MACHINE INTERVAL EXPERIMENTS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Raazesh Sainudiin

May 2005

MACHINE INTERVAL EXPERIMENTS

Raazesh Sainudiin, Ph.D.

Cornell University 2005

A statistical experiment is a mathematical object that provides a framework for statistical inference, including hypothesis testing and parameter estimation, from observations of an empirical phenomenon. When observations in the continuum of real numbers are not empirically measurable to infinite precision and when conventional floating-point computations used in the inference procedure are not exact, the statistical experiment can become epistemologically invalid. The family of measures of the conventional statistical experiment indexed by a compact finite-dimensional continuum is extended to the complete metric space of all compact subsets (of a certain form) of the index set. This is accomplished by the natural interval extension of the likelihood function. The extended experiment allows a statistical decision made with the aid of a computer to be equivalent to a numerical proof of its global optimality. Three open problems in computational statistics were solved using the extended experiment: (1) parametric bootstraps of likelihood ratio test statistics for finite mixture models, (2) rigorous maximum likelihood estimates of the branch lengths of a phylogenetic tree with a fixed topology or shape and (3) Monte Carlo sampling from a multi-modal target density with sharp peaks or witches' hats.

BIOGRAPHICAL SKETCH

Raaz went from one world to another at an impressionable age. He experientially realized that peoples co-believe what ever they wanted to and was deeply troubled by the relativity of locally co-fabricated cosmologies. He looked for epistemological answers in the fields of mathematics, philosophy and religion. He had to drop out and wander the western lands until the platforms for the troubles vanished by themselves. After a network of intense interactions with other earthlings he studied by a seven mile creek that flowed on the blue earth along a river valley near the ecotone of prairies and deciduous woods in the the land of 10,000 lakes. Then he studied more by the southern tip of a long and narrow lake fingered by a recent glacier. He is currently documenting some statistical ideas that have been haphazardly integrated over some recurring thoughts that bounce back and forth through a wavy transect of the twining vines of minds he tries to sustain.

To the twining vines of minds within the geodermodynamically coupled,

DNA-based, heliotrophic, physicochemical continuum in space-time

ACKNOWLEDGEMENTS

This work was supported by the NSF grant DMS-03-06497 to Michael Nussbaum and a joint NSF/NIGMS grant DMS-02-01037 to Durrett, Aquadro, and Nielsen.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Experiments with Machines

1.1 Epistemological Considerations for Experiments

Definition 1 Epistemology *is the study of the nature and grounds of knowledge especially with reference to its limits and validity.*

The epistemological exploration undertaken here is restricted to the domains of the English empirical tradition of Locke, Berkeley, and Hume. The restriction to English empiricism is itself within a frame of reference that includes the language games of Wittgenstein [41], the non-Aristotelean logical traditions of Nāgārjuna's *Mūlamadhyamakakārikā* [21] and Dogen's *Mountains and Waters Sutra* [9], as well as remnants of primarily oral traditions in contemporary American Prairies (personal communications).

Statement 1 *"But where different effects have been found to follow from causes, which are to appearance exactly similar, all these various effects must occur to the mind in transferring the past to the future, and enter into our consideration, when we determine the probability of the event."[17]*

Definition 2 *A* **statistical experiment** $\mathcal{E}_{\mathcal{P}}$ *is the triple* $(\mathbf{X}, \mathcal{F}_{\mathbf{X}}, \mathcal{P})$ *consisting of a sample space* $\mathbf{X}$ *of all possible empirically observable realizations of a natural phenomenon* $\mathbf{\Phi}$*, a sigma-algebra* $\mathcal{F}_{\mathbf{X}}$ *on* $\mathbf{X}$*, and a family of probability measures* $\mathcal{P} = \{P_\theta, \theta \in \mathbf{\Theta}\}$*, where each* P_θ *is a probability measure on the measurable space* $(\mathbf{X}, \mathcal{F}_{\mathbf{X}})$*. The* θ *is an index belonging to the index set* $\mathbf{\Theta}$*. The index map* $d(\theta) = P_\theta : \mathbf{\Theta} \rightarrow \mathcal{P}$ *associates every* $\theta \in \mathbf{\Theta}$ *with* $P_\theta \in \mathcal{P}$*, in an arbitrary manner that even allows for the index map* d *to be the identity map with* $\mathbf{\Theta} = \mathcal{P}$*.*

1.1.1 2 objectives from 3 perspectives and the likelihood

One of the objectives of a statistical experiment is (1) to infer the "true" $\theta \in \mathbf{\Theta}$ with some notion of confidence or (2) to infer a distribution (posterior) over $\mathbf{\Theta}$ that incorporates one's prior knowledge or belief about the empirical phenomenon which is itself expressed through another distribution (prior) over $\mathbf{\Theta}$ or (3) to infer the "closest" θ to the "true" ϑ that may not necessarily belong to $\mathbf{\Theta}$. The three approaches to attain the same objective arise from the different perspectives of some contemporary schools, namely frequentist, Bayesian, and information-theoretic, respectively. The above objective itself may be seen as a parameter estimation problem of a more general framework involving experiments known as decision theory. Another statistical objective of scientific interest is known as hypothesis testing. Hypothesis testing from a frequentist perspective allows for the testing of falsifiable hypotheses. Falsifiable hypotheses demarcate scientific ones from the space of all posable hypotheses. Let us note that there are analogs of hypothesis testing from the Bayesian perspective and from the third perspective that relies on information theory. Hypothesis testing may also be seen as a specific decision theoretic problem. The two statistical objectives from all three perspectives rely on a key concept known as the likelihood and its various derivatives.

Definition 3 *Given an empirically observed $x \in \mathbf{X}$ of an experiment $\mathcal{E}_{\mathcal{P}}$, the* **likelihood** *is a map $\ell(\theta, x) : \mathbf{\Theta} \times \mathbf{X} \to \mathbb{R}$ whose image is proportional to the conditional probability density f of observing x given θ, i.e. $\ell(\theta) \propto f(x|\theta)$. We naturally assume that f is dominated by a σ-finite measure λ, i.e. $f \ll \lambda$.*

1.1.2 Limits on Empirical Resolution (LER)

The empirically observed x is typically supposed to be a point in the sample space $\mathbf{X}$. Thus, an experimenter explicitly or implicitly assumes that the empirical phenomenon $\mathbf{\Phi}$ is observable with infinite precision. Clearly, this assumption need not make the grounds of knowledge about $\mathbf{\Phi}$ epistemologically invalid, provided that the cardinality of $\mathbf{X}$ is finite, i.e. $|\mathbf{X}| < \infty$. However, if $\mathbf{X}$ is an uncountably infinite set, such as the real line $\mathbb{R}$, then inferring θ, for instance, under the assumption that $\mathbf{\Phi}$ with realizations in the continuum is observable with infinite precision may be epistemologically invalid. Statistical decisions made with such epistemologically invalid experiments that do not account for the fundamental physical limits on empirical resolution (LER) can have undesired consequences. The extent to which the consequences are undesirable will depend on the inherent complexity of the probability model providing us with the index map d. Complexity of the index map is mostly meant with reference to Statement 1. Some modern mathematical concepts that make the former reference concrete include; chaos defined in terms of sensitivity to initial conditions, nonlinear dynamical systems that interact across multiple time-scales, and stochastic resonance at the boundaries of phase transitions within our index set $\mathbf{\Theta}$. In order to make the ground of knowledge about such $\mathbf{\Phi}$ with LER epistemologically sound, the empirically indiscernible sets, rather than the infinitely precise points, must be allowed to enter the statistical experiment as data.

Statement 1 succinctly summarizes the fundamental motivation behind the concern for the range of repercussions that result from real-world executions of statistical decisions obtained from experiments that do not account for the physical LER. Therefore one needs to formalize *empirical sufficiency* or its lack thereof

by rigorously accounting for the physical limits on empirical resolution in our statistical experiments. This is a concern similar to inadequate sample size, but apart from and in addition to it. We will revisit this concern and formally account for it in Section 1.3.

1.1.3 Limits on Numerical Resolution (LNR)

When parameter inference is conducted in a maximum likelihood framework from a frequentist perspective, one is interested in the global maximum of the likelihood function over the parameter space. Explicit analytical solutions for the maximum likelihood estimates are typically difficult to obtain. In practice one settles for a local optimization algorithm to numerically approximate the global solution using a computing machine. However, statistical inference procedures that rely on having found some global optimum through any numerical approach may suffer from at least five major sources of errors. To fully appreciate the sources of errors one needs some understanding of a number screen. Computers only support a finite set of numbers that are usually represented in a semi-logarithmic manner as a set of fixed length floating-point numbers of the form

$$x = \pm m \cdot b^e = \pm 0.m_1 m_2 \cdots m_p \cdot b^e$$

where, m is the signed mantissa of precision p, b is the base (usually 2) and e, bounded between $\underline{e}$ and $\overline{e}$, is the exponent. When $b = 2$, the digits of the mantissa $m_1 = 1$ and $m_i \in \{0, 1\}, \forall\, i, 1 < i \leq p$, and the smallest and largest machine-representable numbers in absolute value are $\underline{x} = 0.10 \cdots 0 \cdot 2^{\underline{e}}$ and $\overline{x} = 0.11 \cdots 1 \cdot 2^{\overline{e}}$, respectively [34]. Thus, the binary floating-point system $\mathcal{R} = \mathcal{R}(2, p, \underline{e}, \overline{e})$ is said to form a screen of the real numbers in the interval $[-\overline{x}, +\overline{x}]$ with 0 uniquely represented by $0.00 \cdots 0 \cdot 2^{\underline{e}}$.

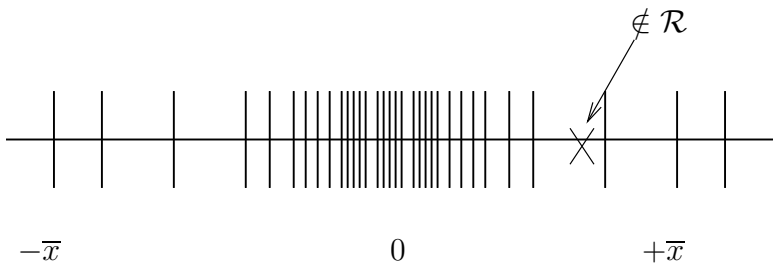


Figure 1.1: A Number Screen $\mathcal{R}$ for the interval $[-\overline{x}, +\overline{x}] \subset \mathbb{R}$

Arithmetic on a machine is performed with such a screen and thus cannot be exact. The computed result of an arithmetic operation is only an approximation to the true result, and this difference between the computed and the actual result is known as the *roundoff error*. Such errors can accumulate catastrophically in a long sequence of imprecise operations [7, 25]. The next major source of error comes from truncating various mathematical expressions, such as, derivatives, integrals, and transcendental functions, that are only defined in terms of the limit of some infinite sequence of operations. Since, only finitely may operations can be performed on a machine, the actual limit is only approximated, and the difference between the actual limit and the computed approximation to it is known as the *truncation error*. Another source of error arises in the *conversion* of constants represented in the decimal format to a binary floating-point format used in a machine. For example, the real number 0.1 has $0.000\overline{1100}_2$ as its unique, but infinite, binary representation. Thus, 0.1 does not have an exact binary representation in any binary floating-point system with a fixed length mantissa, and the truncation of the mantissa to any finite length, in order to keep 0.1 from slipping through the screen $\mathcal{R}$, results in a *conversion error*. Errors in statistical decision may also result from an *ill-posed statistical experiment or model*. For instance, when one has not proved that the model is identifiable, there may exist unknown non-identifiable

subspaces in $\boldsymbol{\Theta}$ that need to be rectified. Finally, experimental procedures are hardly immune to the physical limits on empirical resolution as discussed in Section 1.1.2. Therefore, data as a collection of sets in $\mathbf{X}$, each of which contains *empirically indiscernible points*, cannot always be ignored and substituted with data as a collection of points in $\mathbf{X}$. A *numerically rigorous inference procedure* accounts for all these sources of errors.

Furthermore, traditional nonlinear programming techniques that use local information, such as, clustering methods, generalized descent methods, and other stochastic search methods, including simulated annealing, start from some approximate trial point(s) and iterate by sampling only finitely many points. Therefore, they can neither validate that the objective function has not plunged between the sampled points, nor guarantee escape from a local minimum, albeit they can be made to increase the probability of such desired events. Figure 1.2 shows the trajectories (shaded circles) of two local searches with random initial conditions (white circles) that are attracted to some fixed point (black circle) in a local valley on two different functions. The first function has a sharp valley that is not visited by the local search trajectories. The second function is highly multi-modal with different basins of attraction for each valley. In both cases, the global minimum is missed by the search. Methods that use local information at finitely many points and do not account for all five major sources of errors, cannot be expected to yield anything more than an approximate solution.

There are real-world examples of non-rigorous numerical methods leading to undesired consequences [18, 19]. Numerically-based statistical inference procedures may lead to undesired consequences, especially in parameter-rich models that have not been shown to be identifiable or in models with multi-modal likelihood func-

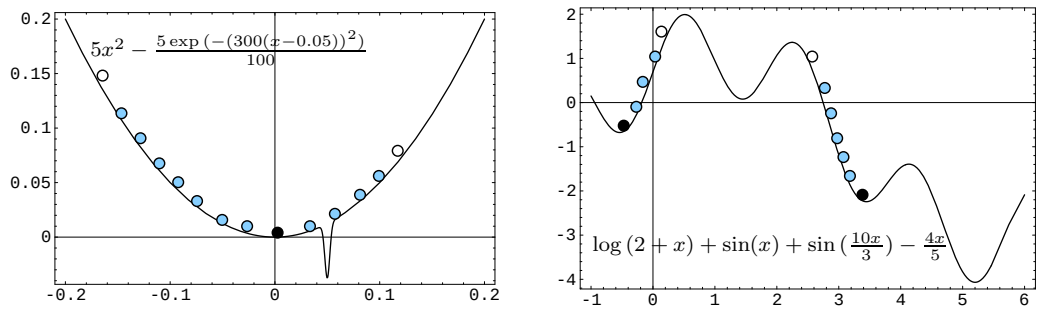


Figure 1.2: Two randomly initialized heuristic searches on two functions.

tions. In some problems, a local approximate solution may be sufficient, but in others one may base statistical decisions that address a real biological problem on an incorrect solution. Therefore we want a statistical experiment tailored for a conventional computing machine that can accommodate numerically rigorous inference procedures. After a brief introduction to interval methods in Section 1.2 we will see in Section 1.3 that such an experiment indeed exists and contains the usual experiment as its special case. In Chapter 2 we apply such ideas to solve some open problems in statistics.

1.1.4 Epistemologically valid experiments

We want an epistemologically valid statistical experiment that accounts for the physical limits on empirical and numerical resolutions. Such an experiment would formalize notions such as empirical sufficiency and at least in principle guarantee the optimality of decisions under a maximum likelihood framework. It would also be more robust to the "curse of locality" that arise in traditional nonlinear optimization and Monte Carlo sampling techniques that are limited to local information. Before we formalize such an experiment we need an introduction to enclosure methods.

1.2 Introduction to Enclosure Methods

The following sections contain a brief introduction to various enclosure arithmetics. For a recent introduction to such arithmetics see [12, 23, 40]. Section 1.2.1 describes the features of intervals, defines an arithmetic on them, and introduces interval analysis. The definitions and theorems in this section are elementary and necessary to appreciate the basic ideas in the sequel. We only prove the most fundamental theorems and merely state others with references in the literature for their proofs. Next we introduce differentiation arithmetic or automatic differentiation. We will need this technology to extend the Newton's method and to rigorously enclose the global maximum of the likelihood function, the subjects of the following two sections. The basic global optimization algorithm based on Hansen's method [13, 14] with Ratz's modifications [37] as implemented in [12], with further extensions that account for non-stationary maxima at the boundaries and increase computational efficiency, is outlined in section 1.2.5.

1.2.1 Interval analysis

Lower case letters denote real numbers, e.g. $x \in \mathbb{R}$, the set of real numbers. Upper case letters represent bounded and closed (compact) *real intervals*:

$$X = [\underline{x}, \overline{x}] = [\inf(X), \sup(X)].$$

Any such compact interval,

$$X \in \mathbb{IR} := \{[a, b] : a \leq b,\ a, b \in \mathbb{R}\},$$

where $\mathbb{IR}$ is the set of all compact real intervals. The *radius*, *diameter* and the *midpoint* of X are given below.

$$\begin{array}{lrcl} \textit{radius} & r(X) & := & (\underline{x} - \overline{x})/2 \\ \textit{diameter} & d(X) & := & \overline{x} - \underline{x} \\ \textit{midpoint} & m(X) & := & (\underline{x} + \overline{x})/2. \end{array}$$

The *smallest* and *largest absolute value* of an interval X are real numbers known as the *mignitude* and *magnitude*, respectively, and the *absolute value* of an interval X are given below.

$$\begin{array}{lrcl} \textit{mignitude} & \langle X \rangle & := & \min\{|x| : x \in X\} = \min\{|\underline{x}|, |\overline{x}|\} \\ \textit{magnitude} & |X| & := & \max\{|x| : x \in X\} = \max\{|\underline{x}|, |\overline{x}|\} \\ \textit{absolute value} & |X|_{[\ ]} & := & \{|x| : x \in X\} = [\langle X \rangle, |X|]. \end{array}$$

The *relative diameter* of an interval X, denoted by d_{rel} is the diameter $d(X)$ itself if $0 \in X$, and $d(X)/\langle X \rangle$, otherwise. An interval X with zero diameter is called a *thin interval* with,

$$\underline{x} = \overline{x} = x.$$

Containment of thin intervals by $\mathbb{IR}$ implies $\mathbb{R} \subseteq \mathbb{IR}$. Elements of $\mathbb{IR}$ inherit various set relations, whose names, notation and definition are as follows:

$$\begin{array}{lrcl} \textit{Equal} & X = Y & \Leftrightarrow & \underline{x} = \underline{y} \text{ and } \overline{x} = \overline{y} \\ \textit{Subset} & X \subseteq Y & \Leftrightarrow & \underline{x} \geq \underline{y} \text{ and } \overline{x} \leq \overline{y} \\ \textit{Completesubset} & X \Subset Y & \Leftrightarrow & \underline{x} > \underline{y} \text{ and } \overline{x} < \overline{y} \end{array}$$

Definition 4 (Partial ordering) *A relation $\sim$ is a partial order on a set S if,*

for all $x, y, z \in S$, it satisfies the following three properties:

$$\begin{aligned}
&(1).\quad \textit{Reflexivity}: && x \sim x \\
&(2).\quad \textit{Antisymmetry}: && (x \sim y)\ \textit{and}\ (y \sim x) \implies x = y \\
&(3).\quad \textit{Transitivity}: && (x \sim y)\ \textit{and}\ (y \sim z) \implies x \sim z
\end{aligned}$$

A set $\mathbf{X}$ w.r.t. a partial ordering $\sim$ is called a *lattice* and is denoted by $\mathbf{X}_{\sim}$. A lattice $\mathbf{X}_{\sim}$ is said to be *complete* if it is closed under the the partial ordering, i.e., the infimum and supremum of any subset of $\mathbf{X}_{\sim}$ w.r.t. $\sim$ exist in $\mathbf{X}_{\sim}$. Note that any $Y \in \mathbb{IR}$ is a complete lattice w.r.t. the two partial orderings:

$$\begin{aligned}
&(A).\quad \textit{Lesser or equal} \quad \leq\!: && X \leq Y \ \Leftrightarrow\ \left(\underline{x} \leq \underline{y}\right) \text{ and } \left(\overline{x} \leq \overline{y}\right), \\
&(B).\quad \textit{Set inclusion} \quad \subseteq\!: && X \subseteq Y \ \Leftrightarrow\ X \subseteq Y.
\end{aligned}$$

These two lattices are denoted by $\mathbb{IR}_{\leq}$ and $\mathbb{IR}_{\subseteq}$, respectively.

Next we look at two set operations in $\mathbb{IR}$. We can make the union operation closed in $\mathbb{IR}$ through the notion of *hull* and make the intersection operation well-defined by adding the empty interval $[\,]$ to $\mathbb{IR}$, as follows:

$$\begin{aligned}
&\textit{Hull} && X \underline{\cup} Y \ := && [\min\{\underline{x}, \underline{y}\}, \min\{\overline{x}, \overline{y}\}] \\
&\textit{Intersection} && X \cap Y \ := && \begin{cases} [\,] & : \text{ if } \overline{x} < \underline{y} \text{ or } \overline{y} < \underline{x}, \\ [max\{\underline{x}, \underline{y}\}, min\{\overline{x}, \overline{y}\}] & : otherwise. \end{cases}
\end{aligned}$$

No notational distinction is made between a real number $x \in \mathbb{R}$ and a vector $x = (x_i, \cdots, x_n)^T \in \mathbb{R}^n$ and between an interval X and an *interval vector* or *box*,

$$X = (X_1, \cdots, X_n)^T \in \mathbb{IR}^n \Leftrightarrow X_i = [\underline{x_i}, \overline{x_i}] = [\inf(X_i), \sup(X_i)] \in \mathbb{IR},\ i = 1, \ldots, n.$$

The dimension n should be clear from the context. For an interval vector X, the radius, diameter, relative diameter, midpoint, and hull operations are defined component-wise to yield vectors, while the maximum over its components is taken

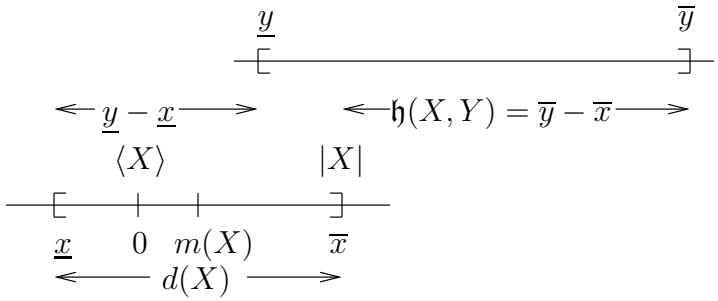


Figure 1.3: Features of intervals

to obtain the maximal diameter and the maximal relative diameter, $d_\infty(X) = \max_i d(X_i)$ and $d_{rel,\infty}(X) = \max_i d_{rel}(X_i)$, respectively.

