0.0206	$[0.0113, 0.0154]$	0.0824	0.0075
		$2\epsilon = 0.1, \delta = 0.1$				
Traditional	$401,316$	0.0071	0.0122	$[0.0060, 0.0084]$	0.0808	0.0042
Sequential	$462,086$	0.0060	0.0093	$[0.0051, 0.0070]$	0.0808	0.0038
MCB	$359,045$	0.0075	0.0120	$[0.0063, 0.0088]$	0.0797	0.0047
MCB heuristic	$249,160$	0.0069	0.0104	$[0.0059, 0.0080]$	0.0827	0.0048
		$2\epsilon = 0.05, \delta = 0.1$				
Traditional	$1,601,805$	0.0064	0.0115	$[0.0053, 0.0075]$	0.0833	0.0037
Sequential	$1,193,611$	0.0060	0.0101	$[0.0051, 0.0071]$	0.0793	0.0038
MCB	$1,426,417$	0.0053	0.0086	$[0.0046, 0.0063]$	0.0790	0.0038
MCB heuristic	$933,320$	0.0074	0.0118	$[0.0062, 0.0087]$	0.0783	0.0048
		$2\epsilon = 0.02, \delta = 0.1$				
Traditional	$10,005,925$	0.0033	0.0056	$[0.0028, 0.0039]$	0.0780	0.0014
Sequential	$4,425,510$	0.0050	0.0083	$[0.0042, 0.0059]$	0.0740	0.0031
MCB	$8,784,848$	0.0037	0.0056	$[0.0032, 0.0043]$	0.0622	0.0021
MCB heuristic	$5,802,750$	0.0068	0.0110	$[0.0058, 0.0080]$	0.0796	0.0043

Table 2.2: Results for IctNeo ID: Experiment 1 (No random numbers shared).

Method	Total # samples	avg. regret	std. dev.	CI	max	median
Random		0.0133	0.0206	$[0.0113, 0.0154]$	0.0824	0.0075
		$(\times 10^{-4})$		$(\times 10^{-3})$		
		$2\epsilon = 0.1, \delta = 0.1$				
Traditional	$401,316$	6.3885	0.0019	$[0.4494, 0.8400]$	0.0131	0
Sequential	$462,859$	6.4026	0.0021	$[0.4555, 0.8752]$	0.0150	0
MCB	$636,895$	5.9030	0.0020	$[0.4132, 0.7884]$	0.0129	0
MCB heuristic	$455,130$	4.5394	0.0016	$[0.3045, 0.6309]$	0.0150	0
		$2\epsilon = 0.05, \delta = 0.1$				
Traditional	$1,601,805$	2.5966	0.0011	$[0.1396, 0.3650]$	0.0108	0
Sequential	$1,200,603$	2.5019	0.0010	$[0.1583, 0.3471]$	0.0107	0
MCB	$2,522,047$	3.0757	0.0013	$[0.1817, 0.4390]$	0.0121	0
MCB heuristic	$1,734,900$	4.3384	0.0018	$[0.2787, 0.6148]$	0.0138	0
		$2\epsilon = 0.02, \delta = 0.1$				
Traditional	$10,005,925$	1.1765	0.0008	$[0.0489, 0.2099]$	0.0102	0
Sequential	$4,420,600$	1.9831	0.0012	$[0.0963, 0.3242]$	0.0152	0
MCB	$15,548,415$	1.1480	0.0009	$[0.0415, 0.2181]$	0.0101	0
MCB heuristic	$10,453,650$	5.7668	0.0020	$[0.3958, 0.7769]$	0.0139	0

Table 2.3: Results for IctNeo ID: Experiment 1 (Random numbers shared only among actions, not observations).

An immediate observation is that the sequential method is very effective in reducing the number of samples for near-optimal selection for this problem, and its effectiveness improves as ϵ decreases. This suggests that the variance of the weight functions used to estimate the unnormalized value is significantly smaller than our bound on the range of the weights.

We also tested a method that shares the random bits *among actions only* (not observations). Recall that this (and more) is allowed by the theoretical results presented in this chapter. The results for this variation are in Table 2.3. Because of the positive correlations, the error is smaller in general, since the estimates of the difference in unnormalized value of actions have smaller variance. Graphical plots of the data are displayed in Figure 2.5 and Figure 2.6. The plots on the left are of the *conditional* error versus the number of samples for each of the 384 evidence scenarios considered. The results for the random method are plotted with squares. (They all fall along the y-axis since we do not take any random samples from the model, just a sample from a uniform distribution over the set of action choices.) The results for the other method are plotted with an 'x'. The right plots are paired versions of the plots on the left, pairing the results for each scenario. The negative slope of the lines indicate the improvement in error over the random method, while the magnitude of the slope indicates the cost in terms of number of samples for the reduction (i.e., a small slope magnitude indicate that we used a large number of samples to reduce the error).

The comparison-based method with uniform allocations did not do well for this problem. This is because the differences in unnormalized value between the best and second best actions are not large enough for this problem. The plots in Figure 2.7 illustrate this point. The plots show the negative ratio of the number of samples taken by the comparison-based method to the number of samples taken by the estimation-based traditional method as a function of the difference between the true unnormalized value of the best and second best actions, for each evidence scenario. We denote this difference by Δ, where we have to keep in mind that this is a function of the evidence scenario. The higher the negative ratio, the better the comparison-based method. A negative ratio equal to -1 means that the methods used the same amount of samples. On average, the number of samples taken by the comparison-based method was about 1.6 times that taken by the traditional method, which corroborates the result for the total number of samples shown in Table 2.3. Note that the predicted values of the negative ratio resulting from the informal analysis in Section 2.3.1 match fairly well the empirical outcomes. (I believe the remaining discrepancies are due to variance and the discreteness of the expressions involving the number of samples which are not matched by the approximations resulting from the informal analysis.) We note that the

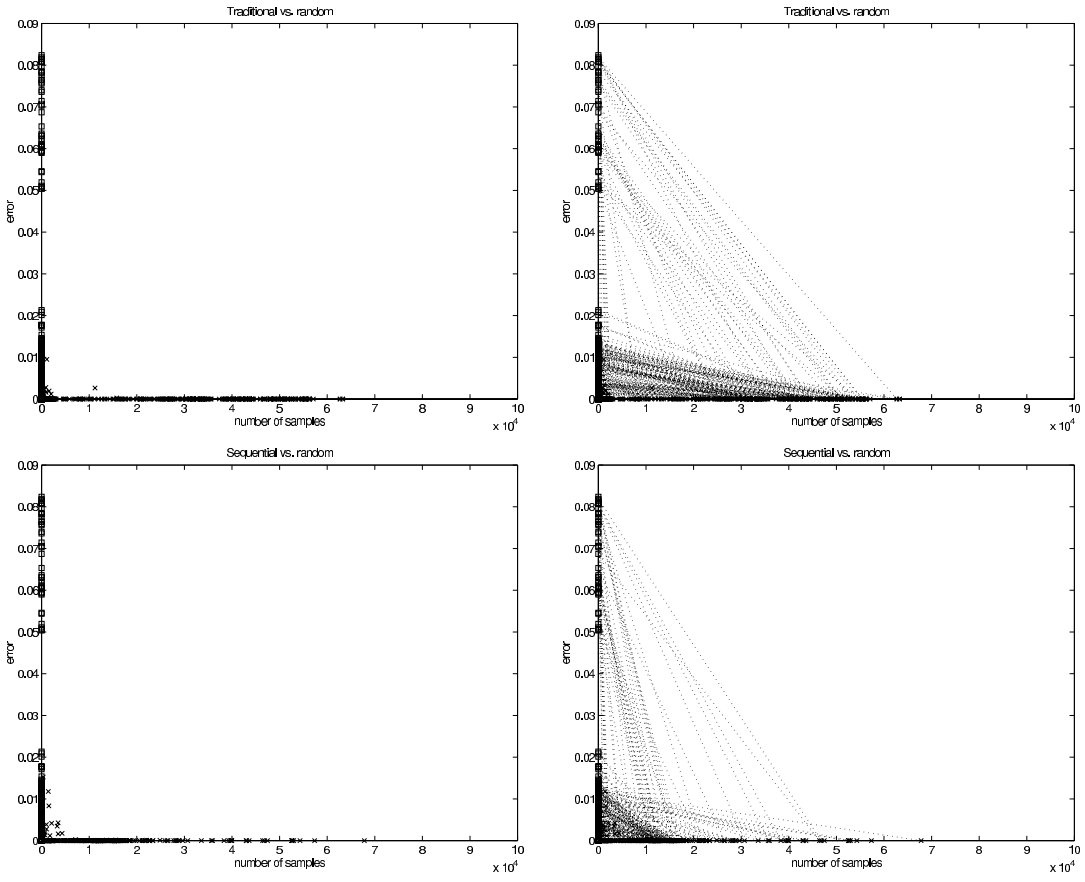


Figure 2.5: Results for IctNeo ID: Experiment 1 ($2\epsilon = 0.02, \delta = 0.1$, shared random numbers). Comparing methods relative to the random method in terms of conditional error and number of samples.

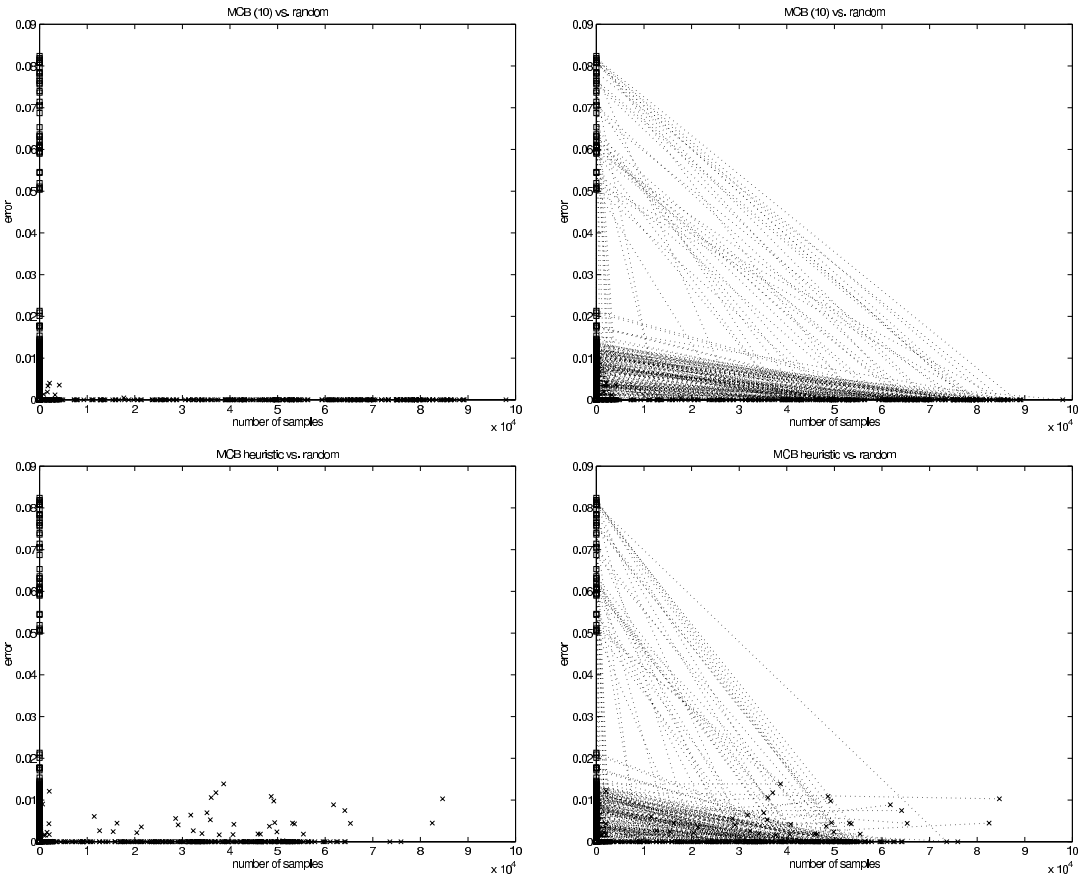


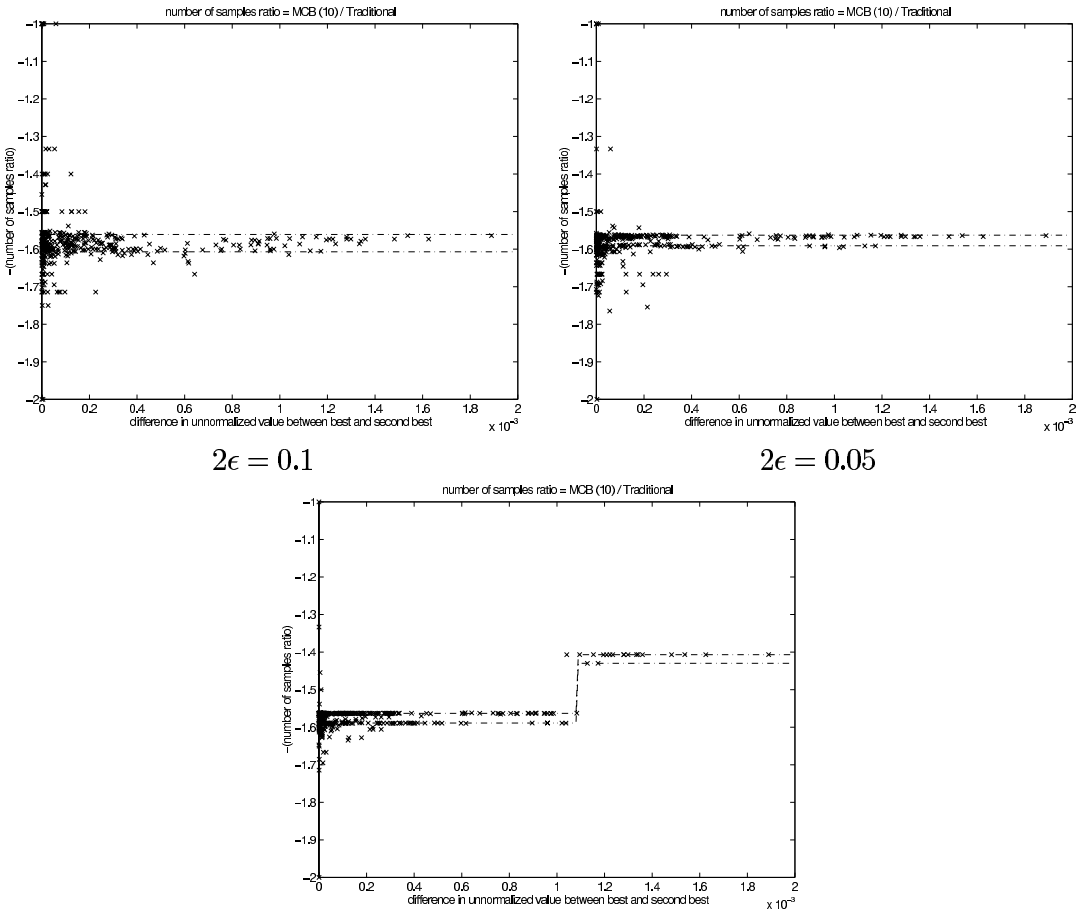
Figure 2.6: Continuation of Figure 2.5.

informal analysis' assumption of equal bounds on the range of the weight functions holds in this problem. For the values of ϵ, δ and $T^{\max}$ used, the informal analysis predicts that Δ has to be larger before the comparison-based method can be more effective than the traditional method.

Finally, the effectiveness of the sequential method for this problem suggests that we should also use variance information in the comparison-based methods. We have started to develop versions of the comparison-based method that use variance information, but we leave the details for future work.

We also studied the effect of the adaptive allocation as compared to uniform allocations in this problem. For this we consider a single evidence scenario. Arbitrarily, we selected the evidence scenario for which Δ was largest. The number of action choices for this evidence scenario is 6. At each stage, the uniform allocation uses 15 samples/action, while the adaptive allocation uses 40 samples/action for the action with largest MCB-confidence-interval lower bound (the value with the potentially smallest regret) and 10 for the rest. Both take 40 samples/action for the first stage. We ran the methods for 50 stages 40 times. Figure 2.8 shows the effect on the MCB-confidence-interval lower bound as a function of the stage. Recall that we use the MCB-confidence-interval lower bound for the stopping condition of the comparison-based methods. The graph shows individual (approximate) 95% confidence intervals of the expected value of the MCB-confidence-interval lower bound at stage $10, 20, \ldots, 50$. Therefore, under the heuristic assumptions regarding the adaptive allocation, the adaptive allocation is increasing the MCB-confidence-interval lower bound faster than the uniform allocation as we wanted, hence, leading to stopping earlier for a fixed error requirement. This comes at some price however. The results presented in Figure 2.9 show that if we selected the action based on the smallest MCB-confidence-interval lower bound at each stage the error is larger than that of the uniform. By error or *regret* here we mean the difference in unnormalized value between the best action and the one we select. (We obtain the *conditional regret* by dividing that value by the probability of the evidence scenario, which is constant for all the actions in a particular scenario.[6]) The adaptive allocation method is willing to trade *regret* for stopping earlier (reducing the number of samples). Since the unnormalized values in this case are very small and 5 of the action choices have values very close to each other, the heuristic adaptive allocation gets fooled often by the lower bound with regard to the best action. In the case of the uniform allocation

[6]To be more specific, in this case, we would divide by the conditional probability of the evidence nodes that have non-evidence nodes as parents, given evidence nodes that do not have non-evidence nodes as parents.



$2\epsilon = 0.1$ $2\epsilon = 0.05$

$2\epsilon = 0.02$

Figure 2.7: Results for IctNeo ID: Experiment 1 ($\delta = 0.1$, shared random numbers among actions). Efficiency of comparison-based method with fixed uniform allocation and maximum number of stages ($T^{max} = 10$) relative to estimation-based traditional method. The dash-dotted lines are the predictions from the informal analysis taking into account discreteness of the number of samples (Top: $k = 6$, Bottom: $k = 5$). See text for further details.

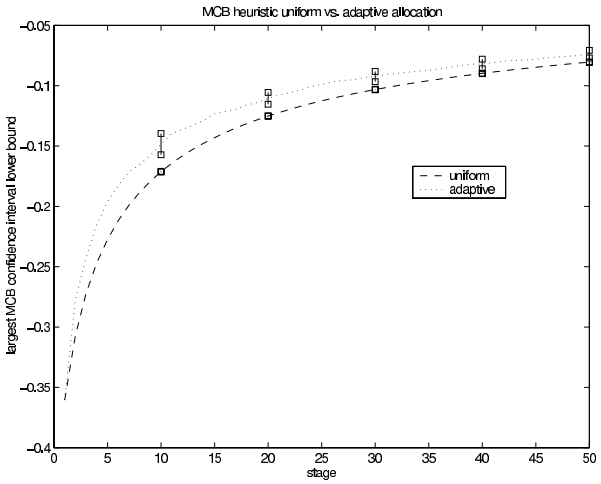


Figure 2.8: Results for IctNeo ID: Experiment 1. Effect of adaptive allocation on the largest MCB-confidence-interval lower bound for the comparison-based method.

and equal bounds on the ranges of the weight functions, selecting the action with the largest MCB-confidence-interval lower bound is equivalent to selecting the action with the largest value estimate. I believe that if we had selected the actions during the adaptive allocation based on the maximum value estimate instead of the largest MCB-confidence-interval lower bound, the regret would have been comparable to that for the uniform allocation. Further analysis of this issue is left for future study.

In summary, the two-stage sequential method produced significant reductions on the number of samples relative to the traditional method, while keeping about the same quality on the selected actions. This shows that allocating samples at an initial stage for variance estimation can be very useful. The comparison-based method did not perform well on this experiment compared to the estimation-based methods. This was so primarily because the difference in (unnormalized) value between the first and second best action was not large enough. In the next experiment, we will see that the comparison-based method can produce signifcant reductions in the number of samples required for selecting actions that are provably near-optimal when this difference is sufficiently large.

2.9.2 Experiment 2: (Partially) solving the first decision stage (assuming random future action sequences)

In this experiment, we consider selecting actions for the first decision stage. We modified the ID by converting the future decision nodes into chance or variable nodes that depended on

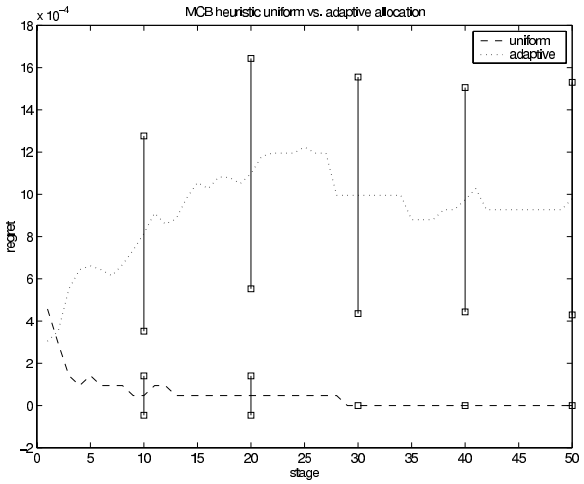


Figure 2.9: Results for IctNeo ID: Experiment 1. Effect of adaptive allocation on quality of the selected action for the comparison-based method.

the previous decision node. Because of the constraint on the actions choices, the conditional probability distribution of each decision node was a uniform distribution over the set of possible action choices for that decision given each action of its parent (previous) decision. We were unable to solve this modified version of the IctNeo ID using our implementation of the exact method, even for a single evidence scenario. This is because performing the summation over the hidden (non-evidence) nodes now involves, at some point, summing over a large number of possibilities at once, corresponding to a large number of variables that become related through the summation process. (The program ran out of space for this case, as it required at least 1.5Gb bytes of memory, maybe more.)

We generated 100 evidence scenarios from the ID model to consider in this experiment. As we could not compute the true values exactly, we estimated them by using importance sampling with 100,000 samples. We use those estimates as the true value of the actions. For this experiment, the methods shared the random numbers between both actions and observations. All the sampling methods tried selected the best action for each of the 100 cases. The table below shows the total number of samples taken by each method. In this table, the rows labeled *MCB (10)*, *MCB (100)*, and *MCB (1000)* corresponds to the comparison-based method with uniform allocation and maximum number of stages $T^{\max} = 10, 100$, and 1000, respectively. The MCB heuristic is as before.

$2\epsilon = 0.01, \delta = 0.05$

Method	Total number of samples
Traditional	7,605,561
Sequential	3,509,029
MCB (10)	2,218,671
MCB (100)	2,058,564
MCB (1000)	2,509,797
MCB heuristic	1,309,440

The estimation-based sequential method is again superior to the traditional method for this problem. Now, however, the comparison-based methods are better than the sequential method. This suggests that the difference in value between the best and second best action is sufficiently large to compensate for the effect of using variance information to reduce the number of samples.

Recall that although we ask for a predetermined error and confidence from our approximations, and the sampling methods presented in this chapter, including the comparison-based method with uniform allocation and fixed maximum number of stages, guarantee such requirements, they can guarantee a better theoretical error and confidence that asked for, since those values are random variables for some of these methods. Figure 2.10 shows plots of the theoretically achieved error for each method as a function of the number of samples taken. The plot for the MCB heuristic method is not really theoretically guaranteed, but the value of the largest (heuristic approximation of) the MCB-confidence-interval lower bounds when it stopped. The comparison-based methods use as much of the error allowed as possible in order to reduce the number of samples until stopping. For the comparison-based method with fixed, uniform allocation and maximum number of stages $T^{\max}$, the larger $T^{\max}$ the more of the error allowed it uses, and for this case, up to a point, the smaller the number of samples. This last point is illustrated in Figure 2.11 where we present a paired representation of the data for MCB (10) and MCB (100). In this plot, the results for MCB (10) are plotted with an 'x' while those for MCB (100) are plotted using a square. A line connect results for the same evidence scenario. Note that, on average, the lines have a negative slope. This improvement is only up to some value of $T^{\max}$ as indicated by the total number of samples shown in table above. We also see why this is the case in a following paragraph where we discuss the effectiveness of the comparison-based methods in this problem.

Figure 2.12 illustrates the effect of the maximum number of stages on the theoretically

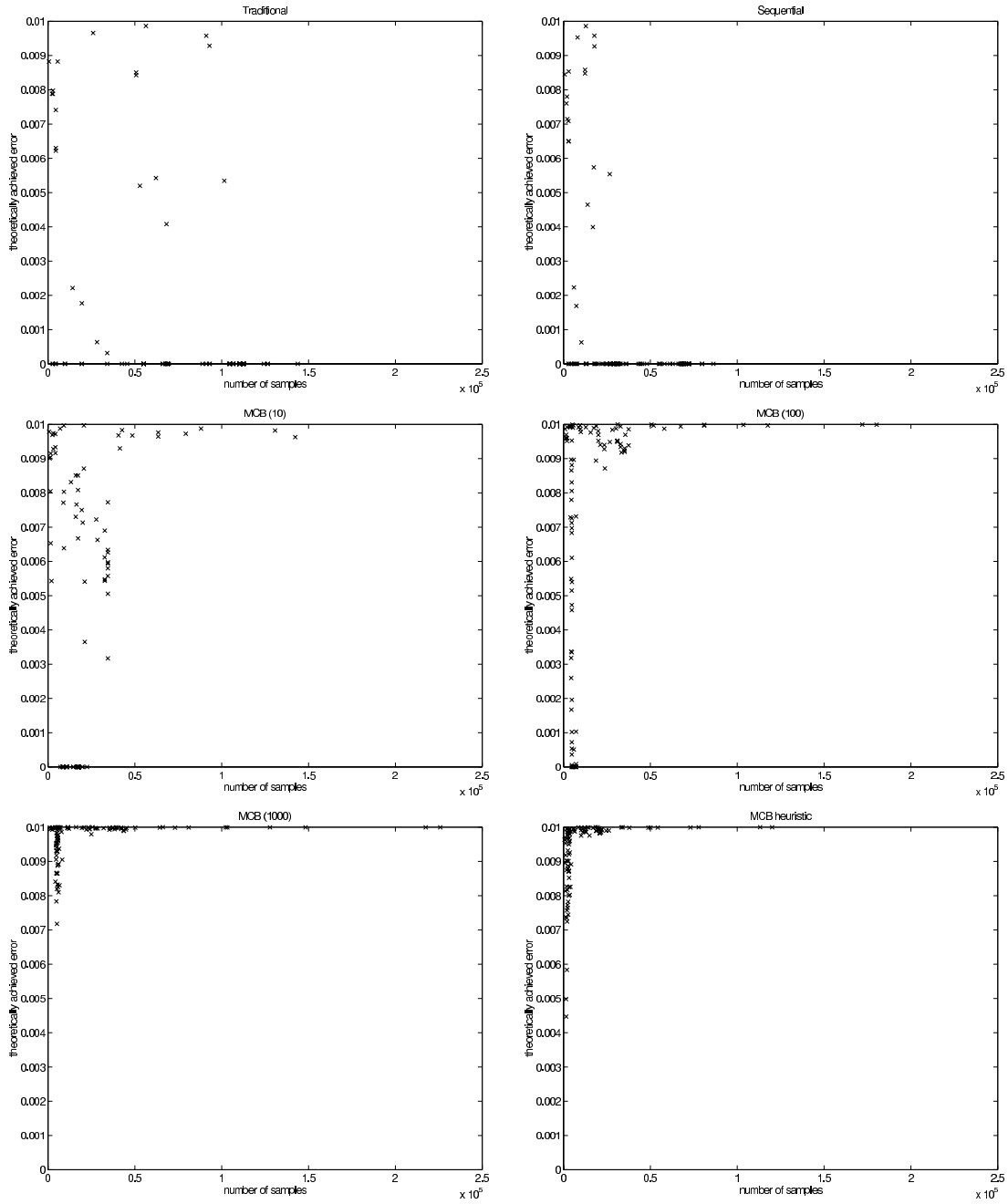


Figure 2.10: Results for IctNeo ID: Experiment 2 ($2\epsilon = 0.01, \delta = 0.05$, shared random numbers among actions and observations). Theoretically achieved error versus the number of samples.

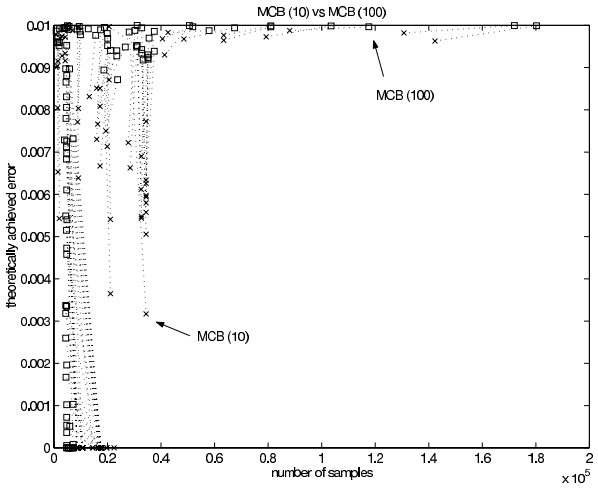


Figure 2.11: Results for IctNeo ID: Experiment 2 ($2\epsilon = 0.01, \delta = 0.05$, shared random numbers among actions and observations). Effect of maximum number of stages ($T^{\max}$) on number of samples and theoretically achieved error.

achieved error and confidence. In the plots, the closer to the left the better the confidence, and the closer to the bottom the smaller the theoretically achieved error. In this case, the larger the maximum number of stages, the larger the theoretically achieved error (within the allowed margin) but the larger the confidence on the error bound.

We evaluated the effect of the adaptive allocation in this problem also. In this case, the number of action choices is 3. At each stage, the uniform allocation uses 20 sample/action while the adaptive allocation uses 40 for the action with largest MCB-confidence-interval lower bound and 10 for the rest. Both methods use 40 samples/action for the first stage. We ran the methods up to stage 50, 40 times. We considered three cases corresponding to the sampled evidence scenario with the maximum, median and minimum value for Δ (the minimum difference in unnormalized value between the best and second best action). Figure 2.13 shows the results. The plots show individual (approximate) 95% confidence intervals at stage $10, 20, \ldots, 50$. For this model, the adaptive allocation does not seem to have a large impact in increasing the MCB-confidence-interval lower bound, at least at early stages. Other factors such as Δ, the bounds on the range of the weights, and the variances seem more important in this case. It might be that the difference in number of samples for the allocations is not large enough.

For this problem, the effectiveness of the comparison-based methods can be explained by the values of Δ for the sampled evidence scenarios. Figure 2.14 shows the results. The

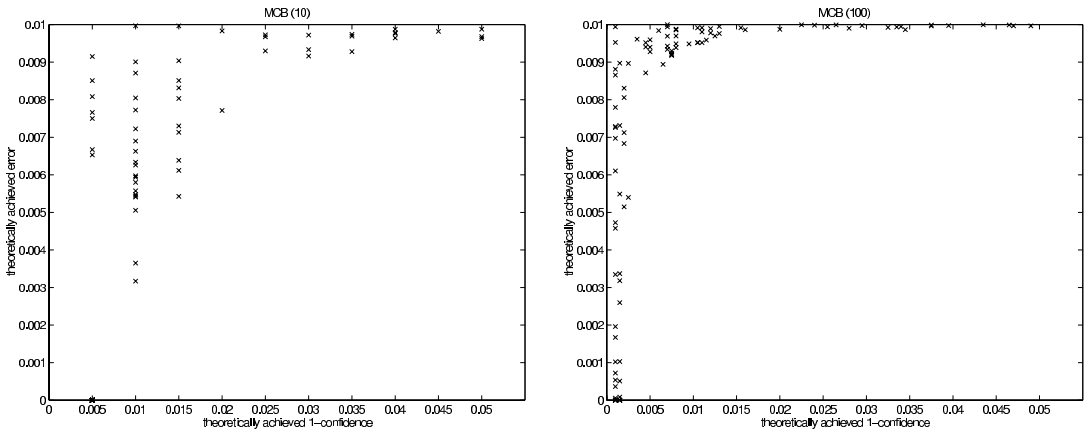
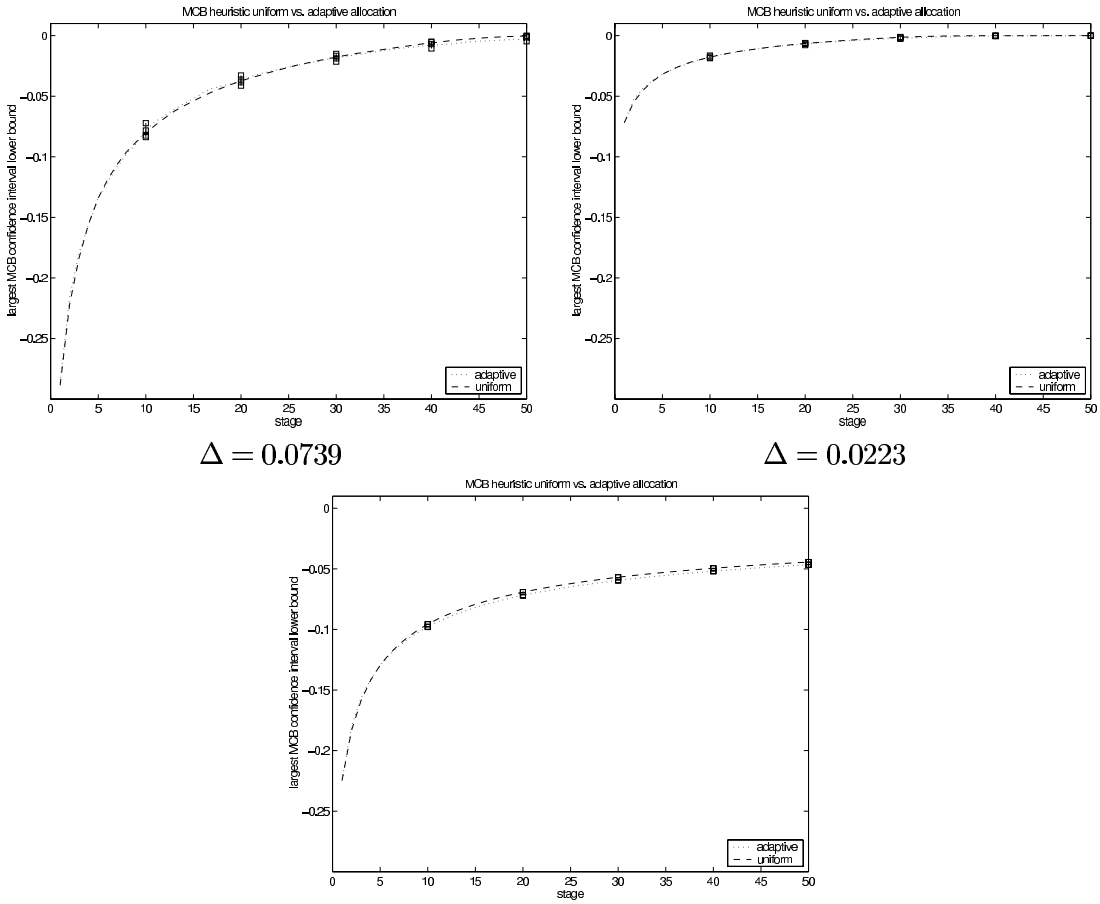


Figure 2.12: Results for IctNeo ID: Experiment 2 ($2\epsilon = 0.01, \delta = 0.05$, shared random numbers among actions and observations). Effect of maximum number of stages ($T^{\max}$) on theoretically achieved error and confidence.

plot shows the negative ratio of the number of samples of the comparison-based methods with uniform allocation and maximum number of stages with respect to the number of samples of the traditional method. In this case the values of the Δ of the evidence scenarios are often relatively large. We note that the informal analysis's assumption of equal bounds on the range of the weight functions holds in this problem. Note also that once again, the predictions from the informal analysis match the outcomes well, suggesting that it is not very far from being correct. Similar results for the adaptive allocation are shown in Figure 2.15

To further emphasize the usefulness of the informal analysis in predicting the behavior of the comparison-based method, we showed the achieved confidence as a function of Δ in Figure 2.16. Note that the theoretically achieved confidence is a linear function of the stopping stage. Again, the predictions from the informal analysis explain the data well. This suggests that we can use information from the informal analysis, after obtaining a rough estimate of Δ to determine which method will be more effective with regard to the number of samples used. This is left for future work.

In summary, we showed that the comparison-based methods can be very effective for the problem of action selection and can be significantly more efficient than the traditional method, specially when the amount of separation between the best and second best action is sufficiently large. The two-stage sequential method can also be very effective at exploiting variance information to produce reductions on the total number of samples for near-optimal



$\Delta = 0.0739$ $\Delta = 0.0223$

$\Delta = 1.1269e - 4$

Figure 2.13: Results for IctNeo ID: Experiment 2. Effect of adaptive allocation on the largest MCB-confidence-interval lower bound for the comparison-based method.