It can be seen that $\mathbb{IR}$ under the metric $\mathfrak{h}$, given by,

$$\mathfrak{h}(X,Y) := \max\{|\underline{x} - \underline{y}|, |\overline{x} - \overline{y}|\},$$

is a *complete metric space*. Convergence of a sequence of intervals $\{X^{(i)}\}$ to an interval X under the metric $\mathfrak{h}$ is equivalent to the sequence $\mathfrak{h}(X^{(i)}, X)$ approaching 0 as i approaches ∞, which in turn is equivalent to both $\underline{x}^{(i)} \to \underline{x}$ and $\overline{x}^{(i)} \to \overline{x}$, i.e.,

$$\begin{aligned} \lim_{i\to\infty} X^{(i)} &\Leftrightarrow \lim_{i\to\infty} \mathfrak{h}(X^{(i)}, X) = 0 \\ &\Leftrightarrow \left(\lim_{i\to\infty} \underline{x}^{(i)} = \underline{x}\right) \text{ and } \left(\lim_{i\to\infty} \overline{x}^{(i)} = \overline{x}\right) \text{ and } \left(\forall i, \quad \underline{x}^{(i)} \le \overline{x}^{(i)}\right) \end{aligned}$$

Continuity and differentiability of a function $F : \mathbb{IR}^n \to \mathbb{IR}^k$ are defined in the usual way. The Figure 1.3 illustrates some of the features of intervals.

Arithmetic on intervals in $\mathbb{IR}$

Definition 5 (Interval arithmetic) *If the binary operator $\star$ is one of the elementary arithmetic operations $\{+, -, \cdot, /\}$, then we define an arithmetic on operands in $\mathbb{IR}$ by*

$$X \star Y := \{x \star y : x \in X, y \in Y\}$$

with the exception that X/Y is undefined if $0 \in Y$.

Theorem 1 *Arithmetic on the pair $X, Y \in \mathbb{IR}$ is given by:*

$$\begin{aligned} X + Y &= [\underline{x} + \underline{y}, \overline{x} + \overline{y}] \\ X - Y &= [\underline{x} - \overline{y}, \overline{x} - \underline{y}] \\ X \cdot Y &= [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}], \\ X/Y &= X \cdot [1/\overline{y}, 1/\underline{y}], \ \textit{provided, } 0 \notin Y. \end{aligned}$$

Proof (cf. [12, 40]): Since any real arithmetic operation $x \star y$, where $\star \in \{+, -, \cdot, /\}$ and $x, y \in \mathbb{R}$, is a continuous function $x \star y := \star(x, y) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, except when $y = 0$ under $/$ operation. Since X and Y are simply connected compact intervals, so is their product $X \times Y$. On such a domain $X \times Y$, the continuity of $\star(x, y)$ (except when $\star = /$ and $0 \in Y$) ensures the attainment of a minimum, a maximum and all intermediate values. Therefore, with the exception of the case when $\star = /$ and $0 \in Y$, the range $X \star Y$ has an interval form $[\min(x \star y), \max(x \star y)]$, where the min and max are taken over all pairs $(x, y) \in X \times Y$. Fortunately, we do not have to evaluate $x \star y$ over every $(x, y) \in X \times Y$ to find the global min and global max of $\star(x, y)$ over $X \times Y$, because the monotonicity of the $\star(x, y^*)$ in terms of $x \in X$ for any fixed $y^* \in Y$ implies that the extremal values are attained on the boundary of $X \times Y$, i.e., the set $\{\underline{x}, \underline{y}, \overline{x}$, and $\overline{y}\}$. Thus the theorem can be verified by examining the finitely many boundary cases. □

Properties of interval arithmetic

It is clear from Theorem 1 that $\mathbb{IR}$, in spite of being a small subset of the power set of $\mathbb{R}$, is closed under the four well-defined elementary operations. The identity elements of $+$ and $\cdot$ are the thin intervals $[0, 0]$ and $[1, 1]$, respectively. Multiplicative and additive inverses do not exist except when X is also thin, since $[0, 0] \subseteq X - X$,

and $[1,1] \subseteq X/X$. Although the commutative and associative laws are satisfied by $+$ and $\cdot$, only a weaker notion of distributivity called *sub-distributivity* is satisfied, i.e.,

$$X \cdot (Y + Z) \subseteq (X \cdot Y) + (X \cdot Z).$$

An extremely useful property of interval arithmetic that is a direct consequence of Definition 5 is summarized by the following theorem.

Theorem 2 (Fundamental property of interval arithmetic) *If $X \subseteq X'$ and $Y \subseteq Y'$ and $\star \in \{+, -, \cdot, /\}$, then*

$$X \star Y \subseteq X' \star Y',$$

where we require that $0 \notin Y'$ when $\star = /$.

Proof:

$$X \star Y = \{x \star y : x \in X, y \in Y\} \subseteq \{x \star y : x \in X', y \in Y'\} = X' \star Y'. \square$$

Note that an immediate implication of Theorem 2 is that when $X = x$ and $Y = y$ are thin intervals (real numbers x and y), then $X' \star Y'$ will contain the result of the real arithmetic operation $x \star y$.

Definition 6 (Range) *Consider a real-valued function $f : D \to \mathbb{R}$ where the domain $D \subseteq \mathbb{R}^n$. The range of f over any $E \subseteq D$ is represented by $Rng(f; E)$ and defined to be the set*

$$Rng(f; E) := \{f(x) : x \in E\}$$

However, when the range of f over any $X \in \mathbb{IR}^n$ such that $X \subseteq D$ is of interest, we will use the short-hand $f(X)$ for $Rng(f; X)$.

Definition 7 (Interval extension of subsets of $\mathbb{R}^n$) *For any Euclidean subset $\mathbf{\Theta} \subseteq \mathbb{R}^n$ let us denote its interval extension by $\mathbb{I}\mathbf{\Theta}$ and define it to be the set*

$$\mathbb{I}\mathbf{\Theta} := \{X \in \mathbb{IR}^n : \underline{x}, \overline{x} \in \mathbf{\Theta}\}$$

Definition 8 (Inclusion isotony) *An box-valued map $F : D \to \mathbb{IR}^m$, where $D \in \mathbb{IR}^n$, is inclusion isotonic if it satisfies the property*

$$\forall\, X \subseteq Y \subseteq D \implies F(X) \subseteq F(Y).$$

Definition 9 (The natural interval extension) *Consider a real-valued function $f : D \to \mathbb{R}$ where the domain $D \in \mathbb{IR}^n$. If real constants, variables, and operations in f are replaced by their interval counterparts, then one obtains*

$$F(X) : \mathbb{I}D \to \mathbb{IR}.$$

F is known as the natural interval extension of f.

Theorem 3 (Inclusion isotony of rational functions) *Consider the rational function $f(x) = p(x)/q(x)$, where p and q are polynomials. Let F be its natural interval extension such that $F(Y)$ is well-defined for some $Y \in \mathbb{IR}$ and let $X, X' \in \mathbb{IR}$. Then we have*

(*i*) Inclusion isotony: $\forall\, X \subseteq X' \subseteq Y \implies F(X) \subseteq F(X')$, and

(*ii*) Range enclosure: $\forall\, X \subseteq Y \implies Rng(f; X) = f(X) \subseteq F(X)$.

Proof (cf. [40]): Since $F(Y)$ is well-defined, we will not run into division by zero, and therefore (i) follows from the repeated invocation of Theorem 2. We can prove (ii) by contradiction. Suppose $Rng(f; X) \nsubseteq F(X)$. Then there exists $x \in X$, such that $f(x) \in Rng(f; X)$ but $f(x) \notin F(X)$. This in turn implies that $f(x) = F([x, x]) \notin F(X)$, which contradicts (i). Therefore, our supposition cannot be true and we have proved (ii) $Rng(f; X) \subseteq F(X)$. $\square$

Definition 10 (Standard functions) *Piece-wise monotone functions, including exponential, logarithm, rational power, absolute value, and trigonometric functions, constitute the set of standard functions*

$$\mathfrak{S} = \{\, a^x, log_b(x), x^{p/q}, |x|, \sin(x), \cos(x), \tan(x), \sinh(x), \ldots, \arcsin(x), \ldots \}.$$

Such functions have well-defined interval extensions that satisfy inclusion isotony and *exact range enclosure*, i.e., $Rng(f; X) = f(X) = F(X)$. Consider the following definitions for the interval extensions for some monotone functions in $\mathfrak{S}$ with $X \in \mathbb{IR}$,

$$\begin{aligned}
\exp(X) &= [\exp(\underline{x}), \exp(\overline{x})] \\
\arctan(X) &= [\arctan(\underline{x}), \arctan(\overline{x})] \\
\sqrt{(X)} &= [\sqrt{(\underline{x})}, \sqrt{(\overline{x})}] \qquad \text{if } 0 \leq \underline{x} \\
\log(X) &= [\log(\underline{x}), \log(\overline{x})] \qquad \text{if } 0 < \underline{x}
\end{aligned}$$

and a piece-wise monotone function in $\mathfrak{S}$ with $\mathbb{Z}^+$ and $\mathbb{Z}^-$ representing the set of positive and negative integers, respectively.

$$X^n = \begin{cases}
[\underline{x}^n, \overline{x}^n] & : \text{if } n \in \mathbb{Z}^+ \text{ is odd}, \\
[\langle X \rangle^n, |X|^n] & : \text{if } n \in \mathbb{Z}^+ \text{ is even}, \\
[1, 1] & : \text{if } n = 0, \\
[1/\overline{x}, 1/\underline{x}]^{-n} & : \text{if } n \in \mathbb{Z}^-; 0 \notin X
\end{cases}$$

Definition 11 (Elementary functions) *A real-valued function that can be expressed as a finite combination of constants, variables, arithmetic operations, standard functions and compositions is called an elementary function. The set of all such elementary functions is referred to as* $\mathfrak{E}$.

One can think of the process by which an elementary function f is computed as the result of a sequence of recursive operations with the subexpressions f_i of f

where, $i = 1, \ldots, n < \infty$. This involves the evaluation of the subexpression f_i at node i with operands s_{i_i}, s_{i_2} from the sub-terminal nodes of i given by the *directed acyclic graph* (DAG) for f

$$s_i = \odot f_i := \begin{cases} f_i(s_{i_1}, s_{i_2}) & : \text{if node } i \text{ has 2 sub-terminal nodes } s_{i_1}, s_{i_2} \\ f_i(s_{i_1}) & : \text{if node } i \text{ has 1 sub-terminal node } s_{i_1} \\ I(s_i) & : \text{if node } i \text{ is a leaf or terminal node, I(x) = x.} \end{cases} \tag{1.1}$$

The leaf or terminal node of the DAG is a constant or a variable and thus the f_i for a leaf i is set equal to the respective constant or variable. The recursion starts at the leaves and terminates at the root of the DAG. For example the elementary function $x \cdot \sin((x-3)/3)$ can be obtained from the terminus $\odot f_6$ of the recursion $\{\odot f_i\}_{i=1}^{6}$ on the DAG for f as shown in Figure 1.4. More generally, for an elementary f with n sub-expressions $f_1, f_2, \ldots, f_n$ and the corresponding DAG

$$\{\odot f_i\}_{i=1}^{n} \quad \rightarrowtail \quad \odot f_n = f(x), \tag{1.2}$$

where each $\odot f_i$ is computed according to Equation 1.1 It would be convenient if guaranteed enclosures of the range $f(X)$ of an elementary f can be obtained by its natural interval extension $F(X)$. We show that inclusion isotony does indeed hold for F, i.e. if $X \subseteq Y$, then $F(X) \subseteq F(Y)$, and in particular, the *inclusion property* that $x \in X \implies f(x) \in F(X)$ does hold.

Theorem 4 (The fundamental theorem of interval analysis) *Consider any elementary function $f \in \mathfrak{E}$. Let $F : Y \rightarrow \mathbb{IR}$ be its natural interval extension such*

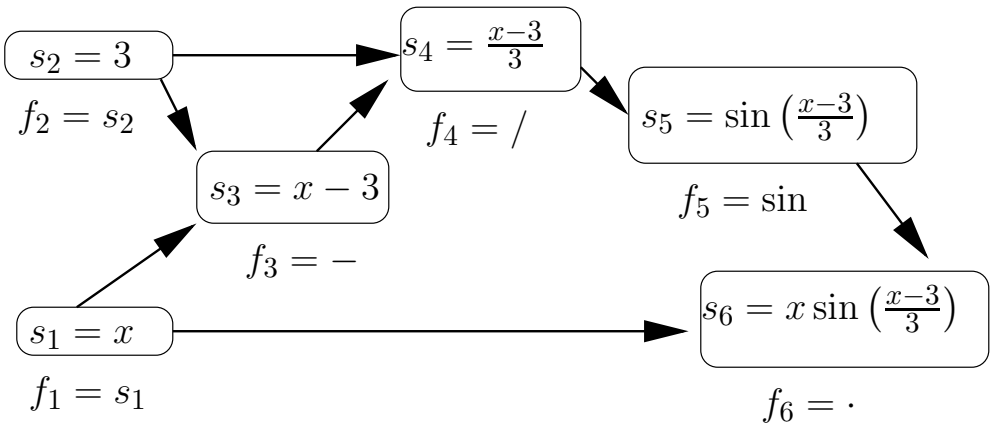


Figure 1.4: Recursive evaluation of the sub-expressions $f_1, \ldots, f_6$ on the DAG of the elementary function $f(x) = \odot f_6 = x \cdot \sin((x-3)/3)$

that $F(Y)$ is well-defined for some $Y \in \mathbb{IR}$ and let $X, X' \in \mathbb{IR}$. Then we have

(i) Inclusion isotony: $\forall\, X \subseteq X' \subseteq Y \implies F(X) \subseteq F(X')$, and

(ii) Range enclosure: $\forall\, X \subseteq Y \implies Rng(f; X) = f(X) \subseteq F(X)$.

Proof (cf. [40]): Any elementary function $f \in \mathfrak{E}$ is defined by the recursion 1.2 on its sub-expressions f_i where $i \in \{1, \ldots, n\}$ according to its DAG. If $f(x) = p(x)/q(x)$ is a rational function, then the theorem already holds by Theorem 3, and if $f \in \mathfrak{S}$ then the theorem holds because the range enclosure is exact for standard functions. Thus it suffices to show that if the theorem holds for $f_1, f_2 \in \mathfrak{E}$, then the theorem also holds for $f_1 \star f_2$, where $\star \in \{+, -, /, \cdot, \circ\}$. By $\circ$ we mean the composition operator. Since the proof is analogous for all five operators, we only focus on the $\circ$ operator. Since F is well-defined on its domain Y, neither the real-valued f nor any of its sub-expressions f_i have singularities in its respective domain Y_i induced by Y. In particular f_2 is continuous on any X_2 and X_2' such that $X_2 \subseteq X_2' \subseteq Y_2$ implying the compactness of $F_2(X_2) =: W_2$ and $F_2(X_2') =: W_2'$, respectively. By our assumption that F_1 and F_2 are inclusion isotonic we have that

$W_2 \subseteq W_2'$ and also that

$$F_1 \circ F_2(X_2) = F_1(F_2(X_2)) = F_1(W_2) \subset F_1(W_2') = F_1(F_2(X_2')) = F_1 \circ F_2(X_2)$$

The range enclosure is a consequence of inclusion isotony by an argument identical to that given in the proof for Theorem 3. □

The fundamental implication of the above theorem is that it allows us to enclose the range of any elementary function and thereby produces an upper bound for the global maximum and a lower bound for the global minimum over any compact subset of the domain upon which the function is well-defined. We will see in the sequel that this is the work-horse of randomized enclosure algorithms that efficiently produce samples even from highly multi-modal target distributions. The contra-positive of the second statement of Theorem 4 states that for any $f \in \mathfrak{E}$ with a well-defined interval extension F on $Y \supseteq X$

$$y \notin F(X) \implies y \notin Rng(f; X).$$

We will see in the sequel that the above contra-positive is the work-horse of rigorous global optimization with interval analysis on computing machines when further augmented by rounding-controlled numerics.

Unfortunately, some interval extensions of $f \in \mathfrak{E}$ are better at enclosing the true range than others. Although the three functions shown in Figure 1.5 are equivalent, their interval extensions yield different range enclosures. The interval extension $F^{(3)}$ is better than $F^{(1)}$ and $F^{(2)}$ as depicted in Figure 1.5. Note that $F^{(3)} \subseteq F^{(2)}$ since $X^2 \subseteq X \cdot X$ in interval arithmetic. If X appears only once in the expression and all parameters are thin intervals, then it was shown by [31] that the natural interval extension does indeed yield a tight enclosure. In general, one can obtain tighter enclosures by minimizing the occurrence of X in the expression.

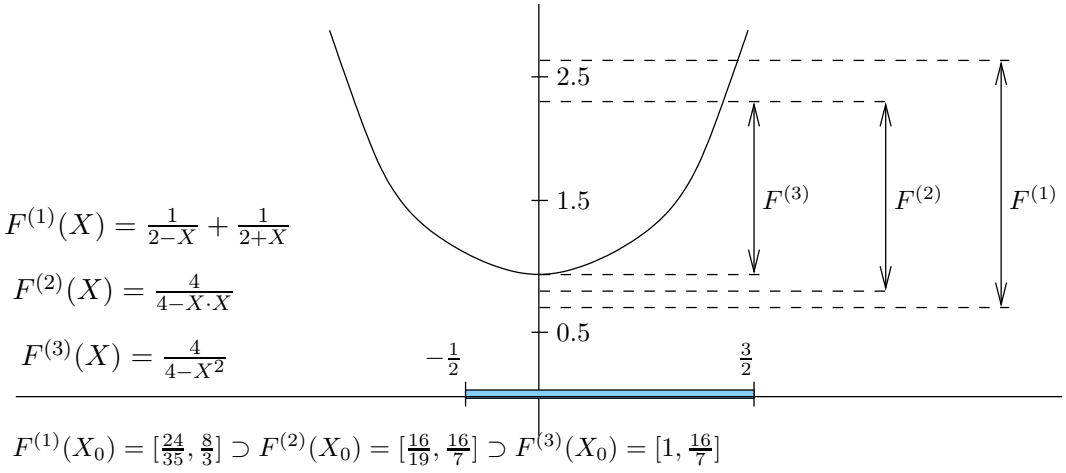


Figure 1.5: Extension-specific dependence of range enclosures

Unlike the natural interval extension of an $f \in \mathfrak{S}$ that produces exact range enclosures, the natural interval extension $F(X)$ of an $f \in \mathfrak{E}$ often overestimates the range $f(X)$, but can be shown under mild conditions to linearly approach the range as the maximal diameter of the box X goes to zero, i.e., $\mathfrak{h}(F(X), f(X)) \leq \alpha \cdot d_\infty(X)$ for some $\alpha \geq 0$. This implies that a partition of X into smaller boxes $\{X^{(1)}, \cdots, X^{(m)}\}$ gives better enclosures of $f(X)$ through the union $\bigcup_{i=1}^{m} F(X^{(i)})$ as illustrated in Figure 1.6. Next we make the above statements precise.

Definition 12 *A function $f : D \to \mathbb{R}$ is Lipschitz if there exists a Lipschitz constant K such that, for all $x, y \in D$, we have $|f(x)-f(y)| \leq K|x-y|$. We define $\mathfrak{E}_{\mathfrak{L}}$ to be the set of elementary functions whose sub-expressions f_i, $i = 1, \ldots, n$ at the nodes of the corresponding DAGs are all Lipschitz.*

Theorem 5 (Range enclosure tightens linearly with mesh) *Consider a function $f : D \to \mathbb{R}$ with $f \in \mathfrak{E}_{\mathfrak{L}}$. Let F be an inclusion isotonic interval extension of f such that $F(X)$ is well-defined for some $X \in \mathbb{IR}, X \subseteq I$. Then there exists a*

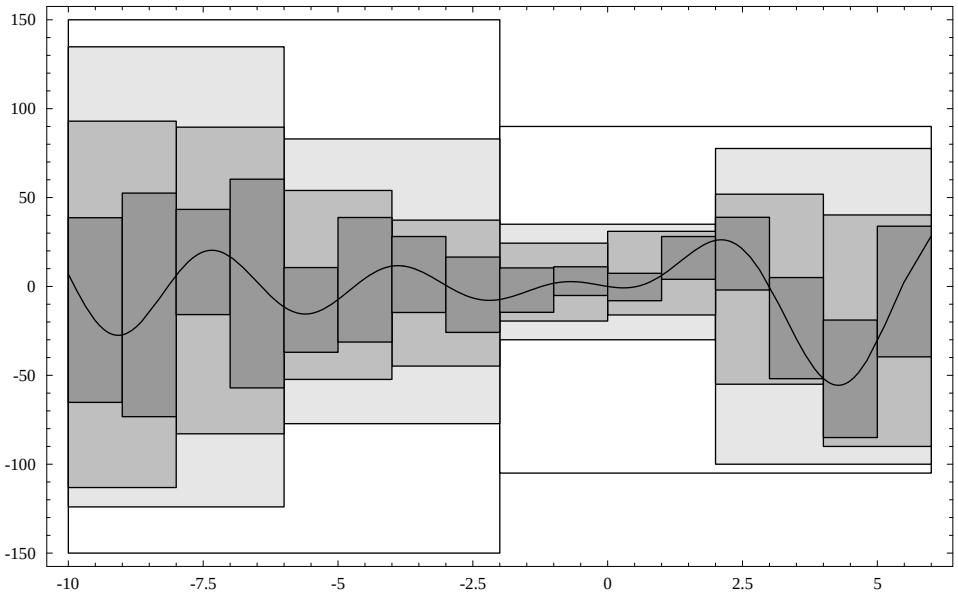


Figure 1.6: Range enclosure of the interval extension of $-\sum_{k=1}^{5} k\,x\,\sin\left(\frac{k(x-3)}{3}\right)$ linearly tightens with the mesh

positive real number K, depending on F and X, such that if $X = \cup_{i=1}^{k} X^{(i)}$, then

$$Rng(f;X) \subseteq \bigcup_{i=1}^{k} F(X^{(i)}) \subseteq F(X)$$

and

$$r\left(\bigcup_{i=1}^{k} F(X^{(i)})\right) \leq r(Rng(f;X)) + K \max_{i=1,\ldots,k} r(X^{(i)})$$

Proof : The proof is given by an induction on the DAG for f similar to the proof of Theorem 4 (See [40]).

Centered forms

Let $\nabla F(x)$ and $\nabla^2 F(x)$ represent the interval extensions of $\nabla f(x)$ and $\nabla^2 f(x)$, the gradient and Hessian of f. A better enclosure of $f(X)$ is possible for an f with the centered form,

$$f(x) = f(c) + \nabla f(b)\cdot(x-c) \in f(c) + \nabla f(X)\cdot(x-c) \subseteq F_c(X) := f(c) + \nabla F(X)\cdot(X-c)\,,$$

where, $b, c, x \in X$ with b between c and x. $F_c(X)$ is the interval extension of the centered form of f with center c and decays quadratically to $f(X)$ as the maximal diameter of $X \to 0$.

1.2.2 Differentiation Arithmetic

When it becomes too cumbersome or impossible to explicitly compute the derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$, or when f itself is only available as an algorithm, one may employ a differentiation arithmetic, often known as automatic differentiation (see for e.g. [36]) to obtain any $\nabla^k f$, the k_{th}-order derivative of f. This approach circumvents the computation of a formal expression for f by defining a differentiation arithmetic on the ordered k-tuples $(f(x), \nabla f(x), \nabla^2 f(x), \cdots, \nabla^k f(x))$ [2]. A brief sketch of such an arithmetic is given for the case when $k = 2$ as it will be used in section 1.2.5.

Consider a twice-continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with the gradient vector and Hessian matrix given by $\nabla f(x) := (\partial f(x)/\partial x_1, \cdots, \partial f(x)/\partial x_n)^T \in \mathbb{R}^n$, and $\nabla^2 f(x) := ((\partial^2 f(x)/\partial x_i \partial x_j))_{i,j=\{1,\cdots,n\}} \in \mathbb{R}^{n\times n}$, respectively. For every, $f(x) : \mathbb{R}^n \to \mathbb{R}$, consider its corresponding ordered triple $(\, f(x),\ \nabla f(x),\ \nabla^2 f(x)\,)$. The ordered triples corresponding to a constant function, $c(x) = c : \mathbb{R}^n \to \mathbb{R}$, and a component identifying function (or variable), $I_j(x) = x_j : \mathbb{R}^n \to \mathbb{R}$, are $(\, c,\ 0,\ 0\,)$ and $(\, x_j,\ e^{(j)},\ 0\,)$, respectively, where, $e^{(j)}$ is the j-th unit vector and the 0's are additive identities in their appropriate spaces. To perform an elementary operation $\star \in \{+, -, \cdot, /\}$ with a pair of such triples to obtain another, the rules of calculus

apply as follows:

$$
\begin{aligned}
&(\, h(x),\ \nabla h(x),\ \nabla^2 h(x)\,) \\
&:= (\, f(x),\ \nabla f(x),\ \nabla^2 f(x)\,) \star (\, g(x),\ \nabla g(x),\ \nabla^2 g(x)\,) \\
&= (\, f(x) \star g(x),\ \nabla f(x) \star \nabla g(x),\ \nabla^2 f(x) \star \nabla^2 g(x)\,), && \text{if } \star \in \{+, -\} \\
&= (\, f(x) \cdot g(x),\ f(x) \cdot \nabla g(x) + g(x) \cdot \nabla f(x), \\
&\qquad g(x) \cdot \nabla^2 f(x) + \nabla f(x) \cdot \nabla g(x)^T + \nabla g(x) \cdot \nabla f(x)^T + f(x) \cdot \nabla^2 g(x)\,), && \text{if } \star = \cdot \\
&= (\, f(x)/g(x),\ 1/g(x) \cdot \{\nabla f(x) - h(x) \cdot \nabla g(x)\}, \\
&\qquad 1/g(x) \cdot \{\nabla^2 f(x) \cdot \nabla h(x) \cdot \nabla h(x)^T - \nabla g(x) \cdot \nabla h(x)^T - h(x) \cdot \nabla^2 h(x)\}, && \text{if } \star = /,
\end{aligned}
$$

where $g(x) \neq 0$ under the division operation. The arithmetic for composition of functions, such as, $h(x) = r(f(x)) : \mathbb{R} \to \mathbb{R}$, with the first and second derivative of r given by r' and r'', on their corresponding triples is given by

$$(\, r(f(x)),\, r'(f(x)) \cdot \nabla f(x),\, r''(f(x)) \cdot \nabla f(x) \cdot \nabla f(x)^T + r'(f(x)) \cdot \nabla^2 f(x)\,).$$

Such compositions are used to obtain the triples for the elementary functions, such as, $\exp(x)$ and $\ln(x)$, which are used in the likelihood computations of section 2.2.2.

For dyadic reasons, the differentiation arithmetic has been explained above only in terms of reals. By replacing the real x's above by interval X's and performing all operations in the real interval arithmetic with the interval extension F of f, as discussed in section 1.2.1, one can rigorously enclose the components of the interval triple $(\, F(X),\ \nabla F(X),\ \nabla^2 F(X)\,)$ through interval differentiation arithmetic, such that, for every $x \in X \in \mathbb{IR}^n$, $f(x) \in F(X) \in \mathbb{IR}$, $\nabla f(x) \in \nabla F(X) \in \mathbb{IR}^n$, and $\nabla^2 f(x) \in \nabla^2 F(X) \in \mathbb{IR}^{n \times n}$.

1.2.3 Interval Newton method and its extension

Newton's method linearly approximates the differentiable real function $f(x)$ in the neighborhood of an initial value $x^{(0)}$ by the tangent,

$$t(x) = f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)}),$$

where, $f'(x)$ is the first derivative of $f(x)$. This tangent equation can be used to solve for an approximation to the zero of $f(x)$, by means of the following discrete dynamical system known as Newton's method:

$$x^{(j+1)} = x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})}, \qquad j = 0, 1, 2, \cdots .$$

Thus, the zero of the tangent to $f(x)$ at the current approximate value $x^{(j)}$ gives the next approximate value $x^{(j+1)}$. The following geometric interpretation shown in Figure 1.7 is useful. During each iteration of the Newton's method, a light beam is shone upon the domain from the point $(x^{(j)}, f(x^{(j)}))$ along the tangent to $f(x)$ at $x^{(j)}$. The intersection of this beam (white line in Figure 1.7) with the domain provides $x^{(j+1)}$, which is where the next iteration is resumed. If x^* is the only root of $f(x)$ in the search interval X containing all iterates $x^{(j)}$, $f(x)$ is twice continuously differentiable, and $x^{(0)}$ is sufficiently close to x^*, then it is well known that Newton's method converges quadratically fast to x^*. Otherwise, it may well diverge or oscillate.

The interval version of Newton's method [30] computes an enclosure of the zero x^* of a continuously differentiable function $f(x)$ in the interval X through the following dynamical system in $\mathbb{IR}$:

$$X^{(j+1)} = \left(m(X^{(j)}) - \frac{f(m(X^{(j)}))}{F'(X^{(j)})} \right) \cap X^{(j)}, \qquad j = 0, 1, 2, \cdots$$

Here, $X^{(0)} = X$, $F'(X^{(j)})$ is the enclosure of $f'(x)$ over $X^{(j)}$, and $m(X^{(j)})$ is the mid-point of $X^{(j)}$. Provided $0 \notin F'(X^{(0)})$ or equivalently a unique zero of f lies in

$X^{(0)}$, the interval Newton method will never diverge. Under natural conditions on f, the sequence of compact sets $X^{(0)} \supseteq X^{(1)} \supseteq X^{(2)} \cdots$ can be shown to converge quadratically to x^* [1]. One can derive the above dynamical system in $\mathbb{IR}$ via the mean value theorem. Let $f(x)$ be continuously differentiable and $f'(x) \neq 0$ for all $x \in X$ such that x^* is the only zero of f in X. Then, by the mean value theorem, for every x, there exists a $c \in (x, x^*)$, such that, $f(x) - f(x^*) = f'(c)(x - x^*)$. Since, $f'(c) \neq 0$, by assumption, and $f(x^*) = 0$, it follows that:

$$x^* = x - \frac{f(x)}{f'(c)} \in x - \frac{f(x)}{F'(X)} =: N(X), \quad \forall x \in X$$

$N(X)$ is called the Newton operator and it contains x^*. Since our root of interest lies in X, $x^* \in N(X) \cap X$. Note that the above dynamical system in $\mathbb{IR}$ is obtained by replacing x with $m(X)$ and X with $X^{(j)}$ in the previous expression. The interval Newton method can also be interpreted geometrically. At the j^{th} iteration, a set of light beams are shone from the point $(x^{(j)}, f(x^{(j)}))$ along the directions of all the tangents to $f(x)$ on the entire interval X. The intersection of these beams (gray floodlight of Figure 1.7) with the domain is $N(X^{(j)})$. The iteration is resumed with the new interval $X^{(j+1)} = N(X^{(j)}) \cap X^{(j)}$. Next we extend the interval Newton method in order to allow $F'(X)$ to contain 0.

By including two ideal points $+\infty$ and $-\infty$ to $\mathbb{R}$, it becomes possible to extend interval arithmetic to $\mathbb{IR}^* := \mathbb{IR} \cup \{(-\infty, \overline{x}] : \overline{x} \in \mathbb{R}\} \cup \{[\underline{x}, +\infty) : \underline{x} \in \mathbb{R}\} \cup (-\infty, +\infty)$, the set of intervals with end points in the complete lattice $\mathbb{R}^* := \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$, with respect to the ordering relation $\leq$. Since division is the inverse operation of multiplication, obtaining any $x/y \in X/Y := \{x/y : x \in X, y \in Y\}$ is equivalent to solving the equation $y \cdot s = x$ for s, i.e., $X/Y := \{s : y \cdot s = x, x \in X, y \in Y\}$. Let [ ] denote the empty interval. With the following rules, division by

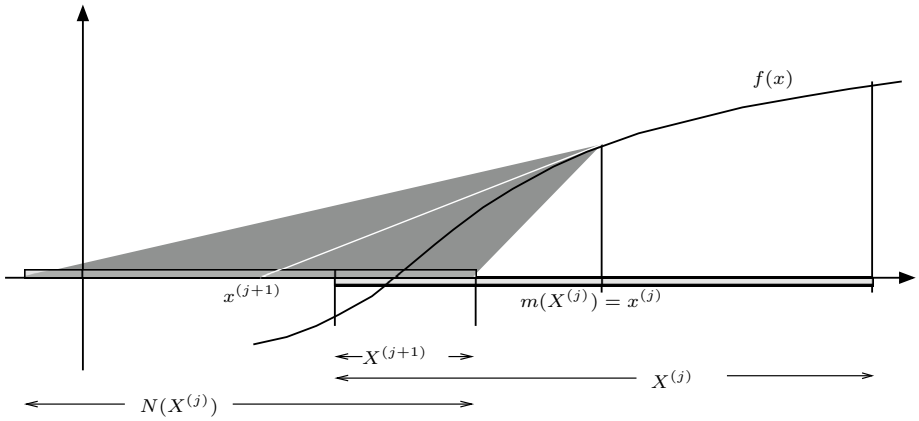


Figure 1.7: Geometric interpretation of the interval Newton method

intervals containing 0 becomes possible.

$$X/Y := \begin{cases} (-\infty, +\infty) & \text{if } 0 \in X, \text{or } Y = [0,0] \\ [\,] & \text{if } 0 \notin X, \text{and } Y = [0,0] \\ [\overline{x}/\underline{y}, +\infty) & \text{if } \overline{x} \le 0, \text{ and } \overline{y} = 0 \\ [\underline{x}/\overline{y}, +\infty) & \text{if } 0 \le \underline{x}, \text{ and } 0 = \underline{y} < \overline{y} \\ (-\infty, \overline{x}/\overline{y}\,] & \text{if } \overline{x} \le 0, \text{ and } 0 = \underline{y} < \overline{y} \\ (-\infty, \underline{x}/\underline{y}\,] & \text{if } 0 \le \underline{x}, \text{ and } \underline{y} < \overline{y} = 0 \\ (-\infty, \overline{x}/\overline{y}\,] \cup [\overline{x}/\underline{y}, +\infty) & \text{if } \overline{x} \le 0, \text{ and } [0,0] \Subset Y \\ (-\infty, \underline{x}/\underline{y}\,] \cup [\underline{x}/\overline{y}, +\infty) & \text{if } 0 \le \underline{x}, \text{ and } [0,0] \Subset Y \end{cases}$$

When X is a thin interval with $x = \underline{x} = \overline{x}$ and Y has $+\infty$ or $-\infty$ as one of its bounds, then extended interval subtraction is also necessary for the extended interval Newton algorithm, and is defined as follows:

$$[\underline{x}, \overline{x}\,] - Y := \begin{cases} (-\infty, +\infty) & \text{if } Y = (-\infty, +\infty) \\ (-\infty, x - \underline{y}] & \text{if } Y = (\underline{y}, +\infty) \\ [x - \overline{y}, +\infty) & \text{if } Y = (-\infty, \overline{y}\,] \end{cases}$$

The extended interval Newton method sketched below uses the extended interval arithmetic described above and is a variant of the method based on [15] with Ratz's modifications [37] as implemented in [12]. It can be used to enclose the roots of a continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}^n$ in a given box $X \in \mathbb{IR}^n$. Let $J_f(x) := ((\partial f_i(x)/\partial x_j))_{i,j=\{1,\cdots,n\}} \in \mathbb{R}^{n\times n}$ denote the Jacobian matrix of f at x. Let $J_F(X) \supset J_f(X)$ denote the Jacobian of the interval extension of f. The Jacobian can be computed via automatic differentiation of section 1.2.2 by computing the gradient of each component f_i of f. By the mean value theorem, $f(m(X)) - f(x^*) = J_f(w) \cdot (m(X) - x^*)$, for some $x^* \in X, w = (w_1, w_2, \cdots, w_n)$, where $w_i \in X, \forall i \in \{1, 2, \cdots, n\}$. Interest in x^* with $f(x^*) = 0$, yields the following relation, provided $\forall x \in X,\ J_F(x)$ is invertible.

$$
\begin{aligned}
f(m(X)) &= J_f(w) \cdot (m(X) - x^*) \\
x^* &= m(X) - (\, J_f(w)\,)^{-1} \cdot f(m(X)) \\
&\in m(X) - (\, J_f(X)\,)^{-1} \cdot f(m(X)) \\
&\subseteq m(X) - (\, J_F(X)\,)^{-1} \cdot F(m(X)) =: \mathcal{N}(X) \\
&\subseteq \mathcal{N}(X) \cap X
\end{aligned}
$$

An iteration scheme $X^{(j+1)} := \mathcal{N}(X^{(j)}) \cap X^{(j)}$, where $j = 0, 1, \cdots$, and $X^{(0)} := X$, will enclose the zeros of f contained in X. To relax the assumption that every matrix in $J_F(X)$ be invertible, the inverse of the midpoint of $J_F(X)$, i.e., $(m(J_F(X)))^{-1} =: p \in \mathbb{R}^{n\times n}$, is used as a matrix preconditioner. The extended interval Gauss-Seidel iteration, which is also applicable to singular systems [32], is used to solve the preconditioned interval linear equation,

$$
\begin{aligned}
p \cdot F(m(X)) &= p \cdot J_F(X) \cdot (m(X) - x^*) \\
a &= G \cdot (c - x^*),
\end{aligned}
$$

where, $a \in A := p \cdot F(m(X)), G := p \cdot J_F(X)$, and, $c := m(X)$. Thus, the solution

set $\mathbf{S} := \{x \in X : g \cdot (c - x) = a, \forall g \in G\}$ of the interval linear equation $a = G \cdot (c - x)$ has the component-wise solution set $\mathbf{S}_i = \{x_i \in X_i : \sum_{j=1}^{n} (\, g_{i,j} \cdot (c_j - x_j)\,) = a_i, \forall g \in G\}$, $\forall i \in \{1, \cdots, n\}$. Now, set $Y = X$, and solve the ith equation for the ith variable, iteratively for each i, as follows:

$$\begin{aligned} y_i &= c_i - \tfrac{1}{g_{i,i}} \left(a_i + \textstyle\sum_{j=1, j \neq i}^{n} (\, g_{i,j} \cdot (y_j - c_j)\,) \right) \\ &\in \left(c_i - \tfrac{1}{G_{i,i}} \left(A_i + \textstyle\sum_{j=1, j \neq i}^{n} (\, G_{i,j} \cdot (Y_j - c_j)\,) \right) \right) \cap Y_i \end{aligned}$$

The interval vector(s) Y obtained at the end of such an iteration is the set, $\mathcal{N}_{GS}(X)$, resulting from one extended interval Newton Gauss-Seidel step, such that, $\mathbf{S} \subseteq \mathcal{N}_{GS}(X) \subseteq X$. Thus, the roots of f are enclosed by the discrete dynamical system $X^{(j)} = \mathcal{N}_{GS}(X^{(j)})$ in $\mathbb{IR}^n$. Every 0 of f that lies in X also lies in $\mathcal{N}_{GS}(X)$. If $\mathcal{N}_{GS}(X) = [\,]$, the empty interval, then f has no solution in X. If $\mathcal{N}_{GS}(X) \Subset X$, then f has a unique solution in X. For proofs of the above three statements see [14]. When $G_{ii} \supset 0$, the method is applicable with extended interval arithmetic that allows for division by 0. In such cases, one may obtain up to two disjoint compact intervals for Y_i subsequent to extended interval arithmetic and intersection with the previous compact interval X_i. In such cases, the iteration is applied to each resulting sub-interval. One can also geometrically interpret the extended interval Newton method in one dimension [22].

1.2.4 Machine interval arithmetic

All interval arithmetic was done above with real intervals. However, there are only finitely many floating-point numbers available on a computing machine. Let R be this set of real floating-point numbers that constitute the real number screen $\mathbb{R}_{Screen}$. A machine interval is a real interval with floating-point bounds. Thus, on a computer, one works with $\mathbb{IR} := \{X \in \mathbb{IR} : \underline{x}, \overline{x} \in \mathsf{R}\}$, the set of all machine

intervals. In spite of the finiteness of $\mathbb{IR}$, the strength of interval arithmetic lies in a machine interval X being able to enclose the entire continuum of reals between its machine-representable boundaries. Through rounding controlled floating-point arithmetic provided by the IEEE arithmetic standard, operations with real intervals can be tightly enclosed by the rounding directed operations with the smallest machine intervals containing them [22]. The errors resulting from converting a decimal number, usually a constant or input data, which in general does not have a finite binary representation, to a binary floating-point number is controlled by first passing the decimal number as a string and then enclosing it with the smallest machine interval by proper outward rounding. The program is written in `C++` using the `C-XSC` class libraries. The differentiation arithmetic of section 1.2.2 is implemented using the `hess_ari` module provided in [12].

1.2.5 Global Optimization

Branch-and-bound

The most basic strategy in global optimization through enclosure methods is to employ rigorous branch-and-bound techniques. Such techniques recursively partition (branch) the original compact space of interest into compact subspaces and discard (bound) those subspaces that are guaranteed to not contain the global optimizer(s). The problem of finding the global maximum of $\ell(\theta)$, the likelihood function, is equivalent to finding the global minimum of $l(\theta) := -\ell(\theta)$. Let $\mathcal{L}(\Theta)$ be the natural interval extension of the negative log likelihood function $l(\theta)$ over Θ. Let $\nabla\mathcal{L}(\Theta)$ and $\nabla^2\mathcal{L}(\Theta)$ be the enclosures of the gradient and the Hessian of $l(\theta)$ over Θ respectively. For the real scalar-valued multi-dimensional objective function $l(\theta)$, the interval branch-and-bound technique can be applied to its natural

interval extension $\mathcal{L}(\Theta)$ to obtain an interval enclosure $\mathcal{L}^*$ of the global minimum value l^* as well as the set of minimizer(s) to a specified accuracy ϵ. Note that this set of minimizer(s) of $\mathcal{L}(\theta)$ is the set of maximizer(s) of the likelihood function for the observed data $\mathcal{D}$. The strength of such methods arises from the algorithmic ability to discard large sub-boxes from the original search region $\subset \mathbb{IR}^b$,

$$\Theta^{(0)} = (\Theta_1^{(0)}, \cdots, \Theta_b^{(0)}) := ([\underline{\theta}_1^{(0)}, \overline{\theta}_1^{(0)}], \cdots, [\underline{\theta}_b^{(0)}, \overline{\theta}_b^{(0)}])$$

that are not candidates for global minimizer(s). Four tests that help discard sub-regions are described below. Let $\mathfrak{L}$ denote a list of ordered pairs of the form $(\Theta^{(i)}, \underline{\mathcal{L}}_{\Theta^{(i)}})$, where, $\Theta^{(i)} \subseteq \Theta^{(0)}$, and $\underline{\mathcal{L}}_{\Theta^{(i)}} := \min (\mathcal{L}(\Theta^{(i)}))$ is a lower bound for the range of the negative log likelihood function l over $\Theta^{(i)}$. Let $\tilde{l}$ be an upper bound for l^* and $\nabla\mathcal{L}(\Theta^{(i)})_k$ denote the k-th interval of the gradient box $\nabla\mathcal{L}(\Theta^{(i)})$. If no information is available for $\tilde{l}$, then $\tilde{l} = \infty$.