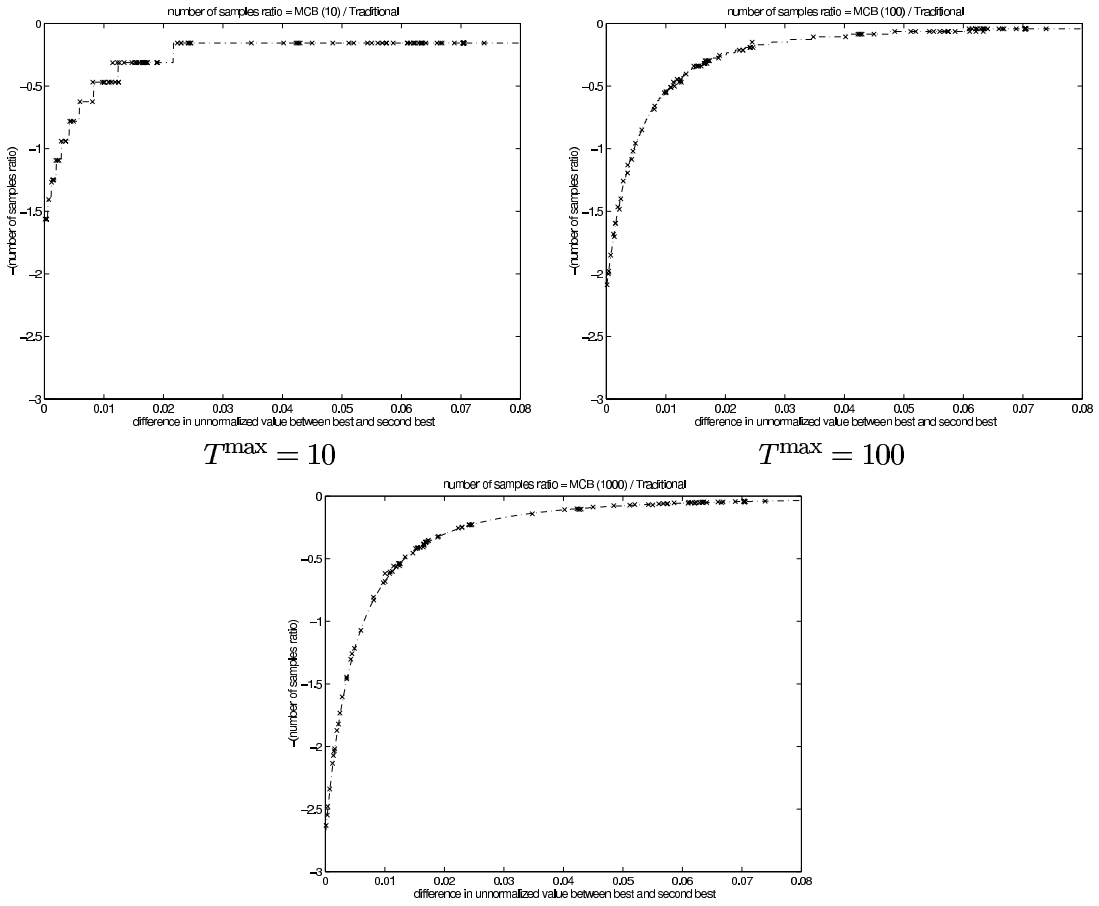


Figure 2.14: Results for IctNeo ID: Experiment 2 ($\delta = 0.05$, shared random numbers among actions and observations). Efficiency of comparison-based method with fixed uniform allocation and maximum number of stages ($T^{\max}$) relative to estimation-based traditional method. The dash-dotted lines are the predictions from the informal analysis taking into account the discreteness of the number of samples. See text for further details.

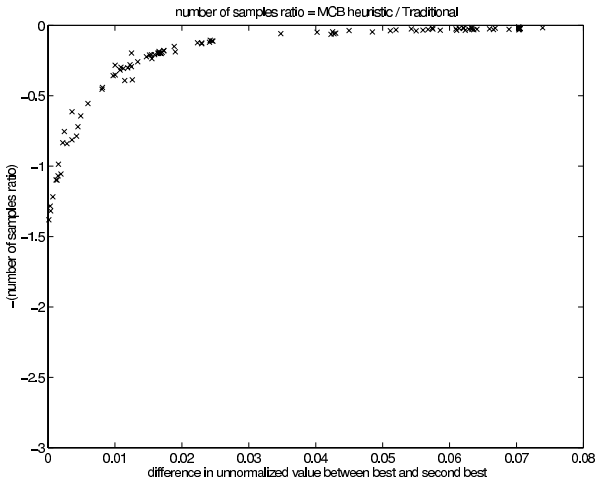


Figure 2.15: Results for IctNeo ID: Experiment 2 ($\delta = 0.05$, shared random numbers among actions and observations). Efficiency of comparison-based method with adaptive allocation relative to estimation-based traditional method.

action selection compared to the traditional method. We also showed that the "back-of-the-envelope" analysis given in Section 2.3.1 for the comparison-based method with uniform allocations and maximum number of stages can describe fairly well the behavior of the method. I believe this suggests that this analysis is not far from being theoretically correct. Finally, the general conclusions for the experiments is that the effectiveness of sampling methods will depend primarily on *numerical* properties of the model (as opposed to exact methods, where the dependency is heavily on *structural* properties of the model).

Before we summarize and conclude this chapter, we discuss some of the open questions for the problem and methods considered in this chapter.

2.10 Open questions

One big issue with the class of comparison-based methods presented in this chapter is that we do not provide interesting bounds on the number of samples required until stopping time. Even in the simpler case of fixed-per-stage sample allocations, for which we have stopping rules that guarantee certain approximation qualities, we could not derive such bounds on expected performance (i.e., total number of samples). Actually, there exists some results of this kind when we assume that the sample weights are normally distributed (See Jennison and Turnbull [2000] for references). Typically, tables have been pre-computed from the (typically Monte-Carlo) approximation of integrals involving normal probability

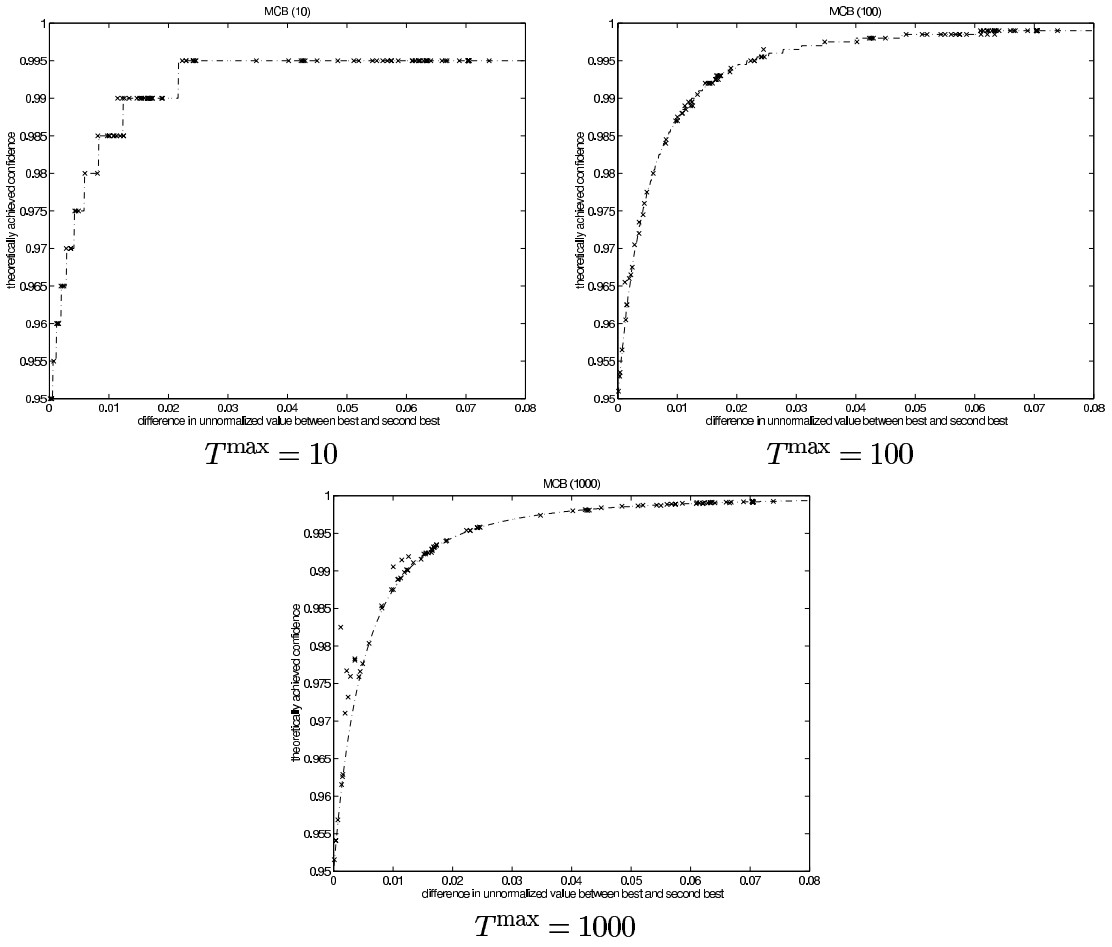


Figure 2.16: Results for IctNeo ID: Experiment 2 ($\delta = 0.05$, shared random numbers among actions and observations). Effect of maximum number of stages ($T^{\max}$) in the comparison-based method with fixed uniform allocation and maximum number of stages on the theoretically achieved confidence. The dash-dotted lines are the predictions from the informal analysis taking into account the discreteness of the number of samples. See text for further details.

density functions. The author is not aware of results of this kind when we are dealing with somewhat general, distribution-free, bounded random variables as we are in the ID model.

On another subject, let us reconsider the heuristic comparison-based method with adaptive allocation presented above. We use an arbitrary allocation strategy: we allocate some arbitrary number of samples for one action and another arbitrary number for the rest. One way to remove this arbitrariness is to turn this question into a sequential decision (theoretic) problem. One of many alternatives we can think of to define a notion of utility is that which rewards allocating more samples to actions whose outcome will result in a larger discrimination between the first and second best actions as soon as possible by penalizing for each sample we take. We can define the state as a k-dimensional continuous variable whose i^{th} dimensional component domain is the range of the weight functions for action i (recall we are considering a particular observation). We can consider simple actions; for instance, take a sample for action i. In this case, the transition probability is the conditional probability over the possible value of the action given the previous values and the action taken. Note that given the simple action structure and independence on the samples for each action, the process has a simple structure: the process governing the estimates in value for each original action component is marginally independent of the others and each new action (take a sample from original action i) only affects a single component value estimate. The reward function should be some empirical function of "the negative of the difference in value between the first and second best value estimate." (i.e., the larger the difference, the larger the reward) and involve some additional cost for each sample and/or stage. By the desription above, we have essentially defined an MDP. Figure 2.17 displays this model graphically. Note however that it seems unlikely that we will find expressions for the transition probabilities for a general model. However, we can always simulate the process (generate samples according to the transition distributions). The randomized method for solving MDPs of Kearns et al. [1999b] is immediately applicable to the type of MDP we have defined above (continuous state space [7], bounded reward, easy to *simulate* from transition probability). (I believe POMDP formulations are also possible.)

Thus, in theory, we have a solution to our MDP. However, it seems like a waste to have to go through such a high computational overhead just to decide where to allocate our samples. As presented, it is not clear that we can combine both problems of allocation and selection and use information from both to solve our original problem of action selection. Also, note that because of our relaxation regarding the quality of the computed strategy,

[7] Actually, in the case of discrete spaces and random utilities, the state space will actually be discrete but grows exponentially with each transition.

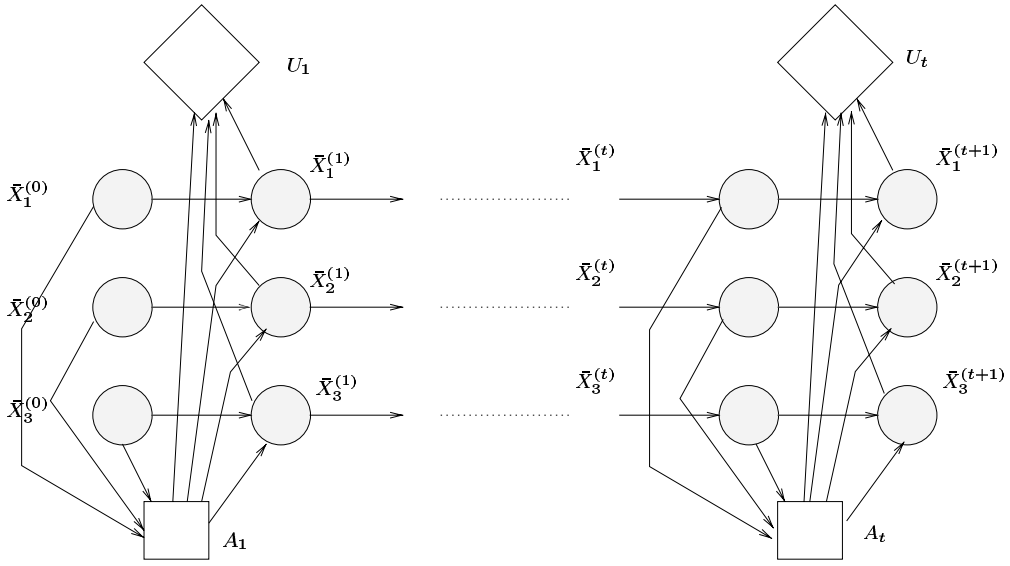


Figure 2.17: Graphical representation of MDP for optimal adaptive allocation.

our main objective is really to determine a near-optimal action as quickly as possible, not necessarily to select the very best. This new objective simplifies the problem and potentially reduces the total number of samples needed for selection. Hardwick [1991], in studying a similar problem, has a nice description of some of the computational issues we have just discussed in this paragraph, but in a completely different context.

2.11 Summary and Conclusion

The methods presented in this chapter are an alternative to exact methods. While the running time of exact methods depends on aspects of the structural decomposition of the ID, the running time of the methods presented in this chapter depends primarily on the range of the weight functions, the variance of the value estimators and the amount of separation between the value of the best action and that of the rest (in addition to the natural dependency on the number of action choices, and the precision and confidence parameters). In some cases, we can know in advance whether they will be faster or not. The methods presented in this chapter can be a useful alternative in those cases where exact methods are intractable. How useful depends on the particular characteristics of the problem.

Sampling is a promising tool for action selection. The empirical results suggest that sampling methods for action selection are more effective when they take advantage of the

intuition that action selection is primarily a comparison task. More experimentation is necessary with IDs large enough that sampling methods are the only potentially efficient alternative. Also, a future empirical comparison study is necessary to determine whether this improvement over the traditional method can make this type of sampling method superior to other kinds of sampling methods like MCMC and stratified sampling, and under what conditions. Also, this work leads to the study of adaptive importance sampling as a way to further improve the effectiveness of sampling methods [Ortiz and Kaelbling, 2000a]. In the next chapter, preliminary work done on adaptive importance sampling is presented.

Chapter 3

Adaptive Sampling

Often, we are interested in computing quantities involving large sums, such as expectations in uncertain, structured domains. For instance, as presented in the introduction, belief inference in *Bayesian networks (BNs)* requires that we sum or marginalize over the remaining variables that are not of interest. Similarly, as presented in the last chapter, in order to solve the problem of action selection in *influence diagrams (IDs)*, we sum over the variables that are not observed at the time of the decision in order to compute the value of different action choices.

We can represent the uncertainty in structured environments using a BN. A BN allows us to compactly define a joint probability distribution over the relevant variables in a domain. It provides a graphical representation of the distribution by means of a directed acyclic graph (DAG). It defines locally a conditional probability distribution for each relevant variable, represented as a node in the graph, given the state of its parents in the graph. This decomposition can help in the evaluation of the sums. However, in general, this is not sufficient to allow an efficient computation of the exact value of the sums of interest.

Sampling provides an alternative tool for approximately computing these sums. Sampling methods have been proposed as an alternative to exact methods for such problems. In particular, *importance sampling* (See Kahn and Marshall [1953], Rubinstein [1981], Geweke [1989], and the references therein) has been applied to the problem of belief inference in BNs [Fung and Chang, 1989, Shachter and Peot, 1989, Fung and Favero, 1994, Cano et al., 1996, Hernández et al., 1998] and action selection in IDs (see Charnes and Shenoy [1999] and the references therein, and Ortiz and Kaelbling [2000b]). In its simpler form, the importance-sampling distribution used is the do-operated distribution of the BN resulting from setting the value of the evidence (See Introduction). It has been noted early on

that this sampling distribution is far from optimal in the sense that it provides estimates with larger variance than necessary [Shachter and Peot, 1989]. For instance, the optimal sampling distribution in the case of belief inference is to sample the unobserved variables from the posterior distribution over them given the observed evidence. If we knew this distribution we would know the answer to the belief inference problem.

Several modifications have been proposed to improve the estimation of the simple importance sampling distribution discussed above, based on information obtained from the samples [Fung and Chang, 1989, Shachter and Peot, 1989, Shwe and Cooper, 1991, Cheng and Druzdzel, 2000, 2001]. In this chapter, methods to systematically and sequentially update the importance-sampling distribution are proposed. The idea is to view the updating process as one of learning a separate BN just for sampling. The learning objective is to minimize some error criterion. A stochastic-gradient method results from the direct minimization of the variance of the estimator with respect to the importance sampling distribution as an error function. Other stochastic-gradient methods result from minimizing error functions based on typical measures of the notion of *distance* between the current sampling distribution and the optimal, or approximations to the optimal, sampling distribution approximations.

We also propose estimators that make more efficient use of all the samples generated. We present some preliminary theoretical analysis of some instantiations of the class of estimators and update rules. We present preliminary empirical results that show the potential of this technique to improve sampling-based estimation in graphical models. In the context of belief inference, we partially study the application of the adaptive importance sampling methods presented to a particular class of BNs (QMR-DT-type BN) with further special structure that allows a more compact representation of the local conditional probability distributions. In particular, inspired by some of the preliminary empirical results, we start to *shed some light* on the theoretical connection between some instantiations of the AIS methods and *variational approximations*, a type of approximations that has been used effectively for problems in this special class of BNs by Jaakkola and Jordan [1999], who were primarily interested in obtaining strict bounds for the likelihoods and posterior marginals in this model.

We now briefly discuss relevant properties of importance-sampling estimators, before we present the main work of this chapter.

3.1 On importance-sampling (IS) estimators

An important property of the importance-sampling estimator $\hat{G}$, presented in the introduction and given in equation 1.4, is the variance of the weights associated with the importance-sampling distribution. This is

$$\mathrm{Var}[\omega(\boldsymbol{Z})] = \sum_{\boldsymbol{Z}} f(\boldsymbol{Z})\omega(\boldsymbol{Z})^2 - G^2.$$

Recall that $G = \sum_{\boldsymbol{Z}} g(\boldsymbol{Z})$ by definition and assume that g is a positive function. From this we can derive that the optimal or minimum-variance importance-sampling distribution is proportional to $g(\boldsymbol{Z})$:

$$f^*(\boldsymbol{Z}) = \frac{g(\boldsymbol{Z})}{\sum_{\boldsymbol{Z}} g(\boldsymbol{Z})}. \tag{3.1}$$

The weights will have zero variance in that case, since the weight function will always output our value of interest G. Note that the normalizing constant for this distribution is the value we want to compute, G. We also note that we need to avoid letting $f(\boldsymbol{Z})$ be too small with respect to $g(\boldsymbol{Z})$, since this will increase the variance. As a matter of fact, $\mathrm{Var}[\omega(\boldsymbol{Z})] \to \infty$ as $f(\boldsymbol{Z}) \to 0$ for at least one value of $\boldsymbol{Z}$. This implies that we should use importance-sampling distributions with sufficiently "fat tails."

Note that for general g (i.e., not necessarily positive), such that $G \neq 0$, $f^* \propto |g|$, where $|g|$ denotes the absolute value of g. (The reader can verify this by solving the optimization problem using Lagrange multipliers [Kahn and Marshall, 1953].) Therefore, in general, the optimal variance will not be zero. However, this can be made zero by combining two different IS distributions: one to deal with the positive part and the other to deal with the negative part [Owen and Zhou, 1999]. From now on we will assume that the target function g is positive. This is a reasonable assumption for the problem of belief inference in BNs. Since, in principle, we can always perform a linear transformation to make the utility functions positive, this is also a reasonable assumption for the problem of action evaluation in IDs.

3.2 Adaptive importance sampling (AIS)

The traditional method presented previously in the introduction uses as the importance-sampling distribution the do-operated distribution of the BN which can be far from optimal in the sense that it can have higher variance than necessary. In the case of evaluating actions in IDs, it also completely ignores potentially useful information about the utility

values. Therefore, the methods proposed try to learn the optimal importance-sampling distribution by adapting the current sampling distribution as they obtain samples from it. We view this adaptive process as one of learning a distribution over the variables the sum is over to use specifically as an importance-sampling distribution. In particular, we will learn a BNs from the samples just for sampling.

From the expression of the optimal importance-sampling distribution we can deduce that in order to be able to represent this distribution graphically using a BN we *at least* need to add arcs that connect every pair of nodes that are parents of observations and/or utility nodes, if they are not already connected. (This will become more evident from the discussion of Section 3.8.5. For now, one way to see this is as follows. First, in a BN, conditioning on the value of a particular variable, the parent variables become dependent. Hence, all those parent variables will be dependent among themselves—as a group. With regard to the utility nodes, note that knowing the output value from a utility function should give us information about the value of the parents variables of the utility node since the utility function for that node is a direct function of those variables. A more technical answer is that, in our context, the optimal importance-sampling distribution is a Gibbs distribution with respect to an undirected graph where, individually for each observation and utility node, their parents form a *clique* in that graph; that is, they are fully connected.) One problem is that the representation of the distributions local to the observation nodes and the values of the utility nodes are often compact, hence, not requiring that we have a value for each possible instance of the parents. Sometimes those representations use smaller parametric forms (i.e., noisy-or's, noisy-and's, etc.) Pearl [1988], Srinivas [1992], Díez [1993]. By connecting them in the traditional (table-like, non-parametric) way, we can significantly increase the size of the model to a degree that renders the BN impossible to use. If the original model uses the traditional representation for the observation and utility nodes, then connecting their parent nodes does not affect the model-size complexity with respect to the original model.

We will start by concentrating on the problem of learning a BN with the same structure as the original BN (or ID). We will revisit the issue of an optimal sampling BN structure in a Section 3.8.5. For now, let us only consider how to update the local conditional probability distributions as we obtain samples for a given fixed structure. Assume that we use the same structure as the one used by the traditional IS method; that is, the do-operated distribution of the BN or ID. We already saw an example in Figure 1.3. We presented another example in Figure 1.2.

Given a particular structure, we can parameterize the importance-sampling distribution

using a set of parameters Θ. Let $\boldsymbol{Z}$ be those variables not involved in the probabilistic query (those that we have to sum over). Let $\mathrm{Pa}(Z_i)$ be the set of parents of Z_i in the importance-sampling BN graph. Let the indicator function $I(Z_i = k, \mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z}) = 1$ if the condition $Z_i = k$ and $\mathrm{Pa}(Z_i) = j$ agrees with the value assigned to $\boldsymbol{Z}$; 0 otherwise. Then, we can express the importance-sampling distribution as

$$f(\boldsymbol{Z} \mid \Theta) = \prod_{i=1}^{n} \prod_{j \in \Omega_{\mathrm{Pa}(Z_i)}} \prod_{k \in \Omega_{Z_i}} \theta_{ijk}^{I(Z_i=k, \mathrm{Pa}(Z_i)=j \mid \boldsymbol{Z})}, \tag{3.2}$$

where for each i, j, k, $\theta_{ijk} = P(Z_i = k \mid \mathrm{Pa}(Z_i) = j, \Theta)$. Hence, for all i, j, $\sum_k \theta_{ijk} = 1$, and for all k, $\theta_{ijk} > 0$. We will refer to this representation as the *traditional* representation of the parameters of the BN. Note that this representation uses the assumptions of *global* and *local parameter independence* typically used in BNs (See, for instance, Heckerman [1995] and Geiger et al. [1996], who refer to Spiegelhalter and Lauritzen [1990]). The weight function is also parameterized and defined as $\omega(\boldsymbol{Z} \mid \Theta) = g(\boldsymbol{Z})/f(\boldsymbol{Z} \mid \Theta)$, where the definition of the target function g depends on the sum we are evaluating (i.e., the problem under consideration). Other representations are possible and will be discussed later on in this document. From now on, we will refer to the BN used as an importance-sampling distribution the *IS BN (importance-sampling BN)*, and the class of IS BNs that use the traditional representation as the *traditional class of IS BNs*.

Now that we have set the BN structure and parameteric representation, it remains to consider how those parameters will be set or adapted in order to provide good IS distributions. We now present how we do this.

3.2.1 Learning criteria and update rules

In the following subsections, different methods are presented for updating the sampling distribution. The update rules are all based on gradient descent. Hence, at each time t, we update the parameters as follows:

$$\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} - \alpha(t) \nabla^p e(\boldsymbol{\theta}^{(t)}). \tag{3.3}$$

In the update rule above, $\alpha(t)$ denotes the learning rate or the step size rule and $\nabla^p e(\Theta)$ denotes the gradient of error function e, appropriately projected (if necessary) to satisfy the constraints on Θ. The methods differ in how they define $\nabla^p e(\boldsymbol{\theta}^{(t)})$. Let us ignore for now the issue of how to initialize the parameters (i.e., how to set $\boldsymbol{\theta}^{(0)}$).

In the discussion below we denote the $N(t)$ i.i.d. samples as $\boldsymbol{z}^{(t,1)}, \ldots, \boldsymbol{z}^{(t,N(t))}$ drawn

according to $\boldsymbol{Z} \sim f(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(t)})$. If we gather samples to estimate G using many different sampling distributions, how can we combine them to get an unbiased estimate? It is sufficient to weight them using any weighting function that is independent of the sub-estimates obtained by using just the samples for one sampling distribution. For instance, the estimator

$$\hat{G}^{(T)} = \sum_{t=1}^{T} W(t)\hat{G}(\boldsymbol{\theta}^{(t)}), \tag{3.4}$$

where $\sum_{t=1}^{T} W(t) = 1$ and $W(t) \geq 0$, for all t, and

$$\hat{G}(\boldsymbol{\theta}^{(t)}) = \frac{1}{N(t)} \sum_{l=1}^{N(t)} \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)}), \tag{3.5}$$

is unbiased as long as $W(t)$ and $\hat{G}(\boldsymbol{\theta}^{(t)})$ are independent for each t. Letting $W(t) = 1/T$ will produce an unbiased estimate. This is the weight we use in the experiments. In general, we would like to give more weight to importance-sampling distributions with smaller variances. Assuming that the variance decreases with t, we would like $W(t)$ to be an increasing sequence of t. Many alternatives are possible.

Note that using $W(t) \propto 1/\hat{\sigma}_t^2$, where $\hat{\sigma}_t^2$ is the sample variance at time t, though appealing, does not necessarily lead to an unbiased estimator since $W(t)$ and $\hat{G}(\boldsymbol{\theta}^{(t)})$ are not independent [Marshall, 1956]. I will later discuss the issue of how to set the weighting scheme "optimally," and how the variance matrix of the sub-estimators comes into play.

Since in our context, $\boldsymbol{\theta}^{(t)}$ results from adapting the IS BN, from now on we will refer to the class of estimators defined by $\hat{G}^{(T)}$ on equation (3.4) as the *AIS estimators (adaptive-importance-sampling estimators)*, and particular estimates as *AIS estimates*. We present some theoretical properties of the AIS estimators in Section 3.8.3.

We will consider three general strategies: minimizing variance directly, minimizing distance to *global approximations* of the optimal sampling distribution, and minimizing distance to the empirical distribution of the optimal sampling distribution based on *local approximations*. (We will see in Section 3.2.3 how we can also minimize distance to the global optimal sampling distribution without the need of a global approximation.) The expressions are derived for the traditional class of IS BNs. For the first two strategies, we will find that we can express the partial derivatives that form the gradient as, for all i, j, k,

$$\frac{\partial e(\boldsymbol{\Theta})}{\partial \theta_{ijk}} = \sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \left[\frac{-I(Z_i = k, \mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z})}{\theta_{ijk}} \varphi(\boldsymbol{Z}, \boldsymbol{\Theta}) \right],$$

where $\varphi(\boldsymbol{Z}, \boldsymbol{\Theta})$ is a function that depends on the error functions. Note that this is an expectation. Then, the methods update the parameters by estimating the value of the

partial derivatives evaluated at the current setting of the parameters $\boldsymbol{\theta}^{(t)}$ as

$$\frac{\partial \widehat{e(\boldsymbol{\theta}^{(t)})}}{\partial \theta_{ijk}} = \frac{1}{N(t)} \sum_{l=1}^{N(t)} \left[\frac{-I(Z_i = k, \mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z} = \boldsymbol{z}^{(t,l)})}{\theta_{ijk}^{(t)}} \varphi(\boldsymbol{z}^{(t,l)}, \boldsymbol{\theta}^{(t)}) \right].$$

Minimizing variance directly As noted above, the optimal importance-sampling distribution for estimating G is that which minimizes the variance of ω. Using that as the objective, we can derive a stochastic-gradient update rule for the parameters of the importance-sampling distribution. Let the error function be

$$\begin{aligned} e_{\mathrm{Var}}(\boldsymbol{\Theta}) &= \mathrm{Var}(\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta})) \\ &= \sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \omega(\boldsymbol{Z} \mid \boldsymbol{\Theta})^2 - G^2 \end{aligned}$$

The corresponding function for the gradient is

$$\varphi_{\mathrm{Var}}(\boldsymbol{Z}, \boldsymbol{\Theta}) = \omega(\boldsymbol{Z} \mid \boldsymbol{\Theta})^2. \tag{3.6}$$

Note that using this definition of φ yields an unbiased estimate of the gradient. This is because the gradient is the expectation of a particular function and, in this case, we can always evaluate the function exactly. Hence, we can obtain an unbiased estimate by sampling from $f(\boldsymbol{Z} \mid \boldsymbol{\Theta})$.

Minimizing variance indirectly via approximate global minimization Recall the optimal importance-sampling distribution f^* for estimating G given in equation 3.1. The update rules of the following subsection are all motivated by the idea of reducing some notion of *distance* between the current sampling distribution and this optimal sampling distribution. Note that we cannot really compute the values of the optimal distribution since that requires knowing the normalizing constant $\sum_{\boldsymbol{Z}} g(\boldsymbol{Z}) = G$ which is exactly the value we want to estimate. (We will see later how this problem can be avoided for some of the error functions considered in this section.) We approximate the optimal distribution using the current estimate of G as follows

$$\hat{f}^t(\boldsymbol{Z}) = \frac{g(\boldsymbol{Z})}{\hat{G}^{(t)}}. \tag{3.7}$$

In the following, we will consider four error functions, one based on the sum-squared-error and three based on versions of the *Kullback-Leibler divergence*.

If we use the L_2 norm or sum-squared-error function as a notion of distance between the distributions, then the error function is

$$e_{L_2}(\Theta) = \frac{1}{2}\sum_{Z}(f(Z \mid \Theta) - f^*(Z))^2 .$$

The corresponding function for the gradient is

$$\begin{aligned} \varphi_{L_2}(Z,\Theta) &= f^*(Z) - f(Z \mid \Theta) \\ &\approx f(Z \mid \Theta) \times \\ &\quad \left(\omega(Z \mid \Theta)/\hat{G}^{(t)} - 1\right), \end{aligned} \tag{3.8}$$

where the approximation results from using $\hat{f}^t(Z)$ as defined in equation 3.7 as an approximation to $f^*(Z)$.

An alternative, commonly-used notion of *distance* between two probability distributions is given by the *Kullback-Leibler (KL) divergence*. This measure is not symmetric. One version of the KL divergence in this context is given by the error function

$$e_{KL_1}(\Theta) = \sum_{Z} f^*(Z)\log\left(f^*(Z)/f(Z \mid \Theta)\right).$$

The corresponding function for the gradient is

$$\begin{aligned} \varphi_{KL_1}(Z,\Theta) &= f^*(Z)/f(Z \mid \Theta) \\ &\approx \omega(Z \mid \Theta)/\hat{G}^{(t)}. \end{aligned} \tag{3.9}$$

Another version of the KL divergence is given by the error function

$$e_{KL_2}(\Theta) = \sum_{Z} f(Z \mid \Theta)\log\left(f(Z \mid \Theta)/f^*(Z)\right).$$

The corresponding function for the gradient is

$$\begin{aligned} \varphi_{KL_2}(Z,\Theta) &= \log\left(f^*(Z)/f(Z \mid \Theta)\right) - 1 \\ &\approx \log\left(\omega(Z \mid \Theta)/\hat{G}^{(t)}\right) - 1. \end{aligned} \tag{3.10}$$

In general, the version of KL given by $e_{KL_1}(\Theta)$ above makes more sense since the error is an expectation taken with respect to the optimal importance-sampling distribution $f^*(Z)$.

A "symmetrized" version of KL sometimes used is given by the error function

$$e_{KL_s}(\Theta) = \frac{1}{2}e_{KL_1}(\Theta) + \frac{1}{2}e_{KL_2}(\Theta).$$

We can obtain the partial derivatives for this error function and their approximation accordingly.

Heuristic local minimization based on empirical distribution The update methods in this subsection are motivated by the idea of minimizing different notions of distance between the current sampling distribution and an empirical distribution of the optimal importance-sampling distribution that we build from the samples. The hope is that the empirical distribution is a good approximation of the optimal sampling distribution. Let the empirical distribution, parameterized by $\widehat{\boldsymbol{\Theta}}$ locally be as follows: for all i, j, k,

$$\hat{\theta}_{ijk}^{(t)} = \frac{\text{(weighted) average number of times sample } \boldsymbol{z}^{(t,l)} \text{ assigned } Z_i = k, \mathrm{Pa}(Z_i) = j}{\text{(weighted) average number of times sample } \boldsymbol{z}^{(t,l)} \text{ assigned } \mathrm{Pa}(Z_i) = j};$$

that is, for all i, j, k,

$$\hat{\theta}_{ijk}^{(t)} = \frac{\sum_{l=1}^{N(t)} I(Z_i = k, \mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z} = \boldsymbol{z}^{(t,l)}) \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)})}{\sum_{l=1}^{N(t)} I(\mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z} = \boldsymbol{z}^{(t,l)}) \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)})}, \tag{3.11}$$

if $\sum_{l=1}^{N(t)} I(\mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z} = \boldsymbol{z}^{(t,l)}) \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)}) \neq 0$; $\hat{\theta}_{ijk}^{(t)} = \theta_{ijk}^{(t)}$ otherwise. The probability estimate $\hat{\theta}_{ijk}^{(t)}$ is proportional to the (weighted) average number of times that the sample $\boldsymbol{z}^{(t,l)}$ assigned $Z_i = k$ when $\mathrm{Pa}(Z_i) = j$. We are essentially defining the empirical distribution using the samples if there are samples that can be used to define it; otherwise, we revert to the current distribution. We try to minimize the distance between the current sampling distribution and the empirical distribution locally.

Similar to the case of the previous strategies, we will find that we can express the partial derivatives that form the gradient of the error functions discussed in this subsection as, for all i, j, k,

$$\frac{\partial e^{\mathrm{loc}}(\boldsymbol{\Theta})}{\partial \theta_{ijk}} = -\varphi^{\mathrm{loc}}(\hat{\theta}_{ijk}, \theta_{ijk}),$$

where $\varphi^{\mathrm{loc}}(\hat{\theta}_{ijk}, \theta_{ijk})$ is a function that depends on the error functions. Then, the methods update the parameters by estimating the value of the partial derivatives evaluated at the current setting of the parameters $\boldsymbol{\theta}^{(t)}$ as

$$\frac{\partial \widehat{e^{\mathrm{loc}}(\boldsymbol{\theta}^{(t)})}}{\partial \theta_{ijk}} = -\varphi^{\mathrm{loc}}(\hat{\theta}_{ijk}^{(t)}, \theta_{ijk}^{(t)}).$$

We define the *local* L_2-norm error function as

$$e_{\mathrm{L}_2}^{\mathrm{loc}}(\boldsymbol{\Theta}) = \frac{1}{2} \sum_{i,j,k} \left(\theta_{ijk} - \hat{\theta}_{ijk}\right)^2, \tag{3.12}$$

the error function for one version of KL as

$$e^{\mathrm{loc}}_{\mathrm{KL}_1}(\boldsymbol{\Theta}) = \sum_{i,j,k} \hat{\theta}_{ijk} \log\left(\hat{\theta}_{ijk}/\theta_{ijk}\right),$$

and the other as

$$e^{\mathrm{loc}}_{\mathrm{KL}_2}(\boldsymbol{\Theta}) = \sum_{i,j,k} \theta_{ijk} \log\left(\theta_{ijk}/\hat{\theta}_{ijk}\right).$$

From this we obtain the corresponding functions for the gradient:

$$\begin{aligned}
\varphi^{\mathrm{loc}}_{\mathrm{L}_2}(\hat{\theta}_{ijk}, \theta_{ijk}) &= \hat{\theta}_{ijk} - \theta_{ijk}, \\
\varphi^{\mathrm{loc}}_{\mathrm{KL}_1}(\hat{\theta}_{ijk}, \theta_{ijk}) &= \hat{\theta}_{ijk}/\theta_{ijk}, \\
\varphi^{\mathrm{loc}}_{\mathrm{KL}_2}(\hat{\theta}_{ijk}, \theta_{ijk}) &= \log\left(\hat{\theta}_{ijk}/\theta_{ijk}\right) - 1.
\end{aligned}$$

We can obtain an update rule based on the "symmetrized" version of KL accordingly.

3.2.2 Discussion of update rules

First, note that of all the update rules presented thus far, only the one derived for e_{Var} clearly uses an unbiased estimate of the gradient. It is not immediately apparent whether the update rules based on e_{L_2}, e_{KL_1} and e_{KL_2}, as presented above, use unbiased estimates. This is because those update rules use an estimate of the gradient that in turn uses the current estimate of the value we are trying to compute, $\hat{G}^{(t)}$.

Note also that the magnitude of the components of the resulting gradients are different, as suggested by their respective φ functions. The function φ_{Var} has magnitude proportional to the squares of the weights. The magnitudes of φ_{L_2} and φ_{KL_1} are linear in the weights. However, the magnitude of φ_{L_2} is potentially smaller since it has the probability of the sample as a factor. The magnitude of φ_{KL_2} is logarithmic in the weights.