Midpoint Cut-off test

The basic idea of the *midpoint cut-off test* is to discard sub-boxes of the search space $\Theta^{(0)}$ with the lower bound for their range enclosures above $\tilde{l}$, the current best estimate of an upper bound for l^*. Figure 1.8 shows a multi-modal l as a function of a scalar valued θ over $\Theta^{(0)} = \cup_{i=1}^{16} \Theta^{(i)}$. For this illustrative example, $\tilde{l}$ is set as the upper bound of the range enclosure of l over the smallest machine interval containing the midpoint of $\Theta^{(15)}$, the interval with the smallest lower bound of its range enclosure. The shaded rectangles show the range enclosures over intervals that lie strictly above $\tilde{l}$. In this example the *midpoint cut-off test* would discard all other intervals except $\Theta^{(1)}$, $\Theta^{(2)}$, and $\Theta^{(4)}$.

- Given a list $\mathfrak{L}$ and $\tilde{l}$

- Choose an element j of $\mathfrak{L}$, such that, $j = \underset{i}{argmin}\, \underline{\mathcal{L}}_{\Theta^{(i)}}$, since $\Theta^{(j)}$ is likely to contain a minimizer.
- Find its midpoint $c = m(\Theta^{(j)})$ and let C be the smallest machine interval containing c.
- Compute a possibly improved $\tilde{l} = \min\left\{\tilde{l}, \overline{\mathcal{L}}_C\right\}$, where, $\overline{\mathcal{L}}_C := \max\left(\mathcal{L}(C)\right)$
- Discard any i-th element of $\mathfrak{L}$ for which $\underline{\mathcal{L}}_{\Theta^{(i)}} > \tilde{l} \geq l^*$

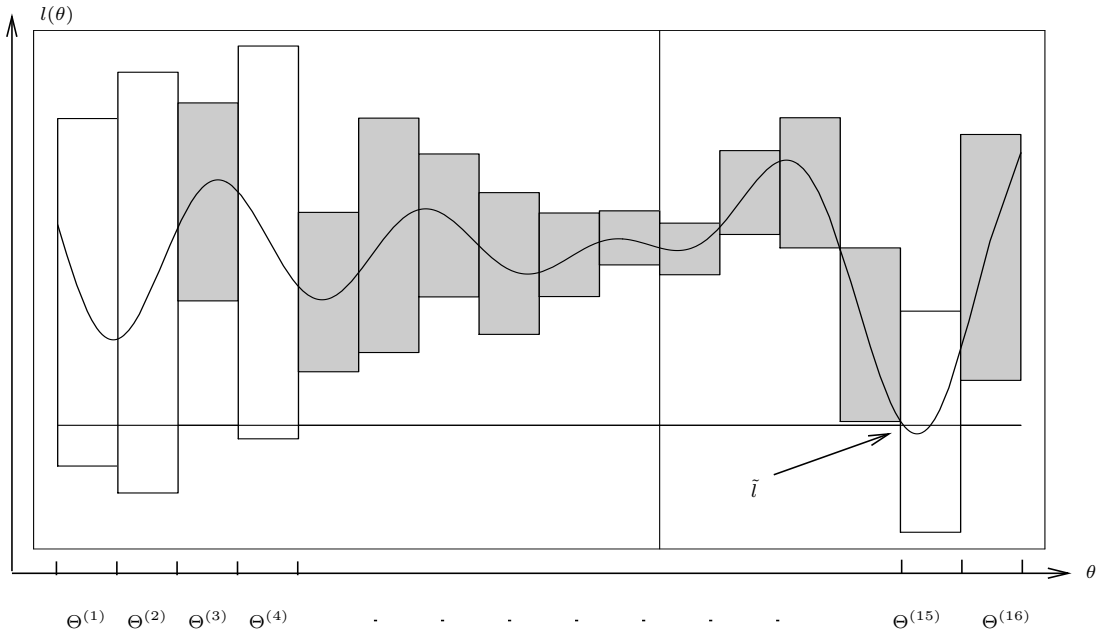


Figure 1.8: Midpoint Cut-off test

Monotonicity test

For a continuously differentiable function $l(\theta)$, the *monotonicity test* determines whether $l(\theta)$ is strictly monotone over an entire sub-box $\Theta^{(i)} \subseteq \Theta^{(0)}$. If l is strictly monotone over $\Theta^{(i)}$, then a global minimizer cannot lie in the interior of $\Theta^{(i)}$. Therefore, $\Theta^{(i)}$ can only contain a global minimizer as a boundary point if this

point also lies in the boundary of $\Theta^{(0)}$. Figure 1.9 illustrates the *monotonicity test* for the one-dimensional case. In this example the search space of interest, $\Theta^{(0)} = [\underline{\theta}^{(0)}, \overline{\theta}^{(0)}] = \cup_{i=1}^{8} \Theta^{(i)}$, can be reduced considerably. In the interior of $\Theta^{(0)}$, one may delete $\Theta^{(2)}$, $\Theta^{(5)}$, and $\Theta^{(7)}$, since $l(\theta)$ is monotone over them as indicated by the enclosure of the derivative $l'(\theta)$ being bounded away from 0. Since $l(\theta)$ is monotonically decreasing over $\Theta^{(1)}$ one can also delete it since we are only interested in minimization. $\Theta^{(8)}$ may be pruned to its right boundary point $\theta^{(8)} = \overline{\theta}^{(8)} = \overline{\theta}^{(0)}$ due to the strictly decreasing nature of $l(\theta)$ over it. Thus, the *monotonicity test* has pruned $\Theta^{(0)}$ to the smaller candidate set $\{\, \overline{\theta}^{(0)}, \Theta^{(3)}, \Theta^{(4)}, \Theta^{(6)} \,\}$ for a global minimizer.

- Given $\Theta^{(0)}$, $\Theta^{(i)}$, and $\nabla \mathcal{L}(\Theta^{(i)})$
- Iterate for $k = 1, \cdots, b$
  - If $0 \in \nabla \mathcal{L}(\Theta^{(i)})_k$, then leave $\Theta_k^{(i)}$ unchanged, as it may contain a stationary point of l.
  - Otherwise, $0 \notin \nabla \mathcal{L}(\Theta^{(i)})_k$. This implies that $\Theta^{(i)}$ can be pruned, since $l^* \notin \Theta^{(i)}$ except possibly at the boundary points, as follows:
    1. if $\min\left(\nabla \mathcal{L}(\Theta^{(i)})_k\right) > 0$ and $\underline{\theta}_k^{(0)} = \underline{\theta}_k^{(i)}$, then $\Theta_k^{(i)} = [\underline{\theta}_k^{(i)}, \underline{\theta}_k^{(i)}]$,
    2. Else if $\max\left(\nabla \mathcal{L}(\Theta^{(i)})_k\right) < 0$ and $\overline{\theta}_k^{(0)} = \overline{\theta}_k^{(i)}$, then $\Theta_k^{(i)} = [\overline{\theta}_k^{(i)}, \overline{\theta}_k^{(i)}]$.
    3. Else, delete the i-th element of $\mathfrak{L}$ and stop the iteration.

Concavity test

Given $\Theta^{(i)} \Subset \Theta^{(0)}$, and the diagonal elements $\left(\nabla^2 \mathcal{L}(\Theta^{(i)})\right)_{kk}$ of $\nabla^2 \mathcal{L}(\Theta^{(i)})$, note that if $\max\left(\left(\nabla^2 \mathcal{L}(\Theta^{(i)})\right)_{kk}\right) < 0$ for some k, then, $\nabla^2 \mathcal{L}(\Theta^{(i)})$ cannot be positive

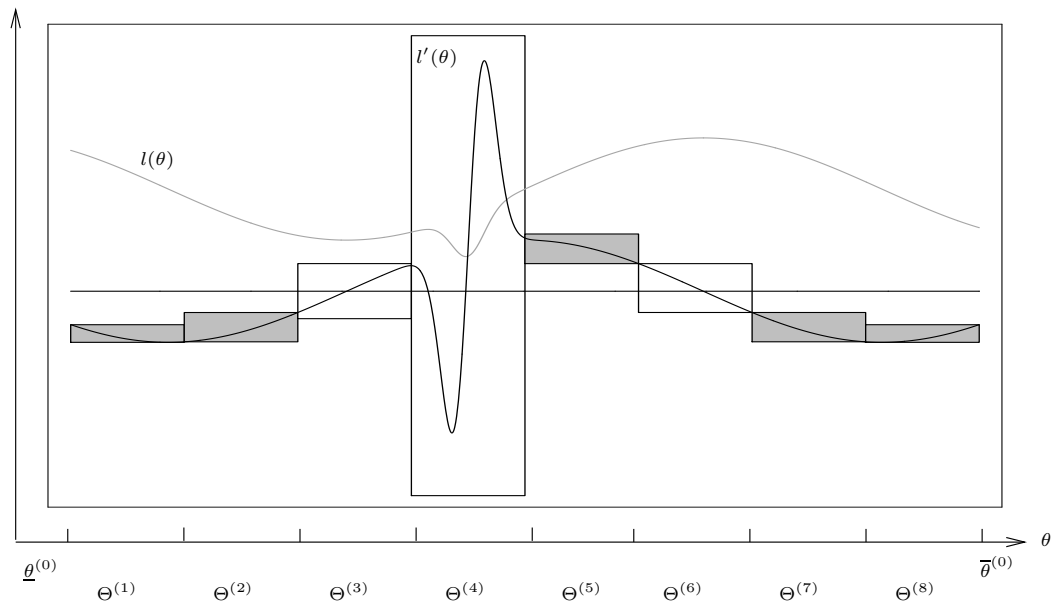


Figure 1.9: Monotonicity test

semidefinite, and therefore $l(\theta)$ cannot be convex over $\Theta^{(i)}$ and thus cannot contain a minimum in its interior. In the one-dimensional example shown in Figure 1.9, an application of the *concavity test* to the candidate set $\{\, \underline{\theta}^{(0)}, \Theta^{(4)}, \Theta^{(6)} \,\}$ for a global minimizer returned by the *monotonicity test*, would result in the deletion of $\Theta^{(6)}$ due to the concavity of $l(\theta)$ over it.

- Given $\Theta^{(i)} \Subset \Theta^{(0)}$ and $\nabla^2 \mathcal{L}(\Theta^{(i)})$
- If $\max\left((\, \nabla^2 \mathcal{L}(\Theta^{(i)}) \,)_{kk} \right) < 0$ for any $k \in \{1, \cdots, b\}$, then delete the i-th element of $\mathfrak{L}$.

Interval Newton test

Given $\Theta^{(i)} \Subset \Theta^{(0)}$, and $\nabla \mathcal{L}(\Theta^{(i)})$, attempt to solve the system, $\nabla \mathcal{L}(\theta) = 0$, in terms of $\theta \in \Theta^{(i)}$.

- Apply one extended interval Newton Gauss-Seidel step of Section 1.2.3 to the linear interval equation $a = G \cdot (c - \theta)$, where, $a := p \cdot \mathcal{L}(m(\Theta^{(i)}))$, $G := p \cdot \nabla^2 \mathcal{L}(\Theta^{(i)})$, $c := m(\Theta^{(i)})$, and $p := (m(\nabla^2 F(X)))^{-1}$, in order to obtain $\mathcal{N}'_{GS}(\Theta^{(i)})$.

- One of the following can happen,

  1. If $\mathcal{N}'_{GS}(\Theta^{(i)})$ is empty, then discard $\Theta^{(i)}$.
  2. If $\mathcal{N}'_{GS}(\Theta^{(i)}) \Subset \Theta^{(i)}$, then replace $\Theta^{(i)}$ by the contraction $\mathcal{N}'_{GS}(\Theta^{(i)}) \cap \Theta^{(i)}$.
  3. If $0 \in G_{jj}$, and the extended interval division splits $\Theta_j^{(i)}$ into a non-empty union of $\Theta_j^{(i),1}$ and $\Theta_j^{(i),2}$, then the iteration is continued on $\Theta_j^{(i),1}$, while $\Theta_j^{(i),2}$, if non-empty, is stored in $\mathfrak{L}$ for future processing. Thus, one extended interval Newton Gauss-Seidel step can add at most $b + 1$ sub-boxes to $\mathfrak{L}$.

Verification

Given a collection of sub-boxes, $\{ \Theta^{(1)}, \cdots, \Theta^{(n)} \}$, each of width $\leq \epsilon$, that could not be discarded by the tests in Section 1.2.5, one can attempt to verify the existence and uniqueness of a local minimizer within each sub-box $\theta^{(i)}$ by checking whether the conditions of the following two theorems are satisfied. For proof of these two theorems see [14] and [37].

1. If $\mathcal{N}'_{GS}(\Theta^{(i)}) \Subset \Theta^{(i)}$, then there exists a unique stationary point of $\mathcal{L}$, i.e., a unique zero of $\nabla \mathcal{L}$ exists in $\Theta^{(i)}$.

2. If $(I + \frac{1}{\kappa} \cdot (\nabla^2 \mathcal{L}(\Theta^{(i)}))) \cdot Z \Subset Z$, where $(\nabla^2 \mathcal{L}(\Theta^{(i)}))_{d,\infty} \leq \kappa \in \mathbb{R}$, for some $Z \in \mathbb{IR}^n$, then, the spectral radius $\rho(s) < 1$ for all $s \in (I - \frac{1}{\kappa} \cdot (\nabla^2 \mathcal{L}(\Theta^{(i)})))$, and all symmetric matrices in $\nabla^2 \mathcal{L}(\Theta^{(i)})$ are positive definite.

If the conditions of the above two theorems are satisfied by some $\Theta^{(i)}$, then a unique stationary point exists in $\Theta^{(i)}$ and this stationary point is a local minimizer. Therefore, if exactly one candidate sub-box for minimizer(s) remained after pruning the search box $\Theta^{(0)}$ with the tests in Section 1.2.5, and if this sub-box satisfies the above two conditions for the existence of a unique local minimizer within it, then one has rigorously enclosed the global minimizer in the search interval. On the other hand, if there are two or more sub-boxes in our candidate list for minimizer(s) that satisfy the above two conditions, then one may conclude that each sub-box contains a candidate for a global minimizer which may not necessarily be unique (disconnected sub-boxes, for example). Observe that failure to verify the uniqueness of a local minimizer in a sub-box can occur if it contains two or more points or even a continuum of points that are stationary (non-identifiable manifolds in the sub-box, for example).

Algorithm

- *Initialization:*

    1. Let the search region be a single box $\Theta^{(0)}$ or a collection of not necessarily connected, but pair-wise disjoint boxes, $\Theta^{(i)}$, $i \in \{1, \cdots, r\}$.
    2. Initialize the list $\mathfrak{L}$ which may just contain one element $(\Theta^{(0)}, \underline{\mathcal{L}}_{\Theta^{(0)}})$ or several elements

    $$\{ (\Theta^{(1)}, \underline{\mathcal{L}}_{\Theta^{(1)}}), (\Theta^{(2)}, \underline{\mathcal{L}}_{\Theta^{(2)}}), \cdots, (\Theta^{(r)}, \underline{\mathcal{L}}_{\Theta^{(r)}}) \}.$$

    3. Let ϵ be a specified tolerance.
    4. Let $\max_{\mathfrak{L}}$ be the maximal length allowed for list $\mathfrak{L}$.
    5. Set the non-informative lower bound for l^*, i.e., $\tilde{l} = \infty$

- *Iteration:*

1. (a) Improve $\tilde{l} = \min\{\, \tilde{l}, \max(\mathcal{L}(m(\Theta^{(j)}))) \,\}$, where $j = \underset{i}{argmin}\{\underline{\mathcal{L}}_{\Theta^{(i)}}\}$.

   (b) Perform the *midpoint cut-off test* to $\mathfrak{L}$.

   (c) Set $\mathcal{L}^* = [\underline{\mathcal{L}}_{\Theta^{(j)}}, \tilde{l}]$.

2. Bisect $\Theta^{(j)}$ along its longest side k, i.e., $d(\Theta_k^{(j)}) = d_\infty(\Theta^{(j)})$, to obtain sub-boxes $\Theta^{(j_q)}$, $q \in \{1, 2\}$.

3. For each sub-box $\Theta^{(j_q)}$, evaluate its triple $(\, \mathcal{L}(\Theta^{(j_q)}), \nabla\mathcal{L}(\Theta^{(j_q)}), \nabla^2\mathcal{L}(\Theta^{(j_q)}) \,)$, and do the following:

   (a) Perform *monotonicity test* to possibly discard $\Theta^{(j_q)}$.

   (b) *Centered form cut-off test:*

   Improve the range enclosure of $\mathcal{L}(\Theta^{(j_q)})$ by replacing it with its centered form $\mathcal{L}_c(\Theta^{(j_q)}) :=$

   $$\{\mathcal{L}(m(\Theta^{(j_q)})) + \nabla\mathcal{L}(\Theta^{(j_q)}) \cdot (\Theta^{(j_q)} - m(\Theta^{(j_q)}))\} \cap \mathcal{L}(\Theta^{(j_q)}),$$

   and then discarding $\Theta^{(j_q)}$, if $\tilde{l} < \underline{\mathcal{L}}_{\Theta^{(j_q)}}$.

   (c) Perform *concavity test* to possibly discard $\Theta^{(j_q)}$.

   (d) Apply an *extended interval Newton Gauss-Seidel step* to $\Theta^{(j_q)}$, in order to either entirely discard it or shrink it into v sub-sub-boxes, where v is at most $2s - 2$.

   (e) For each one of these sub-sub-boxes $\Theta^{(j_{q,u})}$, $u \in \{1, \cdots, v\}$

      i. Perform *monotonicity test* to possibly discard $\Theta^{(j_{q,u})}$.

      ii. Try to discard $\Theta^{(j_{q,u})}$ by applying the *centered form cut-off test* in 3b to it.

iii. Append $(\Theta^{(j_{q,u})}, \underline{\mathcal{L}}_{\Theta^{(j_{q,u})}})$ to $\mathfrak{L}$ if $\Theta^{(j_{q,u})}$ could not be discarded by steps 3(e)i and 3(e)ii.

- *Termination:*

  1. Terminate iteration if $d_{rel,\infty}(\Theta^{(j)}) < \epsilon$, or $d_{rel,\infty}(\mathcal{L}^*) < \epsilon$, or $\mathfrak{L}$ is empty, or $\text{Length}(\mathfrak{L}) > \max_{\mathfrak{L}}$
  2. Verify uniqueness of minimizer(s) in the final list $\mathfrak{L}$ by applying algorithm of section 1.2.5 to each of its elements.

The *extended interval Newton* method in combination with the *midpoint cut-off*, *monotonicity*, and *concavity* tests may be used to study the shape of the likelihood surface itself. For instance, one could rigorously enclose all the local maxima, or search for non-identifiable subspaces, above a given level-set of the likelihood function within any compact subset of the parameter space. Several efficiency increasing steps could be taken. Pre-enclosing the transition probabilities and accessing them through hash functions can save computational effort. Asynchronous parallelization of the algorithm across 6 processors is also observed to increase the rate of convergence to the global maximum. It also provides a natural framework to manage the memory requirements for larger trees through partial likelihood evaluations for non-overlapping subtrees in parallel prior to obtaining the full likelihood.

Most statistical inference today is done on computing machines through numerical methods that do not rigorously account for the physical realities of such machines with finite memory. Possibly sub-optimal decisions may suffice for several decision problems. However, for others, such as, parameter estimation for nonlinear stochastic differential equations, or finding the equilibrium configuration of the

n atoms in a folding protein molecule, or data fitting problems in training neural networks, one may want/have to guarantee the globally optimal decision. The enclosure methods provide statisticians with powerful tools for rigorous numerical inference.

1.3 Machine Interval Experiments

In this section we use the tools of Section 1.2 to address the questions raised in Section 1.1. We will show that the usual experiment can be extended with enclosure methods to account for the limits on empirical and numerical resolutions. This extended experiment is the interval experiment and its machine counterpart is the machine interval experiment.

1.3.1 Two simple examples

First we will see two simple examples to fix ideas. Here we denote points in $\mathbb{R}$ as x_i, and a sequence of such points as a vector $(x_1, ..., x_n)$. We denote a random variable by $\mathtt{X}$, its n independent and infinitely precise realizations by a vector of points $(x_1, ..., x_n)$ and the enclosures of these realizations by a vector of boxes $(X_1, \ldots, X_n)$, where each $X_i = [\underline{x}_i, \overline{x}_i]$. The random variable $\mathtt{X}$ has a single unknown measure $P_{\theta^*} \in \mathcal{P}$ that we are trying to infer under a maximum likelihood framework. If the infinitely precise realizations $(x_1, ..., x_n)$ of $\mathtt{X}$ were indeed measurable then one can find an estimate of the underlying P_{θ^*} with arithmetic on the reals. However, if we only know the realizations up to a sequence of intervals $(X_1, \ldots, X_n)$, in the sense that

$$x_1 \in X_1, \ldots, x_n \in X_n,$$

and nothing more, then we can enclose the underlying measure P_{θ^*} with the smallest possible set (w.r.t. the partial order $\subseteq$) of measures P_{Θ^*} using interval analysis. Naturally, the latter approach uses the set-valued mathematics introduced in Section 1.2 and can therefore rigorously account for the limits on empirical and numerical resolutions.

Exponential Rate Model

The data vector $x = (x_1, ..., x_n)$ is a realization of the random variable $\mathtt{X}$ with values in $\mathbf{X} = (0, \infty)^n$. The distribution of $\mathtt{X}$ is the joint law of n independent and identically distributed Exponential random variables $\mathtt{X}_1, \ldots, \mathtt{X}_n$, each having the law $Exp(\theta)$, where $\theta > 0$ is unknown. This model with unknown rate parameter $\theta^* \in \mathbf{\Theta} = (0, \infty)$ is called the exponential rate model. It is well known that the maximum likelihood estimate of θ^* is the inverse of the sample mean

$$\widehat{\theta}^{(n)} = n(\sum_{i=1}^{n} x_i)^{-1}.$$

To make the example more realistic let us suppose that the model is describing the burn-out time of light bulbs made by the same manufacturer. We want to infer the burn-out rate as precisely as possible. The above inference procedure is more natural if the bulbs are tested with a device that can measure the instant of time (up to seconds) that the bulb burnt out. However, if we have a clock that only makes an hourly alarm, then our measuring scale is much coarser and we only know that the bulb burnt out between two given hours. Let us further suppose that our mean burn-out time is around a few hundred minutes. With such a coarse scale our data vector would be intervals $X = (X_1, \ldots, X_n)$. The interval maximum likelihood estimate (MLE) that contains the MLE of each $x \in X$, such that $x_i \in X_i$, $\forall\, i \in \{1, \ldots, n\}$, is obtained by computing the inverse of the sample

mean with elementary interval operations.

$$\begin{aligned}\widehat{\Theta}^{(n)} &= n \cdot \left(\sum_{i=1}^{n} X_i\right)^{-1} \\ &= n \cdot \left([\sum_{i=1}^{n} \underline{x}_i, \sum_{i=1}^{n} \overline{x}_i]\right)^{-1} \\ &= n \cdot [\left(\sum_{i=1}^{n} \overline{x}_i\right)^{-1}, \left(\sum_{i=1}^{n} \underline{x}_i\right)^{-1}]\end{aligned}$$

Gaussian Location Scale Model

The data vector $x = (x_1, ..., x_n)$ is a realization of a random vector $\mathtt{X}$ with values in $\mathbf{X} = \mathbb{R}^n$. The distribution of $\mathtt{X}$ is the joint law of n independent and identically distributed Gaussian random variables $\mathtt{X}_1, \ldots, \mathtt{X}_n$, each having the law $N(\mu, \sigma^2)$, where $\sigma^2 > 0$ and $\mu \in \mathbb{R}$ is unknown. This model with unknown parameter $\theta = (\mu, \sigma^2) \in \mathbf{\Theta} = \mathbb{R} \times (0, \infty)$ is called the Gaussian location scale model. It is well known that the maximum likelihood estimates $\widehat{\mu}^{(n)}$ and $\widehat{\sigma^2}^{(n)}$ for μ and σ^2 are the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$ and the sample variance $\frac{1}{n}\sum_{i=1}^{n} x_i^2 - n(\widehat{\mu}^{(n)})^2$, respectively.

Now, suppose our data vector were intervals $X = (X_1, \ldots, X_n)$, then the interval MLE $(\widehat{\Theta}_1^{(n)}, \widehat{\Theta}_2^{(n)})$ that contains the MLE $(\widehat{\mu}^{(n)}, \widehat{\sigma^2}^{(n)})$ of each $x \in X$ is obtained by computing with elementary interval operations,

$$\begin{aligned}\widehat{\Theta}_1^{(n)} &= \frac{1}{n} \cdot [\sum_{i=1}^{n} \underline{x}_i, \sum_{i=1}^{n} \overline{x}_i] \\ \widehat{\Theta}_2^{(n)} &= \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^2 - n(\widehat{\Theta}_1^{(n)})^2,\end{aligned}$$

where, $X_i^2 = [min\{\underline{x}_i^2, \overline{x}_i^2\}, max\{\underline{x}_i^2, \overline{x}_i^2\}]$, if $0 \notin X$, and $[0, max\{\underline{x}_i^2, \overline{x}_i^2\}]$, otherwise.

1.3.2 Interval extension of the likelihood function

We generalize the above examples. Let us now suppose that the sample space $\mathbf{X}$ of the experiment $\mathcal{E}_{\mathcal{P}}$ is a subset of the Euclidean reals and its interval extension $\mathbb{I}\mathbf{X}$ is the set of all compact boxes in $\mathbf{X}$. We will assume that the index set $\mathbf{\Theta}$ of $\mathcal{E}_{\mathcal{P}}$ is a compact subset of $\mathbb{R}^b$ and therefore the interval extension of the index set $\mathbb{I}\mathbf{\Theta}$ is contained in $\mathbb{IR}^b$, unless otherwise stated. We refer to such experiments as *compact finite dimensional experiments* in the sequel. This assumption of compactness is partly justified because the number screen of a computer has a largest finite number in absolute value (see Chapter 1.2) and we are fundamentally interested in experiments that will employ computers in the hunt for optimal decisions.

Theorem 6 *If* $L(\Theta, X) : \mathbb{I}\mathbf{\Theta} \times \mathbb{I}\mathbf{X} \to \mathbb{IR}$ *is the natural interval extension of the usual likelihood (or log likelihood) function* $\ell(\theta, x) : \mathbf{\Theta} \times \mathbf{X} \to \mathbb{R}$*, and it is well-defined on* $(Z \times \Xi) \subseteq (\mathbf{\Theta} \times \mathbf{X})$*, then*

(i) Inclusion isotony: $\forall\, (X \times \Theta) \subseteq (Y \times \Theta) \subseteq (Z \times \Xi) \implies L(\Theta, X) \subseteq L(\Theta, Y)$

(ii) Range enclosure: $\forall\, (X \times \Theta) \subseteq (Z \times \Xi) \implies l(\Theta, X) \subseteq L(\Theta, Y)$

Proof: The theorem is merely a restatement of Theorem 4 in terms of the likelihood function.

Definition 13 *A compact finite dimensional experiment with a well-defined interval extension of the likelihood function is the extended interval experiment and its machine counter-part is the machine interval experiment.*

Even if the data enters the experiment as a sequence of intervals or boxes, $X = (X_1, \ldots, X_n), X_i \in \mathbb{I}\mathbf{X}$, instead of a vector of data points, $x = (x_1, \ldots, x_n), x_i \in \mathbf{X}$, the interval extension of the likelihood function allows us to rigorously enclose a

set of measures P_Θ that are all equally likely explanations for the interval data, using the rigorous global optimization methods of section 1.2.5 and operating with machine intervals. Thus the most likely estimate $\widehat{\theta}$ for every vector of data points $x \in X$ can be enclosed in the most likely box estimate $\widehat{\Theta}$ for the box data vector X.

Such an experiment with a well-defined inclusion isotonic likelihood function also allows for statistical consistency of the estimator in the complete metric space of the index set $\mathbb{I}\boldsymbol{\Theta}$ under the Hausdorff metric $\mathfrak{h}$, provided that $\mathcal{E}_\mathcal{P}$ is identifiable. An estimator $\widehat{\Theta}_n$ of Θ^* from n observations is asymptotically consistent if $\mathfrak{h}(\widehat{\Theta}_n, \Theta^*) \xrightarrow{P} 0$.

When $\mathbf{X}$ is a discrete set then the likelihood function may have a well-defined interval extension for every given x. For such discrete experiments LER does not usually pose problems. We can still use the interval extension of the likelihood function to rigorously enclose the MLEs using a computing machine. In fact we exactly do this for a problem in phylogenetics in Section 2.2.

1.3.3 Epistemological validity

The observed data x is usually assumed to be a thin set or point in the sample space $\mathbf{X}$. However, the realization x may be only enclosable by some set $X \in \mathbb{I}\mathbf{X}$, such that $d_\infty(X)$ reflects the physical limits on empirical resolution. It is important to point out that although the enclosure data X of the truly realized point data x is an event in the probability space, i.e. $X \in \mathcal{F}_\mathbf{X}$, it does not enter the statistical experiment as an event but only as an enclosure of the true data point. Therefore we are not interested in the probability of the set X, i.e., $\int_X P_\theta(x) d\lambda(x)$, but rather in the set that encloses the density for each $x \in X$, i.e., $\{P_\theta(x) : x \in X\}$.

We can achieve this through the interval extension of the likelihood function and thereby address empirical sufficiency or its lack there of directly without having to make further assumptions about the nature of the measuring device. Such enclosures of the likelihood may be necessary for making cautious decisions when highly nonlinear probability models in the spirit of Statement 1 are providing the index map for our experiment. Moreover, in a machine interval experiment, the estimator $\widehat{\Theta}_n$ may not converge in probability to a thin box even if the true measure θ^* is a singleton and all observations are measurable exactly, since the machine precision $p < \infty$. The diameter $d(\widehat{\Theta}_n)$ reflects the limit of numerical resolution underlying computations that account for rounding and conversion errors induced by the number screen. Since our machine interval experiment can account for the physical limits on empirical as well as numerical resolutions, we can make decisions that are epistemologically valid at least along the empirical and numerical fronts.

Chapter 2

Applications

2.1 Enclosing the Most Likely Mixtures

Finite mixtures of densities are used routinely to infer a wide variety of random phenomena. We will restrict our analysis of the simplest univariate mixtures to the maximum likelihood (ML) framework of frequentists. Since closed form solutions are generally not available for ML estimate (MLE), one usually finds the MLEs using gradient flow methods that rely on local information, including quasi Newton methods [35] and expectation maximization algorithms. Since these gradient flow algorithms are usually sensitive to the starting and stopping conditions it is possible to obtain a sub-global maximizer and the corresponding local maximum of the log likelihood function [28]. Several such local searches pose acute problems when one is interested in approximating the distribution of the likelihood ratio test statistic (LRTS) for the number of components in our mixture model via parametric bootstraps [28].

In the following sections we study two examples of univariate mixture models: (1) mixture of two Gaussian densities and (2) mixture of two exponential densities. The rigorous global optimization algorithm of section 1.2.5 can be applied in both cases to enclose the MLE since the interval extension of the likelihood function is inclusion isotonic and twice differentiable. For the second example involving a mixture of two exponential densities we can also enclose the distribution of LRTS from parametric bootstraps under the null model of a homogeneous population.

2.1.1 MLE of a Gaussian mixture

Consider the heteroscedastic mixture of two Gaussian densities

$$f(x,\theta) = w_1 \frac{1}{\sqrt{2\pi\sigma_1}} \exp \frac{-(x-\mu_1)^2}{2\sigma_1} + w_2 \frac{1}{\sqrt{2\pi\sigma_2}} \exp \frac{-(x-\mu_2)^2}{2\sigma_2}$$

where, $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, w_1, w_2)$, $w_1 + w_2 = 1$, $x, \mu_1, \mu_2 \in \mathbb{R}$, and $\sigma_1, \sigma_2 > 0$.

We generated 620 independent realizations from the above mixture model under the following parameters.

$$(\mu_1, \mu_2) = (-2, 2),\ (\sigma_1, \sigma_2) = (1, 1),\ (w_1, w_2) = (0.5, 0.5)$$

Since the log likelihood function under the assumption of independence is $\ell(x,\theta) = \sum_{i=1}^{620} \log f(x,\theta)$, it has a well-defined interval extension $L(X,\Theta)$ that is inclusion isotonic by Theorem 4. The negative log likelihood function was extended to the space of triples $(\, F(x), \nabla F(x), \nabla^2 F(x)\,)$ through interval-extended Hessian differentiation arithmetic of Section 1.2.2, then the MLE was enclosed by adaptive branch and bound algorithm of Section 1.2.5. Assuming that the weights are equal and known we obtained the following machine intervals containing the MLEs for the means and variances. We avoid the non-identifiability induced by the symmetry of the likelihood function by posing the problem via mean-ordering, i.e. we apply the additional constraint that $\mu_1 \leq \mu_2$. The notation x_a^b means the interval $[xa, xb]$, e.g. $-2.0944968_5^3 = [-2.09449685, -2.09449683]$.

1. MLEs and the maximum log likelihood value $\hat{\ell}$ are enclosed by the following intervals when the variances are known to be 1, i.e., $\sigma_1 = \sigma_2 = 1$.

$$\begin{aligned} \widehat{\mu_1} &\subset -2.0944968_5^3 \\ \widehat{\mu_2} &\subset 2.018153897_4^8 \\ \hat{\ell} &\subset -316.115255_7^1 \end{aligned}$$

2. MLEs and $\hat{\ell}$ are enclosed by the following intervals under the assumption of homoscedasticity, i.e., $\sigma = \sigma_1 = \sigma_2$. Notice the improvement in $\hat{\ell}$ when the variance is also estimated.

$$\begin{aligned} \widehat{\mu_1} &\subset -2.08813064940^{8}_{9} \\ \widehat{\mu_2} &\subset 2.007393347630^{6}_{5} \\ \widehat{\sigma} &\subset 1.1746484202766^{3}_{1} \\ \hat{\ell} &\subset -312.7529136799^{2}_{6} \end{aligned}$$

3. MLEs are enclosed by the following intervals under the assumption of heteroscedasticity.

$$\begin{aligned} \widehat{\mu_1} &\subset -2.095974^{5}_{6} \\ \widehat{\mu_2} &\subset 1.9980399^{6}_{4} \\ \widehat{\sigma_1} &\subset 1.0725577^{3}_{1} \\ \widehat{\sigma_2} &\subset 1.09735596^{3}_{1} \\ \hat{\ell} &\subset -312.716619^{6}_{8} \end{aligned}$$

Note that the above enclosures of the solutions for the MLEs and the log likelihood values are equivalent to computer assisted proofs of the true MLEs unlike the solutions for MLEs obtained with numerical local searches.

2.1.2 MLE and bootstrap of an exponential mixture

Consider the exponential mixture model

$$f(y, \theta) = \pi_1 \lambda_1 \exp(-\lambda_1 y) + \pi_2 \lambda_2 \exp(-\lambda_2 y), \ \theta = (\lambda_1, \lambda_2, \pi_1)$$

where $\theta = (\lambda_1, \lambda_2, \pi_1) \in \mathbf{\Theta} := (0, \infty) \times (0, \infty) \times [0, 1]$. We are interested in testing the null hypothesis

$$H_0 : \theta \in \mathbf{\Theta_0} := \{\theta : \lambda_1 = \lambda_2\} \cup \{\theta : \pi_1 = 0\} \cup \{\theta : \pi_1 = 1\}$$

against the alternative hypothesis

$$H_1 : \theta \in \mathbf{\Theta} \setminus \mathbf{\Theta_0}$$

This test leads to the breakdown of regularity conditions for the asymptotic distribution of the LRTS since θ under H_0 lies in a non-identifiable subspace of $\mathbf{\Theta}$. One practical solution for obtaining the distribution of the LRTS for finitely many samples $y_1, \ldots, y_n$ is via parametric bootstrap, where one simulates B data-sets from the ML estimate $1/\sum_{i=1}^{n} y_i$ under H_0 and computes the LRTS $2(\ell(\widehat{\theta_1}) - \ell(\widehat{\theta_0}))$ for each of the B bootstrapped datasets. The empirical distribution of LRTS obtained from B bootstraps is of interest as it can be used to approach the critical value C_α of the α-level test. However, $C_{\alpha=0.05}$ which is the critical value for LRTS can be sensitive to the gradient flows of Expectation Maximization (EM), Newton, Quasi-Newton, Simulated annealing and genetic algorithms that may be used to find the LRTS for each of the B datasets.

To study this problem in detail, we further simplified the model and assumed that the weights were known to be equal and thus fixed, i.e., $\pi_1 = \pi_2 = 0.5$. We first obtained a dataset $\mathcal{D}$ of 100 independent samples from the mixture of two exponentials with the following parameters $\theta^* = (\lambda_1, \lambda_2, \pi_1) = (1, 5, 0.5)$. Next we simulated 1000 datasets each of sample size 100 from the ML estimate $\widehat{\theta_0} = 1/\sum_{i=1}^{n} y_i = 0.28$ in $\mathbf{\Theta_0}$. The critical values $C_{\alpha=0.05}$ for the datasets based on Quasi-Newton searches for the MLEs can range in $[0.0001, 2.54]$ depending on the starting point of the local search. For instance, when all the searches are initialized on the line $\lambda_1 = \lambda_2$, then the search trajectory tends to the possibly sub-global maximizer along the line. Figure 2.1.2 shows the log likelihood surface and the corresponding contours for two datasets. Notice that the quasi-Newton search started along $\lambda_1 = \lambda_2$ terminates on the sub-global maximizer along $\lambda_1 = \lambda_2$ and

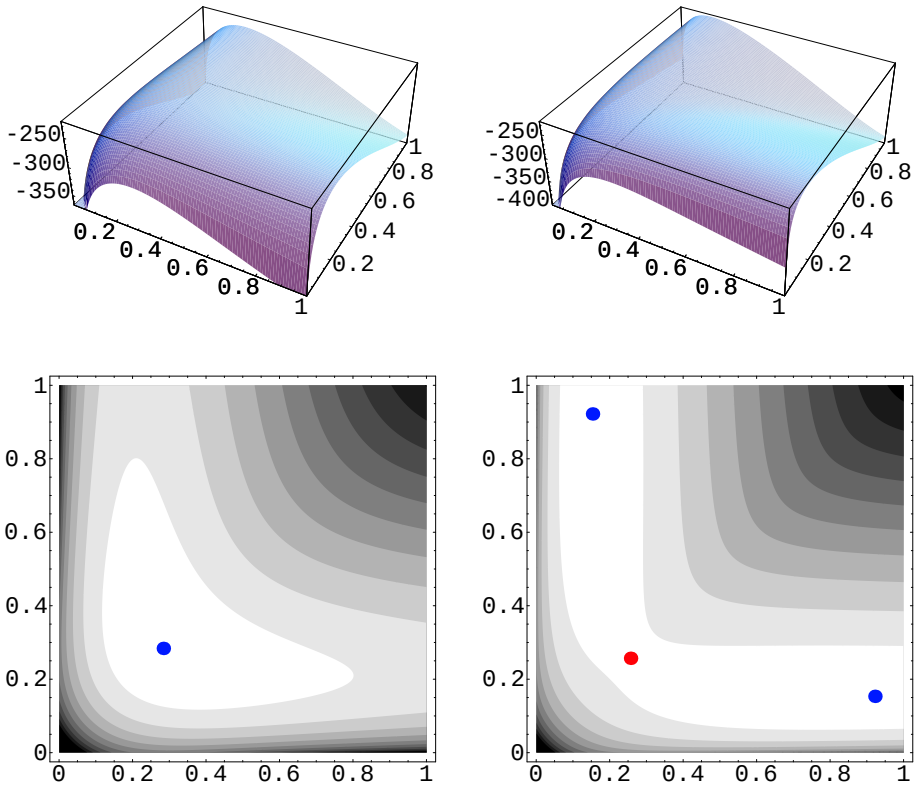


Figure 2.1: Likelihood surface (above) and their respective contours (below) of λ_1 and λ_2 under H_1 for dataset 1 and dataset 5 are plotted

thereby misses the true MLE for dataset 5. Thus the critical value determined by the parametric bootstrap is too close to zero due to the non-rigorous nature of the local searches. The question of interest then is whether we can ever know the true LRTS distribution from a given set of bootstrapped data-sets.

The answer to the above question lies in the ability to enclose the ML value under H_0 and H_1 for each of the B bootstrapped datasets. We pose the model in the following equivalent manner.

$$f(y_j;\Theta) = \pi_1\lambda_1 \exp(\lambda_1 y_j) + \pi_2\lambda_2 \exp(\lambda_2 y_j), \text{ for } \pi_1 = \frac{w_1}{w_1+w_2}, \pi_2 = \frac{w_2}{w_1+w_2}$$

Now, the real-valued negative log likelihood function is extended to the space of triples $(\, F(x), \nabla F(x), \nabla^2 F(x)\,)$ as it is twice differentiable and its natural interval extension is well-defined. Then we operate in interval-extended Hessian differentiation arithmetic to enclose the MLE as well as the LRTS in intervals. The enclosure of the MLEs is given below for the dataset $\mathcal{D}$.

1. When the weights are known and equal we get two solutions by non-identifiably posing the problem. Note that the interval methods will enclose all the maximizers inside the starting search box that is given to the global optimization algorithm.