Because we assume that g is positive, the weights are positive. Hence, φ_{Var} and φ_{KL_1} are always positive.[1] The function φ_{L_2} is positive if $\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta})/G > 1$. Similarly, the function φ_{KL_2} is positive if $\log(\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta})/G) > 1$. If $\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta}) > G$ then the sampling distribution underestimates the value of g, while if $\omega(\boldsymbol{Z} \mid \boldsymbol{\theta}) < G$ then it overestimates the value. Therefore, the sign of φ_{L_2} and φ_{KL_2} depends on whether we under- or over-estimated the value of g. Similarly, the magnitudes of φ_{Var}, φ_{L_2}, φ_{KL_1}, and φ_{KL_2} are related to the amount of under- or over-estimation. For φ_{Var}, φ_{L_2} and φ_{KL_1} the magnitude is larger

[1] Note that once the gradient is projected to take care of the constraints on the parameters, the components will not be all positive.

when the sampling distribution underestimates than when it overestimates. For φ_{KL_2}, the logarithm brings the amount of over- and underestimation to the same scale. Note that for the approximations of φ_{L_2}, φ_{KL_1}, and φ_{KL_2}, $\hat{G}$ cannot be zero, and in addition for φ_{KL_2}, $\omega(\boldsymbol{Z} \mid \boldsymbol{\theta})$ cannot be zero. However, these conditions hold from the assumption that g is positive. Note that using the traditional (tabular) representation, unless we constrain the importance-sampling distribution, all the functions φ_{Var}, φ_{L_2}, φ_{KL_1} and φ_{KL_2} will be unbounded even if g is bounded; we will revisit this issue later.

The local L_2 error function, $e^{\mathrm{loc}}_{\mathrm{L}_2}$, leads to an update rule for which the step size has a very intuitive interpretation as a weighting between the current importance-sampling distribution and the empirical distribution. This can help determine an appropriate setting for the step sizes $\alpha(t)$. In the case of $e^{\mathrm{loc}}_{\mathrm{KL}_1}$, the update direction is proportional to the ratio of the empirical distribution with respect to the current importance-sampling distribution. On the other hand, for $e^{\mathrm{loc}}_{\mathrm{KL}_2}$, the update direction is proportional to the logarithm of the same ratio. Note that φ_{KL_2} is not defined if at least one $\hat{\theta}^{(t)}_{ijk} = 0$. We can fix this by letting, for each i, j, k,

$$\hat{\theta}^{(t)}_{ijk} = \frac{\left(\sum_{l=1}^{N(t)} I(Z_i = k, \mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z} = \boldsymbol{z}^{(t,l)}) \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)})\right) + \theta^{(t)}_{ijk}}{\left(\sum_{l=1}^{N(t)} I(\mathrm{Pa}(Z_i) = j \mid \boldsymbol{Z} = \boldsymbol{z}^{(t,l)}) \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)})\right) + 1}.$$

We can interpret this as imposing a "Dirichlet prior" with parameters equal to the current probability values on the empirical distribution parameters.

We can interpret the update rules based on local KL-divergence as adding weights to the elements of the domain of the importance-sampling distribution and renormalizing. For the version of KL-divergence with respect to the empirical distribution, we are always adding weights. We add values relative to the amount we underestimated or overestimated the magnitude of the distribution for a particular state. If we underestimated, we add weights larger than one. If we overestimated, we add weights smaller than one. For the other version of KL-divergence, due to the logarithm function, we add weight if we underestimated while we subtract weight if we overestimated. Therefore, the logarithm brings the amount of underestimation and overestimation to the same scale and adds or subtracts weight accordingly.

Note that when approximating the gradients for e_{Var}, e_{L_2}, e_{KL_1} and e_{KL_2}, we can use as little as one sample to obtain an estimate of the gradient (i.e., $N(t) = 1$). This is not advisable for the method based on the local heuristic since the empirical distribution of the optimal sampling distribution will be highly inaccurate. It is also fair to say that using a single sample to estimate the gradient leads to a larger variance of its estimate. However,

this will not affect its theoretical properties, as we will see later. Hence, the update rules based on the empirical distribution will work better when we take a larger number of samples between updates. Finally, note that when $t = 1$ and $N(t) = 1$, $\varphi_{\mathrm{L}_2} = 0$, and therefore, the parameters will not change in the first iteration.

Note that, through the application of Taylor-expansion approximations we can establish many relationships between the error measures (i.e., their surfaces), particularly between the KL-based errors and the others, in a local neighborhood around the optimal IS distribution (assuming our parametric class of IS BNs can represent it). The most important questions remain unanswered however: Can we characterize the global relationships between the error measures? Is one "smoother" than another? Does one have fewer local minima? Under what conditions one is "better" than the rest? Is one "always better" (in some sense) than the rest? Answers to these questions are important since they can help us determine online which one will be more effective for a particular problem.

3.2.3 Minimizing KL-based difference from actual (not approximate) optimal distribution

We can actually minimize the KL-based errors between the IS BN distribution and the optimal distribution without having to approximate the optimal distribution. To see how this can be done, note that, because G is a constant,

$$\begin{aligned}
\operatorname*{argmin}_{\boldsymbol{\Theta}} e_{\mathrm{KL}_1}(\boldsymbol{\Theta}) &= \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{Z}} (g(\boldsymbol{Z})/G) \log \left(\frac{g(\boldsymbol{Z})/G}{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})} \right) \\
&= \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{Z}} g(\boldsymbol{Z}) \log \left(\frac{g(\boldsymbol{Z})}{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})} \right).
\end{aligned}$$

Hence minimizing the error function

$$e^{\mathrm{equiv}}_{\mathrm{KL}_1}(\boldsymbol{\Theta}) = \sum_{\boldsymbol{Z}} g(\boldsymbol{Z}) \log \left(\frac{g(\boldsymbol{Z})}{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})} \right) \tag{3.13}$$

is equivalent to minimizing the KL error function e_{KL_1}. The derivatives of $e^{\mathrm{equiv}}_{\mathrm{KL}_1}$ however do not require knowledge of the normalizing constant G. The corresponding function of the gradient expression in equation (3.9) is

$$\varphi^{\mathrm{equiv}}_{\mathrm{KL}_1}(\boldsymbol{Z}, \boldsymbol{\Theta}) = \omega(\boldsymbol{Z}, \boldsymbol{\Theta}). \tag{3.14}$$

Similarly, for e_{KL_2},

$$\begin{aligned}
\operatorname*{argmin}_{\boldsymbol{\Theta}} e_{\mathrm{KL}_2}(\boldsymbol{\Theta}) &= \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \log \left(\frac{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})}{\frac{g(\boldsymbol{Z})}{G}} \right) && (3.15) \\
&= \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \log(\frac{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})}{g(\boldsymbol{Z})}) - \log G && (3.16) \\
&= \operatorname*{argmin}_{\boldsymbol{\Theta}} \sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \log(\frac{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})}{g(\boldsymbol{Z})}). && (3.17)
\end{aligned}$$

The equivalent error function is

$$e^{\mathrm{equiv}}_{\mathrm{KL}_2}(\boldsymbol{\Theta}) = \sum_{\boldsymbol{Z}} f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \log(\frac{f(\boldsymbol{Z} \mid \boldsymbol{\Theta})}{g(\boldsymbol{Z})}). \tag{3.18}$$

The corresponding function of the gradient expression in equation (3.10) is

$$\varphi^{\mathrm{equiv}}_{\mathrm{KL}_2}(\boldsymbol{Z}, \boldsymbol{\Theta}) = \log(\omega(\boldsymbol{Z}, \boldsymbol{\Theta})) - 1. \tag{3.19}$$

Similarly, for e_{KL_s}.

We should use the equation for $\varphi^{\mathrm{equiv}}_{\mathrm{KL}_1}$ and $\varphi^{\mathrm{equiv}}_{\mathrm{KL}_2}$ to replace φ_{KL_1} and φ_{KL_2} presented previously. This equivalence is important because now we can evaluate the gradient exactly for particular assignments to $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$. Therefore, we can obtain unbiased estimates of the gradients to use in our updates for minimizing the KL-based differences to the optimal distributions as the error measures. Obtaining unbiased estimates of the gradients can make the theoretical analysis of the convergence of the resulting stochastic-gradient methods simpler. Except where noted, the versions presented in this subsection are not used in the empirical experiments but we consider them in the theoretical analysis.

We now present related work for the problem of computing summations (marginals, posteriors, etc.) in BNs and connect the methods presented in this chapter with AIS methods for other problems in other fields.

3.3 Related work

Different variations of importance sampling have been used for the problems previously presented involving the evaluation of sums (See Lin and Druzdzel [1999] and the references therein). The methods presented here belong to the class of *forward samplers* since they sample from a distribution based on the original structure of the BN. Of these, *self-importance sampling* [Shachter and Peot, 1989, Shwe and Cooper, 1991] and a method called *AIS-BN* [Cheng and Druzdzel, 2000] are the methods closest to the methods proposed here

since they also update the sampling distribution as they obtain information from the samples. The SIS method has an update rule that is very similar to the one derived for $e^{\text{loc}}_{\text{L}_2}$. It updates the distribution after obtaining the empirical distribution, but the update is a weighting between the empirical distribution and the first sampling distribution used [Shwe and Cooper, 1991]. The update rule is

$$\begin{aligned}\theta^{(t+1)}_{ijk} &\leftarrow (1-\alpha(t))\hat{\theta}^{(t)}_{ijk} + \alpha(t)\theta^{(0)}_{ijk} \\ &= \theta^{(t)}_{ijk} - \alpha(t)\left(\theta^{(t)}_{ijk}/\alpha(t) - (1-\alpha(t))\hat{\theta}^{(t)}_{ijk}/\alpha(t) - \theta^{(0)}_{ijk}\right).\end{aligned}$$

In our framework, we can think of this update rule as resulting from the error function

$$e^{\text{loc}}_{SIS}(\boldsymbol{\Theta}, t) = \frac{1}{2\alpha(t)} \sum_{ijk} \left(\theta_{ijk} - \left((1-\alpha(t))\hat{\theta}_{ijk} + \alpha(t)\theta^{(0)}_{ijk}\right)\right)^2.$$

The update rule of AIS-BN is equivalent to $e^{\text{loc}}_{\text{L}_2}$. AIS-BN also uses a BN as the IS distribution and with the same structure as the original BN. In our framework, their update rules results from the local L_2 error function $e^{\text{loc}}_{\text{L}_2}$ (Equation 3.12). Cheng and Druzdzel [2000, 2001] also suggest estimators similar to the AIS estimator in equation (3.4). They present heuristics for initialization of the parameters, weighting schemes for partial estimates and setting of step sizes for the gradient update. They showed empirically that the combination of their update rules with their heuristics had a big impact on the performance of their method over that of the traditional (non-adaptive) importance-sampling method (*likelihood weighting*) when evaluated in three significantly large and complex real BNs. They later suggested new heuristic methods for setting the step size rules used for the gradients, the number of samples used from each importance-sampling BN for estimation and updating, the weighting scheme for combining partial estimators, and for computing relative-approximations using the resulting estimates [Cheng and Druzdzel, 2001].

Other work in the literature on inference in BNs has attempted to find a different, hopefully better, importance-sampling distribution than that used by the traditional likelihood-weighting method (for instance, see Fung and Favero [1994], Cano et al. [1996], Hernández et al. [1998] and the references therein). Typically, a deterministic process is used to find the distribution and no adaptation is done. Since only one IS function is used, the estimates are from a singe IS distribution.

In what follows, we discuss work on AIS on other areas. The problem considered in the literature we present next does not involve problems in BNs or IDs. Also, most (if not all) of that work is primarily interested in computing integrals of different kinds, not summation, and do not deal with problems in BNs or IDs. We briefly summarize this work here because

many of the problems we encounter are, not surprisingly, similar, and it might provide us with a starting point to deal with the problems faced when applying AIS methods in the context of the problems studied in this thesis.

Adaptive importance sampling has been developed and studied in the literature on statistical multiple integration, most particularly, Monte Carlo integration [Flournoy and Tsutakawa, 1991]. The general idea used is similar to that used here, and the update rules based on local heuristics presented here, particularly those resulting from $e^{\text{loc}}_{\text{L}_2}$, are similar to the typical adaptive importance sampling method presented in that literature. The idea is to use the samples to estimate some important characteristics of the optimal IS distribution and update the parameters of the IS distribution so as to try to match those estimates (see Kloek and van Dijk [1978], Oh [1991], Geweke [1991], and Evans [1991], for example). Their main concern is typically the evaluation of "normalized" integrals as is typically necessary in the Bayesian setting for evaluating expectation with respect to posterior densities. Oh [1991] and Evans [1991] also suggest "pooling" the partial estimators as in the AIS estimator presented here. Oh analyzes the effect of dimensionality on the IS estimators for several typical importance-sampling densities and finds that the variance of the weights is affected the most by the dimensionality of the problem (i.e., the dimensionality of the integral, which in our context translates to the number of variables involved in the summations). Oh found that the variance of the weights increases exponentially with the dimensionality, but the increase is only linear in the dimensionality if the IS density *matches the target function well.* This serves as another motivation to try to reduce the variance of the weight function. Rubinstein [1981] also suggests the minimization of the variance and warns about multi-modality as a potential problem when optimizing the parameters of the IS density as to minimize the variance. Rubinstein goes on to suggests that any general global optimization technique can be used, but does not suggest one in particular. In discussing the problem of trying to find a good importance sampling distribution, Geweke [1991] warns against using the same samples from the sampling process itself to adapt the IS density. Geweke suggests as an alternative that one could apply "steepest ascent or other hill-climbing methods to the log-weight function," (in our case $\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta})$, " which would seek out a maximum." Evans [1991] concentrates on the problem of how to find good IS densities. Evans suggests a method based on setting up a "chain" or sequence of integration problems with the following properties: (1) the sequence starts with a problem for which is easy to find a good IS density, from a parameterized family of IS densities, (2) the change from one problem to the next is small enough that we can adapt the IS distribution well by modifying the parameters, and (3) the last problem in the "chain" is the original integration problem.

The idea is to be able to find a good initial IS distribution and control the updates in such a way that good IS densities could be found up to that for the problem itself. In principle, all of these ideas can be applied to the problems considered in this thesis. We do not do so here however, but leave them for future work.

Annealed importance sampling [Neal, 1998, 2001] is a related technique in that it tries to obtain samples from the optimal sampling distribution. As we understand it, the user sets up a sequence of distributions, the last distribution being the optimal distribution, typically defined by Markov chains. We move from one distribution to another as we "anneal" and the sequence converges to the optimal sampling distribution. The hope is that we can get an independent sample from that distribution, then we restart the process to try to obtain another independent sample, and so on. Finally, it uses those independent samples to obtain an estimate. Notice that each "traversal" of the sequence of distributions (or Markov chains) produces a single sample. The technique is very general and we are unaware of whether it has been applied to the problems considered here. There might be connections between the methods presented here and this technique, which we will leave for future work.

In general, I believe one can also view Markov chain Monte Carlo (MCMC) methods [Neal, 1993, Tierney, 1994, Gilks et al., 1996] and its common variants like Metropolis-Hastings [Metropolis et al., 1953, Hastings, 1970] and Gibbs sampling [Geman and Geman, 1984] as sampling methods that "inherently" or "indirectly adapt" the sampling distribution so as to generate samples from the optimal sampling distribution. It is well known that, under some mild conditions, typically met by BNs, asymptotically the samples from MCMC sampling methods will be from the optimal sampling distribution. Many variants of Gibbs sampling and hybrid methods have been proposed in the literature for problems in graphical models [Neal, 1993].

Another "branch" in the literature on AIS methods follows back from very recent theoretical work of Kollman et al. [1999], and Baggerly et al. [2000], showing some conditions for exponential convergence of an adaptive IS procedure in some specific models. This work points back to work in the applied probability theory community from a physics perspective [Fitzgerald et al., 1999, 2000], work in the nuclear science and engineering community (presenting empirical evidence of exponential convergence of an AIS procedure in a particular model) [Booth, 1986, 1989] , and finally, going all the way back to the very "beginnings" of Monte Carlo methods [Meyer, 1956] (where AIS methods extremely similar to those presented here are discussed and studied). The same problem of seeking variance reduction for the traditional IS methods guides all that work. Preliminary review suggests that most methods in that line of research fall primarily in the heuristic-local approximations,

in the context of this document. The estimators are similar, if not the same, as the AIS estimators here in that they reuse samples by combining sub-estimators from the IS distributions found, and they deal with the same problem of how to optimally combine them. Their solutions are about the same as those suggested here. Finally, Halton [1962] proposes and studies a particular instantiation of estimators of this kind and theoretically establish general convergence conditions.

In the chemical-physics community, Alexandrowicz [1971] suggests a method for lattice problems (Ising models) that uses a simpler distribution that allows efficient exact simulation.[2] Such distributions can be seen graphically as a special class of BNs; hence the connection to some instantiations of the general AIS method suggested here. In that work, stochastic adaptation is avoided as much as possible and several analytic expressions are developed. In our context, the error measure corresponds to one of the KL-based global errors and heuristic local approximation errors for others. He establishes the connection of its "analytical formulations" to Kikuchi's approximation [Kikuchi, 1951] (now popularized in the context of graphical models by Yedidia et al. [2001] in their use for the analysis of a popular deterministic approximation method for belief inference in BNs and other graphical models known as *belief propagation*, which was developed in the context of BNs by Pearl [1988]). In particular, he tries to differentiate its method by noting that Kikuchi's "evaluates the distribution of certain figures formed out of spins alloted to neighboring lattice sites." He argued that his method is more applicable by being in a sense more general and states Kikuchi's "analytical formulation appears to be limited to the simpler stochastic models and as such offers fewer possibilities to deal with intricate physical problems than does the full dress Monte Carlo evaluation of stochastic models." On another interesting note, Alexandrowicz also tries to differentiate his method from Metropolis (now the foundation of most MCMC method), and which he incidentally refers to as "the more sophisticated "importance sampling" Monte Carlo technique." Alexandrowicz's main argument against Metropolis' method is a very common one among some researchers in the graphical models community: it is too slow to converge. He argues that it will be impractical for large lattices, because although optimal in the limit, "it still requires a very large number of repetitions of the n [number of variables] step process, and as such permits to describe quite small lattices only."

In the communications community, Al-Qaq et al. [1995] present a method for adapting the IS density that also uses stochastic gradient descent to optimize a given class of IS

[2] I would like to thank Radford Neal for pointing me to this work.

densities. In that sense, this method is very close to ours. In their context, the IS density is a conditional Gaussian and because of special properties of this class, they use standard stochastic approximation arguments to state that the parameters of the IS density found during the adaptive process will converge to a global optimum in the class, with probability one (i.e., a *strong* convergence result).

We should also point out that there is an immediate connection between our methods and variational methods [Jordan et al., 1997].[3] Once we have selected a particular structure for the IS BN to approximate the optimal IS distribution, and parameterized the IS BN according to that structure, the parameters can be viewed as variational parameters that need to be optimized according to some error function. Typically, for problems of inference in graphical models, the error measure e_{KL_2} presented here is used. For some parameterizations and further approximations, the resulting error surface for the variational problem has some nice properties (i.e., convexity), making the optimization problem easier. Also, typically deterministic methods have been used, while here stochastic methods are used. In principle, there is no reason why stochastic methods could not have been used for variational problems. Similarly, there is no reason, in principle, that in those cases that the parameterization makes it amenable, we could not use deterministic methods instead to minimize the error functions proposed here. We will briefly discuss this connection in Section 3.7.

Before we present the empirical results, in the following two sections, we discuss some practical implementation issues and compare the "run-time complexity" of the AIS methods to the traditional IS method.

3.4 Implementation issues

All of the adaptive importance-sampling methods described above in Section 3.2.1 were implemented for the experiments. Next, we present some of the implementation issues. We describe how we dealt with these issues in the implementations of the methods. Nevertheless, fully satisfactory solutions remain an open problem.

3.4.1 Learning rate

The learning rate used is $\alpha(t) = \beta/t$, where β is a value that depends on the updating method. We need different values of β for the different methods because of the differences

[3] We would like to acknowledge Nir Friedman for hinting to us the connection between AIS and variational approximations.

in magnitude of their gradients.

Cheng and Druzdzel [2000] present alternative heuristics for setting the learning rate. Work from the literature on learning neural networks, machine learning techniques, and general nonlinear optimization techniques can be also used in principle.

3.4.2 Avoiding extreme probabilities

An additional constraint is imposed on the parameters which we call the ϵ-*boundary*. This constraint requires that for all i, j, k, $\theta_{ijk} \geq \epsilon \times (|\Omega_{X_i}|) = \gamma / |\Omega_{X_i}|$, where γ is a constant factor. In the experiments, $\gamma = 0.1$ is used. This constraint is introduced so that the sampling distribution avoids extrema in probability and hence the possibility of infinite variance.

Cheng and Druzdzel [2000] suggests similar heuristics to avoid extrema probabilities in the sampling BN by imposing a constraint on the smallest conditional probability value of a node. In Section 3.8.2, we present alternative ways to avoid extreme probabilities.

3.4.3 Initial importance-sampling distribution

The parameters $\boldsymbol{\theta}^{(0)}$ are initialized such that the starting importance-sampling distribution is the do-operated probability distribution of the original BN. However, if one of the local conditional probability values does not satisfy the ϵ-boundary constraint, we change the distribution so that it does. In the experiments, we do this by forcing each probability value that falls outside the constraint to be at the boundary (i.e., if we are dealing with the conditional probability of variable X_i, we set the value to be $\gamma / |\Omega_{X_i}|$) and subtracting the total amount required to move those probabilities to the boundary uniformly among all the probability values that fall inside the constraint space. If after doing so, other values fall outside the constraint space, we keep moving them until all the probability values satisfy the ϵ-boundary constraint. (Less *ad hoc* ways of doing this are possible.) Note that only probabilities that are strictly greater than zero need to be moved to the constraint space. This is because assignments to the hidden variables $\boldsymbol{Z}$ that agree with assignments for which any of the local conditional (or marginal) probabilities has value zero do not contribute to the sum under consideration (i.e., the value of the target function g for those assignments is zero!). Hence, those probabilities do not need to be learned, reducing the total number of parameters needed to be learned for the sampling BN. Such reductions in the number of parameters are important since, in general, the smaller the search space defined by the class of sampling BNs, the "nicer" the error surface and the easier the optimization problem.

There are many other ways to initialize the parameters of the first IS BN for AIS. We will discuss one of many possible alternatives. For instance, the parameters and structure of the BN or ID under consideration can help us do this. One idea is to have an IS BN such that for variables not directly relevant to the utility nodes uses the original local conditional probability distributions of the ID, while for variables directly relevant to the utility nodes uses a distribution that is proportional to the weighting of the utility values with the original local conditional probability distribution associated with those variables. In other words, in general, one would want those nodes that are hidden nodes parents of evidence and/or utility nodes to have a conditional distribution defined by their *Markov blanket*, such that they could have *knowledge* of the local probability of the evidence and/or utility values. However, computing such conditional distribution can be too computationally intensive in general. One can always obtain simpler (less-computationally intensive), but more naive, initializations.

Other heuristic initialization techniques have been proposed by Cheng and Druzdzel [2000] in the context of belief inference in BNs. They showed in some sufficiently large and complex real BNs that their heuristics provided a significant improvement in quality of the estimates they produced from the adaptive sampler.

There is no doubt that this is a very important part of the potential success (or failure) of the adaptive importance sampling as a viable practical technique. Hence, the initialization problem requires more attention than given in this work. Further study of this problem is left as future work.

3.4.4 Dealing with parameter constraints

In the implementation used in the experiment, in order to satisfy the constraint that for all i, j, $\sum_k \theta_{ijk} = 1$, the approximation of the gradients is projected onto the simplex of the local conditional probability distribution [Bertsekas, 1995, Binder et al., 1997]. This is done by letting, for all i, j, k,

$$\frac{\partial \widehat{p_{e(\theta)}}}{\partial \theta_{ijk}} \leftarrow \frac{\partial \widehat{e(\theta)}}{\partial \theta_{ijk}} - \frac{1}{|\Omega_{Z_i}|} \sum_{k=1}^{|\Omega_{Z_i}|} \frac{\partial \widehat{e(\theta)}}{\partial \theta_{ijk}}. \tag{3.20}$$

Note that this is not enough to guarantee that after taking a step in the projected direction, the parameters will remain in the constraint space. If, when updating a local conditional probability distribution, its respective parameters do not satisfy the constraint, the minimum step α' that will allow them to remain inside the constraint space is found and a step of size $\alpha'/2$ along the gradient direction (i.e., half the distance between the current position

of the parameter we are updating in the simplex and the closest point on the ϵ-boundary along the gradient direction) is taken.

Other potentially better ways of dealing with parameter constraints are possible. In Section 3.8.1, we will see ways to optimize the error function on an unconstrained space by making a parameter transformation leading to a parameterization without constraints.

3.5 Cost for AIS

In this section, we will briefly discuss the "run-time complexity" of AIS as compared to the traditional IS method. The discussion is mainly meant to be an illustrative, rather than rigorous, account of the complexity of the AIS methods.

We now present the approximate cost per stage for AIS methods. Let $N(t) = N_s$ be the number of samples per stage. For simplicity, assume it is constant (the same for all stages). Let n be the total number of nodes in the original ID. Approximately, the basic units of time used in the discussion below are operations like generating a random number, basic math like addition, subtraction, multiplication, and division, accessing conditional distribution probability tables, and evaluating the utility function.

The approximate cost per stage of AIS is:

- Getting samples

$$N_s \times (2 \times \sum_{\text{adaptive-nodes}} \text{\#-node-assigments})$$

- Evaluating target

$$N_s \times n$$

- Evaluating IS probability

$$N_s \times (\text{\#-adaptive-nodes})$$

- Computing weights

$$N_s$$

- Evaluating gradient

$$\sum_{\text{adaptive-nodes}} (\text{\#-parent-assignments}) \times (\text{\#-node-assignments}) \times (4N_s + 1)$$

- Updating IS distribution

$$\sum_{\text{adaptive-nodes}} (\text{\#-parent-assignments}) \times (\text{projection-cost} + \text{constraints-cost})$$

- Computing average

$$N_s$$

- Total

$$N_s(2 \times \sum_{\text{adaptive-nodes}} \text{\#-node-assignments} + n + \text{\#-adaptive-nodes} + 2) + \sum_{\text{adaptive-nodes}} (\text{\#-parent-assignments}) \times ((\text{\#-node-assignments}) \times (4N_s + 3) + \text{projection-cost} + \text{constraints-cost})$$

Let N be the total number of samples. The approximate cost of the traditional IS sampling is:

- Getting samples

$$N \times (2 \times \sum_{\text{\#-adaptive-nodes}} \text{\#-node-assignments})$$

- Getting weights

$$N \times (\text{\#-obs-nodes})$$

- Compute average

$$N$$

- Total

$$N(2 \sum_{\text{\#-adaptive-nodes}} \text{\#-node-assignments} + \text{\#-obs-nodes} + 1)$$

Let

1. $N_s = \lceil N/(\#\text{-AIS-stages}) \rceil$,
2. $m = \#\text{-adaptive-nodes}$,
3. $o = \#\text{-obs-nodes}$,
4. $a = \sum_{\text{adaptive-nodes}} \#\text{-node-assignments}$,
5. $b = \sum_{\text{adaptive-nodes}} \#\text{-parent-assignments}$,
6. $c = \sum_{\text{adaptive-nodes}} (\#\text{-parent-assignments}) \times (\#\text{-node-assignments})$,
7. $d =$ projection-cost,
8. $e =$ constraints-cost.

The ratio of AIS "complexity" with respect to traditional IS is

$$\text{ratio} \approx \frac{2a + n + m + 2 + 4c}{2a + o + 1} + \frac{3(c + b(d + e))}{N_s(2a + o + 1)} \tag{3.21}$$

Hence, the "complexity" hit for AIS decreases as one increases the number of samples per stage.

Unless otherwise noted, we use the *number of samples taken* instead of *CPU times* to compare the different methods. This is to disregard implementation details as a factor in the comparisons. In general, the reader should be aware of the overhead involved in using AIS, as suggested by the discussion above. We argue during the discussion of the empirical results that the improvement of the quality of the estimators as a function of the number of samples more than compensates for the overhead penalty.

3.6 Preliminary empirical results

The main objective for our empirical evaluation is to illustrate the potential of AIS to improve the traditional IS method for estimation problems. The empirical results presented in this section are for two artificial problems. One involved the estimation of action values on a simple ID. The other involves the estimation of the probability of an observation in a somewhat complex BN.

3.6.1 Results on computer-mouse ID problem

The methods were tested on the *computer mouse problem*, a simple made-up ID shown in Figure A.1 and described in Appendix A. All the utility values were increased by one unit to make g positive. The problem considered was to obtain the value $V_{MP_t}(A)$ for the action $A = 2$ and the observation $MP_t = 1$.

Each method was evaluated by computing the *mean-squared-error (MSE)* between the true value of the expectation of interest ($V_{MP_t}(A)$) and the estimate generated using the adaptive sampling method. The first results show how the methods achieve better MSEs with fewer samples for this problem. Only results for those methods that were the most competitive are shown. Let us denote by "Var" the method based on the minimization of the variance, and by "L2 ", "KL1", and "KLS" the methods based on the (approximate) global minimization of L_2, KL_1 and KL_s respectively. For the update methods we use $N(t) = 1$ for all t. We need to take into account that the update methods have to traverse the graph once every iteration to update the parameters relevant to the sample taken. To compensate for this time, the estimate based on LW is allowed to use twice as many samples. Figure 3.1 shows the results. The graph shows the average MSE over 40 runs as a function of the total number of samples taken (times 2 for LW) by the methods. Note that Var and L2 achieve better MSEs than LW and converge to them faster. With significance level 0.005 it can be stated (individually) for each total number of samples $N = 50, 150, 250$, that Var and L2 (individually) are better with respect to MSE than LW. Also, for $N = 250$, KLS is better than LW.

The methods were also tested with $N(t) = 50$, including the local heuristic methods. They were only competitive after a larger total number of samples ($N > 150$). Although further analysis is necessary, some general observations can be conveyed. In general there seems to be a tradeoff in the setting of $N(t)$ and β. We note that, of the updates based on the two KL versions, KL1 typically performs better than KL2. It is believed this is because the error function e_{KL_1} is defined with respect to the optimal sampling distribution while e_{KL_2} is with respect to the current sampling distribution. KLS seems to perform better than both. L2 is more stable than any of the other methods, suggesting further theoretical analysis. Several possible reasons for this behavior are (1) the variance of the gradient might be smaller than in other cases, (2) the error function is bounded, and/or (3) the error surface might be smoother than in other cases. We conjecture that L2 converges to a stationary point of e_{L_2}.

The second result shows that the update methods indeed lead to importance-sampling

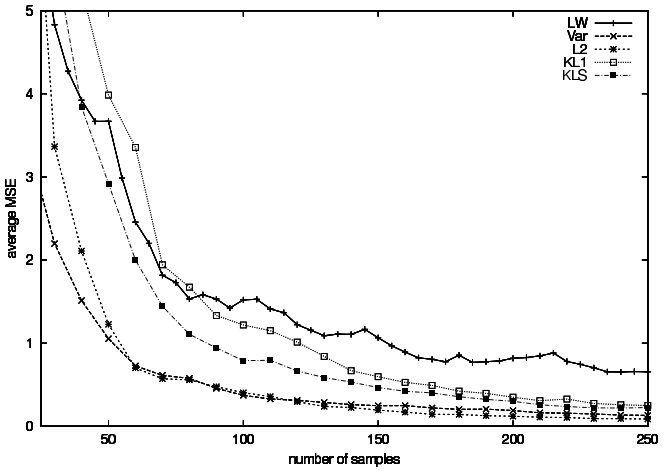


Figure 3.1: Average mean squared error, over 40 runs, as a function of the number of samples taken.

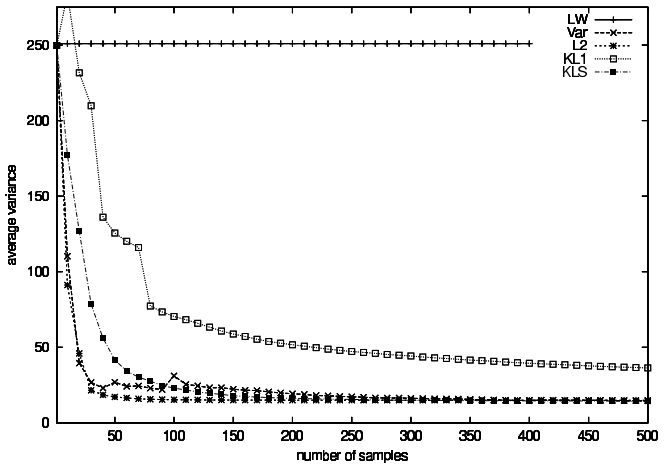


Figure 3.2: Average of the true variance of the weight function, over 40 runs, as a function of the total number of samples taken.

distributions with smaller variance relatively quickly for this problem. Figure 3.2 shows a graph of the true variance of the sampling distribution learned using the different update methods as a function of the total number of samples used. The horizontal line shows the variance associated with the sampling distribution used by LW (i.e., the do-operated distribution of the original BN).

In summary, the results show that AIS methods can indeed provide better estimators than the traditional IS method for this ID problem. We consider a BN problem next.

3.6.2 Results on QMR-DT-type BN

The results in this section are from applying the AIS methods to a randomly generated, synthetic QMR-DT type BN. The QMR-DT is a well known real, large, multiply-connected BN that has been studied in the community (see, for example, Shwe and Cooper [1991] and Jaakkola and Jordan [1999] and the references therein) and has some very particular structure. As typically presented, the QMR-DT model is a two-layer BN, one layer representing diseases and the other findings, with arcs going from diseases to findings and no arcs among variables at the same layer (another layer representing background information is sometimes shown but it has no effect on the complexity of the inference problem, since those background variables are all parents of the disease variables and once instantiated they defined a prior distribution over each disease.) Hence, the disease variables are marginally independent, and the findings are conditionally independent given the diseases. However, in general, given just findings, the diseases become dependent.

The synthetic BN was randomly generated. It has 20 disease nodes and 40 finding nodes. The average number of parents for each finding node is 16 and at least one has all the diseases as parents. The large number of parents is possible because this models assumes a "causally" motivated mode of interaction between the diseases (*causes*) and the finding (*effects*). This model is known as the *noisy-or* [Pearl, 1988]. It defines the conditional probability distribution of each finding given its parent diseases such that the probability of obtaining a *negative* finding decreases monotonically with the number of potentially causing diseases (i.e., parents of the finding variable) active. The model requires, for each causing parent disease, the probability that its child finding will be negative given that just that disease is active and all the other parent diseases are inactive. The conditional probability that a finding is negative is then the product, for each active disease parent, of the probabilities the finding is negative given that that particular disease is the only one active (recall, this is given as a parameter of the model). Also, the model typically uses

a *leak* probability of having a positive finding even in the absence of any cause, maybe as a result of some other conditions, which are always assumed to be active. Note that the number of parameters to define the conditional probability distribution for each finding node using a noisy-or is exactly the number of parents plus one (i.e., linear in the number of parents, as opposed to being the number of all possible assignments of the parents which is exponential in the number of parents).

For our synthetic model, the probabilities were generated uniformly at random in $(0, 1)$. The leak probabilities were generated uniformly at random in $(0.9, 1)$. The parents of each finding node were selected at random. We generated a random evidence case which had 29 positive findings, and its probability was 3.6884^{-16}. The probability was computed using the traditional junction-tree exact method [Lauritzen and Spiegelhalter, 1988, Jensen et al., 1989]. We used the implementation by Murphy [1999]. There are other exact methods that exploit the local properties of the QMR-DT model that we could have used to compute the exact probability. In particular, Heckerman [1989] developed a method called *Quickscore* that is exponential on the number of positive findings. Takikawa and D'Ambrosio [1999] also developed a method based on a local transformation of the noisy-or model, which can be seen as a generalization of Quickscore in the context of the QMR-DT model, and can potentially be superior to Quickscore.

Our task was to estimate the probability of the evidence. We tried a variety of AIS methods with different settings. We tried the traditional IS methods to use as a baseline comparison. Some of the results for the AIS methods along with the corresponding descriptions are found in Figures 3.3 and 3.4 (See Appendix C for all the results). First, some basic general descriptions of the graphs. The y-axis in all the figures is in natural-log scale (for readability purposes). The natural-log of the probability of the random evidence case considered is -35.5362. The x-axis is the number of samples. The y-axis is the estimates generated by each estimator (as a function of the number of samples, and in natural-log scale). Each figure present estimates for each of 10 runs of each method under the particular settings of the methods' meta-parameters β and $N(t)$ (i.e., the number of samples per iteration update, was set constant—did not depend on t). The results from the 10 runs of the traditional (LW) method is displayed in the top-left graph on each figure.