Solution 1 :	Solution 2 :
$\widehat{\lambda_1} \subset 1.01835517318^{2}_{0}$	$\widehat{\lambda_1} \subset 0.163926463752796^{6}_{4}$
$\widehat{\lambda_2} \subset 0.163926463752^{8}_{7}$	$\widehat{\lambda_2} \subset 1.0183551731806^{5}_{2}$
$\hat{\ell} \subset -214.92234835201^{1}_{4}$	$\hat{\ell} \subset -214.92234835201^{1}_{4}$

2. Now we assume that the weights are possibly distinct.

$$\begin{aligned}
\widehat{\lambda_1} &\subset 0.16638506096^{3}_{1} \\
\widehat{\lambda_2} &\subset 1.07381585708^{8}_{5} \\
\widehat{w_2} &\subset 0.925011553506^{9}_{6} \\
\hat{\ell} &\subset -213.5185156105^{5}_{6}
\end{aligned}$$

Since our global optimization algorithm returns intervals that contain the MLEs under H_0 and H_1 for each one of the B bootstrapped datasets, the LRTS obtained for the B datasets are a sequence of intervals $\{X_i\}_1^B$. Thus we need to enclose the empirical cumulative distribution function (ECDF) $\hat{f}_{\{x_i\}_1^B}(x)$,

$$\hat{f}_{\{x_i\}_1^B}(x) := \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}_{x_i \leq x},$$

of every sequence in the set

$$\{\{x_i\}_1^B : x_i \in X_i = [\underline{x}_i, \overline{x}_i], \forall\, i \in \{1, \ldots, B\}\}.$$

To avoid notational clutter we will use $\{x_i\}$ for $\{x_i\}_1^B$ and $\{X_i\}$ for $\{X_i\}_1^B$. We can easily enclose the CDFs of the above set of sequences by noting that they are contained in the lower and upper ECDFs

$$\begin{aligned}
\hat{f}_{\{\underline{x}_i\}}(x) &:= \tfrac{1}{B}\textstyle\sum_{i=1}^{B} \mathbf{1}_{\underline{x}_i \leq x}, \text{ and} \\
\hat{f}_{\{\overline{x}_i\}}(x) &:= \tfrac{1}{B}\textstyle\sum_{i=1}^{B} \mathbf{1}_{\overline{x}_i \leq x}.
\end{aligned}$$

Let us recall that the Kolmogorov-Smirnov metric between $\hat{f}, \hat{g}$ is given by

$$KS(\hat{f}, \hat{g}) := \sup_x |\hat{f}(x) - \hat{g}(x)|.$$

By slightly abusing the interval notation, we refer to the set of ECDFs obtained with a sequence of intervals $\{X_i\} = \{\underline{x}_i, \overline{x}_i\}, \forall\, i \in \{1, \ldots, B\}$ by

$$\hat{F}_{\{X_i\}_1^B} = [\hat{f}_{\{\underline{x}_i\}}, \hat{f}_{\{\overline{x}_i\}}(x)].$$

Once again we use the shorthand notation $\hat{F}_{\{X_i\}}$ for $\hat{F}_{\{X_i\}_1^B}$ unless necessary. Now consider the collection of such compact sets of ECDFs

$$\mathbb{IF} := \{\hat{F}_{\{X_i\}} : X_i \in \mathbb{IR}, \forall\, i \in \{1, \ldots, B\}\}$$

Let $\hat{F}_{\{X_i\}}$ and $\hat{G}_{\{Y_i\}}$ be elements of $\mathbb{IF}$, then we can define the Hausdorff metric between them as

$$\mathfrak{h}_{KS}(\hat{F}_{\{X_i\}}, \hat{G}_{\{Y_i\}}) := \max\{KS(\hat{f}_{\{\underline{x}_i\}}, \hat{g}_{\{\underline{y}_i\}}), KS(\hat{f}_{\{\overline{x}_i\}}, \hat{g}_{\{\overline{y}_i\}})\}.$$

This makes $\mathbb{IF}$ a complete metric space and allows us enclose the limiting $\hat{F}$ by applying the Gilvenko-Cantelli Theorem (cf. [10]) to the the lower and upper ECDFs, i.e.,

$$\lim_{B\to\infty} \mathfrak{h}_{KS}(\hat{F}_{\{X_i\}_1^B}, \hat{F}) = 0.$$

Recall that this argument is identical in spirit to that used to enclose the sample mean of i.i.d. samples contained within interval-valued observations X_i by applying the law of large numbers to the lower and upper bounds of the intervals. Figure 2.2 shows the ECDF of the LRTS for the dataset $\mathcal{D}$ under the null hypothesis H_0 from bootstrapped datasets of size 100 (magenta), 500 (azure), and 1000 (blue). The ECDFs are actually interval valued enclosures of the true ECDFs. They appear as lines at the resolution of Figure 2.2. However, when we blow up the image and look closely at zero in Figure 2.3, we can see that the actual ECDFs for each level of bootstrap replications are enclosed by its corresponding lower and upper ECDFs. Thus this method also allows us to study the accumulation of "Point-mass" at 0 for the finite sample size of $n = 100$ for our dataset $\mathcal{D}$.

Next we produce the enclosures of the critical value C_α for an α-level test for the dataset $\mathcal{D}$ with finite sample size n from B bootstrap replicates. Recall that $C_{\alpha=0.05}$ from non-rigorous estimates based on Quasi-Newton ranged in $[0.0001, 2.54]$.

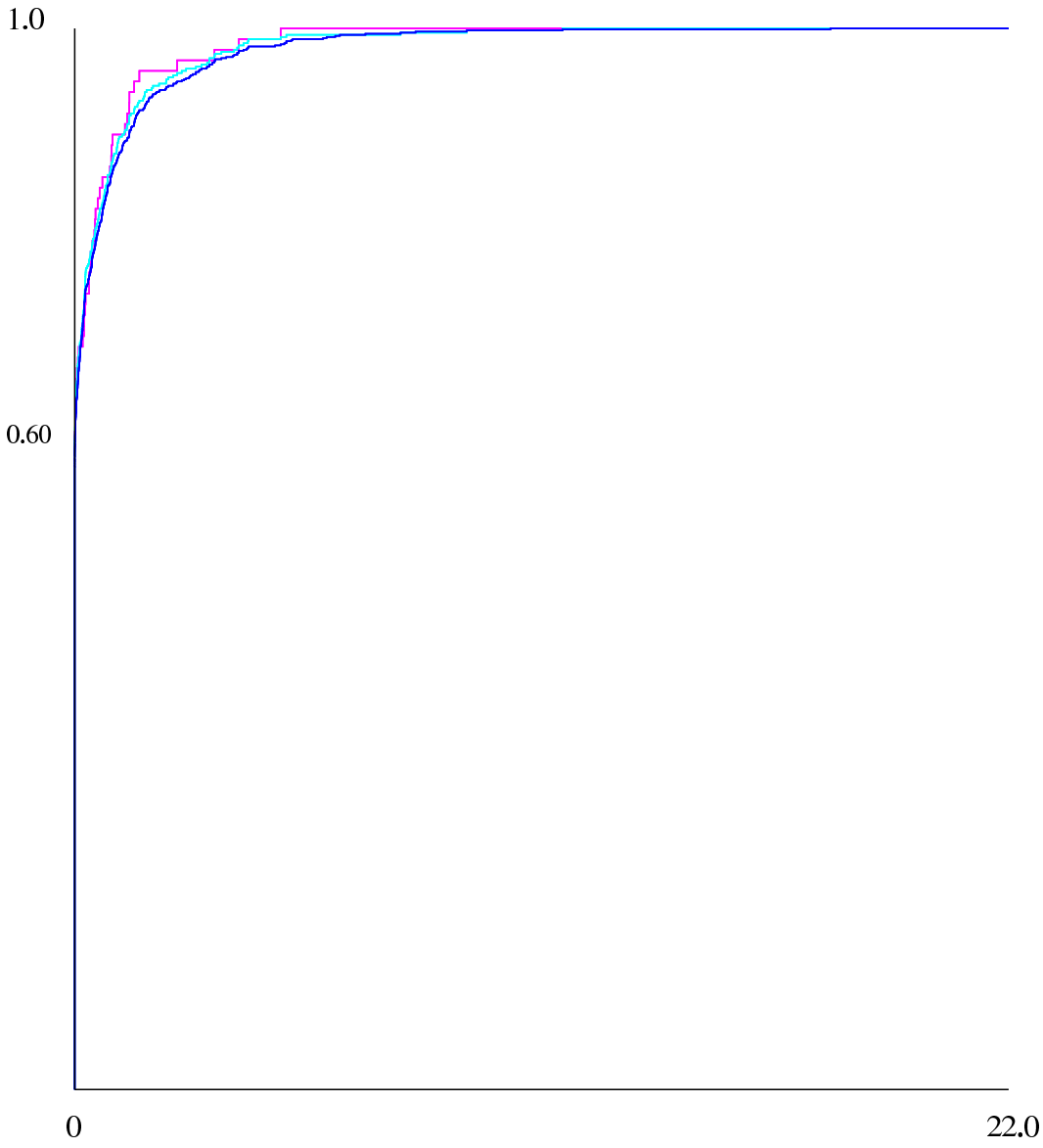


Figure 2.2: Enclosing the empirical CDF of the LRTS for dataset $\mathcal{D}$ under H_0 from bootstrapped datasets of size 100 (magenta), 500 (azure), and 1000 (blue).

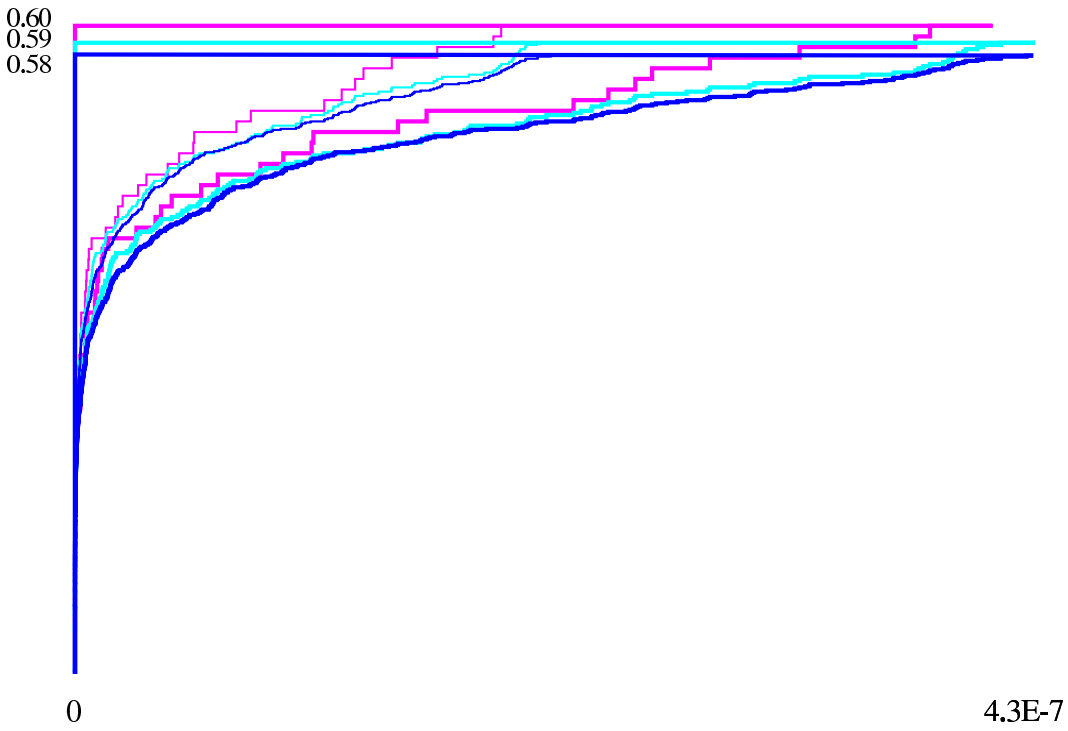


Figure 2.3: Accumulation of "Point-mass" at 0 for sample size $n = 100$ and bootstrap replicates $B = 100$ (magenta), $B = 500$ (azure), and $B = 1000$ (blue). The thin lines are the ECDFs obtained from the midpoints of the interval valued LRTS and the thicker lines are the respective upper and lower ECDFs.

```
-------------------------------------------------------

n = 100   :   alpha = 0.05   :-

-------------------------------------------------------


B = 100   C is in [1.52284897140 , 1.5228493000]

B = 500   C is in [2.16059526182 , 2.1605952618]

B = 1000  C is in [2.52103675001 , 2.5210369663]

-------------------------------------------------------
```

Although we study one of the simplest cases here, similar problems of sensitivity of the critical values to the heuristic optimization algorithms arise in mixtures of more complex models as well as mixtures of other univariate and multivariate densities. For example, when the parametric bootstrap to obtain $C_{\alpha=0.05}$ was based on EM algorithm with different starting and stopping rules, to test the number of components in this exponential mixture problem with unknown rates and unknown weights, $C_{\alpha=0.05}$ was found to range in $[2.55, 4.99]$ [4]. When the dimension of the parameter space gets larger, one can imagine more complicated dependence of the bootstrapped critical values on the particularities of local searches.

Other natural applications include mixtures of multivariate Gaussian distributions as "interval classifiers". This method is particularly suited for certain types of massive data ($\{x_i\}_1^n$) problems, where each observation $x_i \in \mathbb{R}^m$, with $m \approxeq 10$ and $n \approxeq 10^{10}$. Also this method immediately allows for the observations to be intervals and therefore can naturally model the physical limits on empirical resolution of observations in the finite-dimensional continuum.

2.2 Enclosing the Most Likely Trees

A general procedure that guarantees to solve for the maximum likelihood estimates (MLEs) of the branch-lengths of a phylogenetic tree with a fixed topology does not exist. Thus, even for small trees with three and four taxa, one is not certain whether the MLEs obtained from heuristic local searches are indeed the actual MLEs. An interval extension of the recursive formulation for the likelihood function of the simplest Markov model of DNA evolution on unrooted phylogenetic trees with a fixed topology is used to obtain rigorous enclosure(s) of the global maximum likelihood value and the maximum likelihood estimates of all branch lengths of the tree. The algorithm is an adaptation of a widely applied global optimization method using interval analysis. Using such an interval-extended likelihood function we rigorously enclose the MLEs of three and four taxa trees based on primate mitochondrial DNA sequences.

2.2.1 The most likely phylogenetic tree problem

When one is given a homologous set of distinct deoxyribonucleic acid (DNA) sequences of length v from s species and asked for an estimate of their interrelationships back through time under some model of DNA evolution, a phylogenetic tree estimation problem arises. This problem is two-fold. First, one has to estimate the shape or topology of the tree, which captures the set of "who is related to whom and in what order? and whose ancestors are related to whose and in what order?"questions. Second, one has to estimate the lengths of the branches when given a particular topology. The branch lengths of a tree usually represent a scaled product of mutation rate and number of generations between the nodes.

The s extant species are represented by the external nodes or leaves and their ancestors are represented by the internal nodes of the tree. A rooted tree always has a bifurcation at the root, typically the most recent common ancestor of all s leaves, where as, an unrooted tree has m-furcations at all internal nodes with $m \geq 3$. This section focuses on the second problem, namely, estimating the branch lengths for a given topology in a maximum likelihood framework.

When statistical inference is conducted in a maximum likelihood (ML) framework, one is interested in the global maximum of the likelihood function over the parameter space. Explicit analytical solutions for the maximum likelihood estimates of the branch lengths for a specified unrooted topology with more than 2 leaves are not available even for the simplest model of DNA evolution [20] without assuming a molecular clock. See 5.(b) of [42] for results on clocked 3-leaved rooted trees. Results are known for models with two character states superimposed on 3-leaved trees [42], as well as for specific observations on 4-leaved trees [6].

In practice one settles for a local optimization algorithm to numerically approximate the global solution. However, statistical inference procedures that rely on having found some global optimum through any numerical approach may suffer from the five major sources of errors described earlier.

The global optimization method [13] sketched below rigorously encloses the global maximum of the likelihood function through interval analysis [30]. Such interval methods, in contrast to local search methods with floating-point arithmetic, evaluate the likelihood function over a continuum of points including those that are not machine-representable and account for all sources of errors described earlier. We first enclose the likelihood function over a compact set of trees. The global optimization algorithm described in section 1.2.5 is finally applied to enclose the

MLEs for the branch lengths of phylogenetic trees with two, three and four leaves.

2.2.2 Enclosing the log likelihood over a box of trees

Let $\mathcal{D}$ denote a homologous set of distinct DNA sequences of length v from n species. We want the maximum likelihood estimates of branch lengths for the most likely tree under a particular topology. Recall that the branch lengths usually represent a scaled product of mutation rate and number of generations. Let b denote the number of branches and s denote the number of nodes of a tree with topology τ. Thus, for a given unrooted topology τ with n leaves and b branches, the unknown parameter $\theta = (\,\theta_1, \cdots, \theta_b\,)$ is the real vector of branch lengths in the positive orthant ($\theta_q \in \mathbb{R}_+$). An explicit model of DNA evolution is needed to construct the likelihood function which gives the probability of observing data $\mathcal{D}$ as a function of the parameter θ. The simplest such continuous time Markov chain model (JC69) on the nucleotide state space $\Sigma := \{A, G, C, T\}$ is due to [20].One may compute $\ell^{(k)}(\theta)$, the log likelihood at site $k \in \{1, \cdots, v\}$, through the following post-order traversal [11]:

1. Associate with each node $q \in \{1, \cdots, s\}$ with m descendants, a partial likelihood vector, $\mathbf{l}_q := (\mathbf{l}_q^A, \mathbf{l}_q^C, \mathbf{l}_q^G, \mathbf{l}_q^T) \in \mathbb{R}^4$, and let the length of the branch leading to its ancestor be θ_q.

2. For a leaf node q with nucleotide i, set $\mathbf{l}_q^i = 1$ and $\mathbf{l}_q^j = 0$ for all $j \neq i$. For any internal node q, set $\mathbf{l}_q := (1, 1, 1, 1)$.

3. For an internal node q with descendants $s_1, \cdots, s_m$,

$$\mathbf{l}_q^i = \sum_{j_1, \cdots, j_m \in \Sigma} \{\, \mathbf{l}_{s_1}^{j_1} \cdot P_{i,j_1}(\theta_{s_1}) \cdot \ \cdots \ \cdot \mathbf{l}_{s_m}^{j_m} \cdot P_{i,j_m}(\theta_{s_m}) \,\}$$

4. Compute $\mathbf{l}_q$ for each sub-terminal node q, then those of their ancestors recursively to finally compute $\mathbf{l}_r$ for the root node r and obtain the log likelihood for site k, $\ell^{(k)}(\theta) = \log \sum_{i \in \Sigma} (\pi_i \cdot \mathbf{l}_r^i)$.

Assuming independence across sites one obtains $\ell(\theta) = \sum_{k=1}^{v} \ell^{(k)}(\theta)$, the natural logarithm of the likelihood function for the data $\mathcal{D}$ by multiplying the site-specific likelihoods. The problem of finding the global maximum of this likelihood function is equivalent to finding the global minimum of $l(\theta) := -\ell(\theta)$. Replacing θ, a positive real vector of branch lengths, in the above algorithm by a positive real interval vector or box Θ and all real operations by their interval counterparts, yields $\mathcal{L}(\Theta)$, the natural interval extension of the negative log likelihood function $l(\theta)$ over Θ. Since $\nabla\mathcal{L}(\Theta)$ and $\nabla^2\mathcal{L}(\Theta)$, the enclosures of the gradient and the Hessian of $l(\theta)$ over Θ, respectively, are needed in Section 1.2.5, one may use the constant triples, $(C, 0, 0)$, variable triples, $(\Theta_j, e^{(j)}, 0)$, appropriate triples for the elementary functions, exp and log, and perform all operations in the interval differentiation arithmetic of Section 1.2.2, in order to obtain the negative log likelihood triple $(\mathcal{L}(\Theta), \nabla\mathcal{L}(\Theta), \nabla^2\mathcal{L}(\Theta))$.

2.2.3 Examples

Example 1 *Unrooted 3-leaved Tree*

The global maximum of the log likelihood function for the JC69 model of DNA evolution on the three taxa unrooted tree with data from the mitochondria of Chimpanzee, Gorilla, and Orangutan [5] is enclosed. There is only one unrooted multifurcating topology for three species with all three branches emanating from the root of the star tree τ_*. This data set has 29 data patterns.

Table 2.1: Chimpanzee (1), Gorilla (2), and Orangutan (3).

$\Theta^{(0)}$, Tree, $-\mathcal{L}(\Theta^*) \supset -l(\theta^*)$	$\Theta^* \supset \theta^*$
$[1.0 \times 10^{-11}, 10.0]^{\otimes 3}$	$5.9816221384_0^2 \times 10^{-2}$
$\tau_* = (1,2,3)$	$5.4167416794_0^2 \times 10^{-2}$
$-2.150318065856_6^5 \times 10^3$	$1.3299089685_8^9 \times 10^{-1}$

```
PATTERN COUNTS :
232 71 229 168 13 31 16 18 9 20 1 8 22 3 10 8 7 1 9 2 4 2 1 2 1 1 2 1 3
PATTERNS:
agctatcacccatctgccgtactaagcgt
agctgttatcaacacgcaaaatccggtat
agctaccgttcccataataataaagcgca
```

The parameter space is three dimensional corresponding to the three branch lengths of the 3-leaved star tree τ_1. The algorithm is given a large search box $\Theta^{(0)}$. The results are summarized in Table 2.1. Recall that the notation x_a^b means the interval $[xa, xb]$, e.g. $5.9816221384_0^2 \times 10^{-2} = [5.98162213840 \times 10^{-2}, 5.98162213842 \times 10^{-2}]$. Figure 2.4 shows the the parameter space being rigorously pruned as the algorithm progresses according to section 1.2.5.

Example 2 *Four Unrooted 4-leaved Trees*

By adding the homologous mitochondrial sequence from Gibbon to the previous problem, one obtains a simple phylogeny estimation problem with 61 data patterns [5].

```
PATTERN COUNTS :
```

```
209 71 192 157 28 5 11 20 2 10 10 15 5 15 1 5 1 2 15 3 14 3 4 2 5 4 5
4 9 2 4 5 1 1 7 6 3 1 1 4 2 3 3 1 1 2 3 1 1 1 4 1 1 1 1 1 1 1 1 2 1
PATTERNS:
agctccatatcacctacaaatccatatctgtccgggattacccctatttacgcgtcgcttc
agctccgtgttatctaaaaacatacacctgccaaagaatactttcccatgcgtattacctt
agctccacaccgttcaccaacatctccctatataaagattaccaaacatgtcgcagattaa
agcttaacgtcaccactcgtcatgcatgcgtcctaagtgaatacaaaaactgaaagcaata
```

Four topologies need to be considered for a tree with four leaves. The star tree τ_* has all four lineages coalescing at the same time, while the other three trees have an additional parameter θ_5 representing the only internal branch length. They differ due to the order in which the leaves relate to one another as shown in Table 2.2. The algorithm is given a large search box $\Theta^{(0)}$ for each topology and the results are summarized in Table 2.2. Within each one of the four topologies there exists a unique global maximum. However, the global maximizer over all five topologies falls under topology τ_2 with the global maximum $-l^*$ contained in the interval $-\mathcal{L}^* = -2.656936470946^5_6 \times 10^3$.