The general observation from the results is that the AIS methods can be very effective in estimating the marginal quantity—the probability of the random evidence case. They can obtain significantly better estimates with many fewer samples: the estimates converge quickly to the answer. The step-function behavior observed in some of the graphs is due to the relatively sharp changes on the estimates resulting from observing many samples

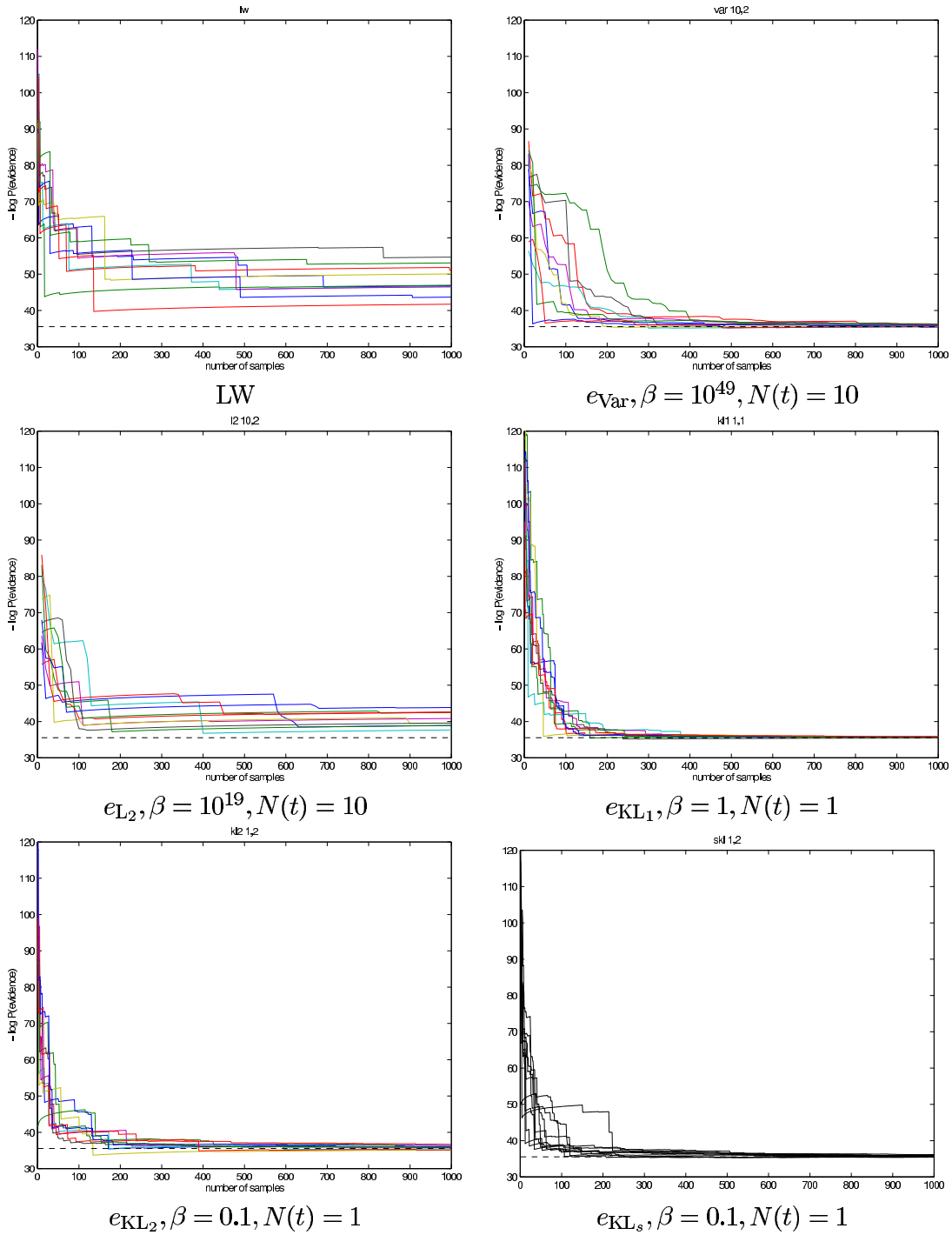
with small weights, sporadically followed by a sample with large weight. This behavior (particularly for LW) suggests that the IS distribution being used is not good (produces estimates with large variance). The sharp changes are further magnified by the log function.

Although we could not compute the actual variance of the estimators, a visual inspection of the sample weights suggested that the variance of the estimators was being reduced. Hence, we believe that the improvement of the AIS methods was indeed produced by a reduction in the variance of the IS estimators. We cannot really argue that the methods were faster than LW since our implementations were not good enough to assess running times in a fair way. However, we believe even though AIS methods have larger overhead than LW (particularly when the number of samples between updates is small), the improvements in the quality of the estimate more than compensate for such overhead. (Cheng and Druzdzel [2000] found that this was the case for a special type of AIS method.) Note that even after a significantly larger number of samples than those of the AIS methods the LW estimates are still orders of magnitude away from the true value.

We believe the empirical results for the synthetic QMR-DT model can be explained through the connection of the IS BN class we use in the experiments and the mean-field approximations which have been used before for this model [Jaakkola and Jordan, 1999, Jordan et al., 1997]. We briefly discuss this connection further in the next section. Through that connection and the results obtained in this experiment, we believe that the problem considered for this experiment is relatively easy, because the posterior distribution is probably *closely* unimodal and therefore, the mean-field approximation provides a *good* approximation. It is interesting, however, that the estimate of the probability of evidence given by LW is not very good. We suspect that the estimates of the posterior marginals for each "disease" node are if not all good, at least *reasonable* in that they point us in the right direction (the "modes" of the estimated posterior marginals fall in the correct side, at least most of the time). We did not keep the estimates of the posterior marginals generated by LW however, which would have helped us test this claim.

3.6.3 Preliminary conclusions

These experiments are all carried out on synthetic problems. Although they must clearly be extended to a variety of larger problems, they indicate that adaptive importance-sampling methods, particularly those that minimize variance and the L_2 norm, can lead to significant improvements in the efficiency of sampling as a method for computing large expectations.



LW

$e_{\mathrm{Var}}, \beta = 10^{49}, N(t) = 10$

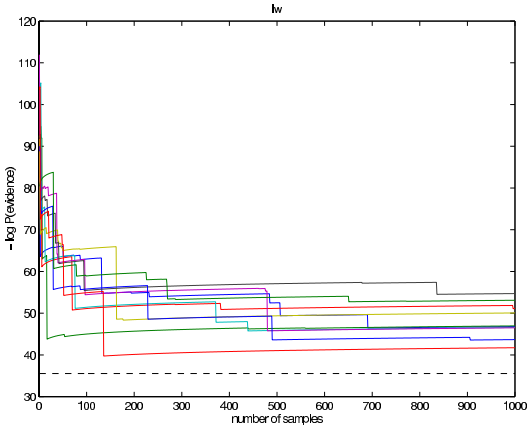
$e_{\mathrm{L}_2}, \beta = 10^{19}, N(t) = 10$

$e_{\mathrm{KL}_1}, \beta = 1, N(t) = 1$

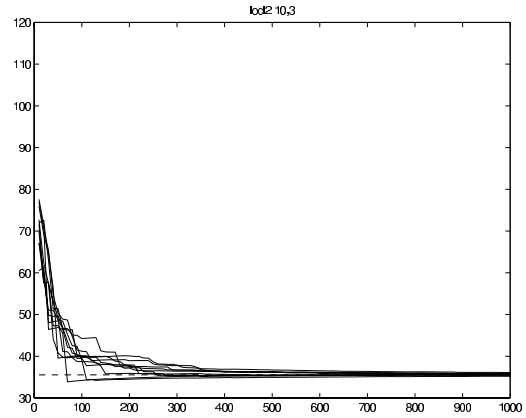
$e_{\mathrm{KL}_2}, \beta = 0.1, N(t) = 1$

$e_{\mathrm{KL}_s}, \beta = 0.1, N(t) = 1$

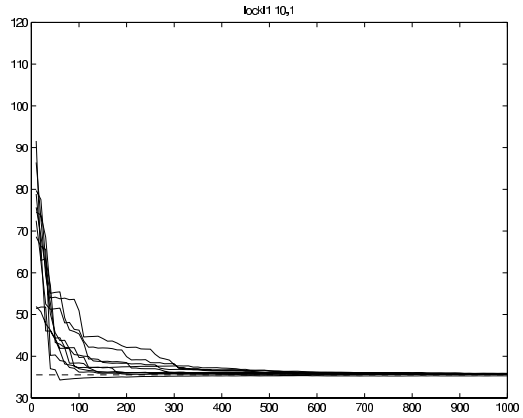
Figure 3.3: Results for AIS method based for estimating the probability of a random evidence in the synthetic QMR-DT model. Refer to the text for basic general descriptions. Continue in Figure 3.4.



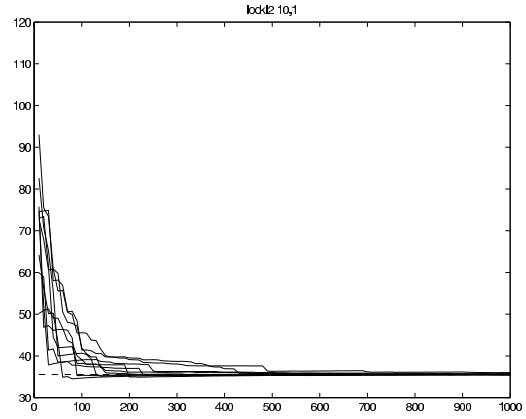
LW



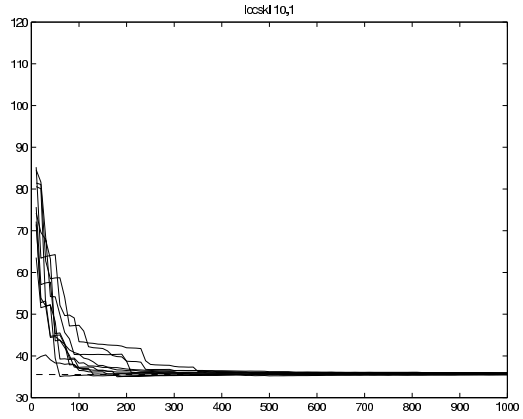
$e^{\text{loc}}_{\text{L}_2}, \beta = 10, N(t) = 1$



$e^{\text{loc}}_{\text{KL}_1}, \beta = 1, N(t) = 10$



$e^{\text{loc}}_{\text{KL}_2}, \beta = 1, N(t) = 10$



$e^{\text{loc}}_{\text{KL}_s}, \beta = 1, N(t) = 10$

Figure 3.4: Results for AIS method for estimating the probability of a random evidence in the synthetic QMR-DT model. Refer to the text for basic general descriptions. Continue in Figure 3.4.

3.7 On AIS with mean-field approximations

In this section, we briefly discuss some properties of AIS when applied with IS BN structure for a mean-field approximation. Th effectiveness AIS methods showed on the empirical test for the QMR-DT-type BN, for which we use a what we will call a *mean-field IS BN*, motivated the following discusion. (The discussion is presented on a high level and the details are left for a future document.) The objective is to shed some light into the properties of running AIS with the special mean-field IS BNs for general BN and ID problems with binary hidden variables $\boldsymbol{Z}$, and by doing so, start to establish the connection between AIS and variational methods.

For simplicity, let us assume that the (hidden) variables $Z_1, \ldots, Z_n$ are binary valued. Then, the mean-field IS BN distribution discussed in this section can be more simply expressed as

$$f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{n} \theta_i^{Z_i} \, (1-\theta_i)^{1-Z_i} \,, \tag{3.22}$$

where, for all i, the parameter $\theta_i = f(Z_i = 1 \mid \boldsymbol{\Theta}) = \sum_{\boldsymbol{Z_{-i}}} f(\boldsymbol{Z_{-i}}, Z_i = 1 \mid \boldsymbol{\Theta})$ and $\boldsymbol{Z_{-i}} = \boldsymbol{Z} - \{Z_i\}$. We refer to this class of IS BNs as the "mean-field" class of IS BNs. It corresponds to a "degenerate BN" where all the nodes are disconnected (i.e., all the Z_i are marginally independent).

The error function resulting from using this class of IS BN distributions have some interesting convexity properties. If a function is convex, then it has a single stationary point which corresponds to the global minimum. In our context, the global minimum of the error function used is the best IS BN for that error function (i.e., the mean-field IS BN for which the error function attains its smallest value). In general, results establishing conditions for the convergence of stochastic-gradient methods similar to those used here would typically lead to convergence to stationary points only, not necessarily minima. But, if there is a single stationary point which is a global minimum, then we establish convergence to the global minimum. Finally, the global and local versions of the respective error functions become very similar.

In particular, for the mean-field IS BN class, the error function e_{KL_1} is convex on $\boldsymbol{\Theta}$ and for the case of estimating likelihoods of evidence ($P(\boldsymbol{O} = \boldsymbol{o})$) in BNs, the (global) minimum is attained at the posterior marginals. Hence, if the parameters $\boldsymbol{\theta}^{(t)}$ resulting from the update process converge, each individual parameter $\theta_i^{(t)}$ will converge to its respective posterior marginal ($P(Z_i = 1 \mid \boldsymbol{O} = \boldsymbol{o})$), as expected from the nature of the mean-field

approximation. This is useful since those quantities are often of interest (specially in QMR-DT-type BN problems) and we obtain an approximation for them as a by-product of the AIS update process.

The error function e_{KL_2} is often used in variational methods for belief inference in BNs. For the special mean-field IS BN, for each i, each subfunction resulting from fixing all parameters but θ_i in e_{KL_2} is convex in θ_i (its only parameter). Similarly, for each i, each subfunction resulting from fixing all parameters but θ_i in e_{L_2} is convex *and quadratic* in θ_i. This suggests that e_{L_2} is a "nicer" function than e_{KL_2}. Finally, for each i, each subfunction resulting from fixing all parameters but θ_i in e_{Var} is convex and has the form $a_i/\theta_i + b_i/(1-\theta_i)$ for constants a_i and b_i.

Note also, that although what we have stated above holds in general for any problem where we use the mean-field IS BN class, we can exploit further special properties of the particular problem to provide stronger statements about the characteristics of the resulting error functions. In particular, I believe this is the case for the QMR-DT-type BN problem. Also, I believe that the analysis above, combined with the discussion of the empirical results for the QMR-DT-type BN problem given in the previous section, better explain the stability of the AIS methods and the rapid convergence of the estimates seen in the empirical results for that problem.

In the next section, we study theoretical properties of the AIS methods using more general classes of IS BNs than that considered in this section.

3.8 On theoretical properties of AIS

Given the potential effectiveness of the AIS methods, in this section, we analyze some of the AIS methods and estimators from a theoretical perspective and present some results. In the process, we consider some variants of the AIS methods which involve a new parameterization of the IS BN, and another subclass of IS BN distributions that when used with AIS, are more amenable to theoretical analysis. The variants are not only of theoretical interest but can easily be used in practice, and be effective. I also discuss the issue of optimal IS BN structures and optimal weighting schemes for AIS estimators.

3.8.1 Canonical (re)parameterization of the IS BN

An alternative parameterization that will prove to be useful later when we consider some theoretical properties of the adaptive sampling process is based on the canonical representation of the local conditional probability distributions. Let n be the number of variables

in the IS BN. In this representation [4], we let, for each $i \in \{1, \ldots, n\}$,$j \in \Omega_{\mathrm{Pa}(Z_i)}$, and a new set of parameters, $\tau_{ijl} \in \mathcal{R}$, for $l = 1, \ldots, |\Omega_{Z_i}| - 1$,

$$\theta_{ijk}(\boldsymbol{\tau}_{ij}) = \frac{e^{\tau_{ijk}}}{\sum_{k=1}^{|\Omega_{Z_i}|-1} e^{\tau_{ijk}} + 1}. \tag{3.23}$$

for $k = 1, \ldots, |\Omega_{Z_i}| - 1$, and

$$\theta_{ij|\Omega_{Z_i}|}(\boldsymbol{\tau}_{ij}) = \frac{1}{\sum_{k=1}^{|\Omega_{Z_i}|-1} e^{\tau_{ijk}} + 1}. \tag{3.24}$$

The expression of the joint distribution can be expresses as in equation (3.2), but with $\boldsymbol{\Theta}$ replaced by $\boldsymbol{\Theta}(\boldsymbol{\tau})$ (and similarly $\boldsymbol{\theta}_{ijk}$ by $\theta_{ijk}(\boldsymbol{\tau}_{ij})$). Note now that the sampling BNs are parameterized by $\boldsymbol{\tau}$, which are unconstrained parameters over all $\mathcal{R}$. So any assignment to $\boldsymbol{\tau}$ produces a setting of $\boldsymbol{\Theta}$ that satisfy their constraints automatically. This reduces the optimization problem from a constrained to an unconstrained one.

The partial derivatives of $f \equiv f(\boldsymbol{Z} \mid \boldsymbol{\Theta}(\boldsymbol{\tau}))$ with respect to $\boldsymbol{\tau}$ are as follows: for each i, j, l,

$$\begin{aligned}
\frac{\partial f}{\partial \tau_{ijl}} &= \sum_{k=1}^{M_i} \frac{\partial f}{\partial \theta_{ijk}} \frac{\partial \theta_{ijk}}{\partial \tau_{ijl}} \\
&= \sum_{k=1, k \neq l}^{M_i} (f \times I[Z_i = k, \mathrm{Pa}(Z_i) = j]/\theta_{ijk}) (-\theta_{ijk}\theta_{ijl}) + \\
&\quad (f \times I[Z_i = l, \mathrm{Pa}(Z_i) = j]/\theta_{ijl}) (-\theta_{ijl}\theta_{ijl} + \theta_{ijl}) \\
&= f \times I[\mathrm{Pa}(Z_i) = j] \left(\sum_{k=1}^{M_i} I[Z_i = k](-\theta_{ijl}) + I[Z_i = l] \right) \\
&= f \times I[\mathrm{Pa}(Z_i) = j](I[Z_i = l] - \theta_{ijl}).
\end{aligned}$$

3.8.2 How to bound the smallest IS BN probability

In what follows, I present two ways in which we can define a class of IS BNs for which we can lower bound the probability of any event, and that will be useful for theoretically analyzing the behavior of the AIS methods.

Global mixing

This class results from mixing a BN with a uniform distribution globally. That is, the globally mixing sampling distribution class is the set of all distributions f^{gmix} such that

$$f^{\mathrm{gmix}}(\boldsymbol{Z} \mid \boldsymbol{\Theta}) = (1 - \Delta) f(\boldsymbol{Z} \mid \boldsymbol{\Theta}) + \Delta \frac{1}{M} \tag{3.25}$$

[4] This representation was originally suggested to me by Thomas Hofmann.

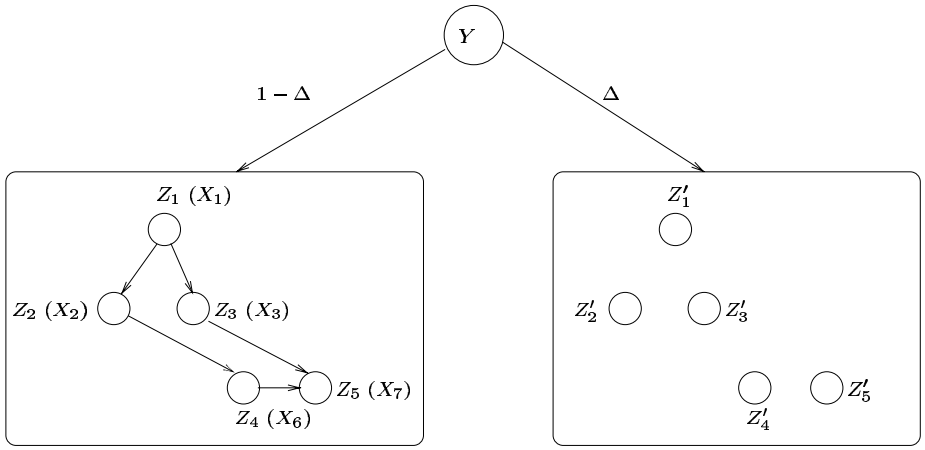


Figure 3.5: Graphical representation of global-mixing IS BN class.

where $\Delta \in (0,1)$ (presumably less than $1/2$), and $M \equiv |\Omega_{\boldsymbol{Z}}|$. Similarly, for the case where the actual parameters are $\boldsymbol{\tau}$. We can interpret this sampling BN class as the (global) mixing of the probability distributions of two BNs (See Figure 3.5). With probability $1-\Delta$ we sample from the original parameterized BN. With probability Δ we sample from a uniform BN (a BN with no arcs and uniform probability distributions on each node) which will not be adapted.

Local mixing

This class results from mixing each local conditional probability distribution of a BN with a uniform distribution, for each assignment of the parents. That is, the set of all $f^{\text{lmix}}(\boldsymbol{Z} \mid \boldsymbol{\Theta})$ such that

$$f^{\text{lmix}}(\boldsymbol{Z} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{n} \prod_{j=1}^{|\Omega_{\text{Pa}(Z_i)}|} \prod_{k=1}^{M_i} \left((1-\Delta)\theta_{ijk} + \Delta \frac{1}{M_i} \right) \tag{3.26}$$

where n is the number of nodes in the IS BN, and $M_i \equiv |\Omega_{Z_i}|$. Again, similarly for the case where the parameters are $\boldsymbol{\tau}$. We can interpret this model graphically (See Figure 3.6). For each node we introduce a new "locally mixing" node as parent. The new mixing random variable represented by this node is binary: in one state, which is attained with probability $1-\Delta$, it tells its child node to sample its own state from its "original" IS BN local (conditional) distribution, considering the value of its other parents (if any); in the other state, which it attains with probability Δ, it tells its child node to sample from a uniform probability distribution, and hence ignoring the values of its other parents. These

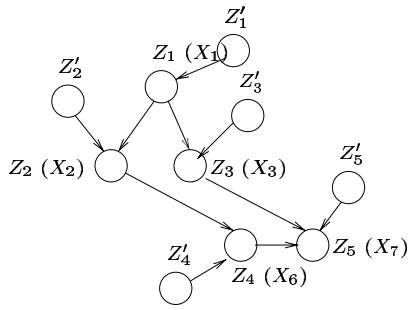


Figure 3.6: Graphical representation of local-mixing IS BN class

nodes need to be marginalized to obtain the joint probability distribution over the original set of nodes $\boldsymbol{Z}$. The marginalization is simple, since the "mixing nodes" are all roots in the "expanded" BN, leading to the expression above for f^{lmix}.

If mixing is used, the class of sampling BNs considered during the adaptive sampling process is different; that is, we need to define our sampling BNs and weight functions using f^{gmix} or f^{lmix} instead of f. The derivatives needed for the update rules with respect to this class can be as easily obtained as for the previous traditional IS BN class discussed earlier.

Partial derivatives for "local-mixing" sampling BN class

This derivatives will be useful when we discuss the convergence properties of the resulting update rules. Since everything we say is with respect to each i, j, to simplify notation let us denote

1. $\tau_l \equiv \tau_{ijl}$,

2. $\theta_k \equiv \theta_{ijk}$ (recall that Θ is really a function of $\boldsymbol{\tau}$ if the canonical representation is used),

3. $\phi_k \equiv (1-\Delta)\theta_k + \Delta\frac{1}{M_l}$.

The partial derivatives of f^{lmix} are

$$\frac{\partial f^{\text{lmix}}}{\partial \theta_k} = f^{\text{lmix}} \times (1-\Delta) I[Z_i = k, \text{Pa}(Z_i) = j]/\phi_l. \tag{3.27}$$

and

$$\begin{aligned}
\frac{\partial f^{\text{lmix}}}{\partial \tau_l} &= \sum_{k=1}^{M_i} \frac{\partial f^{\text{lmix}}}{\partial \theta_k} \frac{\partial \theta_k}{\partial \tau_l} \\
&= \sum_{k=1,k\neq l}^{M_i} f^{\text{lmix}} \times (1-\Delta) I[Z_i = k, \text{Pa}(Z_i) = j] \frac{1}{\phi_k}(-\theta_k \theta_l) + \\
&\quad f^{\text{lmix}} \times (1-\Delta) I[Z_i = l, \text{Pa}(Z_i) = j] \frac{1}{\phi_l}(-\theta_l \theta_l + \theta_l) \\
&= f^{\text{lmix}} \times (1-\Delta) I[\text{Pa}(Z_i) = j] \theta_l \left(I[Z_i = l] \frac{1}{\phi_l} - \sum_{k=1}^{M_i} I[Z_i = k] \frac{\theta_k}{\phi_k} \right) \\
&= f^{\text{lmix}} \times (1-\Delta) I[\text{Pa}(Z_i) = j] \theta_l \left(I[Z_i = l] \frac{1}{\phi_l} - \prod_{k=1}^{M_i} \left(\frac{\theta_k}{\phi_k} \right)^{I[Z_i = k]} \right).
\end{aligned}$$

The expressions for this derivatives will be useful when we study the convergence of the AIS updates in Section 3.8.4. But before we we consider the convergence of the AIS estimates to the true value of the summation under consideration (G).

3.8.3 Convergence of AIS estimate

In this section, we study some theoretical properties of the AIS estimator, in particular, whether the estimates converges to their true value as the number of samples increases (i.e., the *consistency* of the AIS estimator).

If we use a sampling BN class that has a non-zero lower-bound on the smallest probability assigned to any outcome of $\boldsymbol{Z}$, then for any possible sampling BN in that class, the weight function is bounded (recall we have assumed that the target function g is bounded). Now, for simplicity, assume that we use an equal weighting of the sub-estimator generated for each stage (i.e., $W(t) = 1/T$), and that each sub-estimators is the result of a single sample of the IS BN (i.e., $N(t) = 1$, $N \equiv \sum_{t=1}^{T} N(t) = T$). This assumption can be relaxed. To simplify notation, let $\mu \equiv G$, and $\overline{X} \equiv \hat{G}^{(T)}$. Under all these conditions, we can apply Hoeffding's strengthened bounds to get

$$\Pr(\overline{X} - \mu \geq \epsilon) \leq \exp(-2N\epsilon^2/(b-a)^2) \tag{3.28}$$

and

$$\Pr(\overline{X} - \mu \leq -\epsilon) \leq \exp(-2N\epsilon^2/(b-a)^2) \tag{3.29}$$

where N is the total number of samples used for estimation, $0 < \epsilon < b - a$, and a and b are such that they satisfy $\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta}) \in [a, b]$ for all possible $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$ (or $\boldsymbol{\tau}$) in the sampling BN

class (recall that we are dealing in the "lower-bounding" sampling BN class; for instance, if we are using f^{lmix} then $\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta}) = g(Z)/f^{\text{lmix}}(\boldsymbol{Z} \mid \boldsymbol{\Theta})$). The reason we can use Hoeffding's strengthened bound is that the sum of the weights in this case form a martingale. The dependence between the weights of the samples at different time stages is "weak" in the sense that it does not affect their individual expectation.

Theorem 10 *For fixed N, and for $t = 1, \ldots, N$, let*

1. $X_t \equiv \omega(\boldsymbol{z}^{(t,l)} \mid \boldsymbol{\theta}^{(t)})$,
2. *for all $m \leq N$, $S_m = \sum_{t=1}^{m} X_t$, $S'_m = S_m - \mathrm{E}[S_m]$.*

S'_m forms a martingale.

Proof: For all $m \leq N$, $\mathrm{E}[S_m] = \mathrm{E}[\sum_{t=1}^{m} X_t = \sum_{t=1}^{m} \mathrm{E}[X_t] = \sum_{t=1}^{m} \mu = m\mu$. Hence, for all $m \leq N$, $S'_m = S_m - m\mu$. Now, for all $m \leq N$,

$$\begin{aligned}
\mathrm{E}[S'_m \mid S'_1, \ldots, S'_j] &= \mathrm{E}[S_m \mid S'_1, \ldots, S'_j] - m\mu \\
&= S'_j + \sum_{t=j+1}^{m} \mathrm{E}[X_t \mid S'_1, \ldots, S'_j] - (m-j)\mu \\
&= S'_j + \sum_{t=j+1}^{m} \mathrm{E}[X_t] - (m-j)\mu \\
&= S'_j + (m-j)\mu - (m-j)\mu \\
&= S'_j.
\end{aligned}$$

□

From this, we get that, the estimator will converge *in probability* to the true value (i.e., a *weak law of large numbers (WLLN)* for the AIS estimator). We can use the bound expressions above and apply a *Borel-Cantelli lemma* (see Lemma 6.1 of Durrett [1996]), to obtain convergence *with probability one* (i.e., a *strong law of large numbers* for the AIS estimator). Let $0 < \epsilon < b - a$, and for fixed n, $A_n = \{|\overline{X} - \mu| > \epsilon\}$. From the expressions above, $\Pr(A_n) < 2\exp(-2n\epsilon^2/(b-a)^2)$. Let $\gamma = 2\exp(-2\epsilon^2/(b-a)^2) < 1$. Then, for all $\epsilon > 0$, $\sum_{n=1}^{\infty} \Pr(A_n) < \sum_{n=1}^{\infty} \gamma^n = 1/(1-\gamma) < \infty$. By the Borel-Cantelli lemma, for all $\epsilon > 0$, $\Pr(A_n, \text{ infinitely often}) = 0$. Hence, $\overline{X} \to \mu$ *almost surely* (i.e., with probability one) as $N \to \infty$. In other words, the AIS estimate (denoted here by $\overline{X}$) is *consistent*.

Also, the probability that it deviates more than a fixed value ϵ decreases exponentially with the number of samples. The expression above gives us confidence bounds on the true

value from our estimates. In principle, we would use those confidence bounds in order to combine AIS with the methods described in the previous chapter.

Let us consider the case that g is defined for an estimation problem in an ID. (The expressions can be easily adapted to the case of a BN.) A simple way to compute a and b from quantities that can be computed almost immediately from the parameters of the model is as follows: let

1. $M_r \equiv \left|\Omega_{\mathrm{Pa}(Z_r)}\right|$ be the number of all possible assignments to the parents of Z_r in the original ID,
2. $p_i^{\max} = \max_{j,k} P(Z_i = k \mid \mathrm{Pa}(Z_i) = j)$, $p_i^{\min} = \min_{j,k} P(Z_i = k \mid \mathrm{Pa}(Z_i) = j)$ be the maximum and minumum probability values in the conditional probability distribution table for node Z_r in the original ID, respectively,
3. $u_i^{\max} = \max_j U_i(\mathrm{Pa}(U_i) = j)$, $u_i^{\min} = \min_j U_i(\mathrm{Pa}(U_i) = j)$ be the maximum and minumum utility values in the utility function for utility node U_i in the original ID, respectively, and
4. $\theta_i^{\max} = \max_{jk} \theta_{ijk}$, $\theta_i^{\min} = \min_{jk} \theta_{ijk}$ be the maximum and minumum probability values in the conditional probability distribution table for node Z_i in the IS BN (recall that depending on the representation θ_{ijk} can be actually a function of $\boldsymbol{\tau}_{ij}$), and
5. $0 < \Delta < 1/2$ define the local mixing parameter for the sampling BN class.

From this we can get

$$b = \frac{(\prod_{i=1}^{n} p_i^{\max})(\sum_{l=1}^{m} u_l^{\max})}{\prod_{r=1}^{s}((1-\Delta)\theta_r^{\min} + \Delta/M_r)} \tag{3.30}$$

$$a = \frac{(\prod_{i=1}^{n} p_i^{\min})(\sum_{l=1}^{m} u_l^{\min})}{\prod_{r=1}^{s}((1-\Delta)\theta_r^{\max} + \Delta/M_r)} \tag{3.31}$$

(The "$r = 1, \ldots, s$" are the indices to the sampling BN nodes, not – necessarily – the original nodes.) Needless to say, these bounds can be (and typically are) very loose.

It might be possible in special cases to obtain better (even strict) upper and lower bounds if we spend additional computation time. In general, however, we believe it is hard to obtain strict bounds (say by a form of dynamic programming). This is because we believe this problem is equivalent to that of computing maximum and/or minimum probability assignments in a BN, a problem known to be hard [Shimony, 1994].

In conclusion, we have just presented ways to apply adaptive importance sampling such that the resulting estimators converge to the true value with probability one. Also, confidence interval (bound) expressions results from bound expressions for the weight functions.

3.8.4 On the convergence of AIS updates

In the previous section we studied some properties of the AIS estimator, helping us characterize the behavior of the resulting estimates. In this section, we discuss some properties of the AIS update rules, in particular, whether the changes to the parameters of the IS BN converge, and if so, in what sense. This will help us characterize the behavior of the *learning process*.

Let us consider the variance error function (the analysis for the KL-based error functions is similar). Let the sampling BN class be $f \equiv f^{\text{lmix}}(\boldsymbol{Z} \mid \boldsymbol{\Theta}(\boldsymbol{\tau}))$. Recall that for this class $\omega \equiv \omega(\boldsymbol{Z} \mid \boldsymbol{\Theta}(\boldsymbol{\tau}))$ is bounded. First note all the following partial derivatives are bounded: $\partial f/\partial\tau_l$, $\partial\omega/\partial\tau_l = -\omega^2\partial f/\partial\tau_l$, $\partial\theta_k/\partial\tau_l$, $\partial\phi_k/\partial\tau_l = (1-\Delta)\partial\theta_k/\partial\tau_l$, $\partial\phi_k^{-1}/\partial\tau_l = -(1/\phi_k^2)\partial\phi_k/\partial\tau_l$. For the variance error function

$$e \equiv e(\boldsymbol{\Theta}(\boldsymbol{\tau})) = \sum_{\boldsymbol{Z}} f(\boldsymbol{\Theta}(\boldsymbol{\tau}))\omega(\boldsymbol{Z} \mid \boldsymbol{\Theta}(\boldsymbol{\tau}))^2 - G^2 \equiv \sum_{\boldsymbol{Z}} f\omega^2 - G^2,$$

the first derivatives

$$\frac{\partial e}{\partial\tau_l} = -\sum_{\boldsymbol{Z}} \omega^2 \frac{\partial f}{\partial\tau_l} \tag{3.32}$$

are bounded. Note that the second partial derivatives of the sampling BN are

$$\begin{aligned}\frac{\partial^2 f}{\partial\tau_r\partial\tau_l} \;=\; & (1-\Delta)I[\mathrm{Pa}(Z_i)=j]\left(\left(\frac{\partial f}{\partial\tau_r}\theta_l + f\frac{\partial\theta_l}{\partial\tau_r}\right)\left(\frac{I[Z_i=l]}{\phi_l} - \sum_{k=1}^{M_i}\frac{I[Z_i=k]\theta_k}{\phi_k} + \right.\right. \\ & \left.\left. f\times\theta_l\left(I[Z_i=l]\frac{\partial\phi_l^{-1}}{\partial\tau_r} - \sum_{k=1}^{M_i} I[Z_i=k]\left(\frac{1}{\phi_k}\frac{\partial\theta_k}{\partial\tau_r} + \theta_k\frac{\partial\phi_k^{-1}}{\partial\tau_r}\right)\right)\right)\right),\end{aligned}$$

which are also bounded. Hence, the second derivatives of the error function

$$\frac{\partial^2 e}{\partial\tau_r\partial\tau_l} = -\sum_{\boldsymbol{Z}} 2\omega\frac{\partial\omega}{\partial\tau_r}\frac{\partial f}{\partial\tau_l} + \omega^2\frac{\partial^2 f}{\partial\tau_r\partial\tau_l} \tag{3.33}$$

are also bounded.

According to Bertsekas and Tsitsiklis [1996], the boundedness of the first and second derivatives of e implies a *Lipschitz continuity of* ∇e. Since also e is non-negative, conditions (a) and (b) of Assumption 4.2 (Page 140) in their book hold. Also, condition (c) and (d) of the same assumption hold since the estimate of the gradients used in the update rules are unbiased and have bounded variance (since the gradients themselves are bounded). Hence, Assumption 4.2 holds. If in addition we let the step sizes $\alpha_t \equiv \alpha(t) = \beta/t, \beta > 0$, then they are nonnegative and satisfy the conditions $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$. Therefore, all the conditions for Proposition 4.1 (Page 141) in their book hold and therefore, the following hold with probability 1:

1. The sequence $e(\boldsymbol{\Theta}(\boldsymbol{\tau}^{(t)}))$ converges,

2. $\lim_{t\to\infty} \nabla e(\boldsymbol{\Theta}(\boldsymbol{\tau}^{(t)})) = 0$,

3. every limit point of $\boldsymbol{\tau}^{(t)}$ is a stationary point of e.

As Bertsekas and Tsitsiklis warns, this does not mean that the sequence of paratemers will converge. For that the error function has to satisfy additional conditions and I am not sure it does. What it does mean is that the process will converge in *error-function space*, and the gradients will also converge to zero *almost surely*. We can apply a similar analysis to get to the same results if we use the KL-based (global) error functions presented earlier.

The immediate problem with obtaining convergence results when the more traditional parameterization Θ is used, is the constraints in the parameters. The space defining the sampling BN class is a constrained space in this case. To do the updates, gradient projections or other methods to handle the constraints could be used [Bertsekas, 1995]. If gradient projections are used, there is still the need to correct the step size so that the constraints are still satisfied. We are aware of results in stochastic approximation for this case [Kushner and Clark, 1978] but do not know how to apply them.

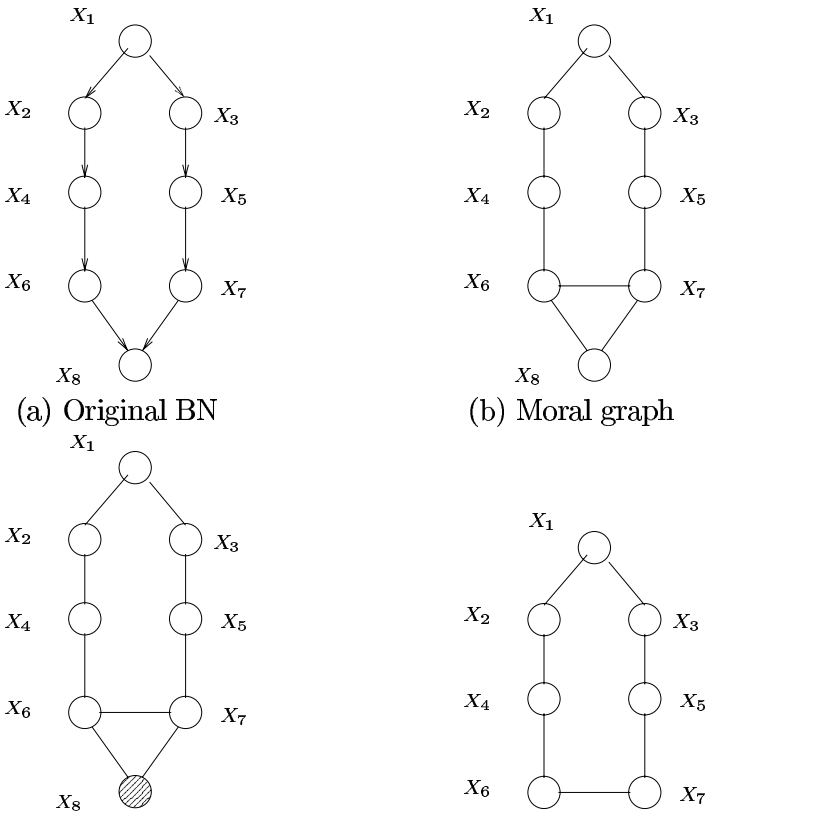
What we really want is to be able to guarantee convergence to a (at least local) minimum. In general, such guarantees can be hard to establish theoretically. However, it might be possible to obtain results for special classes of IS BN, such as mean-field approximations, considered in Section 3.7. This problem requires further analysis.

3.8.5 On the optimal IS BN structure

In Section 3.2, we made a somewhat arbitrary decision to use the do-operated BN structure for the IS BN, avoiding the discussion of how to select a BN structure that is optimal in some sense. In this section, we return to discuss this problem further.

In general, one objective we would like to have is to be able to represent the optimal sampling distribution using a BN. Recall that in the context of estimating marginal probabilities in a BN, the optimal sampling distribution is the posterior distribution. Even in the BN problem, however, we are not familiar with any result that tries to determine the "optimal" BN that can represent the posterior distribution (where presumably the notion of optimality is a BN that has no more conditional independencies than that of the posterior and has the minimum possible number of conditional independences in the posterior missing). We will not go much into the details in the discussion that follows, but refer the reader to Pearl [1988], Chapter 3, for a rigorous treatment of the concepts discussed below.

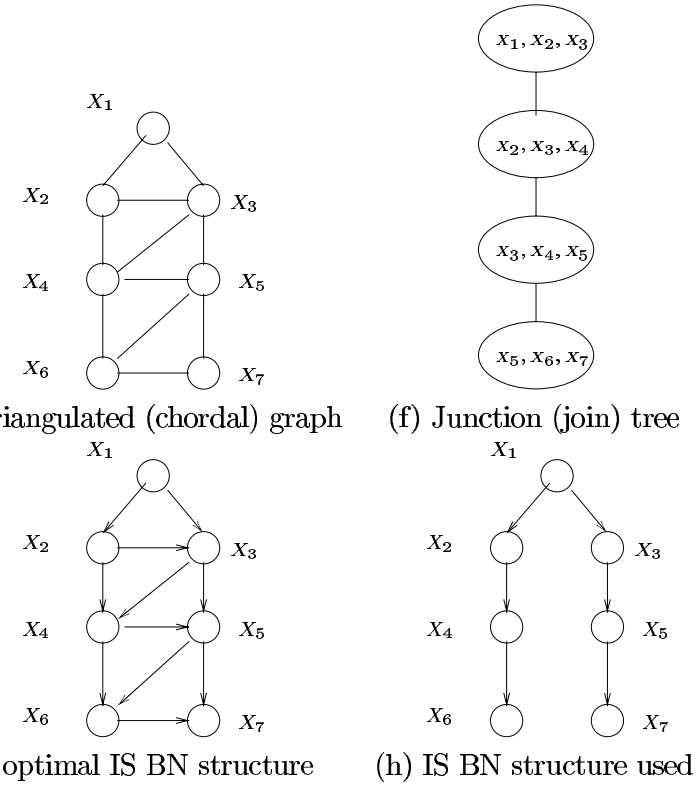
The moral graph of a BN graph is the undirected graph that results from joining all the parents of a node by adding undirected edges between them to form a clique and removing the direction of the arcs in the original graph of the BN (See Figure 3.7 (b) for an example). A BN is a Gibbs distribution with respect to the moral graph. If we remove the observed nodes from the moral graph, we obtain an undirected graph over the remaining (hidden) nodes (See Figure 3.7 (d) for an example). This graph provides an (undirected) graphical representation of a subset of conditional independencies that must be present in the posterior). Hence, it might have fewer conditional independencies than the posterior. If we would like to have a BN that can represent the posterior distribution, we might need to remove conditional independencies from the moral graph by adding edges to it. This is because undirected and directed-acyclic graphs in general represent different sets of conditional independences. However, the sets of conditional independencies are the same in the case of chordal graphs (decomposable models). One way (we are not sure it is the only way) to move from the undirected to the directed (BN) representation is as follows. As Pearl states in his book (Section 3.3.3, page 127), " every chordal graph can be oriented so that the tails of every pair of converging arrows are adjacent." This can be done, for instance, by going through a chordal graph, with the help of another graph called a *join* or *junction tree* (See Figure 3.8 (f) for an example). From a join tree we can build a Bayesian network that has at most the conditional independencies of the undirected graph. The process of getting a chordal graph that has at most the conditional independencies of the moral graph is typically achieved through a graph triangulation process. However, in general, bear in mind that performing optimal graph triangulation (obtaining a chordal graph that minimizes some objective function – minimum number of additional edges or minimum largest clique or minimum largest clique state-space – by possibly adding extra edges to the graph) is computationally intractable. The triangulated graph is chordal (by definition). Once we have the graph triangulated, we can get a join tree (there might be many that are equivalent) by a process that requires the computation of a maximum-weight spanning tree (which is computationally tractable). Once we have a join tree, there is a way to direct the edges so as to get a directed-acyclic graph (there might also be many that are equivalent) to use as our sampling BN structure and which represents the same dependencies as the chordal graph, and hence at most the conditional independencies of the the moral graph. In any case, the point is that there is a *sufficient* way to obtain a sampling BN that can represent the optimal sampling distribution. (See Figures 3.7 and 3.8for an example of the process.) But is it necessary? We do not know.



(a) Original BN (b) Moral graph

(c) Moral graph with evidence instantiated (d) Undirected graph (Markov network) representation of optimal IS (posterior)

Figure 3.7: How to obtain a BN representation of the optimal IS distribution: Example 1 (Continue in Figure 3.8)



(e) Triangulated (chordal) graph (f) Junction (join) tree

(g) optimal IS BN structure (h) IS BN structure used

Figure 3.8: How to obtain a BN representation of the optimal IS distribution: Example 1 (Continuation of Figure 3.7)

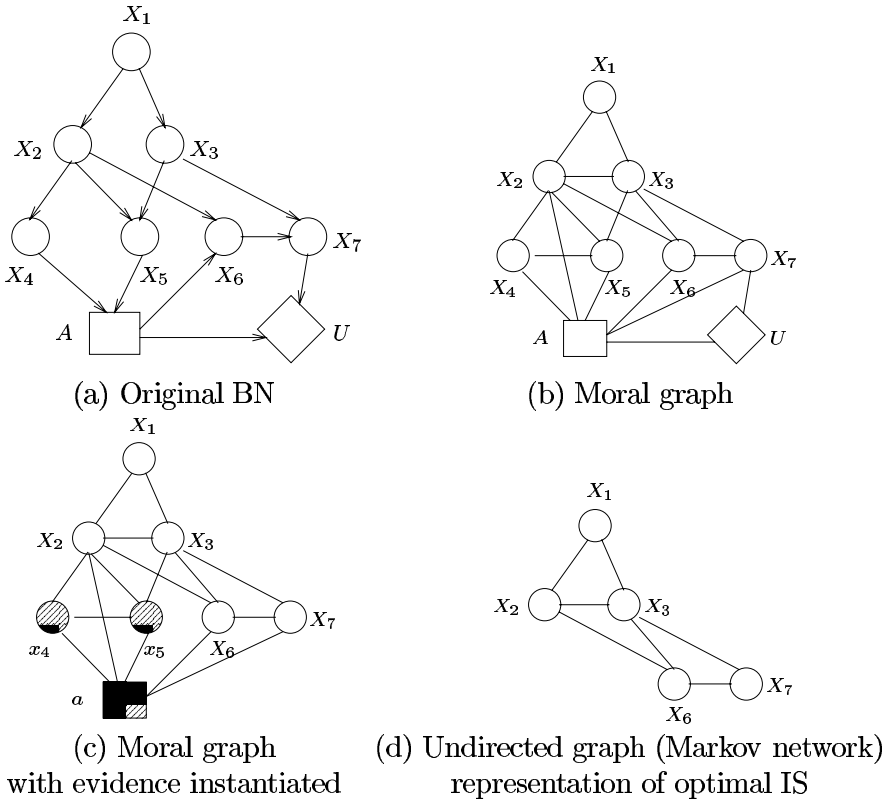
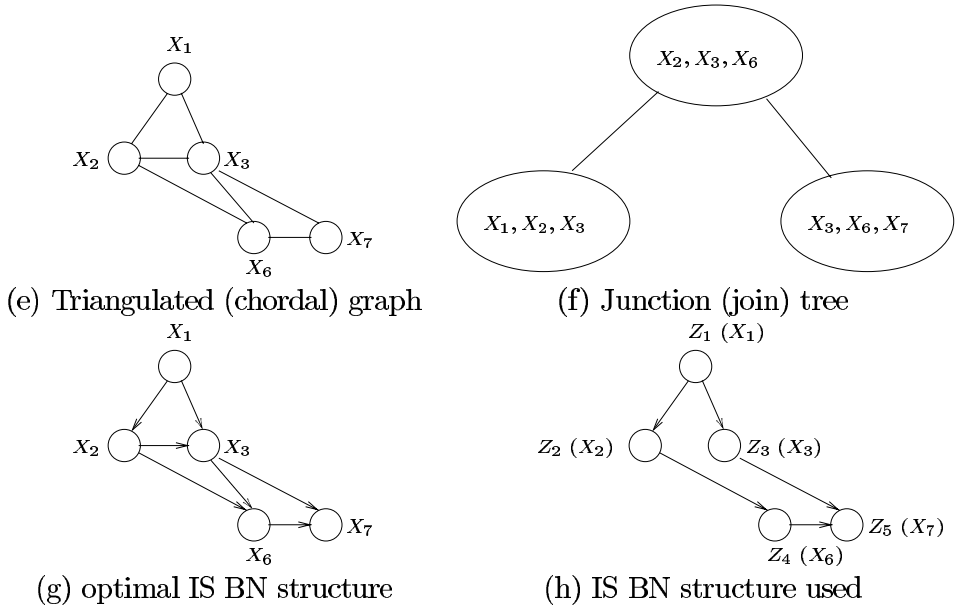


Figure 3.9: How to obtain a BN representation of the optimal IS distribution: Example 2 (Continue in Figure 3.10)

The discussion above also apply to ID problems, except that we not only remove observations from the moralized graph, but also actions and utility nodes. Note that the optimal sampling distribution in this case is not necessarily the posterior, as it depends on both the action and utility functions. (One could think of it as a posterior if we transform the ID into a BN as mentioned in the introduction (see also [Shachter and Peot, 1992, Zhang, 1998] and the references therein) or just a special "unnormalized" posterior.) See Figures 3.9 and 3.11 for examples. Note that when we constructed the moralized graph of the ID graph, we connected the parents of the utility nodes, not only those of the observations. This is because the value of the utility functions of which those nodes are parents will in general create dependecies between the parent nodes (i.e., knowing the value of the utility node gives information about the value of the parents, since the utility is a function of the parents). Hence the optimal sampling distribution is Gibbs with respect to the moral graph.

There are several problems with the approach above. One is that we have to do the triangulation which is in general intractable and typically only simple heuristics are used in



(e) Triangulated (chordal) graph (f) Junction (join) tree

(g) optimal IS BN structure (h) IS BN structure used

Figure 3.10: How to obtain a BN representation of the posterior: Example 2 (Continuation of Figure 3.9)

practice. Most importantly, the triangulation can lead to a sampling BN that has a node with many parents, in particular, many more parents that any node in the original BN. Hence, the sampling BN will require a very large number of parameters just to represent this node. As a matter of fact, if we can represent that BN, we could have done the exact computation of the marginal efficiently. This leads one to believe that the optimal structure for the sampling BN will be intractable in general. Right now, this is just a conjecture, since what we have described is a *sufficient* way to get a sampling BN able to represent the optimal BN, but it might not be necessary. In the case it were indeed necessary, we believe the intractability result will follow, connecting the complexity of the sampling BN with optimal structure (and the process of finding it) to that of performing exact computations.

For now, in our implementations we have just used a sampling BN structure that is the same as that of the original BN, maybe adding extra arcs between the nodes that are parents of observation and/or utility nodes, to model the strong dependency typically imposed among them by the observations and/or the utility values. The discussion on this section suggests interesting ways in which we can adapt *both the structure and the parameters* simultaneously, but we leave the details for future work.

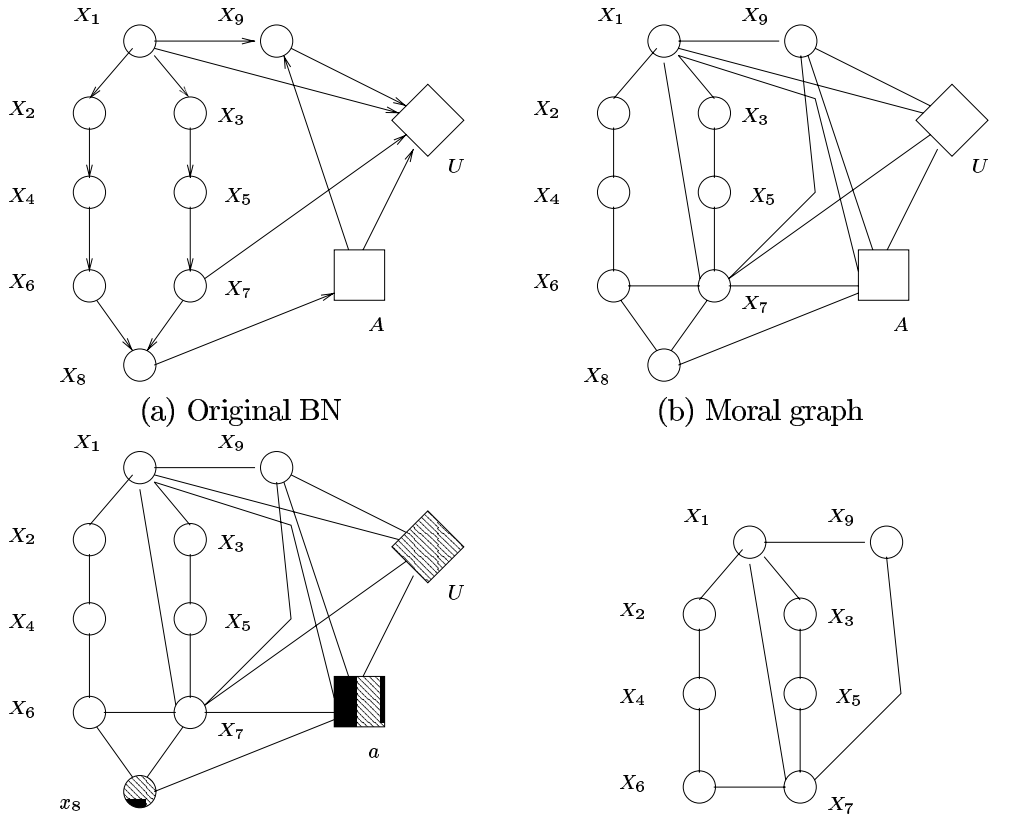
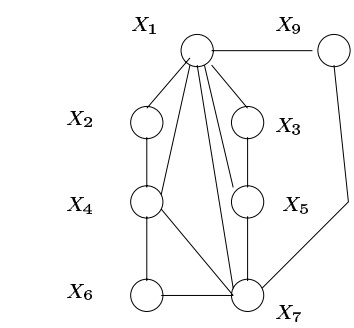
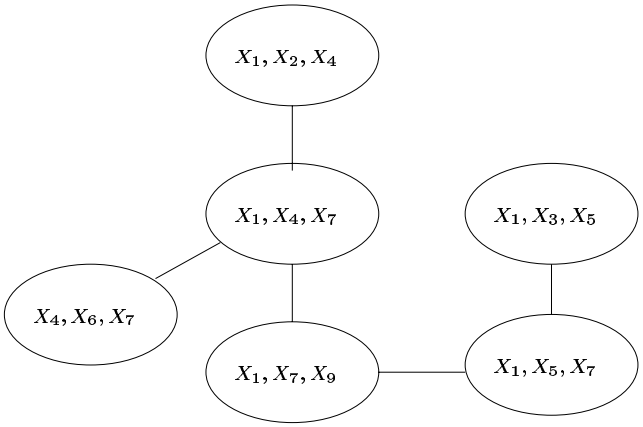


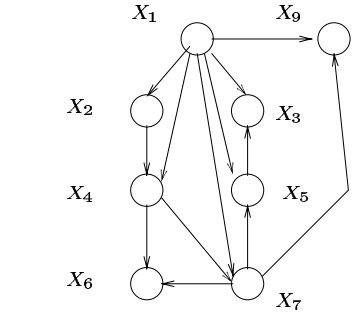
Figure 3.11: How to obtain a BN representation of the optimal IS distribution: Example 3 (Continue in Figure 3.12)



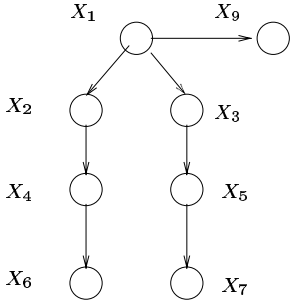
(e) Triangulated (chordal) graph



(f) Junction (join) tree



(g) optimal IS BN structure



(h) IS BN structure used

Figure 3.12: How to obtain a BN representation of the posterior: Example 3 (Continuation of Figure 3.11)

3.8.6 Theoretically optimal weighting

In this section, we return to the problem of how to set the weighting for the sub-estimators originally presented in equation (3.4) in Section 3.2.1. We establish what would be an *optimal* weighting scheme in this context (i.e., a weighting yielding estimates with smallest variance).

Recall that the problem is to select values for weighting the outcome of the estimates from T separate IS estimators that will be used to get the final AIS estimate. The objective is to produce a final AIS estimator with smallest error. Assume those weights are constant and independent of the estimators. The estimators themselves can be dependent. Recall that the weights need to sum to 1.

We can cast this problem as an optimization problem: find the weighting that produces a final estimator with the smallest variance. To simplify notation, let

1. $\overline{X} \equiv \hat{G}^{(T)}$ be the (global) final estimate at time T,

2. $\boldsymbol{w} \equiv (W(1), \ldots, W(T))$ be the vector formed from the weights $W(t)$ used for the partial estimate at t, for all $t = 1, \ldots, T$,

3. $\overline{\boldsymbol{y}} \equiv (\hat{G}(\boldsymbol{\theta^{(1)}}), \ldots, \hat{G}(\boldsymbol{\theta^{(T)}}))$ be the vector formed from the partial estimators $\hat{G}(\boldsymbol{\theta^{(t)}})$ at time t, and

4. $\mu \equiv G = E[\overline{X}]$ be the true value we are estimating. (Recall that the AIS estimator is unbiased in the case of constant nonnegative weights summing to 1.)

The variance of the AIS estimator in Equation (3.4) is

$$\begin{aligned} \operatorname{Var}\overline{X} &= \mathrm{E}\,\overline{X}^2 - (\mathrm{E}\,\overline{X})^2 \qquad (3.34)\\ &= \mathrm{E}\,\overline{X}^2 - \mu^2 \\ &= \mathrm{E}(\boldsymbol{w}^T\boldsymbol{y})^2 - \mu^2 \\ &= \mathrm{E}(\boldsymbol{w}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{w}) - \mu^2 \\ &= \boldsymbol{w}^T\left(\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)\right)\boldsymbol{w} - \mu^2 \\ &= \boldsymbol{w}^T(\boldsymbol{\Sigma} + \mu^2\overline{\overline{\mathbf{1}}})\boldsymbol{w} - \mu^2. \qquad (3.35) \end{aligned}$$

We have let $\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T) \equiv \boldsymbol{\Sigma} + \mu^2\overline{\overline{\mathbf{1}}}$, where $\overline{\overline{\mathbf{1}}}$ is the matrix with all its elements being 1; that

is, the symmetric random matrix, resulting from the expectation of $\boldsymbol{y}\boldsymbol{y}^T$,

$$\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T) = \begin{bmatrix} \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(1)}})^2) & \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(1)}})\hat{G}(\boldsymbol{\theta^{(2)}})) & \cdots & \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(1)}})\hat{G}(\boldsymbol{\theta}^{(T)})) \\ \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(1)}})\hat{G}(\boldsymbol{\theta^{(2)}})) & \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(2)}})^2) & \cdots & \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(2)}})\hat{G}(\boldsymbol{\theta}^{(T)})) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(1)}})\hat{G}(\boldsymbol{\theta}^{(T)})) & \mathrm{E}(\hat{G}(\boldsymbol{\theta^{(2)}})\hat{G}(\boldsymbol{\theta}^{(T)})) & \cdots & \mathrm{E}(\hat{G}(\boldsymbol{\theta}^{(T)})^2) \end{bmatrix}, \tag{3.36}$$

and therefore $\boldsymbol{\Sigma}$ is the covariance matrix of the estimators,

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Var}(\hat{G}(\boldsymbol{\theta^{(1)}})) & \mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(1)}}), \hat{G}(\boldsymbol{\theta^{(2)}})) & \cdots & \mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(1)}}), \hat{G}(\boldsymbol{\theta}^{(T)})) \\ \mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(1)}}), \hat{G}(\boldsymbol{\theta^{(2)}})) & \mathrm{Var}(\hat{G}(\boldsymbol{\theta^{(2)}})) & \cdots & \mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(2)}}), \hat{G}(\boldsymbol{\theta}^{(T)})) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(1)}}), \hat{G}(\boldsymbol{\theta}^{(T)})) & \mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(2)}}), \hat{G}(\boldsymbol{\theta}^{(T)})) & \cdots & \mathrm{Var}(\hat{G}(\boldsymbol{\theta}^{(T)})) \end{bmatrix}. \tag{3.37}$$

Also note that, $\mathrm{Var}(\hat{G}(\boldsymbol{\theta}^{(t)})) = \mathrm{Var}(\omega(\boldsymbol{Z}^{(t)} \mid \boldsymbol{\theta}^{(t)}))/N(t)$ and similarly,

$$\mathrm{Cov}(\hat{G}(\boldsymbol{\theta^{(i)}}), \hat{G}(\boldsymbol{\theta^{(j)}})) = \mathrm{Cov}(\omega(\boldsymbol{Z^{(i)}} \mid \boldsymbol{\theta}^{(i)}), \omega(\boldsymbol{Z^{(j)}} \mid \boldsymbol{\theta}^{(j)}))$$

for $i \neq j$. (Recall that $\mathrm{Var}(\omega(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(t)}))$ is the variance of the weight function for the IS distribution defined by $\boldsymbol{\theta}^{(t)}$, and $\mathrm{Cov}(\omega(\boldsymbol{Z^{(i}} \mid \boldsymbol{\theta}^{(i)}), \omega(\boldsymbol{Z}^{j} \mid \boldsymbol{\theta}^{(j)}))$ is the respective covariance between IS distributions defined by $\boldsymbol{\theta}^{(i)}$ and $\boldsymbol{\theta}^{(j)}$ for $i \neq j$. We are assuming that both are bounded for all $\boldsymbol{\theta}^{(t)}$.) For simplicity, denote $\boldsymbol{M} \equiv \boldsymbol{\Sigma} + \mu^2$, and $\bar{\mathbf{1}} \equiv (1, 1, \ldots, 1)^T$ (the vector of all ones).[5] We will soon argue that the optimal weights are given by

$$\boldsymbol{w}^* = \frac{\boldsymbol{M}^{-1}\bar{\mathbf{1}}}{\bar{\mathbf{1}}^T\boldsymbol{M}^{-1}\bar{\mathbf{1}}}. \tag{3.38}$$

Before we go into the argument, let us state some immediate properties. One immediate observation is that the sum of the components of $\boldsymbol{w}^*$ is 1 as it should. Let us consider the special case that the individual estimators are independent. In this case, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_T^2)$, where $\sigma_i^2 \equiv \mathrm{Var}(\hat{G}(\boldsymbol{\theta^{(i)}}))$. If $T = 2$, the optimal weighting is

$$w_i^* = \frac{\frac{1}{\sigma_i^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

[5] Actually, we are defining the matrix $\boldsymbol{M}$ this way to make its connection with the covariance matrix of the estimators, but we do not believe the dependency on μ is strong.

for $i = 1, 2$. More generally, for arbitrary T, and $i = 1, \ldots, T$,

$$\begin{aligned} w_i^* &= \frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{T} \frac{1}{\sigma_i^2}} \\ w_i^* &\propto \frac{1}{\sigma_i^2} \end{aligned} \tag{3.39}$$

This is as expected: the smaller the variance of an estimator, the larger its weight should be. Let us us now see why this is so in the case of independent estimators. A result on matrix inversions [Roweis, 1999] is, for matrices $\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{B}$,

$$(\boldsymbol{A} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}(\boldsymbol{B}^{-1} + \boldsymbol{X}^T\boldsymbol{A}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{A}^{-1}, \tag{3.40}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are square and invertible. Using this result with $\boldsymbol{A} = \boldsymbol{\Sigma}$, $\boldsymbol{X} = \bar{\mathbf{1}}$, and $\boldsymbol{B} = \mu^2$, we get,

$$\boldsymbol{M}^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\bar{\mathbf{1}}(\mu^{-2} + \bar{\mathbf{1}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{1}})^{-1}\bar{\mathbf{1}}^T\boldsymbol{\Sigma}^{-1}. \tag{3.41}$$

Now, note that

$$\boldsymbol{\Sigma}^{-1}\bar{\mathbf{1}} = \left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_T^2}\right),$$

$$\bar{\mathbf{1}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{1}} = \sum_{i=1}^{T} \frac{1}{\sigma_i^2},$$

and

$$\boldsymbol{\Sigma}^{-1}\bar{\mathbf{1}}\left(\mu^{-2} + \bar{\mathbf{1}}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{1}}\right)^{-1}\bar{\mathbf{1}}^T\boldsymbol{\Sigma}^{-1} = (\frac{1}{\mu^2} + \sum_{i=1}^{T} \frac{1}{\sigma_i^2})^{-1} \begin{bmatrix} \frac{1}{\sigma_1^4} & \frac{1}{\sigma_1^2\sigma_2^2} & \cdots & \frac{1}{\sigma_1^2\sigma_T^2} \\ \frac{1}{\sigma_1^2\sigma_2^2} & \frac{1}{\sigma_2^4} & \cdots & \frac{1}{\sigma_2^2\sigma_T^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sigma_1^2\sigma_T^2} & \frac{1}{\sigma_2^2\sigma_T^2} & \cdots & \frac{1}{\sigma_T^4} \end{bmatrix}.$$

This yields

$$\boldsymbol{M}^{-1} = \left(\frac{1}{\mu^2} + \sum_{i=1}^{T} \frac{1}{\sigma_i^2}\right)^{-1} \begin{bmatrix} \frac{1}{\sigma_1^2}\left(1 - \frac{1}{\sigma_1^2}\right) & -\frac{1}{\sigma_1^2\sigma_2^2} & \cdots & -\frac{1}{\sigma_1^2\sigma_T^2} \\ -\frac{1}{\sigma_1^2\sigma_2^2} & \frac{1}{\sigma_2^2}\left(1 - \frac{1}{\sigma_2^2}\right) & \cdots & -\frac{1}{\sigma_2^2\sigma_T^2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{\sigma_1^2\sigma_T^2} & -\frac{1}{\sigma_2^2\sigma_T^2} & \cdots & \frac{1}{\sigma_T^2}\left(1 - \frac{1}{\sigma_T^2}\right) \end{bmatrix} \tag{3.42}$$

and

$$\begin{aligned} \boldsymbol{M}^{-1}\bar{\mathbf{1}} &= \left(\frac{1}{\mu^2}+\sum_{i=1}^{T}\frac{1}{\sigma_i^2}\right)^{-1}\begin{bmatrix} \frac{1}{\sigma_1^2}\left(1-\sum_{i=1}^{T}\frac{1}{\sigma_i^2}\right) \\ \frac{1}{\sigma_2^2}\left(1-\sum_{i=1}^{T}\frac{1}{\sigma_i^2}\right) \\ \vdots \\ \frac{1}{\sigma_T^2}\left(1-\sum_{i=1}^{T}\frac{1}{\sigma_i^2}\right) \end{bmatrix} \\ &= \left(\frac{1}{\mu^2}+\sum_{i=1}^{T}\frac{1}{\sigma_i^2}\right)^{-1}\left(1-\sum_{i=1}^{T}\frac{1}{\sigma_i^2}\right)\begin{bmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_T^2} \end{bmatrix}. \end{aligned} \tag{3.43}$$

Substituting the last expression in the expression for the optimal weights $\boldsymbol{w}^*$ (Equation (3.38)), we get

$$\boldsymbol{w}^* = \frac{1}{\sum_{i=1}^{T}\frac{1}{\sigma_i^2}}\begin{bmatrix} \frac{1}{\sigma_1^2} \\ \frac{1}{\sigma_2^2} \\ \vdots \\ \frac{1}{\sigma_T^2} \end{bmatrix}. \tag{3.44}$$

Now we present the argument leading to the optimal weighting $\boldsymbol{w}^*$. We introduce a Lagrange multiplier λ to deal with the constraint $\sum_{i=1}^{T} w_i = 1$. The derivative with respect to the weights of the Lagrangian associated with the variance of the estimator is

$$\frac{\partial(\mathrm{E}(\overline{X}^2) - \lambda(\bar{\mathbf{1}}^T\boldsymbol{w} - 1))}{\partial \boldsymbol{w}} = 2\boldsymbol{M}\boldsymbol{w} - \lambda. \tag{3.45}$$

Setting the gradient to zero, we get

$$\boldsymbol{w} = \frac{1}{2}\lambda\boldsymbol{M}^{-1}\bar{\mathbf{1}}. \tag{3.46}$$

We have assumed for now that the matrix $\boldsymbol{M}$ is invertible. We will get back to this assumption soon. Now, since the components of $\boldsymbol{w}$ need to sum to 1 (i.e., $\bar{\mathbf{1}}^T\boldsymbol{w} = 1$),

$$\lambda^* = \left(\frac{1}{2}\bar{\mathbf{1}}^T\boldsymbol{M}^{-1}\bar{\mathbf{1}}\right)^{-1}. \tag{3.47}$$

Substituting λ^* into the expression for $\boldsymbol{w}$ above we get the result

$$\boldsymbol{w}^* = \frac{\boldsymbol{M}^{-1}\bar{\mathbf{1}}}{\bar{\mathbf{1}}^T\boldsymbol{M}^{-1}\bar{\mathbf{1}}}. \tag{3.48}$$

Now, we argue that the matrix $\boldsymbol{M}$ is invertible. We know that the matrix $\boldsymbol{M}$ is positive semi-definite: by definition, $\boldsymbol{\Sigma}$ is positive semi-definite and u^2 is non-negative. Thus, $\boldsymbol{M}$ is strictly positive-definite if no two estimators are deterministically related and no estimator has zero variance (i.e., it is *perfect* in that it gives the answer μ all the time). It seems that we also need for $\mu > 0$, but we believe this is just an artifact of the way we have expressed $\boldsymbol{M}$. Again, the analysis assumes that the covariance matrix for pairs of individual estimators is positive-definite. We argue this is true if the sampling distribution for any pair of estimators is not exactly the same.

Other researchers have previously suggested that we can obtain good results in practice by using the same samples used for the individual estimators to estimate the variance of the estimators (See for example Marshall [1956] and Cheng and Druzdzel [2000]). However, since there is no theoretical guarantee that the estimates of the variances are independent of those for the estimators themselves, the resulting AIS estimator can be biased in general. Also, by using only individual variance terms (no covariance terms between estimators), they are assuming that the estimators themselves are independent. In general, this will only be true if we estimate the gradients using an independent set of samples from those used for the actual individual estimators. The expression given above is more general, as it takes into account potential dependencies between the individual estimators.

Although the expression above is theoretically attractive, in order to use it in practice it seems that one has to ignore some of the assumptions behind it. In particular, we are assuming that the weights are constants, not random variables as would be the case if we use samples to estimate the optimal weighting scheme. I believe one can partially remove this assumption, by introducing another, looser set of assumptions, (i.e., interchangeability of differentiation and expectation for functions of the weights – linear and quadratic), and the result is an optimal weighting for the expectation of the weights, not the weights themselves. In practice, it would be a waste not to reuse as many of the samples as we can, even if no theoretical guarantees hold anymore. The risk of things going terribly wrong might be smaller than the potential gain in computational speed provided by the heuristic implementation of the estimator. Also, the bias of the estimator is not bad as long as it is somewhat controlled, and decreases asymptotically. Hence, in practice, one can estimate the matrix $\boldsymbol{M}$ from the same samples used to estimate the actual final AIS estimate.

3.9 Summary and conclusions

In this chapter, we presented adaptive importance sampling methods for problems in Bayesian networks and influence diagrams. We presented preliminary empirical results on synthetic Bayesian network and influence diagram problems that suggests that the method can be a very significantly effective alternative to the traditional importance-sampling method used for problems in these models. Motivated by the empirical results, we studied some characteristics of the methods and established some theoretical results. Although further work remains, I believe that using adaptive importance-sampling methods in the context of action selection will prove useful and provide an improvement over the sampling methods for action selection in influence diagrams presented in the previous chapter. In conclusion, AIS can be a very effective, yet still relatively simple sampling method for handling hard problems in graphical models for complex structured domains.

Chapter 4

Conclusions

Our objective in this work was to develop methods that attempt to exploit, in an online fashion, particular characteristics of the models used and the particular problem under consideration. In particular, we exploited the inherent comparative nature of the problem of action selection, and used results and ideas already developed for similar problems presented in the statistical literature and applied them to the problem considered in this thesis. In proposing the adaptive importance sampling methods, we also exploited the structure already available in the models for problems of estimating large-dimensional summations typically required for problems in graphical models. The adaptive nature of the methods is to gather information about the particular problem being considered, online, as we perform the estimation, and use that information to improve the solutions we provide (i.e. our estimates).

We studied methods for action selection in single-decision influence diagrams based on simple (forward) importance sampling. We provided bounds on the number of samples needed by the simple estimation-based traditional method to guarantee that the action selected will be approximately optimal with high probability. In trying to reduce the number of samples required to make near-optimal action selection with high probability, we proposed an estimation-based two-stage sequential method and a general framework for comparison-based methods.

For the estimation-based two-stage sequential method, we gave a result stating that by allocating a pre-determined initial number of samples to estimate the variance of the estimators and using the resulting variance estimates to determine the number of samples to estimate the (unnormalized) value of each action, we can reduce, both in expectation and with high probability, the total number of samples required to make good action selections.

On the other hand, the comparison-based method are multi-stage or group sequential methods that try to directly exploit the fact that the problem of action selection is primarily a comparison problem. We suggested several instantiations of the general framework and stopping rules, and argued that the resulting methods guarantee the correct approximation requirement for the action selected. We also considered a heuristic-based framework of the comparison-based method based on adaptive allocation of samples. We presented preliminary empirical evidence that the comparison-based method can be effective in reducing the total number of samples needed to make good action selections, as compared to those needed by the estimation-based methods.

We also studied the problem of reducing the variance of estimators for quantities involving summations in graphical models such as Bayesian networks and influence diagrams. We approached this problem as a learning problem; that is, we suggested representing the importance sampling distribution as a BN and learning the BN from information obtained from the samples generated during the sampling process. We proposed new adaptive-importance-sampling update rules based on directly relevant error measures. Also, we suggested corresponding estimators that take advantage of the samples resulting from different importance sampling distribution used during the adaptation process. We studied some of the theoretical issues of this class of estimators and showed that, under a restricted but reasonable class of importance-sampling Bayesian networks and sub-estimators weighting schemes, we can guarantee convergence of the estimators to their true value (with probability one as the number of samples goes to infinity), and provide theoretical confidence intervals for the estimators as a function of the total number of samples used for the estimate.

We theoretically analyzed the behavior of the adaptive process and showed that for an interesting set of error measures and under a particular, but significantly general and interesting class of sampling BN distributions, the adaptive process converges to a stationary point of the error function (i.e., the sequence of parameters for the sampling BN found during the update process are such that, with probability one, their error converges, the gradients converge to zero, and if the parameters themselves converge, they will converge to a point in the error measure such that the gradient is zero (the error measure characterizes the notion of difference of any sampling BN in the class to the optimal distribution).

Many theoretical problems remain open. Bounds on the number of samples for the comparison-based method could not be provided, particularly when we use a sophisticated adaptive sample reallocation schedule. Also, how to obtain *optimal* allocation schedules is not yet clear (although it is clear that we can benefit from approaching this problem as a special class of a multi-armed bandit problem). As for the adaptive importance sampler,

there are obvious theoretical problems yet to be resolved. For instance, how do (or can) we assess the quality of the converging sampling BN distributions and the rate of convergence? In what cases can we represent the optimal sampling BN efficiently? Is it possible to adapt the structure of the sampling BN along with the parameters in an interesting, effective, and efficient way?

From a practical perspective, we believe most of the theoretical results in this thesis are too loose to be of direct practical use for the average model we are likely to find or develop in practice. Hence, apart from developing methods with provably better bounds on *sample complexity*, we should also consider adapting methods that use heuristic approximations based on asymptotic normality which have been effectively used in practice in the statistical literature. The work of Charnes and Shenoy [1999], for instance, follows this direction, but it is mainly an estimation-based approach. We believe we can extend their method easily to use a comparison-based approach instead. Heuristic approximations methods based on normality assumptions could be very effective in practice even though they provide no theoretical guarantees, since the assumptions they are typically based on do not hold for the general ID model and are hard to corroborate or establish empirically.

4.1 Contributions

We now summarize the main contributions made in the thesis.