Example 3 *Enclosing non-identifiable subspaces*

For time reversible Markov chains, such as JC69, evolving on a rooted tree, only the sum of the branch lengths emanating from the root is identifiable. Identifiability is a prerequisite for statistical consistency of estimators. To demonstrate the ability of interval methods, unlike the local search methods, to enclose the non-identifiable ridge along $\theta_1+\theta_2$, in the simplest case of a 2-leaved tree, a non-identifiable negative log likelihood function $l(\theta)$ is formulated and its global minimizers along $\theta_1 + \theta_2 = \frac{3}{4}\log(45/17) = 0.730087$ are enclosed as shown in Figure 2.5 for a fictitious

Table 2.2: Chimpanzee (1), Gorilla (2), Orangutan (3), and Gibbon (4)

$\Theta^{(0)}$, Tree, $-\mathcal{L}(\Theta^*) \supset -l(\theta^*)$	$\Theta^* \supset \theta^*$
$[1.0 \times 10^{-11}, 10.0]^{\otimes 4}$	$6.57882493333^{5}_{4} \times 10^{-2}$
$\tau_1 = (1, 2, 3, 4)$	$6.2361625124038^{8}_{2} \times 10^{-2}$
	$1.3248749022485^{5}_{4} \times 10^{-1}$
$-2.70274345019641^{1}_{4} \times 10^{3}$	$1.6359125624763^{4}_{3} \times 10^{-1}$
$[1.0 \times 10^{-11}, 10.0]^{\otimes 5}$	$4.962819343269^{9}_{8} \times 10^{-2}$
	$5.899264246908^{8}_{7} \times 10^{-2}$
$\tau_2 = ((1, 2), (3, 4))$	$5.518490773874^{4}_{3} \times 10^{-2}$
	$9.097140075962^{3}_{2} \times 10^{-2}$
$-2.6569364709466^{5}_{6} \times 10^{3}$	$1.2315160183102^{2}_{1} \times 10^{-1}$
$[1.0 \times 10^{-11}, 10.0]^{\otimes 5}$	$9.07177047^{7}_{6} \times 10^{-3}$
	$6.142391114^{4}_{3} \times 10^{-2}$
$\tau_3 = ((1, 3), (2, 4))$	$1.2963838225^{5}_{4} \times 10^{-1}$
	$5.6506921813^{3}_{0} \times 10^{-2}$
$-2.699878136170^{0}_{5} \times 10^{3}$	$1.600054316565^{5}_{1} \times 10^{-1}$
$[1.0 \times 10^{-11}, 10.0]^{\otimes 5}$	$1.1495164302969^{9}_{5} \times 10^{-2}$
	$5.825806134318^{8}_{7} \times 10^{-2}$
$\tau_4 = ((1, 4), (2, 3))$	$1.5888166092523^{3}_{1} \times 10^{-1}$
	$5.7069581801999^{9}_{2} \times 10^{-2}$
$-2.69855862854055^{5}_{9} \times 10^{3}$	$1.2932141694891^{1}_{0} \times 10^{-1}$

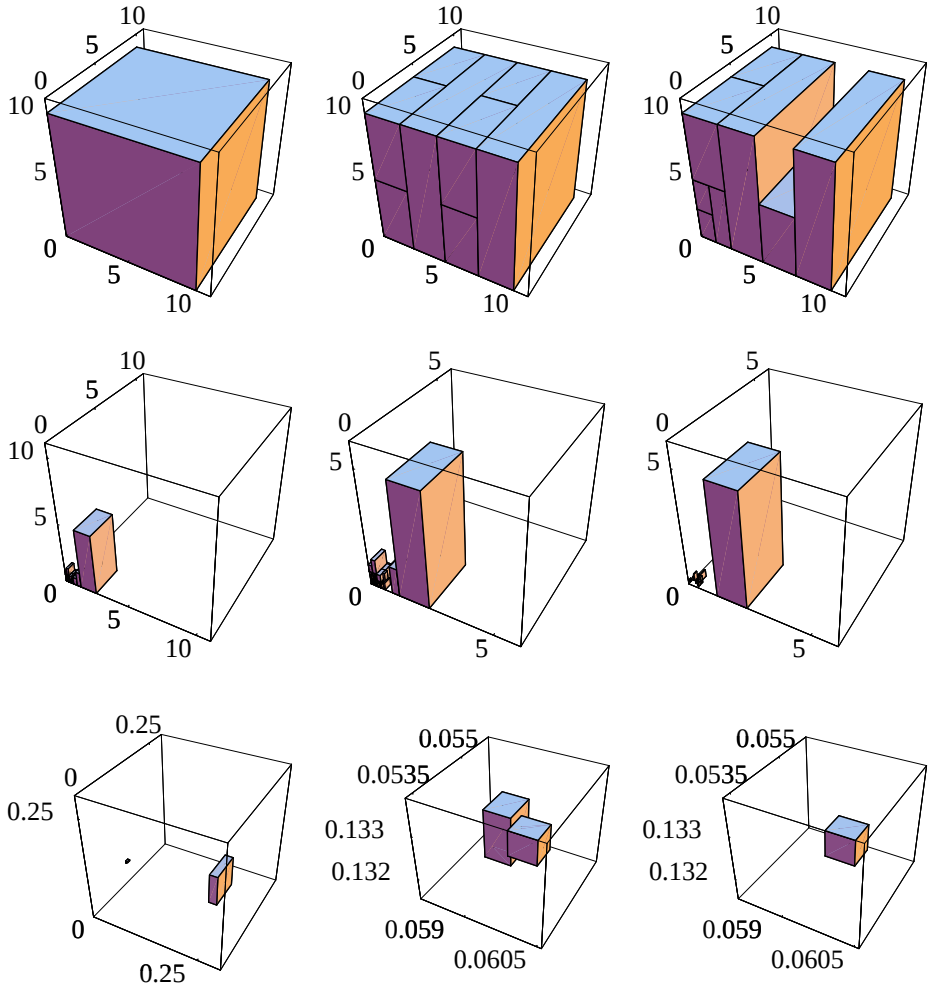


Figure 2.4: Progress of the algorithm (left to right starting from top row) as it prunes $[0.001, 10.0]^{\otimes 3}$.

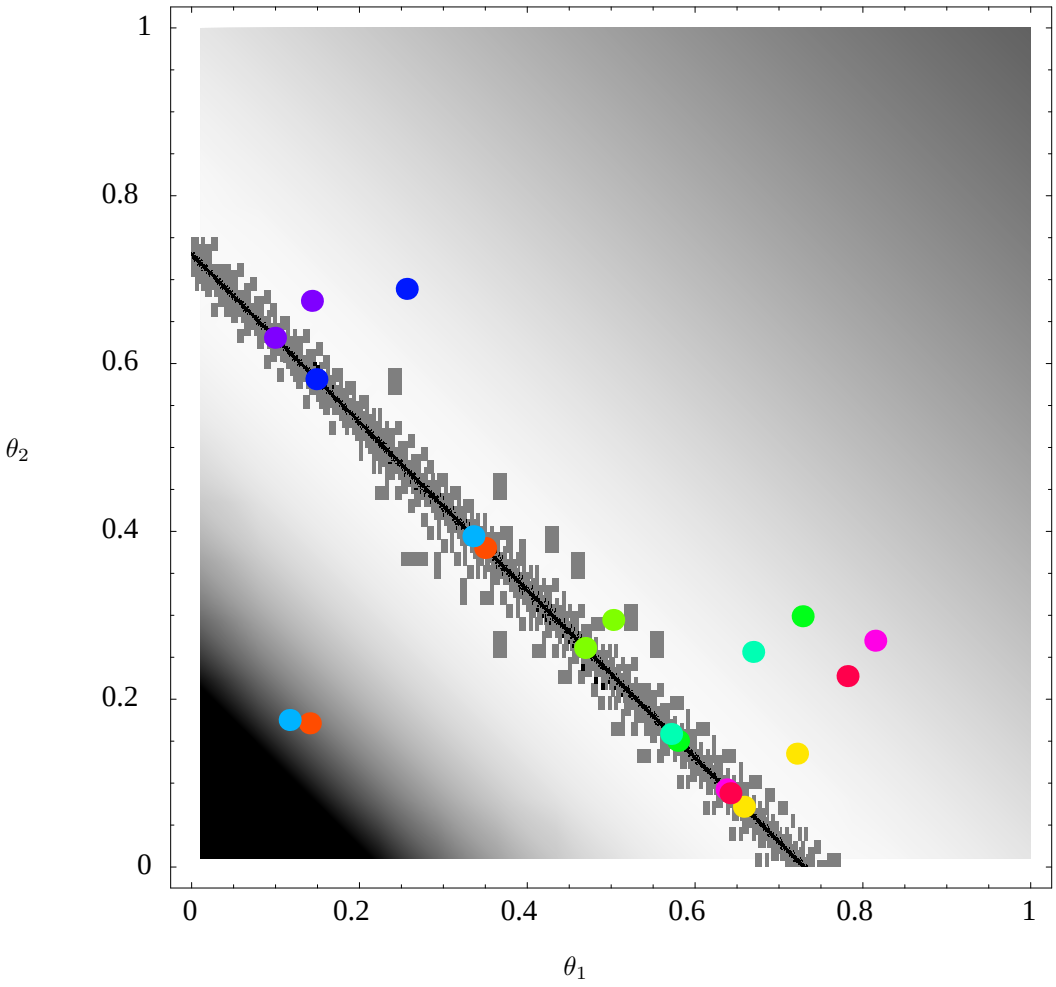


Figure 2.5: For a pair of homologous sequences of 600 nucleotides out of which 280 sites are polymorphic, the non-identifiable subspace of minimizers $\theta_1 + \theta_2 = \frac{3}{4}\log(45/17) = 0.730087$ of the negative log likelihood function under the JC69 model evolving on a rooted two-leaved tree is enclosed by a union of up to 30,000 boxes. The larger gray, and smaller black boxes have tolerances of $\epsilon = 1.0 \times 10^{-4}$ and $\epsilon = 1.0 \times 10^{-6}$, respectively. The 10 pairs of colored circles are the initial and final points of 10 Quasi-Newton searches with random initializations.

dataset for which 280 out of 600 sites were polymorphic. Observe that the basin of attraction for each point on $\theta_1+\theta_2 = 0.730087$ under the Quasi-Newton local search algorithm is the line running orthogonal to it. This trivial example is only chosen for pedantic reasons. Enclosing possibly non-identifiable sub-manifolds, that may not even be simply connected, within any compact subset of higher dimensional parameter spaces, may be accomplished, at least partly, by studying the rates of decay of the hyper-volume of the union of all pending boxes as the algorithm progresses, for instance.

2.2.4 Discussion

Modifications of this algorithm are also applicable to Markov models with unknown parameters under constraints [14], even in the absence of analytical spectral decompositions [27]. The running time of the algorithm on phylogenetic trees is sensitive to the shape of the likelihood function. In general, flatter likelihood functions due to small sample sizes have longer convergence times [38]. One can perform *rigorous generalized neighbor joining* [38] to construct phylogenetic trees with many leaves in polynomial time. The algorithm can enclose all the maximizers even in a non-identifiable model and is therefore robust to *non-identifiably-posed models*. Several efficiency increasing steps could be taken. Pre-enclosing the transition probabilities and accessing them through hash functions can save computational effort. Asynchronous parallelization of the algorithm across 6 processors is also observed to increase the rate of convergence to the global maximum. It also provides a natural framework to manage the memory requirements for larger trees through partial likelihood evaluations for non-overlapping subtrees in parallel, prior to obtaining the full likelihood. Simpler versions of the algorithm can be carried out without

automatic differentiation by means of range enclosures. Finally, it is worth noting that inclusion isotony does indeed hold by the continuity of the likelihood function in the CAT(0) space of trees [3], and thus in conjunction with interval analysis may be made to provide a rigorous numerical framework for global maximization of the likelihood over compact sets containing distinct topologies. Preliminary results indicate that efficiency increases when one starts with a disjoint union of compact subsets of branch lengths from finitely many topologies, i.e, $\Theta^{(0)} = \cup_i \Theta^{(0,\tau_i)}$ and simultaneously prunes away sub-boxes from distinct $\Theta^{(0,\tau_i)}$ with a variant of the above algorithm that allows for compact sets contained in each $\Theta^{(0,\tau_i)}$ with its corresponding τ_i-specific likelihood function. Thus, interval methods may be able to enclose the global maximum more efficiently when several topologies are considered simultaneously than when the global maximum is enclosed for each member of a finite set of topologies, one at a time, and finally compared.

The ability to enclose the range, gradient, and Hessian of the log likelihood function over a compact set of trees, not only allows for the rigorous estimation of the Fisher Information at the MLE, but also provides a natural framework for efficiently sampling branch-lengths in the continuous part of the parameter space of several Monte Carlo algorithms on tree spaces. For instance, one may use enclosures of the likelihood function to produce an almost optimal importance sampler over an adaptively partitioned compact box of branch-lengths. In this section a general procedure has been provided to rigorously enclose the maximum likelihood value, as well as the most likely branch lengths of a tree (with a specified topology) upon which the simplest Markov model of DNA evolution is superimposed. This partly addresses an open problem in phylogenetics [39] (p. 207).

2.3 Randomized Enclosure Algorithms

In Bayesian statistical inference and computationally intensive frequentist inference, one is interested in obtaining samples from $p(\theta) := p^*(\theta)/N_p$, a high dimensional, and possibly multi-modal density, where $\theta \in \mathbf{\Theta} \Subset \mathbb{R}^n$. The challenge is to obtain samples from p without any knowledge of the normalizing constant $N_p := \int_{\mathbf{\Theta}} p^*(\theta)\, \partial\theta$. Several approaches to this problem rely on Monte Carlo methods. We will concentrate on (i) the rejection sampler due to John Von Neumann [33] and (ii) the Metropolis-Hastings (M-H) sampler [29, 16]. We will introduce interval versions of these samplers in honor of Ramon E. Moore who was one of the influential founders of interval analysis.

2.3.1 Rejection Sampler

Rejection sampling is a Monte Carlo method to draw independent samples from a probability distribution $p(\theta) := p^*(\theta)/N_p$, where $\theta \in \mathbf{\Theta} \Subset \mathbb{R}$. In general p is any density that is typically dominated by the counting measure or the Lebesgue measure. We are under the constraint that the normalizing constant N_p is unknown. However, we can compute $p^*(\theta)$ for any $\theta \in \mathbf{\Theta}$.

Rejection sampling Algorithm

1. Choose a proposal density $q(\theta) = q^*(\theta)/N_q$ from which independent samples can be drawn and $q^*(\theta)$ is computable for any $\theta \in \mathbf{\Theta}$.

2. Find some proposal function $f(x)$, such that the inequality

$$f(q(\theta)) \geq p^*(\theta), \forall\, \theta \in \mathbf{\Theta}$$

is satisfied. Usually, a linear form for f, i.e., $f(q(\theta)) = cq^*(\theta)$, is chosen and then one tries to find the smallest c for which the inequality

$$f(q(\theta)) = cq^*(\theta) \geq p^*(\theta), \forall\, \theta \in \mathbf{\Theta}$$

is satisfied. This smallest possible value of c when f takes a linear form is said to be optimal and denoted by $\hat{c}$, i.e.,

$$\hat{c} := \inf\{c : cq^*(\theta) \geq p^*(\theta), \forall\, \theta \in \mathbf{\Theta}\}.$$

3. Given (i) a target shape $p^*(\theta)$, (ii) a proposal density $q(\theta)$, and (iii) a proposal function $f(q(\theta))$, $\forall\, \theta \in \mathbf{\Theta}$, that satisfy the above conditions, we can draw independent samples from the target $p(\theta)$ as follows:

   (a) GENERATE $T \sim q$.

   (b) DRAW $H \sim Uniform[0, f(q(T))]$, where $f(q(T)) \geq p^*(T)$.

   (c) IF $H \leq p^*(T)$, THEN accept T AND set $U \leftarrow T$,

   (d) IF $H > p^*(T)$, THEN reject AND return to Step 3a.

See [26] for a proof that U generated by the above algorithm is distributed according to p, provided $f(q(\theta)) = cq^*(\theta)$. Observe that the probability $\mathbf{A}^p_{f(q)}$ that a point proposed according to q gets accepted as an independent sample from p is the ratio of the integrals

$$\mathbf{A}^p_{f(q)} = \frac{N_p}{N_{f(q)}} := \frac{\int_{\mathbf{\Theta}} p^*(\theta)\, \partial\theta}{\int_{\mathbf{\Theta}} f(q(\theta))\, \partial\theta}.$$

Therefore, for a given p we want to minimize $N_{f(q)}$ over the allowed possibilities for q and $f(q)$, since the probability distribution over the number of samples from q to obtain one sample from p is geometrically distributed with mean $1/\mathbf{A}^p_{f(q)}$.

2.3.2 Moore-Rejection Sampler

We introduce the Moore-rejection samplers as a class of rejection samplers that can be applied to target a density over a compact domain with a well-defined interval extension. These samplers obtain the possible q's and $f(q)$'s needed to optimize a rejection sampler through enclosure methods. We will consider one of the simplest such samplers here to target p in one dimension with the following characteristics:

Compact domain	$\mathbf{\Theta} = [\underline{\theta}, \overline{\theta}]$
Target shape	$p^*(\theta) : \mathbf{\Theta} \to \mathbb{R}$
Target integral	$N_p := \int_{\mathbf{\Theta}} p^*(\theta)\, \partial\theta$
Target density	$p(\theta) := \frac{p^*(\theta)}{N_p} : \mathbf{\Theta} \to \mathbb{R}$
Interval extension of p^*	$P^*(\Theta) : \mathbb{I}\mathbf{\Theta} \to \mathbb{IR}$

If $p^* \in \mathfrak{E}$, the class of elementary functions, and the interval extension of P^* is well-defined on $\mathbf{\Theta}$ then by Theorem 4

$$Rng(p^*; \mathbf{\Theta}) \;=:\; p^*(\mathbf{\Theta}) \;\subseteq\; P^*(\mathbf{\Theta}) \;=:\; [\underline{P}^*(\mathbf{\Theta}), \overline{P}^*(\mathbf{\Theta})]$$

which implies that

$$\underline{P}^*(\mathbf{\Theta}) \;\leq\; p^*(\theta) \;\leq\; \overline{P}^*(\mathbf{\Theta}), \quad \forall\, \theta \in \mathbf{\Theta}$$

Bearing in mind that $[\underline{P}^*(\mathbf{\Theta}), \overline{P}^*(\mathbf{\Theta})]$ is often a terrible over-estimate of the range $p^*(\mathbf{\Theta})$, we may nonetheless construct a naive Moore-rejection sampler for p with

$$\begin{aligned} q(\theta) &= \frac{\overline{P}^*(\mathbf{\Theta})}{d(\mathbf{\Theta}) \cdot \overline{P}^*(\mathbf{\Theta})} = (d(\mathbf{\Theta}))^{-1}, \text{ and} \\ f(q(\theta)) &= \overline{P}^*(\mathbf{\Theta}), \end{aligned}$$

where, $d(\mathbf{\Theta}) = d([\underline{\theta}, \overline{\theta}]) = \overline{\theta} - \underline{\theta}$ is the diameter of $\mathbf{\Theta}$. We can crudely bound the acceptance probability $\mathbf{A}^p_{\overline{P}^*(\mathbf{\Theta})}$ for this naive sampler from below

$$\mathbf{A}^p_{\overline{P}^*(\mathbf{\Theta})} = \frac{N_p}{N_{\overline{P}^*(\mathbf{\Theta})}} = \frac{N_p}{d(\mathbf{\Theta}) \cdot \overline{P}^*(\mathbf{\Theta})} \geq \frac{d(\mathbf{\Theta}) \cdot \underline{P}^*(\mathbf{\Theta})}{d(\mathbf{\Theta}) \cdot \overline{P}^*(\mathbf{\Theta})}$$

Although this naive rejection sampler is extremely inefficient for targets whose shapes are far from being a constant function on $\mathbf{\Theta}$, one has the assurance that

$$f(q(\theta)) = \overline{P}^*(\mathbf{\Theta}) \geq p^*(\theta),\ \forall\, \theta \in \mathbf{\Theta},$$

a necessary condition for the rejection sampler. So one can use the constant $\overline{P}^*(\mathbf{\Theta})$ as a guaranteed upper bound for any parametric family of $f(q(\theta))$, including unimodal parametric families of the form $cq^*(\theta)$ which are commonly used in standard rejection sampling.