- Provided sample-complexity bounds for traditional method.
- Proposed method for action selection based on sequential estimation. Provided sample-complexity bounds for that method and showed that it reduces the number of samples with respect to the traditional method both on expectation and with high probability.
- Proposed a general framework for comparison-based methods. Theoretically analyzed some instances. Provided empirical evidence for its potential in improving the effectiveness of sampling methods for action selection.
- Proposed stochastic-gradient methods (update rules) for adaptive importance sampling in problem in BNs and IDs, in which we systematically update the importance-sampling distribution.
- Empirically showed the potential of the proposed adaptive importance-sampling methods for estimation problems.

- Presented preliminary theoretical results on the quality of the estimators and the behavior of the update rules of the adaptive importance-sampling methods.

4.2 Future work

Despite the progress already made in analyzing the sampling methods presented in this thesis and demonstrating their practicality, there are still many practical and theoretical questions left unanswered. Most of these questions were already posed throughout the document. Here, we list the most pressing problems for future work:

- Theoretical analysis of the comparison-based method with regard to bounds on the number of samples using both simple allocation rules and adaptive reallocation. Also, study the problem of how to optimally allocate samples in a sequential (i.e., adaptive) way. We believe we can borrow significantly from the work on multi-armed bandit problems.

- Theoretically study additional properties of the adaptive importance-sampling methods proposed, such as convergence quality and rate of convergence.

- Connect AIS and variational approximation methods.

- Study how to integrate the methods for adaptive importance sampling and action selection in an effective way. Theoretically study the properties of the resulting combined method.

- Study ways to reduce the model complexity by using model-approximation techniques based on the relevance of certain domain variables when making particular action decisions. This is important since reducing the dimensionality of the summations apart from simplifying the problem leads, in general, to estimators with smaller variance.

- Empirical evaluation of all the methods in a large problem for which exact methods are intractable and sampling methods can provide a reasonable alternative. Typically, this can be assessed *a priori* by looking at some general characteristics of the problem. For instance, it is generally understood that the characteristics of an ID that make it hard to solve have to do with the numerical parameters defining the local conditional probability distributions and the utilities. On the other hand, the effectiveness of exact methods depend more on the characteristics of the graph which represents the global structure of the problem.

- Extensions to IDs with multiple decisions, and MDP and POMDPs, in the spirit of Charnes and Shenoy [1999], Kearns et al. [1999b] and Kearns et al. [1999a].
- Better ways of dealing with problems having large numbers of observations leading to large optimal policy descriptions. Maybe combine sampling methods with other approximations techniques that attempt to represent the optimal policy more compactly or approximate the value functions computed at each stage of dynamic programming [Dearden and Boutilier, 1997, Boutilier et al., 2000, Koller and Parr, 1999, 2000, Kim and Dean, 2001]. Maybe develop methods that try to compute near-optimal randomized policies in a forward manner (as opposed to the backward induction typically used in dynamic programming, as applied to these problems) in the spirit of Kearns et al. [1999b] and Kearns et al. [2000] (see also Kearns et al. [1999a]).

4.3 Final remarks

Sampling methods can be a very effective alternative for action selection in influence diagrams. I believe this is specially true for models large enough that exact methods are simply not an option. Adaptive sampling improves over naive applications of sampling methods while keeping their simplicity. Further extensions are necessary to deal with more complex models, but the hope is that the approach taken in this thesis provides a sound foundation.

Bibliography

Srinivas M. Aji and Robert J. McEliece. The generalized distributed laws. *IEEE Transactions on Information Theory*, 46(2):325–343, March 2000. Can also be downloaded from `http://www.systems.caltech.edu/EE/Faculty/rjm/papers/GDL.ps`.

Wael A. Al-Qaq, Michael Devetsikiotis, and J. Keith Townsend. Stochastic gradient optimization of importance sampling for the efficient simulation of digital communications systems. *IEEE Transactions on Communications*, 43(12):2975–2985, December 1995.

Z. Alexandrowicz. Stochastic models for the statistical description of lattice systems. *Journal of Chemical Physics*, 55(6):2765–2779, September 1971.

F. J. Anscombe. Fixed-sample-size analysis of sequential observartions. *Biometrics*, 10(1): 89–100, March 1954.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969.

Keith Baggerly, Dennis Cox, and Rick Picard. Exponential convergence of adaptive importance sampling for Markov chains. *Journal of Applied Probability*, 37(2):342–358, June 2000.

Robert E. Bechhofer, Thomas J. Santner, and David M. Goldsman. *Design and analysis of experiments for statistical selection, screening and multiple comparisons.* Wiley, 1995.

Richard Ernest Bellman. *Dynamic Programming.* Princeton University Press, Princeton, 1957.

S. Bernstein. *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow, 1946.

Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Massachusetts, 1995.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

C. Bielza, M. Gómez, S. Ríos-Insua, and J.A. Fernández del Pozo. Structural, elicitation and computational issues faced when solving complex decision making problems with influence diagrams. *Computers & Operations Research*, 27(7-8):725–740, June 2000.

Concha Bielza, Peter Müller, and David Ríos Insua. Monte Carlo methods for decision analysis with applications to influence diagrams. *Management Science*, 1999. Forthcoming.

John Binder, Daphne Koller, Stuart Russell, and Keiji Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 1997.

Thomas E. Booth. A Monte Carlo learning/biasing experiment with intelligent random numbers. *Nuclear Science and Engineering*, 92(3):465–481, March 1986.

Thomas E. Booth. Zero-variance solutions for linear Monte Carlo. *Nuclear Science and Engineering*, 102(4):332–340, August 1989.

Craig Boutilier, Tom Dean, and Steve Hanks. Decision theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11: 1–94, 1999.

Craig Boutilier, Richard Dearden, and Moisés Goldszmidt. Stochastic dynamic programming with factored representations. *Artificial Intelligence*, 121:49–107, 2000.

Jose E. Cano, Luis D. Hernández, and Seraín Moral. Importance sampling algorithms for the propagation of probabilities in belief networks. *International Journal of Approximate Reasoning*, 15(1):77–92, July 1996.

J. Y. Chang and J. C. Hsu. Optimal designs for multiple comparisons with the best. *Journal of Statistical Planning and Inference*, 30:45–62, 1992.

John M. Charnes and Prakash P. Shenoy. A forward Monte Carlo method for solving influence diagrams using local computation. School of Business, University of Kansas, Working Paper No. 273, August 1999.

Jian Cheng and Marek J. Druzdzel. AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research*, 13:155–188, 2000.

Jian Cheng and Marek J. Druzdzel. Confidence inference in Bayesian networks. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-2001), August 2nd-5th, Seattle, Washington.*, 2001. To appear.

Paul R. Cohen. *Empirical Methods for Artificial Intelligence.* MIT Press, 1995.

Gregory F. Cooper. A method for using belief networks as influence diagrams. In *Proceedings of the Workshop on Uncertainty in Artificial Intelligence*, pages 55–63, Minneapolis, Minnesota, 1988.

Gregory F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

Paul Dagum, Richard Karp, Michael Luby, and Sheldon Ross. An optimal algorithm for Monte Carlo estimation. *SIAM Journal of Computing*, 29(5):1484–1496, 2000.

Paul Dagum and Michael Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.

Richard Dearden and Craig Boutilier. Abstraction adn approximate decision-theoretic planning. *Artificial Intelligence*, 89(1):219–283, 1997.

Luc Devroye, Lásló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

F. J. Díez. Parameter adjustment in Bayes networks. the generalized noisy OR-gate. In David Heckerman and Abe Mamdani, editors, *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence, July 9-11, 1993, The Catholic University of America, Providence, Washington D. C. , USA*, pages 99–105, San Francisco, CA, 1993. Morgan Kaufmann Publishers, Inc.

Richard Durrett. *Probability: Theory and Examples.* Duxbury Press, second edition, 1996.

Michael Evans. Adaptive importance sampling and chaining. In Nancy Flournoy and Robert K. Tsutakawa, editors, *Statistical Multiple Integration: Proceedings of a Joint Summer Research Conference held at Humboldt University, June 17-23, 1989*, volume 115 of *Computational Mathematics*, pages 137–143, Providence, Rhode Island, 1991. American Mathematical Society.

M. Fitzgerald, R. R. Picard, and R. N. Silver. Canonical transition probabilities for adaptive Metropolis simulation. *Europhysics Letters*, 46(3):282–287, May 1999.

M. Fitzgerald, R. R. Picard, and R. N. Silver. Monte Carlo transition dynamics and variance reduction. *Journal of Statistical Physics*, 98(1/2):321–345, January 2000.

Nancy Flournoy and Robert K. Tsutakawa, editors. *Statistical Multiple Integration : Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference held at Humboldt University, Arcata, CA, on June 17-23, 1989*, volume 113 of *Contemporary Mathematics*. American Mathematical Society, Providence, RI, 1991.

Robert Fung and Kuo-Chu Chang. Weighting and integrating evidence for stochastic simulation in Bayesian networks. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pages 112–117, 1989.

Robert Fung and Brendan Del Favero. Backward simulation in Bayesian networks. In Ramon López de Mantaras and David Poole, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 227–234, San Francisco, CA, 1994. Morgan Kaufmann Publishers.

Dan Geiger, David Heckerman, and Christopher Meek. Asymptotic model selection for directed networks with hidden variables. Technical Report MSR-TR-96-07, Microsoft Research, Advanced Technology Division, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, May 1996.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, November 1989.

John Geweke. Generic, algorithmic approaches to Monte Carlo integration in Bayesian inference. In Nancy Flournoy and Robert K. Tsutakawa, editors, *Statistical Multiple Integration: Proceedings of a Joint Summer Research Conference held at Humboldt University, June 17-23, 1989*, volume 115 of *Computational Mathematics*, pages 117–135, Providence, Rhode Island, 1991. American Mathematical Society.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Chapman & Hall, 1996.

M. Gómez, S. Ríos-Insua, C. Bielza, and J. A. Frnández del Pozo. Multiattribute utility analysis in the IctNeo system. In Yacov Y. Haimes and Ralph E. Steuer, editors,

Research and Practice in Multiple Criteria Decision Making: Proceedings of the XIVth International Conference on Multiple Criteria Decision Making (MCDM), Charlottesville, Virginia, USA, June 8-12, 1998, Berlin; New York, 2000. Springer.

J. H. Halton. Sequential Monte Carlo. *Proceeding of the Cambridge Philosophical Society (Mathematical and Physical Sciences)*, 58(1):57–78, January 1962.

J. P. Hardwick. Computational problems associated with minimizing the risk in a simple clinical trial. In Nancy Flournoy and Robert K. Tsutakawa, editors, *Statistical Multiple Integration: Proceedings of a Joint Summer Research Conference held at Humboldt University, June 17-23, 1989*, volume 115 of *Computational Mathematics*, pages 239–256, Providence, Rhode Island, 1991. American Mathematical Society.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

David Heckerman. A tractable inference algorithm for diagnosing multiple diseases. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence, August 18-20, Windsor, Ontario*, pages 174–181, August 1989.

David Heckerman. A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, March 1995.

Luis D. Hernández, Serafín Moral, and Antonio Salmerón. A Monte Carlo algorithm for probabilistic propagation in belief networks based on importance sampling and stratified simulation techniques. *International Journal of Approximate Reasoning*, 18(1-2):53–91, January 1998.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

Ronald A. Howard and James E. Matheson. Influence diagrams. In Ronald A. Howard and James E. Matheson, editors, *Readings on the Principles and Applications of Decision Analysis II*, pages 721–762. Strategic Decision Group, 1981.

Jason C. Hsu. Simultaneous confidence intervals for all distances from the "best". *Annals of Statistics*, 9(5):1026–1034, September 1981.

Jason C. Hsu. *Multiple Comparisons: Theory and Methods*. Chapman and Hall, 1996.

Tommi S. Jaakkola and Michael I. Jordan. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, 10:291–322, January-June 1999.

Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, 2000.

Finn V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, London, England, 1996.

Finn V. Jensen, Steffen L. Lauritzen, and Kristian G. Olesen. Bayesian updating in recursive graphical models by local computations. Technical Report R 89-15, Institute for Electronic Systems, Department of Mathematics and Computer Science, University of Aalborg, June 1989.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, Adaptive Computation and Machine Learning. MIT Press, first edition, 1997.

Leslie Pack Kaelbling. *Learning in Embedded Systems*. A Bradford Book. The MIT Press, Cambridge, Massachusetts; London, England, 1993.

H. Kahn and A. W. Marshall. Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, November 1953.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. Approximate planning in large POMDPs via reusable trajectories. Can be downloaded from `http://www.cs.berkeley.edu/~ang/papers/pomdp-long.ps`, May 1999a.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1324–1331, Menlo Park, Calif., 1999b. International Joint Conference on Artificial Intelligence, Inc., Morgan Kaufmann.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. Approximate planning in large POMDPs via reusable trajectories. In *Advances in Neural Information Processing Systems 12*. MIT Press, 2000.

Ralph L. Keeney and Howard Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Probability and Mathematical Statistics. John wiley & Sons, Inc., 1976.

Ryoichi Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81(6):988–1003, March 1951.

Kee-Eung Kim and Thomas Dean. Solving factored MDPs via non-homogeneous partitioning. In *International Joint Conference on Artificial Intelligence (IJCAI-01)*, Seattle, Washington, August 2001.

T. Kloek and H. K. van Dijk. Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica*, 46(1):1–19, January 1978.

Daphne Koller and Ronald Parr. Computing factored value functions for policies in structured MDPs. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*, volume 2, pages 1332–1339, Stockholm, Sweden, August 1999. Morgan Kaufmann Publishers.

Daphne Koller and Ronald Parr. Policy iteration for factored MDPs. In Craig Boutilier and Moisés Goldszmidt, editors, *Proceedings of the Sixteenth Annual Conference on Artificial Intelligence (UAI-00)*, pages 326–334, Stanford, California, June 2000. Morgan Kaufmann Publishers.

Craig Kollman, Keith Baggerly, Dennis Cox, and Rick Picard. Adaptive importance sampling on discrete Markov chains. *The Annals of Applied Probability*, 9(2):391–412, May 1999.

Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, volume 26 of *Applied Mathematical Sciences*. Springer-Verlag New York Inc., New York, 1978.

Jaimyoung Kwon and Kevin Murphy. Modeling freeway traffic using coupled HMMs. Technical report, University of California - Berkeley, May 2000. Can be downloaded from `http://www.cs.berkeley.edu/~murphyk/Papers/traffic.ps.gz`.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50(2):157–224, 1988.

Yan Lin and Marek Druzdzel. Stochastic sampling and search in belief updating algorithms for very large Bayesian networks. In *Working Notes of the AAAI Spring Symposium on Search Techniques for Problem Solving Under Uncertainty and Incomplete Information*, pages 77–82, Stanford, California, March 1999. Stanford University. Available from `http://www.pitt.edu/~druzdzel/publ.html`.

Oded Maron and Andrew W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6, pages 59–66. Morgan Kaufmann Publishers, Inc., 1994.

Andrew W. Marshall. The use of multi-stage sampling schemes in Monte Carlo computations. In Herbert A. Meyer, editor, *Symposium on Monte Carlo Methods: Held at the University of Florida*, A Wiley Publication in Applied Statistics, pages 123–140. Statistical Laboratory, John Wiley & Sons, Inc., 1956.

Frank J. Matejcik and Barry L. Nelson. Two-stage multiple comparisons with the best for computer simulation. *Operations Research*, 43(4):633–640, July-August 1995.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

Herbert A. Meyer, editor. *Symposium on Monte Carlo Methods: Held at the University of Florida, March 16 and 17, 1954*, A Wiley Publication in Applied Statistics. Statistical Laboratory, John Wiley & Sons, Inc., 1956.

Kevin P. Murphy. Bayes net toolbox for Matlab, 1999. Available from `http://www.cs.berkeley.edu/~murphyk/Bayes/bnt.html`.

Radford M. Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. PhD thesis, Department of Computer Science, University of Toronto, 1993.

Radford M. Neal. Annealed importance sampling. Technical Report 9805, Department of Statistics, University of Toronto, Toronto, Ontario, Canada, September 1998. Available from `http://www.cs.utoronto.ca/~radford/`.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Dennis Nilsson and Steffen L. Lauritzen. Evaluating influence diagrams using LIMIDs. In Craig Boutilier and Moisés Goldszmidt, editors, *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, June 30–July 3, Stanford Univeristy, Stanford, California.*, pages 436–445, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

Man-Suk Oh. Monte Carlo integration via importance sampling: Dimensionality effect and an adaptive algorithm. In Nancy Flournoy and Robert K. Tsutakawa, editors, *Statistical Multiple Integration: Proceedings of a Joint Summer Research Conference held at Humboldt University, June 17-23, 1989*, volume 115 of *Computational Mathematics*, pages 165–187, Providence, Rhode Island, 1991. American Mathematical Society.

Luis E. Ortiz. Selecting approximately-optimal actions in complex structured domains. Technical Report CS-00-05, Computer Science Department, Brown University, 2000.

Luis E. Ortiz and Leslie Pack Kaelbling. Adaptive importance sampling for estimation in structured domains. In Craig Boutilier and Moisés Goldszmidt, editors, *Proceeding of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 446–454, 2000a.

Luis E. Ortiz and Leslie Pack Kaelbling. Sampling methods for action selection in influence diagrams. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 378–385, 2000b.

Art Owen and Yi Zhou. Safe and effective importance sampling. Can be downloaded from `http://www-stat.stanford.edu/~owen/reports/`, March 1999.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Representation and Reasoning. Morgan Kaufmann, San Mateo, California, revised second edition, 1988.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, first edition, 2000.

Malcolm Pradhan and Paul Dagum. Optimal Monte Carlo estimation of belief network inference. In *Proc. Twelfth Conf. on Uncertainty in Artificial Intelligence*, pages 446–453, 1996.

John W. Pratt, Howard Raiffa, and Robert Schlaifer. *Introduction to Statistical Decision Theory*. The MIT Press, Cambridge, Massachusetts; London, England, 1995.

Sam Roweis. Matrix identities. Can be downloaded from `http://www.gatsby.ucl.ac.uk/~roweis/notes.html`, 1999.

R. Y. Rubinstein. *Simulation and the Monte Carlo Method.* New York: Wiley, 1981.

Ross Shachter. Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98)*, pages 480–487, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

Ross D. Shachter and Mark A. Peot. Simulation approaches to general probabilistic inference on belief networks. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, pages 311–318, 1989.

Ross D. Shachter and Mark A. Peot. Decision making using probabilistic inference methods. In Didier Dubois, Michael P. Wellman, Bruce D'Ambrosio, and Phillipe Smets, editors, *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence, July 17-19, Stanford University*, pages 276–283, San Mateo, California, July 1992. Morgan Kaufmann Publishers.

Solomon Eyal Shimony. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, 68(2):399–410, August 1994.

Michael Shwe and Gregory Cooper. An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. *Computers and Biomedical Research*, 24:453–475, 1991.

David J. Spiegelhalter and Steffen L. Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20:579–605, 1990.

Sampath Srinivas. Generalizing the noisy or model to *n*ary variables. Technical Memorandum 79, Rockwell International Science Center, Palo Alto Laboratory, Palo Alto, CA, April 1992.

Masami Takikawa and Bruce D'Ambrosio. Multiplicative factorization of noisy-max. In Kathryn B. Laskey and Henri Prade, editors, *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 622–630, San Francisco, California, July 1999. Morgan Kaufmann.

Joseph A. Tatman and Ross D. Shachter. Dynamic programming and influence diagrams. *IEEE Transactions on Systems, Man and Cybernetics*, 20(2):365–379, March/April 1990.

Luke Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22 (4):1701–1728, December 1994.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

Richard L. Wheeden and Antoni Zygmund. *Measure and Integral: An Introduction to Real Analysis*, volume 43 of *Monographs and textbooks in pure and applied mathemathics*. Marcel Dekker, Inc., New York, New York, 1977.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Bethe free energy, Kikuchi approximations and belief propagation algorithms. Technical Report TR2001-16, MERL, May 2001. Can be downloaded from `http://www.merl.com/papers/TR2001-16/`.

Nevin Lianwen Zhang. Probabilistic inference in influence diagrams. In Gregory F. Cooper and Serafín Moral, editors, *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–98), July 24-26, University of Wisconsin Business School, Madison, Wisconsin, USA*, pages 514–522, San Francisco, CA, 1998. Morgan Kaufmann Publishers.

Appendix A

Computer-mouse problem

In this appendix, we present a simple made-up ID. We use this model for preliminary empirical evaluation of some of the methods presented in this thesis. We understand this problem is very simplistic and exact computation is easy. However, we use this example for the purpose of illustration and ease of analysis.

Figure A.1 gives a graphical representation of the ID for the *computer mouse problem*. The idea is to select an optimal strategy of whether to *buy* a new mouse ($A = 1$), *upgrade* the operating system ($A = 2$), or take *no action* ($A = 3$). The observation is whether the mouse pointer is working ($MP_t = 1$) or not ($MP_t = 0$). The variables of the problem are the status of the operating system (OS), the status of the driver (D), the status of the mouse hardware (MH), and the status of the mouse pointer (MP), all at the current and future time (subscripted by t and $t + 1$). The variables are all binary.

Table A.1 shows the probabilistic values of the local (conditional) probability distributions for this problem. The variables in the tables in the second row do not have a subscript. By this we mean, $P(MD_t \mid OS_t) = P(MD_{t+1} \mid OS_{t+1})$ and $P(MP_t \mid MD_t, MH_t) = P(MP_{t+1} \mid MD_{t+1}, MH_{t+1})$. The probabilistic model encodes the following information about the system. The mouse is old and somewhat unreliable. The operating system is reliable. It is very likely that the mouse pointer will not work if either the driver or the mouse hardware has failed. The utility model is such that, for all actions, states associated with the mouse pointer working in the future have larger values than those associated with the mouse pointer not working in the future. The utility associated with the actions *buy*, *upgrade*, and *no action* increases in that order. The range of the utility values is from 0 to 50. Table A.2 shows the value of the utilities. Table A.3 shows the values of the actions and observations $V_{\boldsymbol{O}}(A)$. From Table A.3 we conclude that the optimal strategy is: buy a

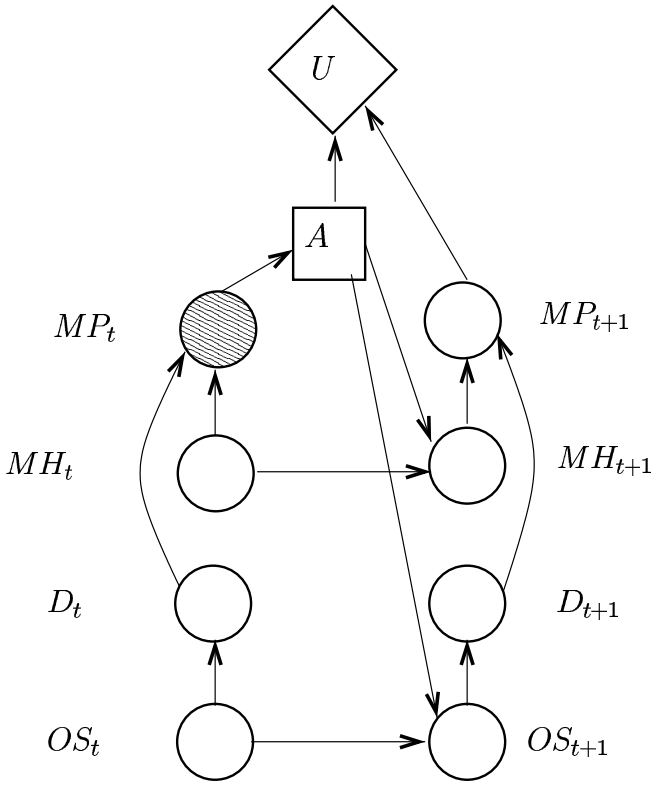


Figure A.1: Graphical representation of the ID for the computer mouse problem.

new mouse ($A = 1$) if the mouse pointer is not working ($MP_t = 0$); take no action ($A = 3$) if the mouse pointer is working ($MP_t = 1$). This strategy has value 26.50.

$P(OS_t)$	OS_t	
	0	1
	0.9	0.1

$P(MH_t)$	MH_t	
	0	1
	0.65	0.35

$P(MD \mid OS)$	MD	
OS	0	1
0	0.85	0.15
1	0.2	0.8

$P(MP \mid MD, MH)$		MP	
MD	MH	0	1
0	0	0.99	0.01
0	1	0.99	0.01
1	0	0.99	0.01
1	1	0.95	0.05

$P(OS_{t+1} \mid OS_t, A)$		OS_{t+1}	
OS_t	A	0	1
0	1	0.9	0.1
0	2	0.05	0.95
0	3	0.9	0.1
1	1	0.1	0.9
1	2	0.05	0.95
1	3	0.1	0.9

$P(MH_{t+1} \mid MH_t, A)$		MH_{t+1}	
MH_t	A	0	1
0	1	0.05	0.95
0	2	0.95	0.05
0	3	0.95	0.05
1	1	0.05	0.95
1	2	0.15	0.85
1	3	0.15	0.85

Table A.1: Probability values for the computer-mouse ID.

$U(MP_{t+1}, A)$	MP_{t+1}	
A	0	1
1	0	40
2	5	45
3	10	50

Table A.2: Utility values for the computer-mouse ID.

$V_{MP_t}(A)$	MP_t	
A	0	1
1	**18.20**	6.60
2	7.54	7.39
3	10.57	**8.30**

Table A.3: Value of actions and observations for the computer-mouse ID problem.

Appendix B

Motivating example for large complex model

We present a "story" behind a large model represented graphically in Figure B.1. Imagine we are drivers that face the same problem every day. As we get out of work, we have the choice of one of many parallel routes (like highways). We would like to take one route (or lane) and stick to it until we arrive to our destination. The only information we would have available at the time we make our decision of which lane to take is some information given on the radio about the general status of the routes (a summary of the state of the first section of the route). This information is limited as it does not exactly tell us which routes might be congested at the start of our trip, but gives us a general statement about the traffic conditions. Instead, it might say something like "There is heavy traffic in routes to 'our destination' " or "There is medium to light traffic" etc.

We have a model of the "dynamics" of the stretches we have to go through to get to our destinations. The routes are "spatially" related, as other traffic can move in and out of adjacent lanes or routes into or out of our lane. Remember, we cannot move once we decided the route we are going to take. We care about the "state of the lane stretches" (i.e., is it congested?). Depending on the lane we select we can have an effect (however small) on the state of the lanes at the next stretch of road for the different routes. Also, our utility is a "global" and potentially non-linear, but compactly represented, function of the state of the stretches of the routes up to our destination and the route we selected (i.e., If we selected route i then we attach some value to each non-congested stretch and a cost to congested stretches; maybe congested stretches early on are not so painful to us as congested stretches later; maybe it hurts us to know afterwards that that there was an alternative route that

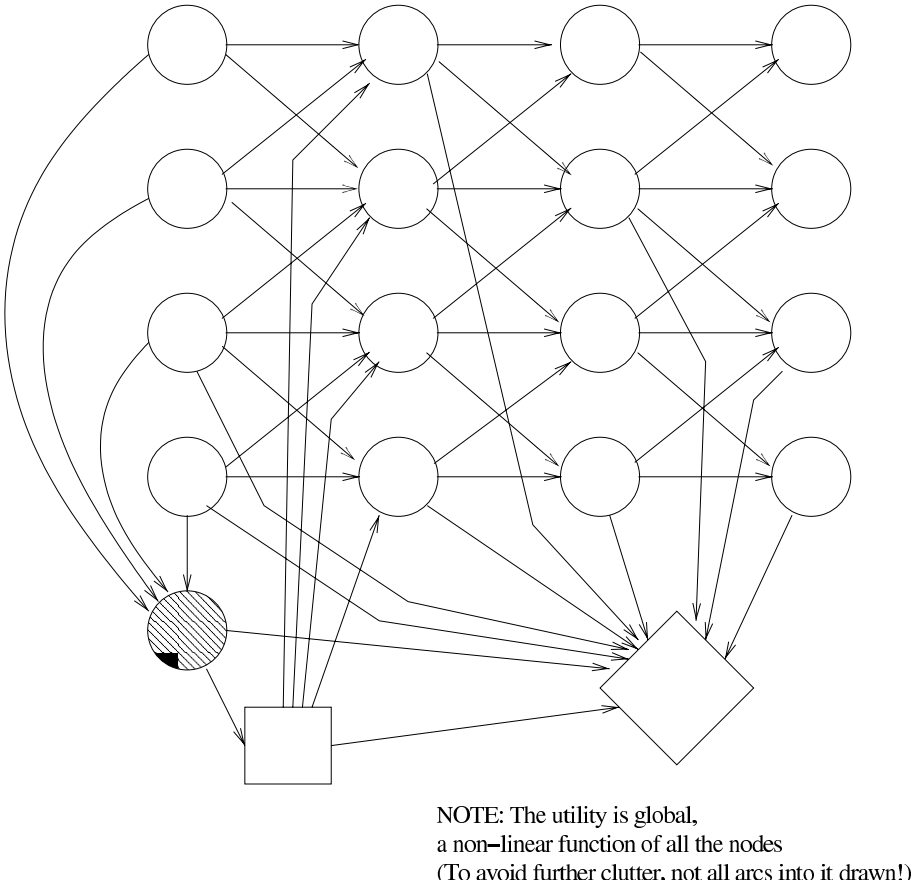


Figure B.1: Graphical representation of a large complex model.

was a lot less congested than the one we selected.). In addition, each stretch might have different characteristics, even though there is a sort of Markovian "spatial" structure, it is not necessarily an homogeneous one (not all the transition probabilities are the same from one stretch to the next).

The "hidden" structure of the problem might seem too complex for such a limited amount of information. However, that might be the level of granularity for which we can have reasonably accurate assessments of the interactions between the state of the stretches at different parts of the route. One might argue that if the process is not "too chaotic," even this limited amount of information is good enough to get an assessment of the potential "trajectories" the process will take.

I believe that the "state-of-the-art" exact methods for solving influence diagrams will have problem with a model such as the one I just described. First, as the number of lanes increases the "width" of the graph becomes large (and it becomes more dense). Also, the fact that the utility function might involve non-linear interactions among all the variables in the system is also a problem as it creates a direct dependency about the states of *all* the variables in the system (i.e., it creates a large clique containing all the variables!). On the other hand, note that sampling methods might fare better. The "spatial" decomposition allows for simple local models of interaction for the states at two subsequent stretches; that is, whether the next stretch for a lane will be congested depends only on whether the current and adjacent lanes are congested. This allows simple forward simulation of the process. Also, we just need to be able to evaluate the utility efficiently given outcomes for the variables. I believe the effectiveness of sampling methods will be directly tied to the actual parameters defining the interactions and the utility, not to the actual "structural" decomposition (i.e., graphical representation); the more chaotic the system, the less effective the sampling method will be.

I have yet to define the actual parameters and utility function for this model.

By the way, I should note that I have seen a similar structure used by Kwon and Murphy [2000] and others to model highway traffic using a special class of factored HMMs.

Appendix C

Additional experimental results for adaptive importance sampling on QMR-DT-type BN

In this appendix, we include all the remaining results obtained for applying several adaptive importance sampling (AIS) methods with different settings. Figures C.1- C.27 show the results. Please refer to to Section 3.6.2 for a description of the plots.

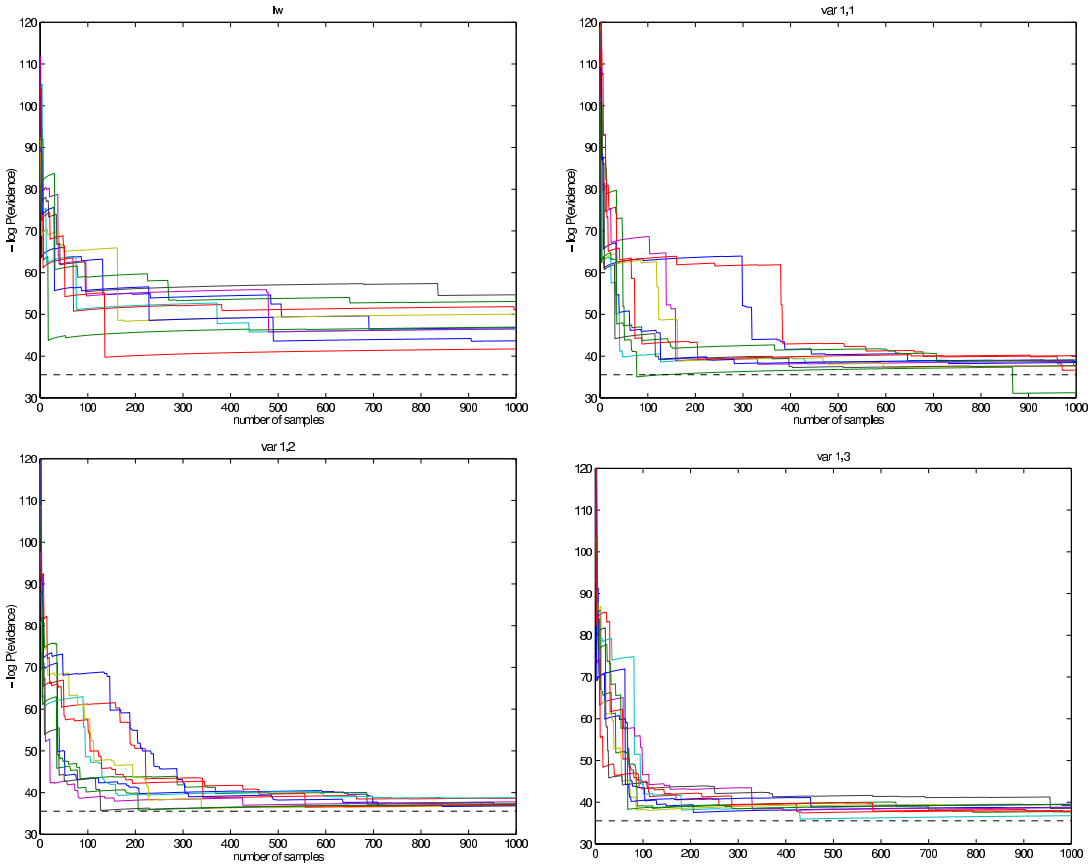


Figure C.1: Results for AIS method based on minimizing e_{Var} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 10^{50}, 10^{49}, and 10^{51}, respectively. Refer to the text for other basic general descriptions.

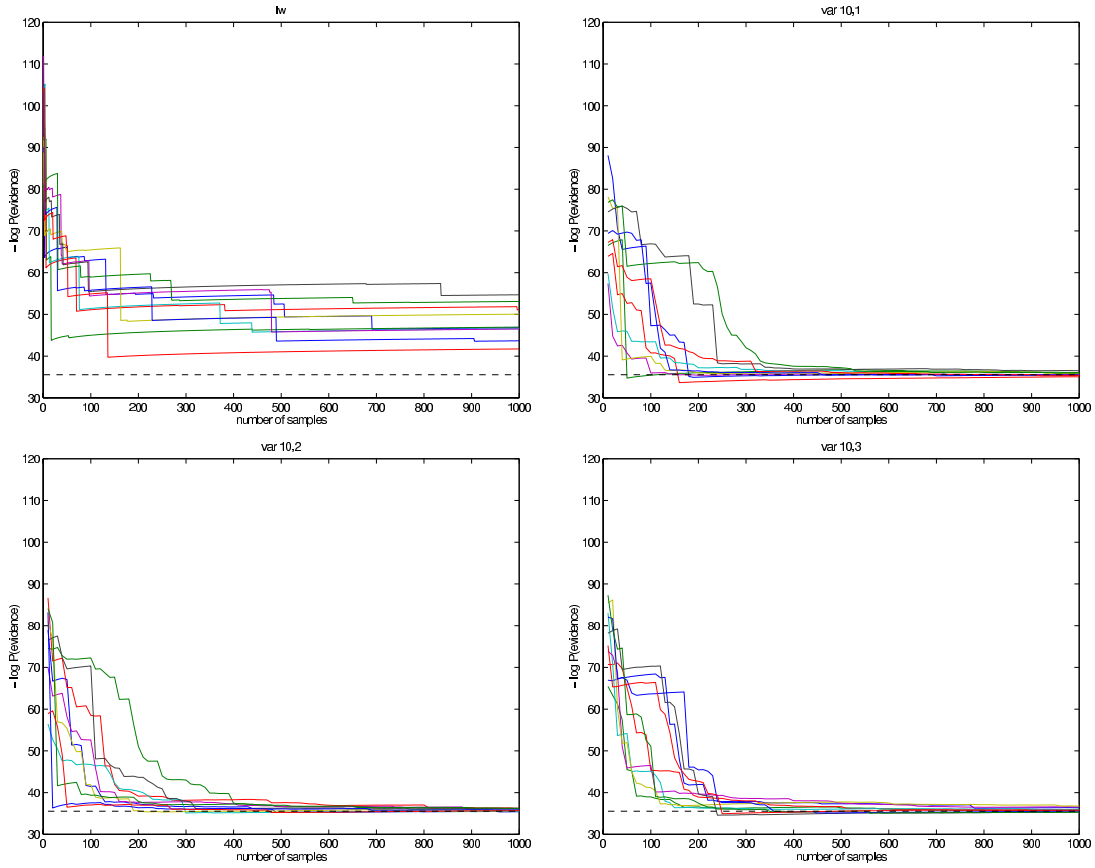


Figure C.2: Results for AIS method based on minimizing e_{Var} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 10^{50}, 10^{49}, and 10^{51}, respectively. Refer to the text for other basic general descriptions.

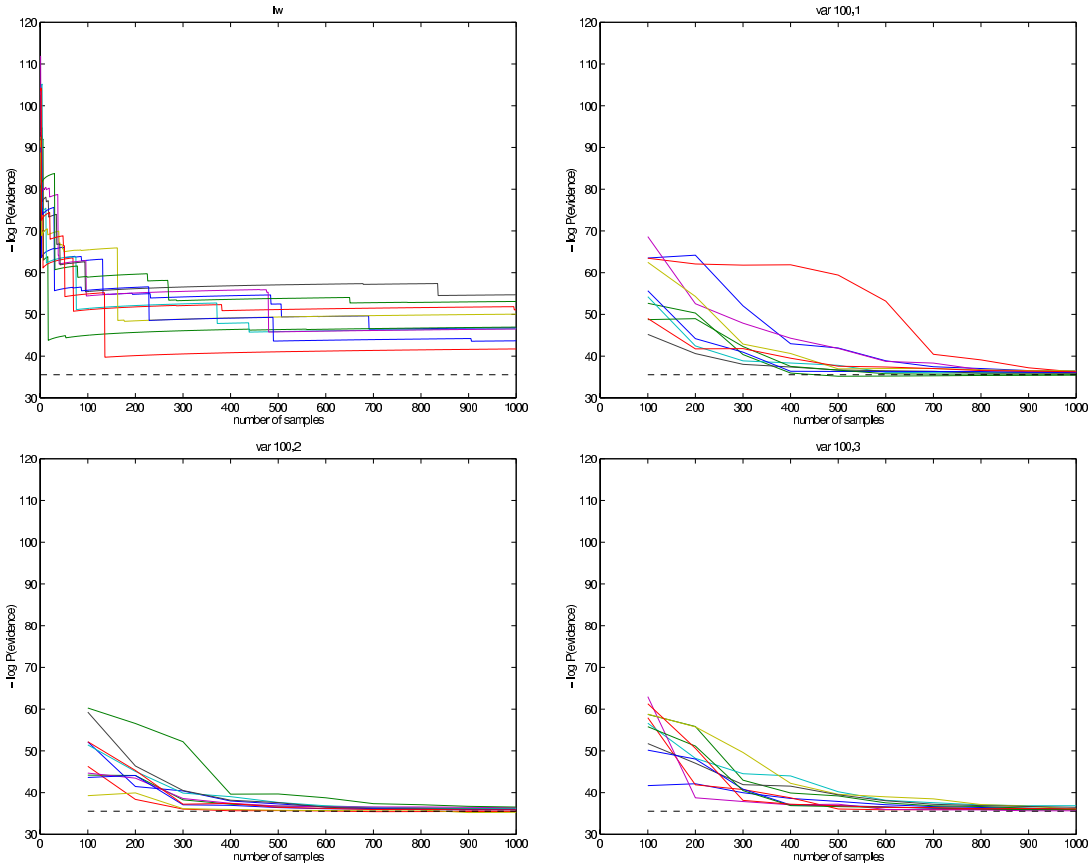


Figure C.3: Results for AIS method based on minimizing e_{Var} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 10^{50}, 10^{49}, and 10^{51}, respectively. Refer to the text for other basic general descriptions.

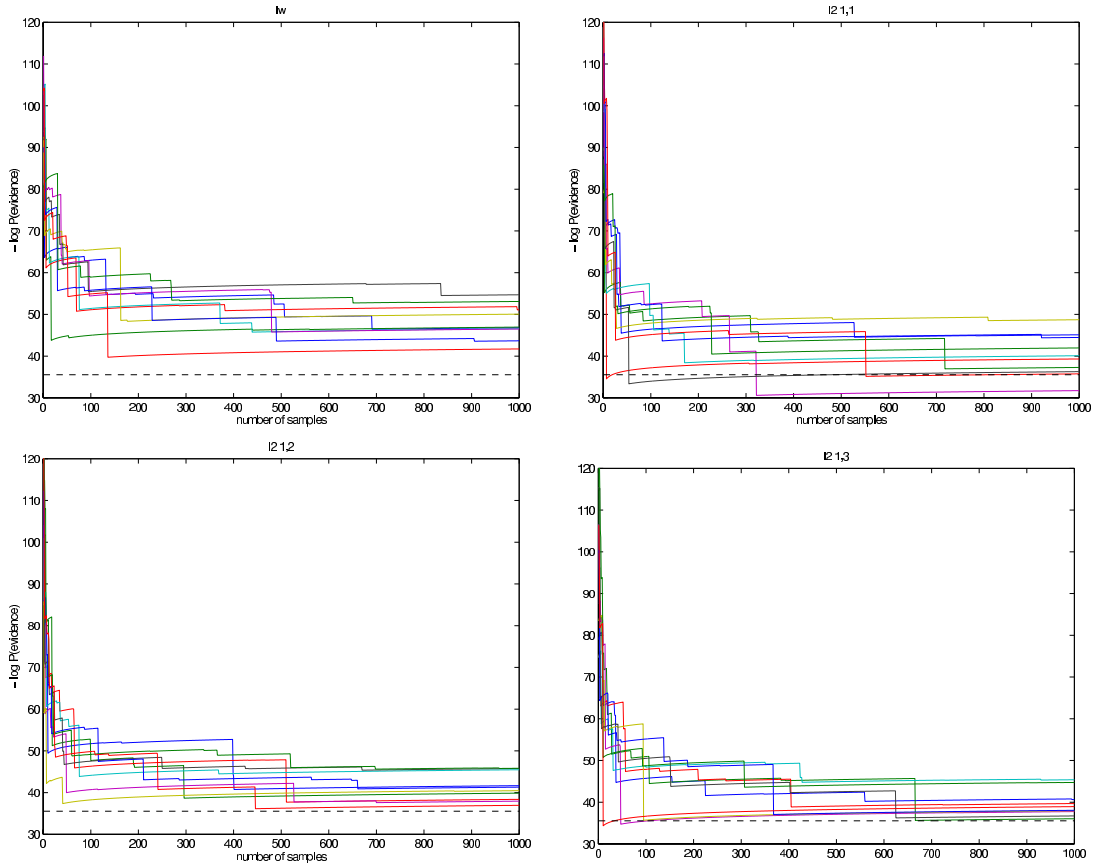


Figure C.4: Results for AIS method based on minimizing e_{L_2} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 10^{20}, 10^{19}, and 10^{21}, respectively. Refer to the text for other basic general descriptions.

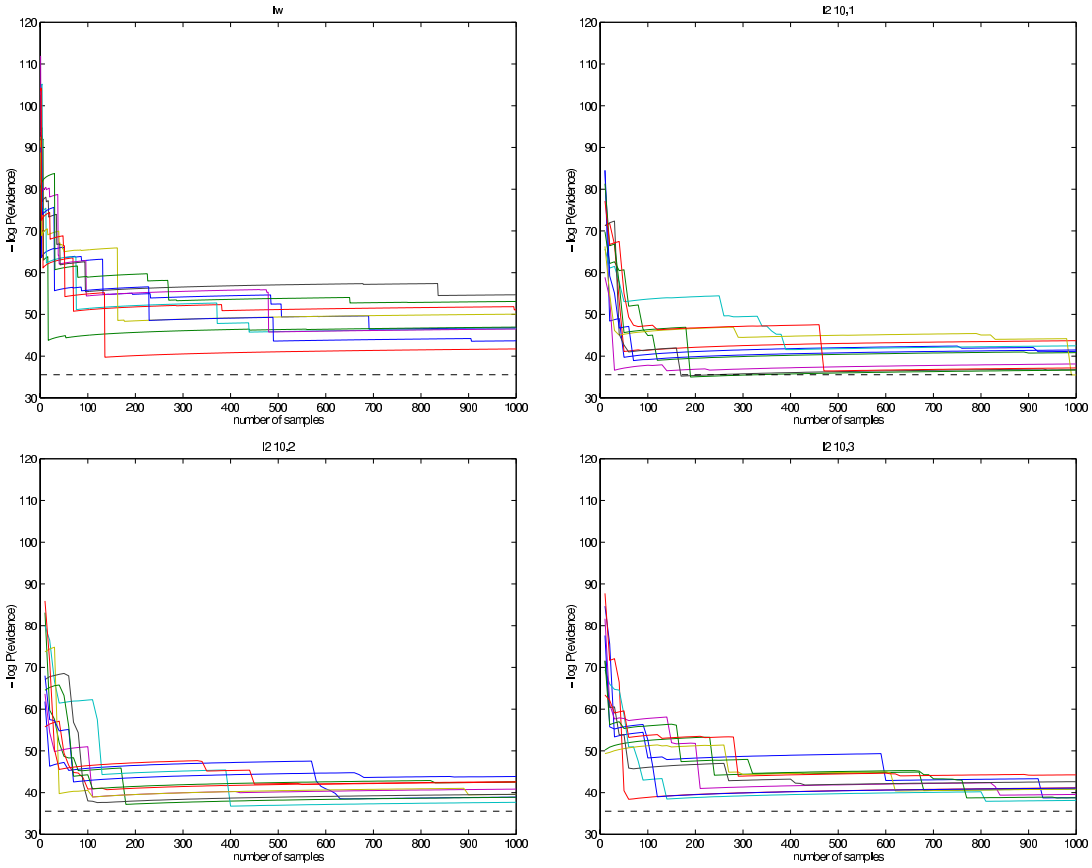


Figure C.5: Results for AIS method based on minimizing e_{L_2} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 10^{20}, 10^{19}, and 10^{21}, respectively. Refer to the text for other basic general descriptions.

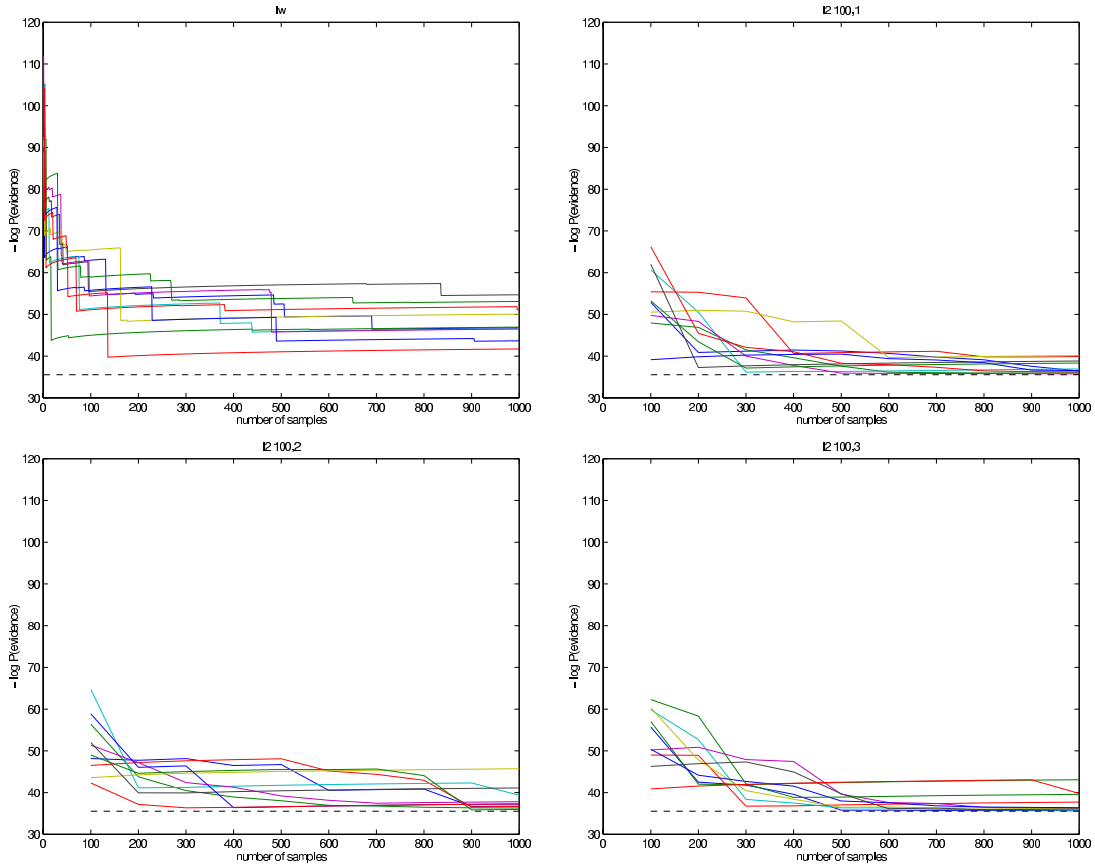


Figure C.6: Results for AIS method based on minimizing e_{L_2} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 10^{20}, 10^{19}, and 10^{21}, respectively. Refer to the text for other basic general descriptions.

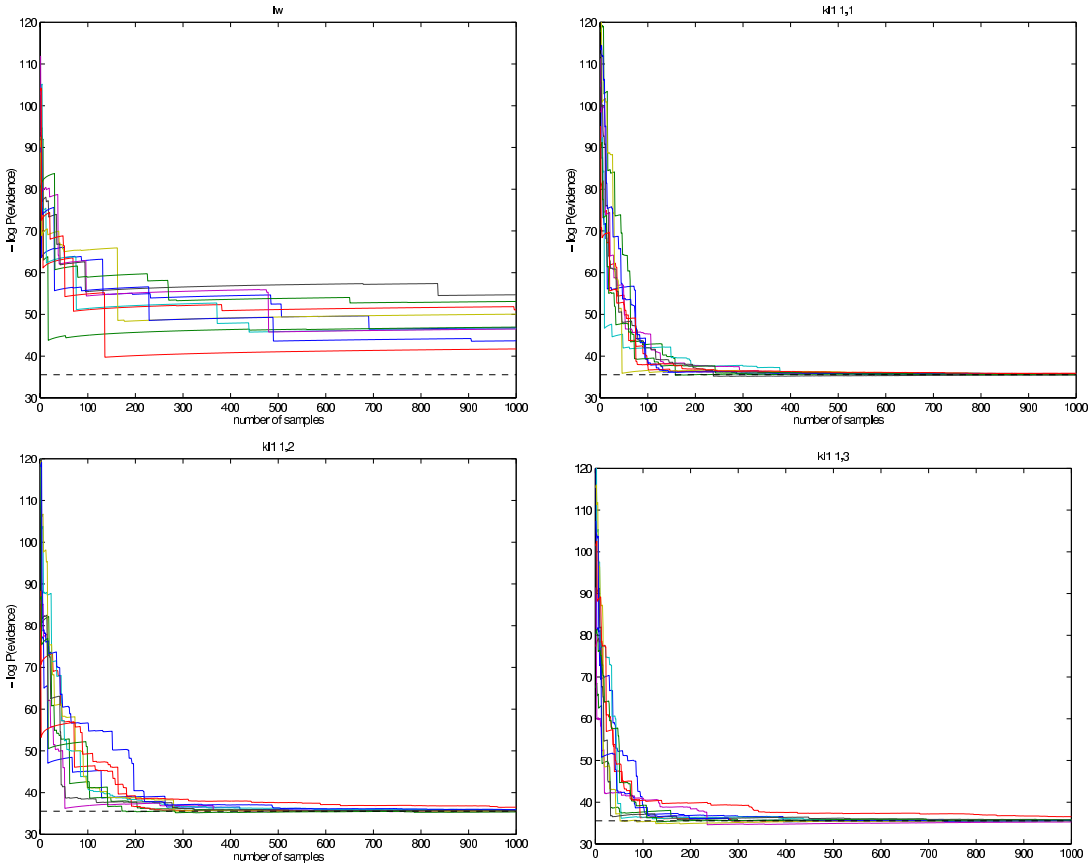


Figure C.7: Results for AIS method based on minimizing e_{KL_1} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

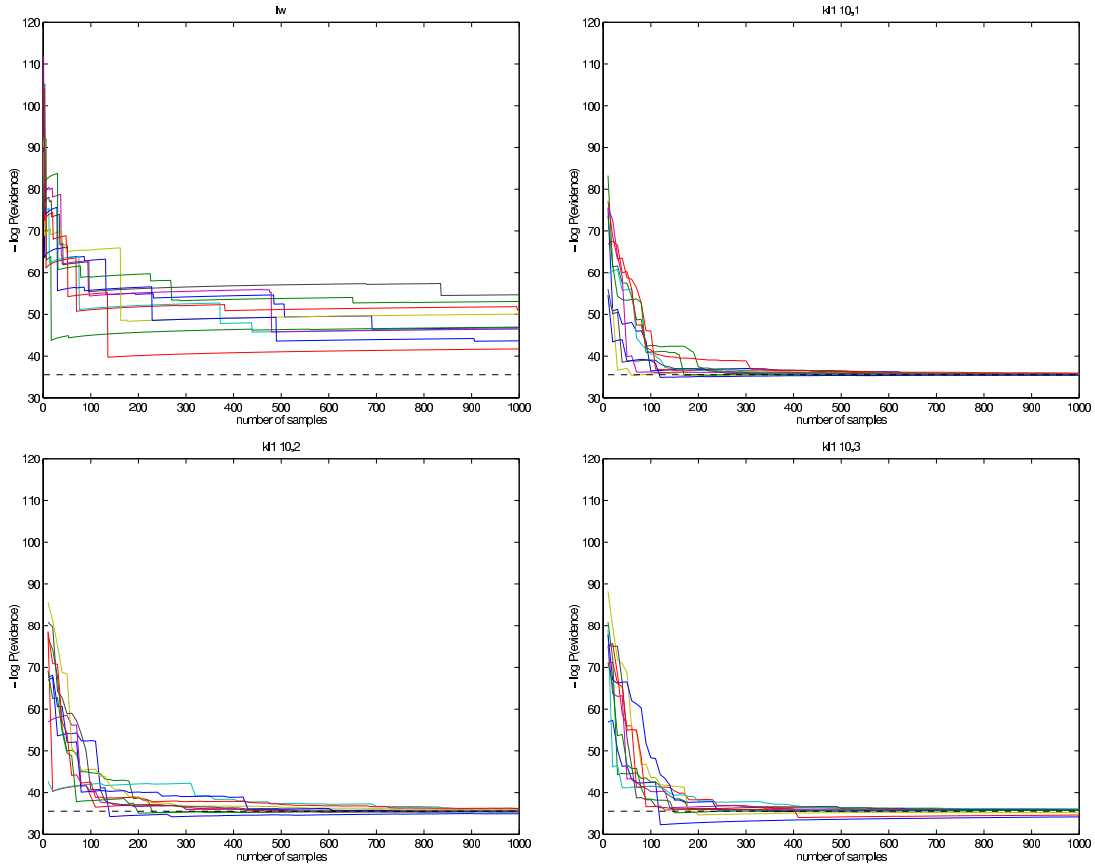


Figure C.8: Results for AIS method based on minimizing e_{KL_1} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

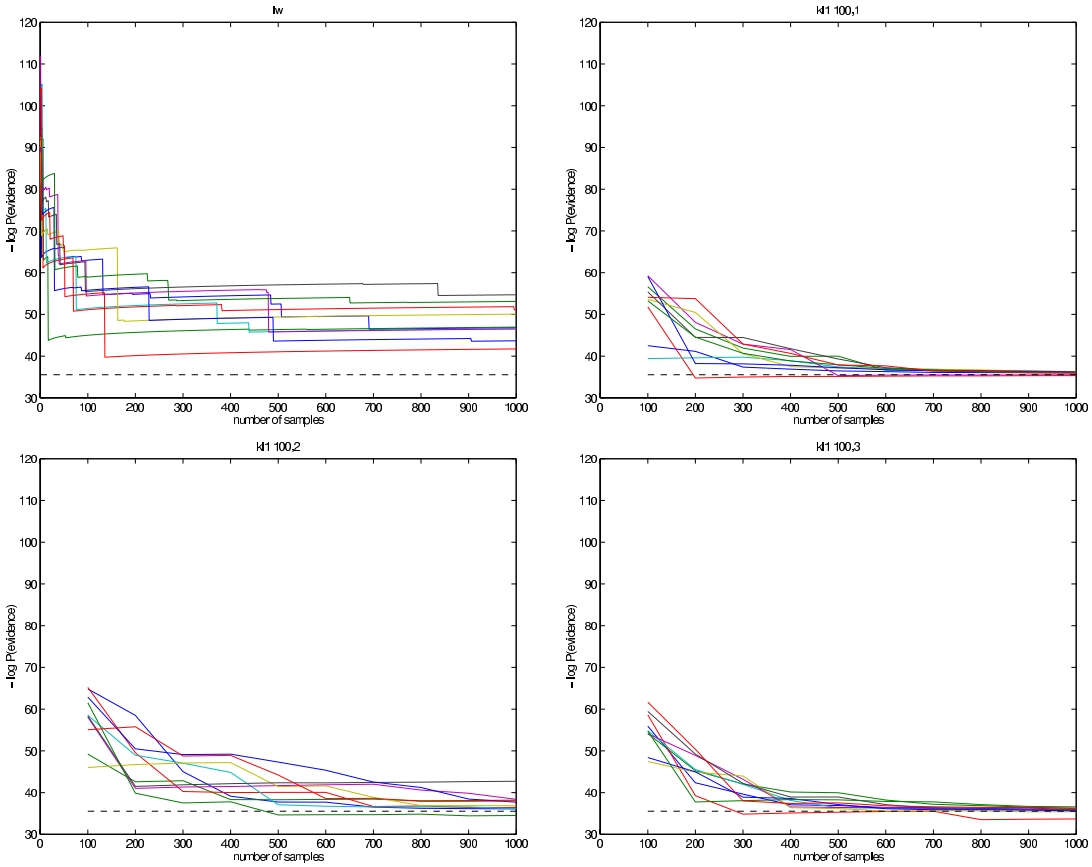


Figure C.9: Results for AIS method based on minimizing e_{KL_1} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

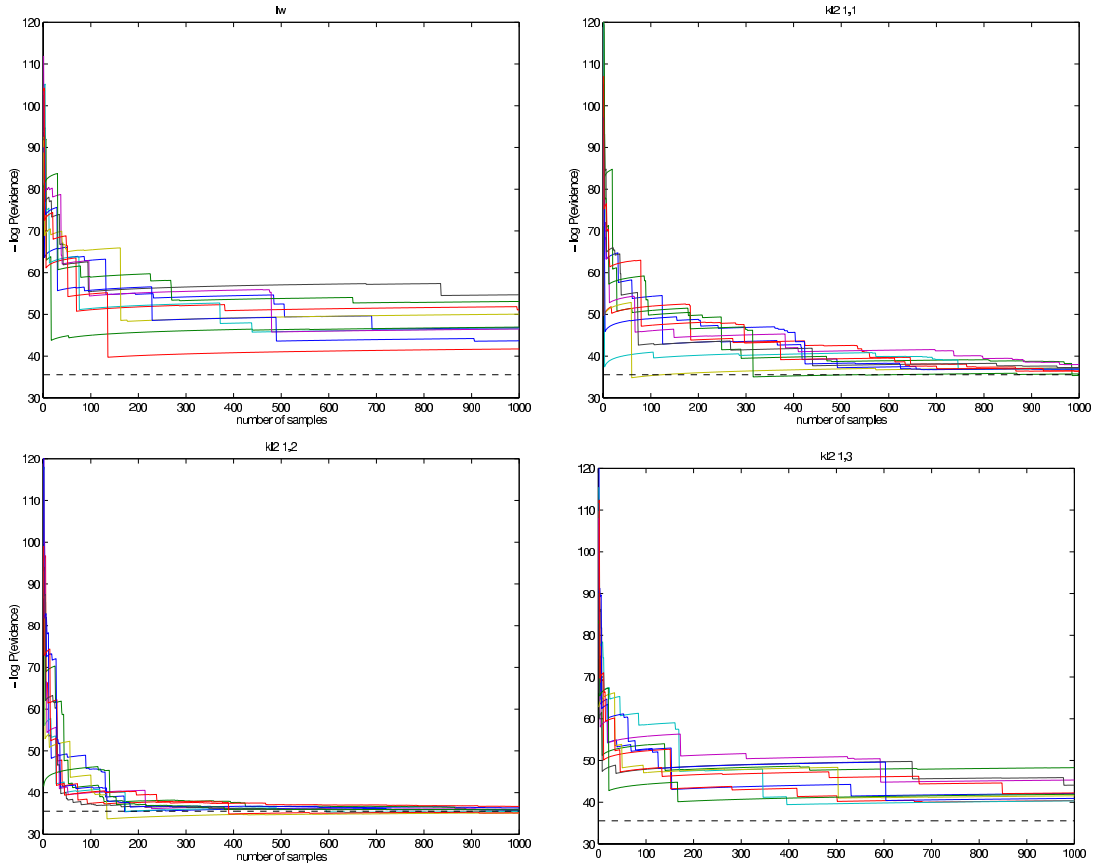


Figure C.10: Results for AIS method based on minimizing e_{KL_2} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

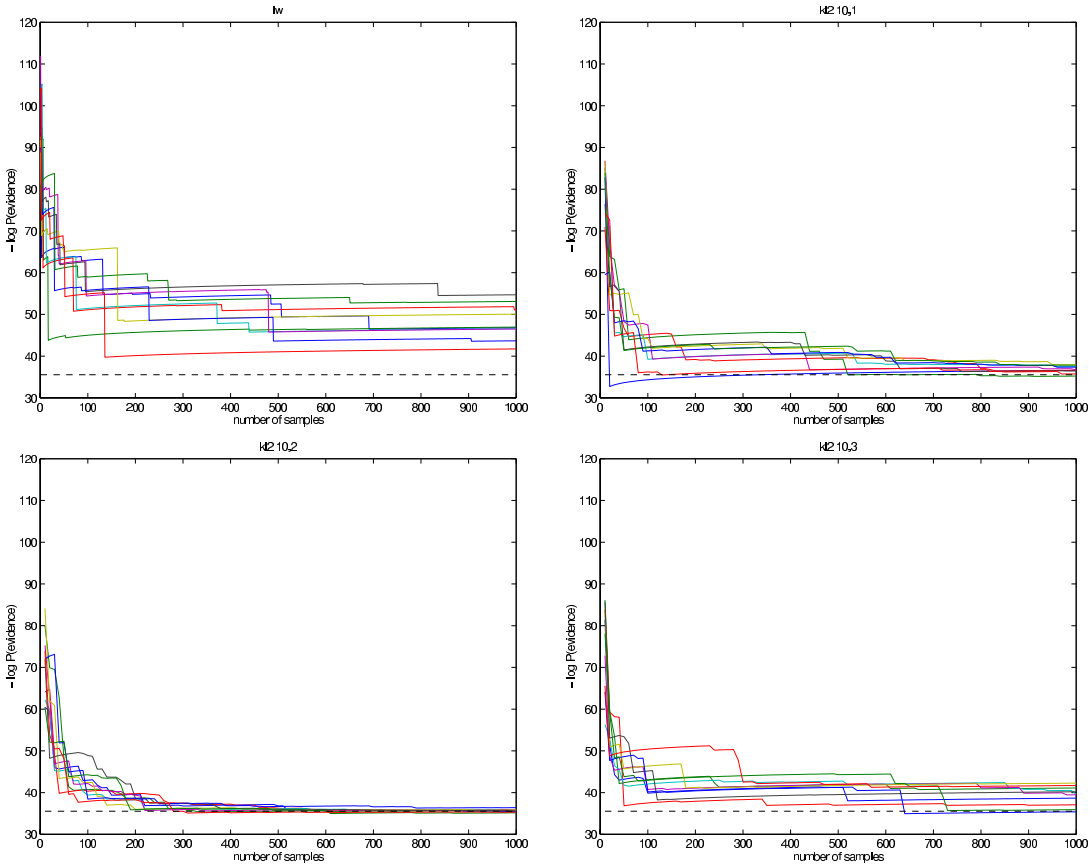


Figure C.11: Results for AIS method based on minimizing e_{KL_2} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

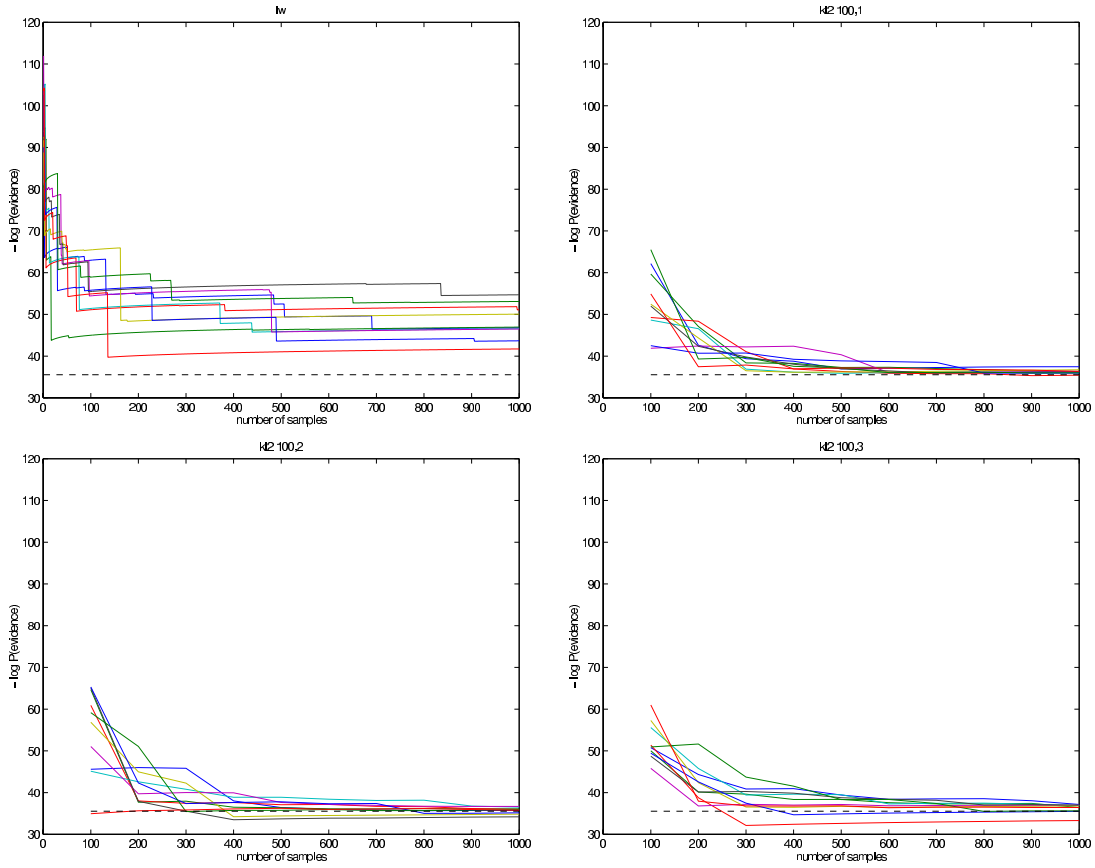


Figure C.12: Results for AIS method based on minimizing e_{KL_2} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

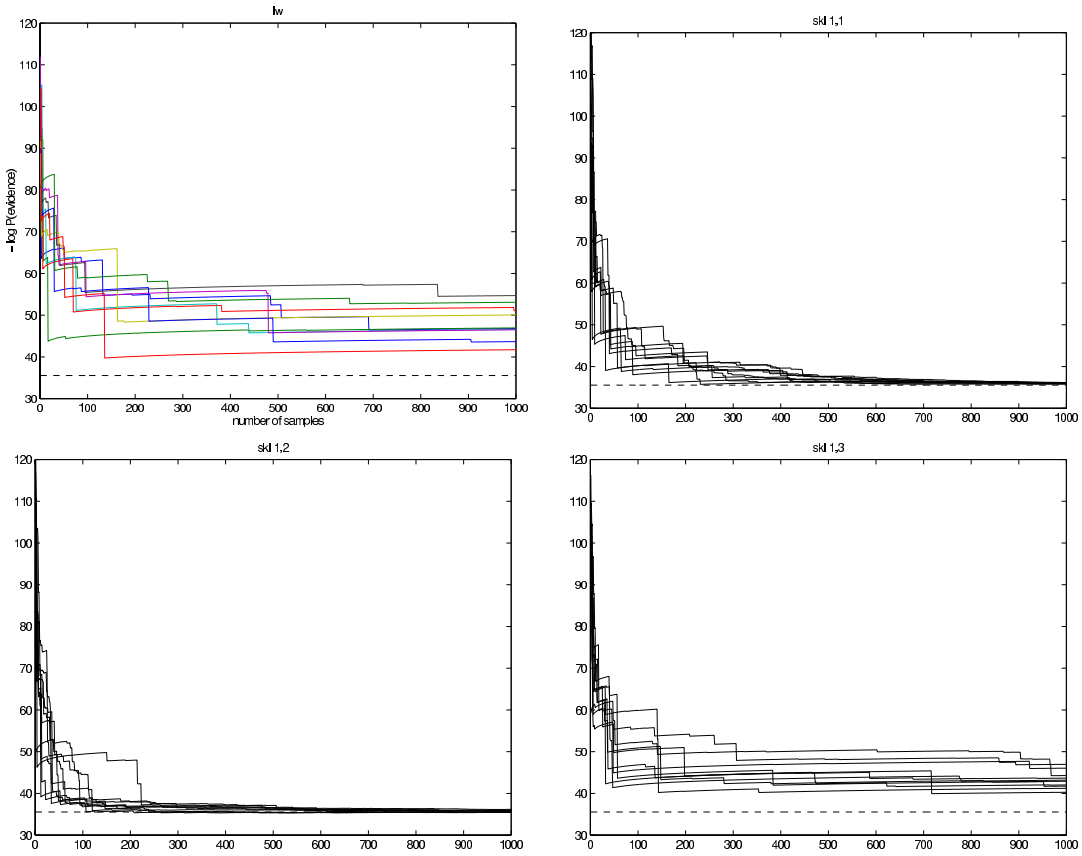


Figure C.13: Results for AIS method based on minimizing e_{KL_s} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

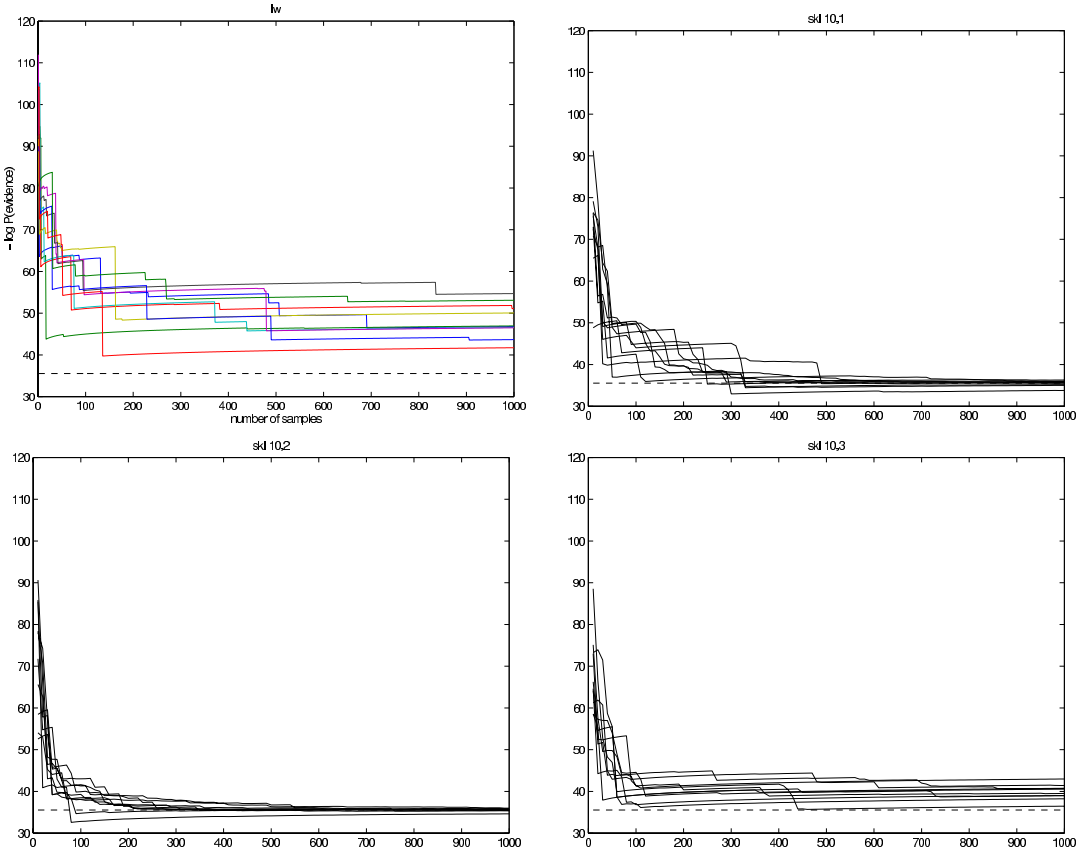


Figure C.14: Results for AIS method based on minimizing e_{KL_s} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

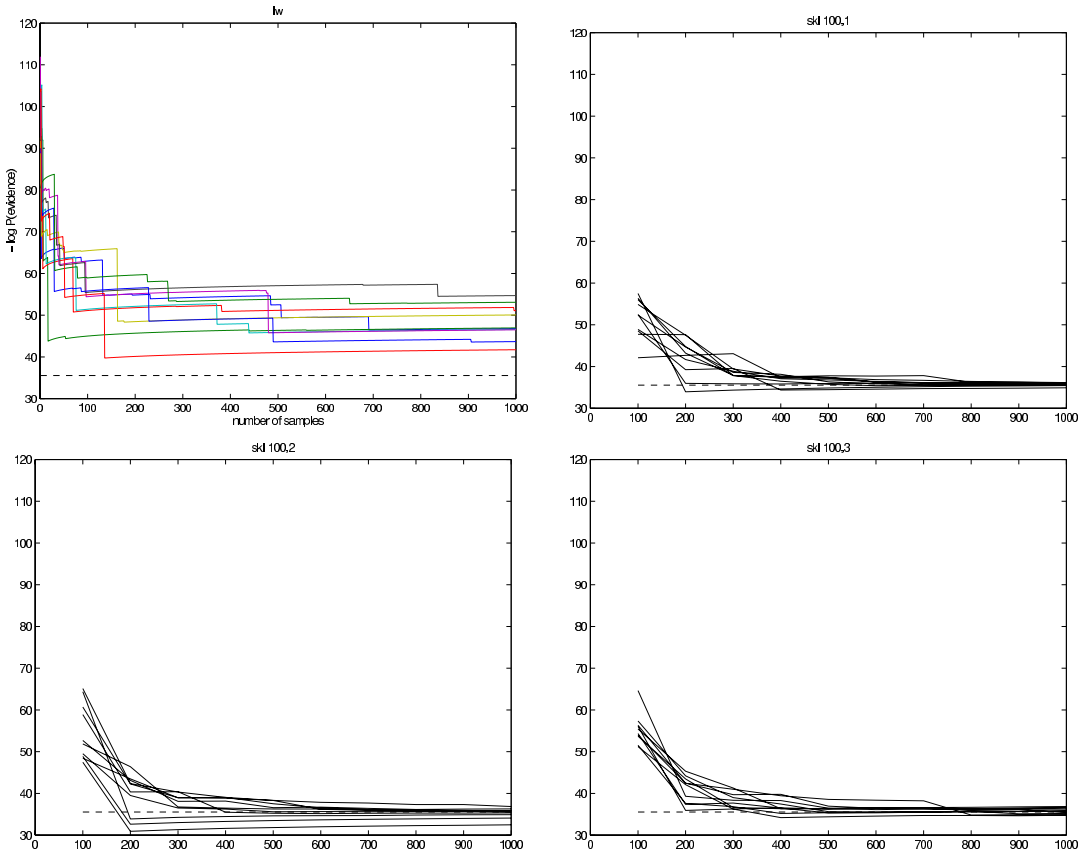


Figure C.15: Results for AIS method based on minimizing e_{KL_s} for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

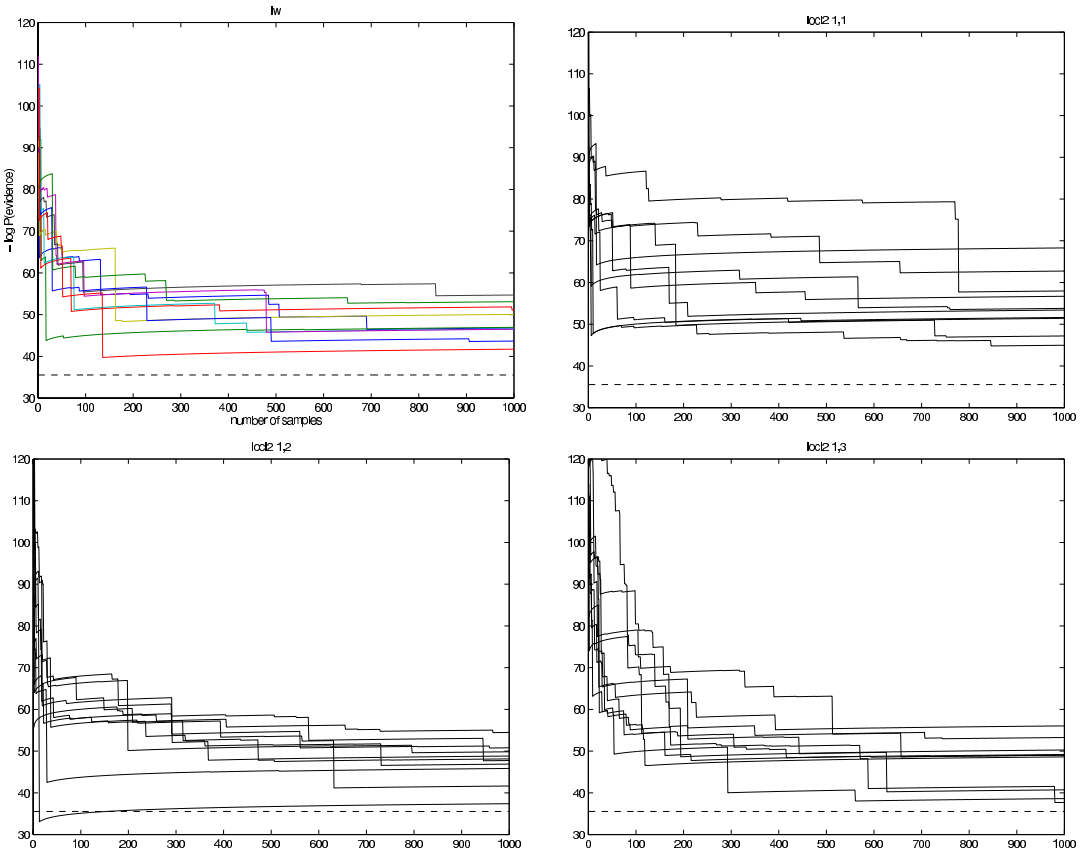


Figure C.16: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{L}_2}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

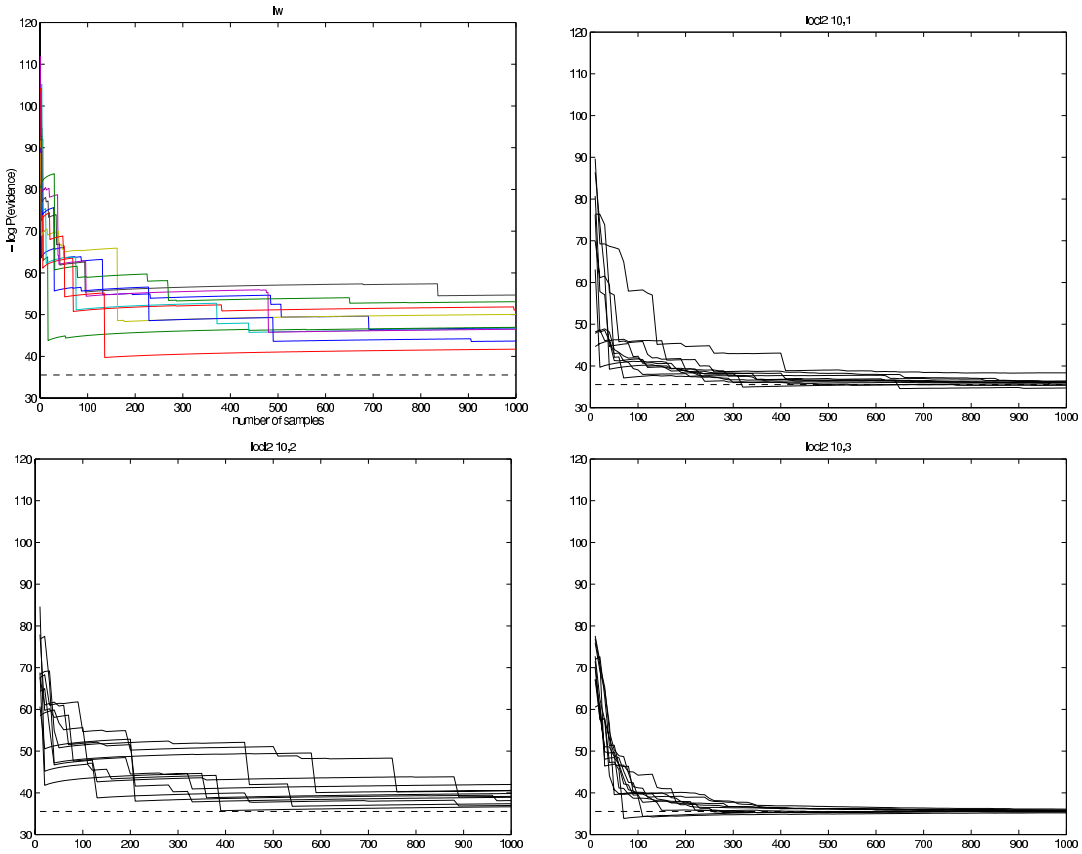


Figure C.17: Results for AIS method based on minimizing $e^{\mathrm{loc}}_{\mathrm{L2}}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

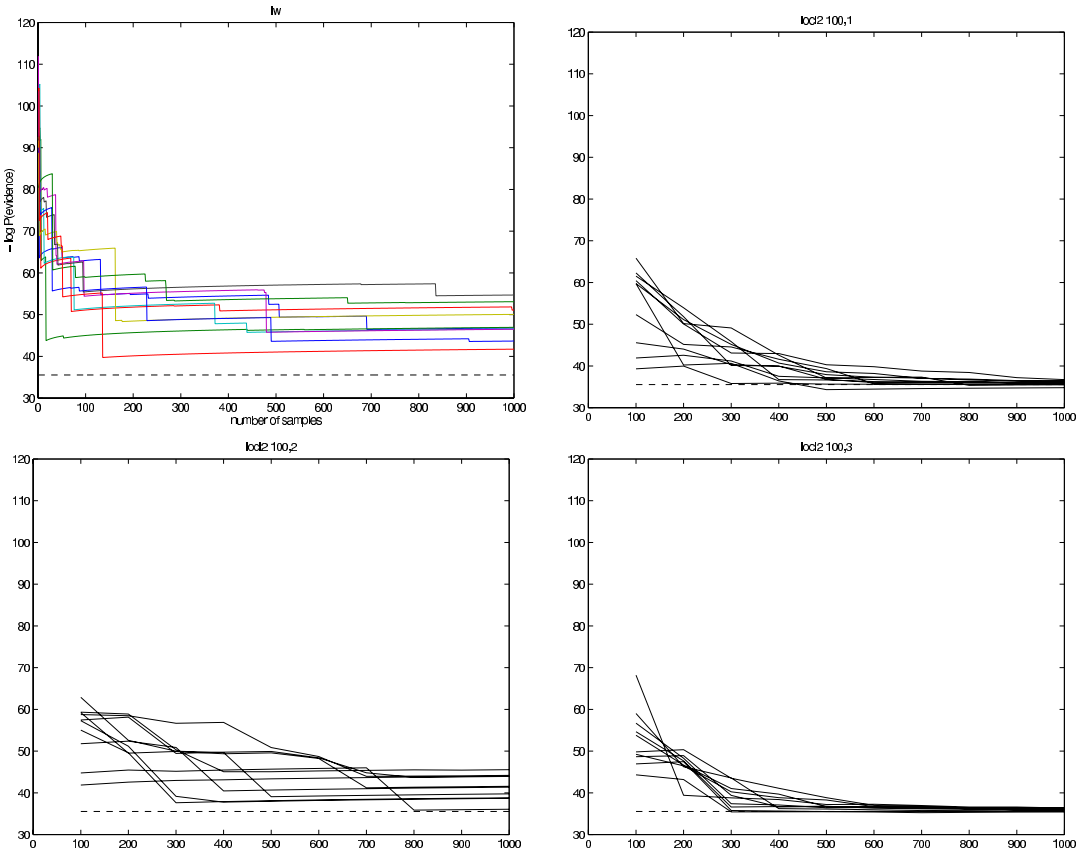


Figure C.18: Results for AIS method based on minimizing $e^{\mathrm{loc}}_{\mathrm{L}_2}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

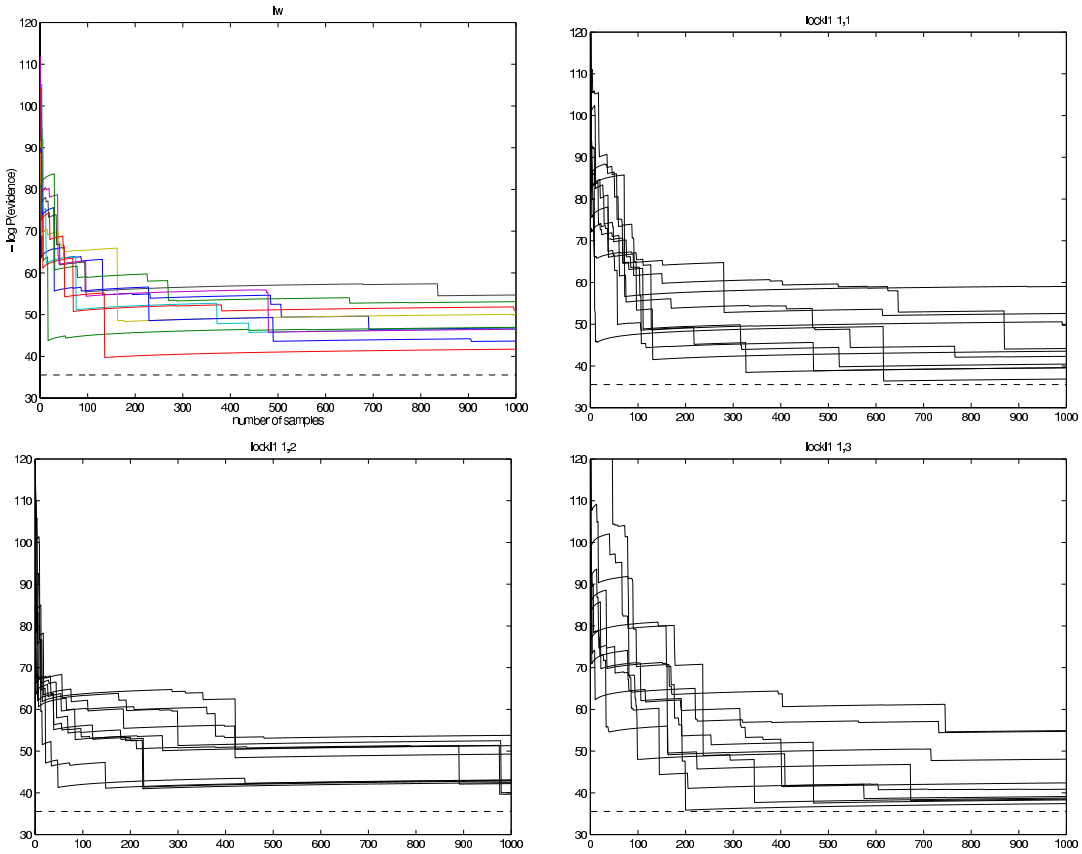


Figure C.19: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_1}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

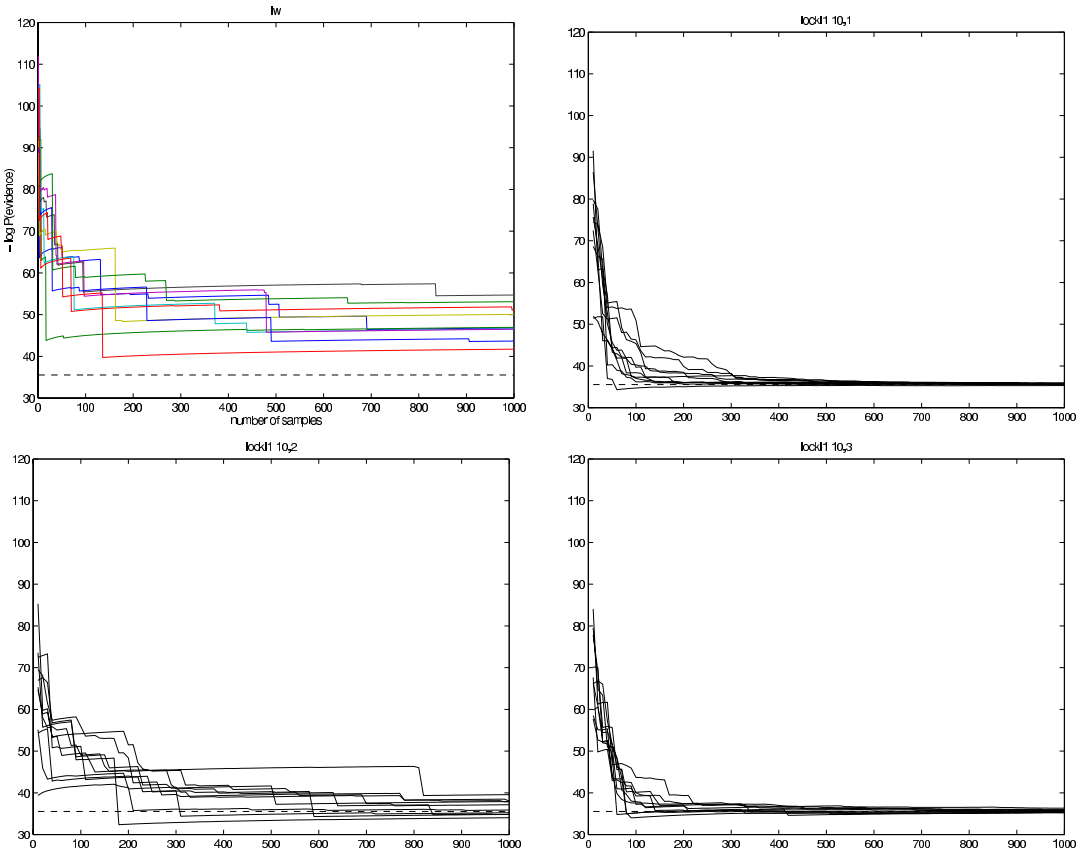


Figure C.20: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_1}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

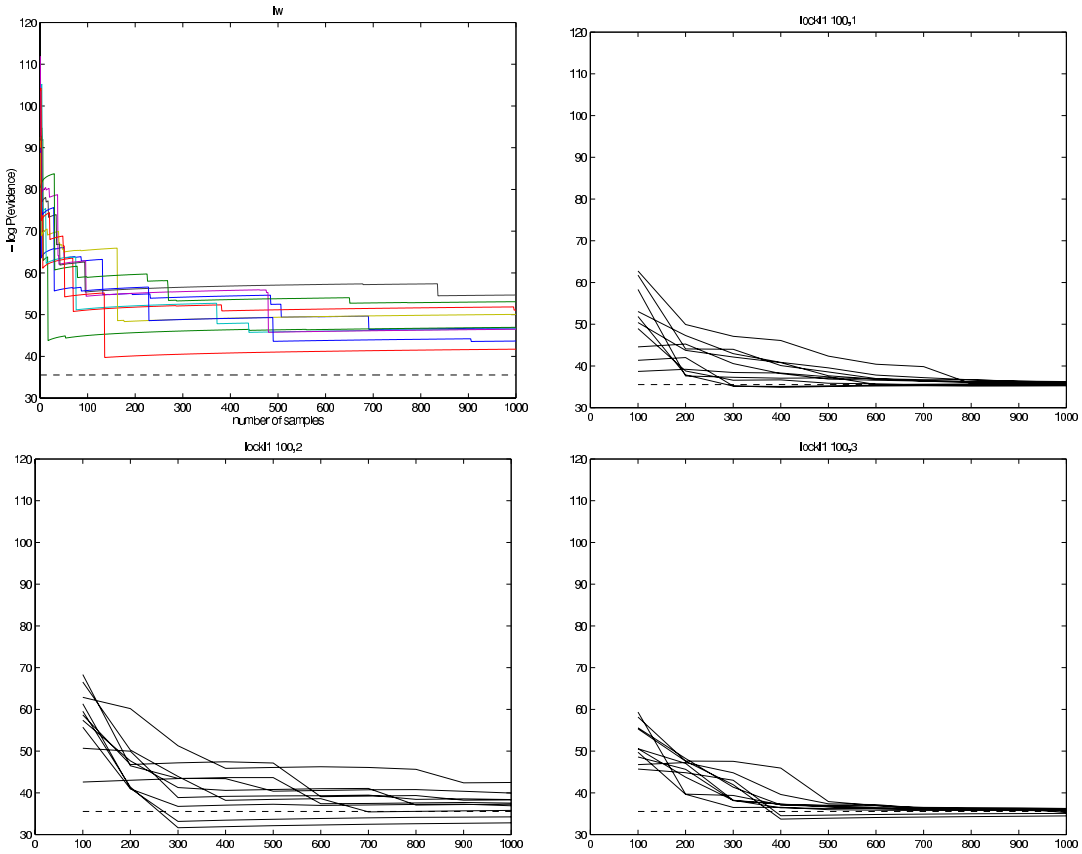


Figure C.21: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_1}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

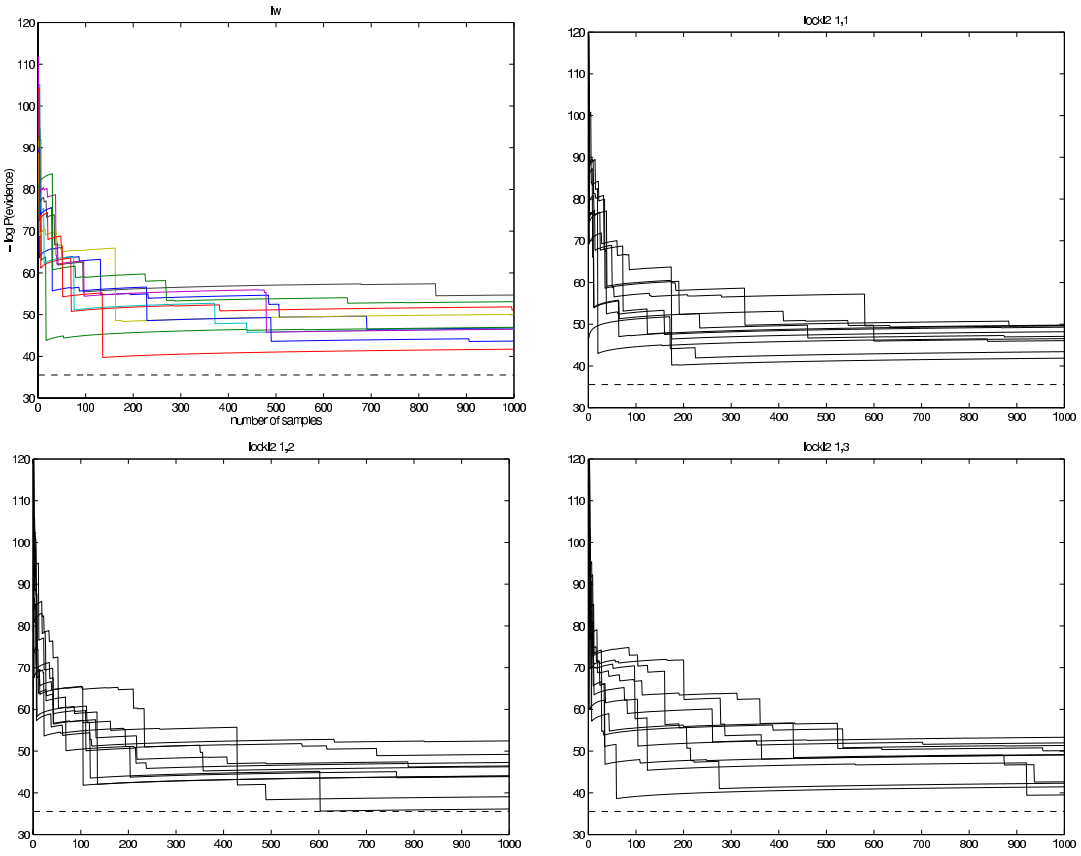


Figure C.22: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_2}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

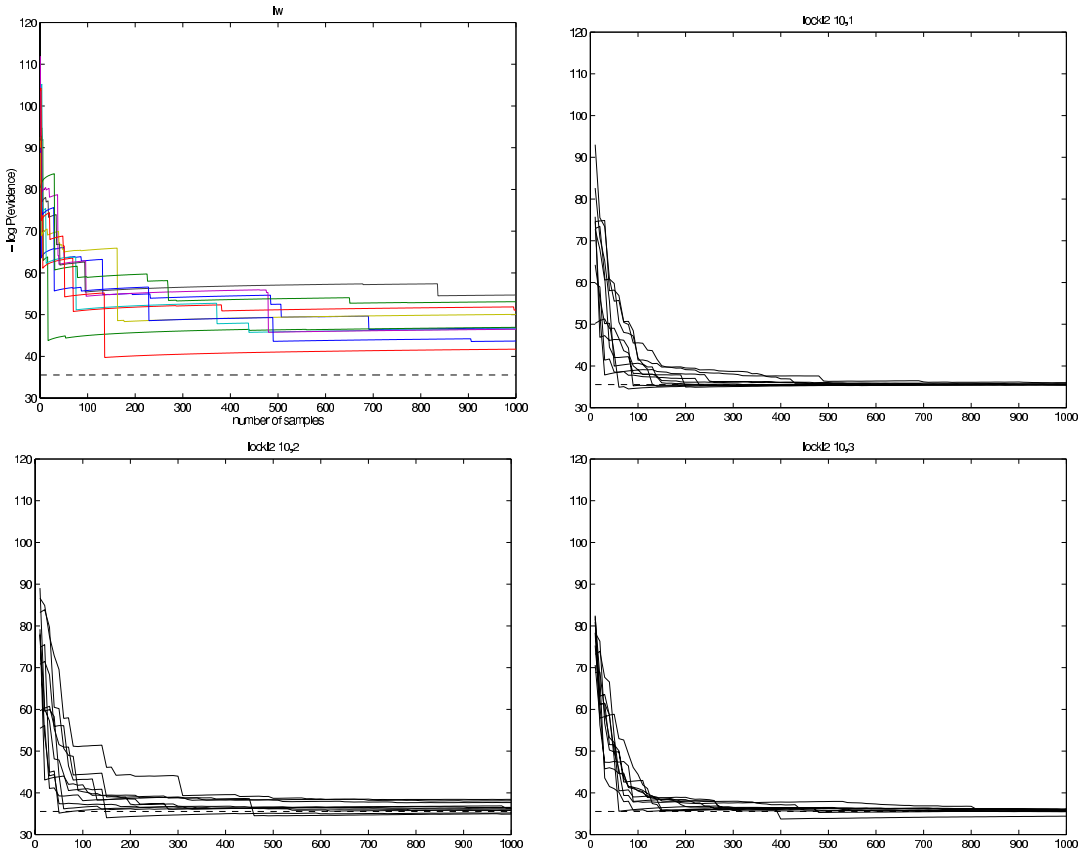


Figure C.23: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_2}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

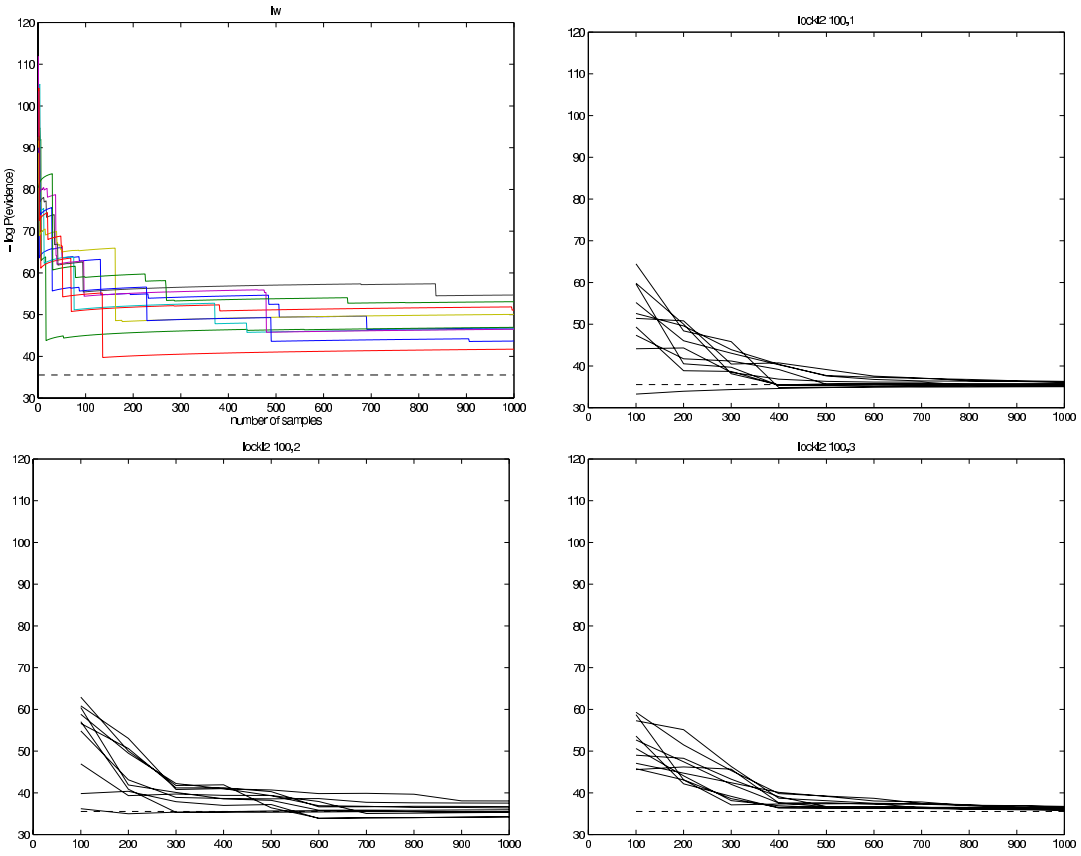


Figure C.24: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_2}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

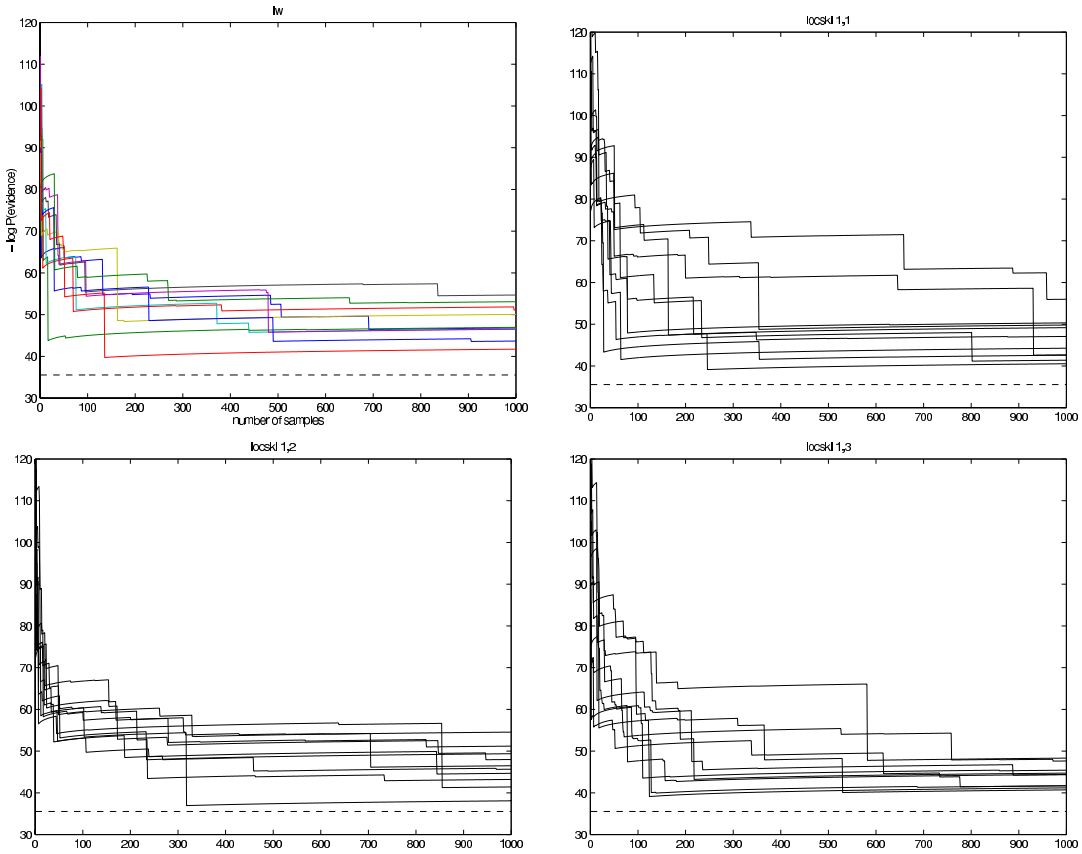


Figure C.25: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_s}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 1. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

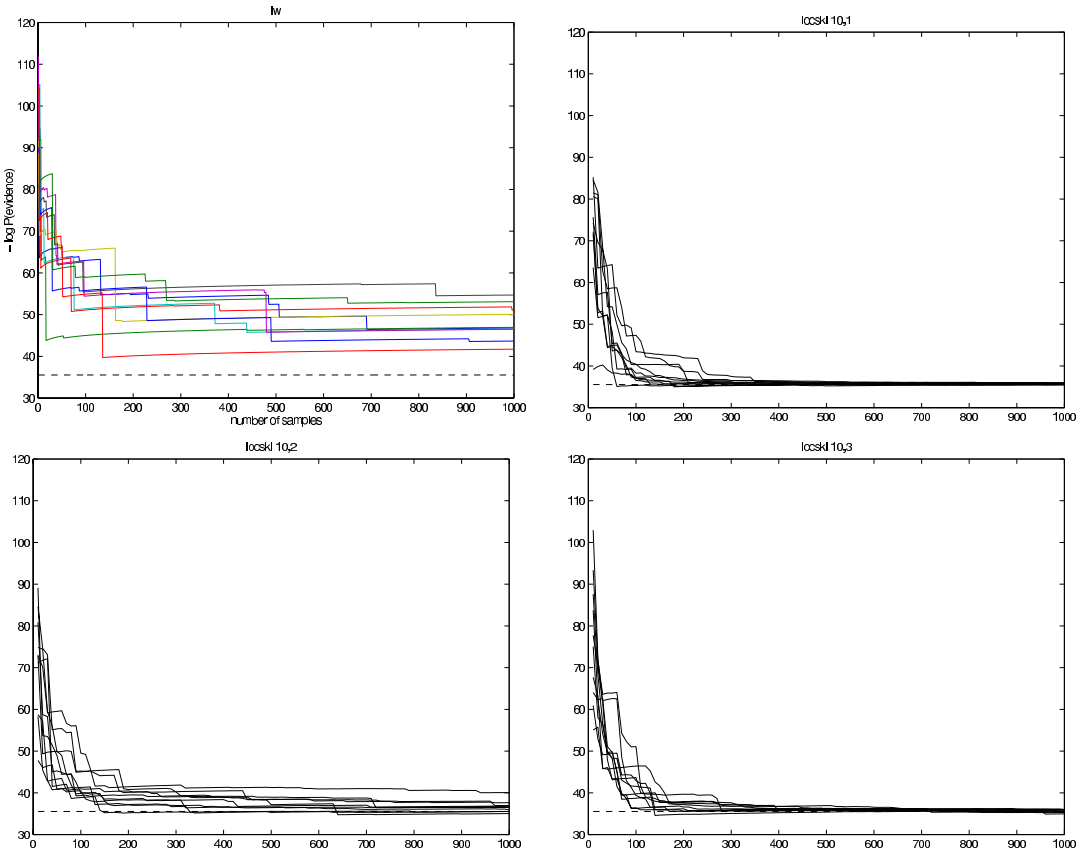


Figure C.26: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_s}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 10. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.

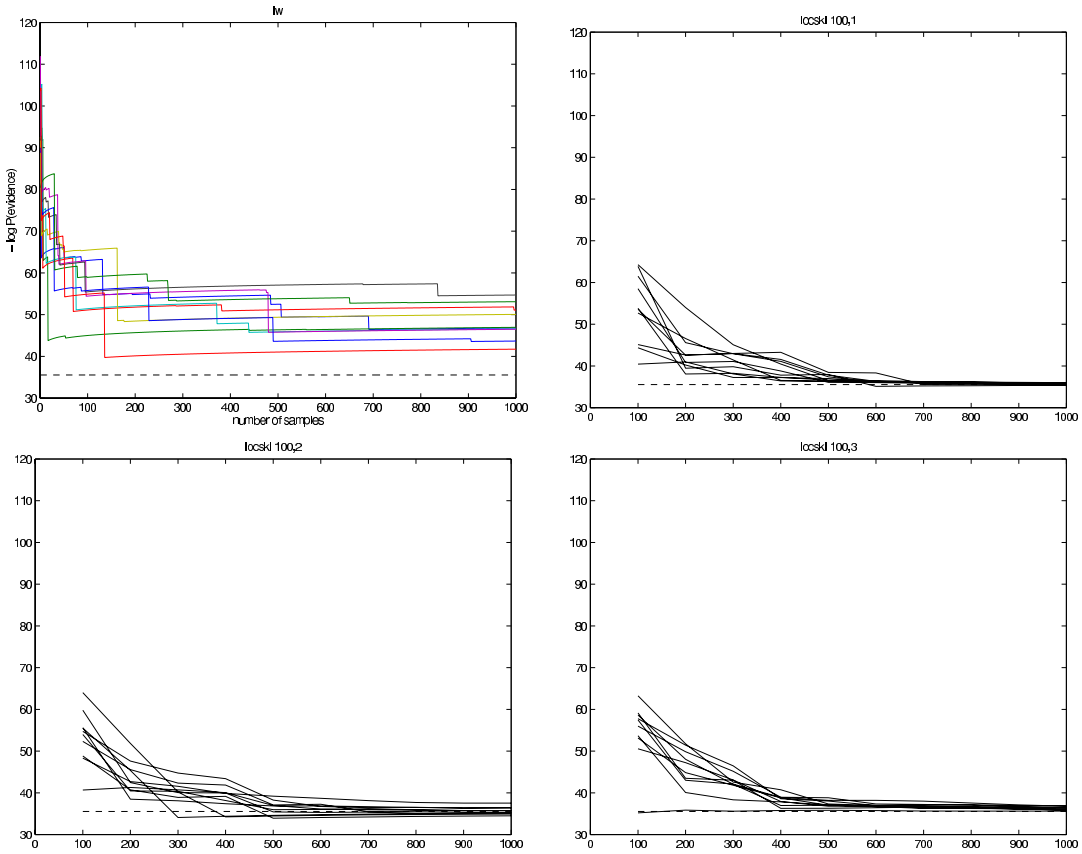


Figure C.27: Results for AIS method based on minimizing $e^{\text{loc}}_{\text{KL}_s}$ for estimating the probability of a random evidence in the synthetic QMR-DT model. The number of samples per stage $N(t)$ was set to 100. The top-right, bottom-left, and bottom-right graphs correspond to setting of β equal to 1, 0.1, and 10, respectively. Refer to the text for other basic general descriptions.