A natural way to improve the efficiency of the rejection sampler for targets with non-constant shapes, including multi-modal shapes, is by partitioning the domain. Let $\mathfrak{T} := \{\, \Theta^{(1)}, \Theta^{(2)}, ..., \Theta^{(|\mathfrak{T}|)} \,\}$ be a finite partition of $\mathbf{\Theta}$. Then by Theorem 4 we can enclose $p^*(\Theta^{(i)})$, the range of p^* over the i-th element of $\mathfrak{T}$, with the well-defined interval extension P^* of p^* over $\mathbf{\Theta}$

$$p^*(\Theta^{(i)}) \subseteq P^*(\Theta^{(i)}) := [\underline{P}^*(\Theta^{(i)}), \overline{P}^*(\Theta^{(i)})],\ \forall\, i \in \{1, 2, ..., |\mathfrak{T}|\}.$$

For the given partition $\mathfrak{T}$ we can construct a proposal $q(\theta) = q^{\mathfrak{T}}(\theta)$ as a normalized simple function over $\mathbf{\Theta}$

$$q^{\mathfrak{T}}(\theta) = \left(N_{q^{\mathfrak{T}}}\right)^{-1} \sum_{i=1}^{|\mathfrak{T}|} \overline{P}^*(\Theta^{(i)})\, \mathbf{1}_{\{\theta\, \in\, \Theta^{(i)}\}} \tag{2.1}$$

where the normalizing constant is obtained from the sum

$$N_{q^{\mathfrak{T}}} := \sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)}) \right).$$

The next ingredient $f(q^{\mathfrak{T}}(\theta))$ for our rejection sampler can simply be

$$f(q^{\mathfrak{T}}(\theta)) = \sum_{i=1}^{|\mathfrak{T}|} \overline{P}^*(\Theta^{(i)})\, \mathbf{1}_{\{\theta\, \in\, \Theta^{(i)}\}} \tag{2.2}$$

Note that our $f(q^{\mathfrak{T}}(\theta))$ satisfies the necessary condition for the rejection sampler

$$f(q^{\mathfrak{T}}(\theta)) \geq p^*(\theta),\ \forall\, \theta \in \mathbf{\Theta} \tag{2.3}$$

Theorem 7 *Suppose that the target shape p^* has a well-defined natural interval extension P^*. If U is generated according to the steps in part 3 of the rejection sampling algorithm, and if the proposal density $q^{\mathfrak{T}}(\theta)$ and the proposal function $f(q^{\mathfrak{T}}(\theta))$ are given by Equations 2.1 and 2.2, respectively, then U is distributed according to the target p.*

Proof: From Equations 2.1 and 2.2 observe that $f(q^{\mathfrak{T}}(t)) = q^{\mathfrak{T}}(t)N_{q^{\mathfrak{T}}}$. Let us define the following two subsets of $\mathbb{R}^2$,

$$\mathcal{B}_q = \{(t,h) : 0 \leq h \leq f(q^{\mathfrak{T}}(t))\}, \text{ and } \mathcal{B}_p = \{(t,h) : 0 \leq h \leq p^*(t)\}.$$

First let us agree that steps 3a and 3b of part 3 of the rejection sampling algorithm produce a pair (T,H) that is uniformly distributed on $\mathcal{B}_q$. We can see this by letting $k(t,h)$ denote the joint density of (T,H) and $k(h|t)$ denote the conditional density of H given $T = t$. Then,

$$k(t,h) = \begin{cases} q^{\mathfrak{T}}(t)\, k(h|t) & \text{if } (t,h) \in \mathcal{B}_q \\ 0 & \text{otherwise .} \end{cases}$$

Since we sample a uniform height h for a given t in Step 3b of the algorithm

$$k(h|t) = \begin{cases} (f(q^{\mathfrak{T}}(t)))^{-1} = (q^{\mathfrak{T}}(t)N_{q^{\mathfrak{T}}})^{-1} & \text{if } h \in [0, f(q^{\mathfrak{T}}(t))] \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$k(t,h) = \begin{cases} q^{\mathfrak{T}}(t)\, k(h|t) = q^{\mathfrak{T}}(t)/(q^{\mathfrak{T}}(t)\, N_{q^{\mathfrak{T}}}) = (N_{q^{\mathfrak{T}}})^{-1} & \text{if } (t,h) \in \mathcal{B}_q \\ 0 & \text{otherwise .} \end{cases}$$

Thus we have shown that the joint density of (T,H) is a uniformly distribution on $\mathcal{B}_q$. The above relationship also makes geometric sense since the volume of $\mathcal{B}_q$

is exactly $N_{q^{\mathfrak{T}}}$. Now, let (T^*, H^*) be an accepted point, i.e., $(T^*, H^*) \in \mathcal{B}_p \subseteq \mathcal{B}_q$. Then, the uniform distribution of (T, H) on $\mathcal{B}_q$ implies the uniform distribution of (T^*, H^*) on $\mathcal{B}_p$. Since the volume of $\mathcal{B}_p$ is N_p, the p.d.f. of (T^*, H^*) is identically $1/N_p$ on $\mathcal{B}_p$ and 0 elsewhere. Hence, the marginal p.d.f. of $U = T^*$ is

$$\begin{aligned} w(u) &= \textstyle\int_0^{p^*(u)} 1/N_p \, \partial h \\ &= 1/N_p \textstyle\int_0^{p^*(u)} 1 \, \partial h \\ &= 1/N_p \textstyle\int_0^{N_p p(u)} 1 \, \partial h, \quad \because p(u) = p^*(u)/N_p \\ &= p(u). \quad \square \end{aligned}$$

Note that the above argument applies even when $\boldsymbol{\Theta} \Subset \mathbb{R}^n$ for $n > 1$. Now, we have all the ingredients to perform a more efficient partition specific Moore-rejection sampling. Once again we can bound the acceptance probability $\mathbf{A}^p_{f(q^{\mathfrak{T}})}$ for this sampler from below

$$\mathbf{A}^p_{f(q^{\mathfrak{T}})} = \frac{N_p}{N_{q^{\mathfrak{T}}}} = \frac{N_p}{\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)}) \right)} \geq \frac{\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \underline{P}^*(\Theta^{(i)}) \right)}{\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)}) \right)}$$

due to the linearity of the integral operator

$$\begin{aligned} N_p &:= \textstyle\int_{\boldsymbol{\Theta}} p^*(\theta) \, \partial\theta \\ &= \textstyle\sum_{i=1}^{|\mathfrak{T}|} \int_{\Theta^{(i)}} p^*(\theta) \, \partial\theta \\ &\in \textstyle\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot P^*(\Theta^{(i)}) \right) \\ &= \left[\textstyle\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \underline{P}^*(\Theta^{(i)}) \right), \textstyle\sum_{i=1}^{|\mathfrak{T}|} \left(d(\Theta^{(i)}) \cdot \overline{P}^*(\Theta^{(i)}) \right) \right] \end{aligned}$$

It is possible to say something more about the lower bound for $\mathbf{A}^p_{f(q^{\mathfrak{T}})}$ by limiting ourselves to target shapes within $\mathfrak{E}_{\mathfrak{L}}$, the Lipschitz class of elementary functions. If $p^* \in \mathfrak{E}_{\mathfrak{L}}$ then we might expect the enclosure of N_p to be proportional to the mesh of the partition $\mathfrak{T}$

$$w = \max_{i \in \{1, \ldots, \mathfrak{T}\}} d(\Theta^{(i)}).$$

Theorem 8 *Let $\mathfrak{T}_W$ be the uniform partition of $\mathbf{\Theta} = [\underline{\theta}, \overline{\theta}]$ into W intervals each of diameter w*

$$\begin{aligned} w &= \tfrac{(\overline{\theta}-\underline{\theta})}{W} \\ \Theta_W^{(i)} &= [\ \underline{\theta} + (i-1)w,\ \underline{\theta} + iw\ ], i = 1, \ldots, W \\ \mathfrak{T}_W &= \{\Theta_W^{(i)}, i = 1, \ldots, W\}. \end{aligned}$$

and let $p^* \in \mathfrak{E}_{\mathfrak{L}}$, then

$$\mathbf{A}^p_{f(q^{\mathfrak{T}_W})} = 1 - \mathcal{O}(1/W)$$

Proof

Then by means of Theorem 5

$$\begin{aligned} d(\Theta_W^{(i)}) = \mathcal{O}(1/W) \quad &\Longrightarrow \quad \mathfrak{h}(\ p^*(\Theta_W^{(i)}), P^*(\Theta_W^{(i)})\ ) = \mathcal{O}(1/W) \\ &\Longrightarrow \quad d(P^*(\Theta_W^{(i)})) = \mathcal{O}(1/W), \qquad \because p^* \in \mathfrak{E}_{\mathfrak{L}} \end{aligned}$$

Therefore

$$\sum_{i=1}^{|\mathfrak{T}_W|} \left(d(\Theta_W^{(i)}) \cdot P^*(\Theta_W^{(i)}) \right) = w \sum_{i=1}^{W} P^* \left([\ \underline{\theta} + (i-1)w,\ \underline{\theta} + iw\ ] \right),$$

and we have

$$d(w \textstyle\sum_{i=1}^{W} P^*(\Theta_W^i)) = \mathcal{O}(1/W) \quad \Longrightarrow \quad \mathbf{A}^p_{f(q^{\mathfrak{T}_W})} = 1 - \mathcal{O}(1/W)$$

Therefore the lower bound for the acceptance probability $\mathbf{A}^p_{f(q^{\mathfrak{T}_W})}$ of the Moore-rejection sampler approaches 1 no slower than linearly with the refinement of $\mathbf{\Theta}$ by $\mathfrak{T}_W$. $\square$

One can also think of this particular family of Moore-rejection samplers constructed from the invoking family of uniform partitions of size W

$$\{\ \mathfrak{T}_W : W = 1, 2, 3, \ldots\ \}$$

as a family of rejection samplers whose proposals descend necessarily from above on p^* in the form of a simple function with a uniform partition over $\mathbf{\Theta}$ at each

refinement level W. More generally, any family of Moore-rejection samplers that construct their proposals with $\overline{P}^*$ from the invoking family of refining partitions

$$\{ \mathfrak{T}_\alpha : \alpha \in \mathcal{A} \}$$

can also be thought of as a family of rejection samplers whose proposals descend necessarily from above on p^* in the form of simple splines whose knots may have more freedom depending on the choice of $\mathcal{A}$. In any case, the acceptance probability approaches 1 at a rate that is no slower than linearly with the mesh.

We can do better, in terms of the rate at which $\mathbf{A}^p_{f(q^{\mathfrak{T}_\alpha})}$ approaches 1 in two major pathways:

1. Better family of refining partitions $\mathcal{A}$, and

2. Tighter than $\mathcal{O}(1/W)$ range enclosures of $p^*(\Theta^{(i)})$ with $P^*(\Theta^{(i)})$.

We can improve the sampler by the first pathway through an adaptive partition of $\mathbf{\Theta}$. For example, we can bisect a box $\Theta^{(i)}$ of the current partition to obtain the subsequent refinement of $\Theta^{(i)}$, along the side with the maximal diameter, only if $d(P^*[\Theta^{(i)}])$, the diameter of the range enclosure of p^* over $\Theta^{(i)}$, is greater than some pre-determined tolerance `TOL`. Improvement by the second pathway can come from the enclosures of Taylor expansions of p^* around the midpoint $m(\Theta^{(i)})$ through automatic differentiation, provided p^* is differentiable enough. Of course, one can gain more efficiency by simultaneously applying improvements in both pathways. We will use uniform refinement and adaptive refinement in the following examples.

Examples

Next we apply the Moore-rejection sampler to target shapes that are mixtures of Gaussian densities truncated over $\mathbf{\Theta}$. The first two target shapes g_a and g_b are

mixtures of two Gaussian shapes, while the third target shape g_c is a mixture of five Gaussian shapes. The means (μ_i's), standard deviations (σ's), weights (w_i's), and domains ($\mathbf{\Theta}$'s) for each of the three targets g_a, g_b, and g_c are shown in Table 2.3 and their graphs are shown in Figure 2.6. Although we only apply M-R sampling to three target shapes, the sampler is applicable to Lipschitz targets in $\mathfrak{E}_{\mathfrak{L}}$ with any finite number of modes on any compact set where the modes are allowed to be as spiky as desired.

Table 2.3: Moore-rejection sampling from three Gaussian mixture shapes

$g_a(x) := \sum_{i=1}^{2} w_i \ \exp -((x-\mu_i)/(\sqrt{2}\sigma_i))^2$	$\mu_1 = -5, \sigma_1 = 1, \ w_1 = 0.25, \ \mu_2 = 5,$
$\mathbf{\Theta} = [-10, 10]$	$\sigma_2 = 0.25,$ and $w_2 = 0.75$
$g_b(x) := \sum_{i=1}^{2} w_i \ \exp -((x-\mu_i)/(\sqrt{2}\sigma_i))^2$	$\mu_1 = -5, \sigma_1 = 1, \ w_1 = 0.25, \ \mu_2 = 50,$
$\mathbf{\Theta} = [-10, 100]$	$\sigma_2 = 0.25,$ and $w_2 = 0.75$
$g_c(x) := \sum_{i=1}^{5} w_i \ \exp -((x-\mu_i)/(\sqrt{2}\sigma_i))^2$	$\mu_1 = -15, \sigma_1 = 1, w_1 = 0.15, \mu_2 = -2,$
$\mathbf{\Theta} = [-30, 80]$	$\sigma_2 = 1, w_2 = 0.2, \mu_3 = 3, \sigma_3 = 0.5,$
	$w_3 = 0.05, \mu_4 = 6, \sigma_4 = 1, w_4 = 0.1,$
	$\mu_5 = 50, \sigma_5 = 0.1, w_5 = 0.5$

We tried to draw as many as $10,000$ samples from a maximum of $100,000$ trials using the M-R sampler from each of the three target shapes g_a, g_b, and g_c with proposals given by Equation 2.3. We used two families of refining partitions to construct two proposals at each refinement level. The first one is the family of uniform partitions $\mathfrak{T}_W$ of size $W = |\mathfrak{T}_W|$ in powers of 2 that is independent of the target shape. The second family $\mathfrak{T}_\alpha^{g_i}$ is formed by adaptive bisections that depend on the target shape $g_i, \ i = a, b, c$. In the adaptive case we obtain a refinement of a partition only by bisecting those intervals whose diameter of range enclosure is above some desired tolerance `TOL`. We set `TOL` $= 0.001$ for the three examples.

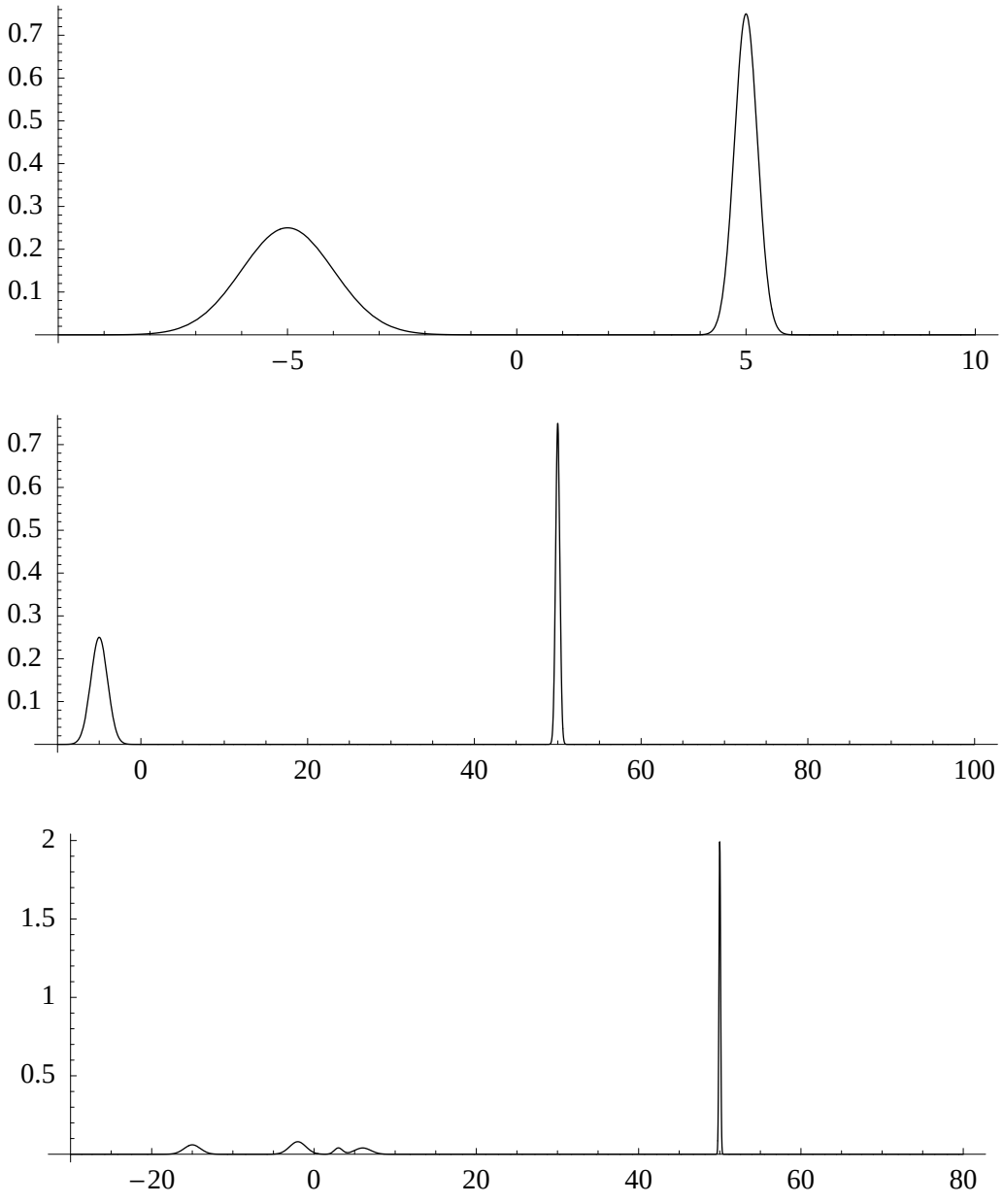


Figure 2.6: Three target shapes g_a, g_b, and g_c and their respective domains $\mathbf{\Theta}$ are depicted. All three are mixtures of Gaussian densities. The top two targets are g_a and g_b respectively and have two components, while the third target at the bottom is g_c and has five components. The parameters of these three targets are shown in Table 2.3.

Table 2.4: Moore-rejection sampling from three Gaussian mixture shapes

$\lvert\mathfrak{T}_W\rvert$	$\mathbf{A}^{g_a}_{f(q^{\mathfrak{T}}W)}$	$\mathbf{A}^{g_b}_{f(q^{\mathfrak{T}}W)}$	$\mathbf{A}^{g_c}_{f(q^{\mathfrak{T}}W)}$	$\lvert\mathfrak{T}^{g_a}_{\alpha}\rvert$	$\mathbf{A}^{g_a}_{f(q^{\mathfrak{T}^{g_a}_{\alpha}})}$	$\lvert\mathfrak{T}^{g_b}_{\alpha}\rvert$	$\mathbf{A}^{g_b}_{f(q^{\mathfrak{T}^{g_b}_{\alpha}})}$	$\lvert\mathfrak{T}^{g_c}_{\alpha}\rvert$	$\mathbf{A}^{g_c}_{f(q^{\mathfrak{T}^{g_c}_{\alpha}})}$
2	0.11	0.02	0.008	2	0.11	2	0.02	2	0.008
4	0.11	0.04	0.02	4	0.11	4	0.04	4	0.02
8	0.21	0.08	0.03	8	0.22	6	0.08	7	0.03
16	0.39	0.15	0.06	16	0.39	8	0.15	11	0.06
32	0.62	0.29	0.12	28	0.62	11	0.28	17	0.12
64	0.78	0.32	0.23	52	0.78	15	0.32	26	0.23
128	0.88	0.53	0.27	96	0.88	22	0.54	40	0.27
256	0.93	0.71	0.47	180	0.94	34	0.72	66	0.47
512	0.97	0.83	0.68	340	0.97	55	0.84	114	0.68
1024	0.98	0.91	0.81	628	0.98	94	0.91	199	0.81
2048	0.99	0.95	0.89	1128	0.99	165	0.95	345	0.89
4096	0.99	0.98	0.94	1936	0.99	297	0.97	568	0.94
8192	0.99	0.99	0.97	2940	0.99	523	0.99	717	0.97

Table 2.4 shows the results of the numerical experiment. Observe that the acceptance probability quickly approaches 1 for all cases. We save a lot of interval evaluations of the target shape when the refinement is done adaptively. Also, observe that for a typical rejection sampler with a family of proposals given in the linear form $f(q) = cq^*$, the acceptance probability N_p/cN_q may never reach 1 even for the optimal c unless q^* belongs to the parametric family of the target shape. Moreover, in the typical rejection sampler, unlike the Moore-rejection sampler, there is no guarantee that $q^*(x)$ dominates $p^*(x)$ for every x in the domain, a necessary condition for rejection sampling.

Of course, the family of Moore-rejection samplers with adaptive refinements of the domain can produce independent samples from a Lipschitz target shape on a compact $\mathbf{\Theta} \Subset \mathbb{IR}^b$, where $b > 1$. However, now we have to think in terms of the hyper-volumes of the partition over $\mathbf{\Theta}$ and the acceptance rate will only approach 1 linearly with the mesh in terms of hyper-volumes. Our next target shape g_d shown in Figure 2.7 is the following bivariate Gaussian mixture

$$\sum_{i=1}^{2} w_i \ \exp \frac{-1}{2(1-c_i^2)} \left(\left(\frac{x-\mu_i}{\sigma_i} \right)^2 - 2c_i \left(\frac{x-\mu_i}{\sigma_i} \right) \left(\frac{y-\nu_i}{\tau_i} \right) + \left(\frac{y-\nu}{\tau_i} \right)^2 \right)$$

with $\mu_1 = -2$, $\nu_1 = -1$, $\sigma_1 = 0.1$, $\tau_1 = 0.1$, $c_1 = 0$, $\mu_2 = 3$, $\nu_2 = 3$, $\sigma_2 = 0.1$, $\tau_2 = 0.1$, $c_2 = 0$, $w_1 = 0.9$, and $w_2 = 0.1$. We next draw as many as $10,000$ samples from a maximum of $100,000$ trials. We can achieve an acceptance probability of 0.5 and 0.75 after adaptively partitioning the domain $[-100, 100] \times [-100, 100]$ into 150 rectangles (see Figure 2.8) and 924 rectangles, respectively. Mixtures for more than two bivariate Gaussian shapes yield comparable results.

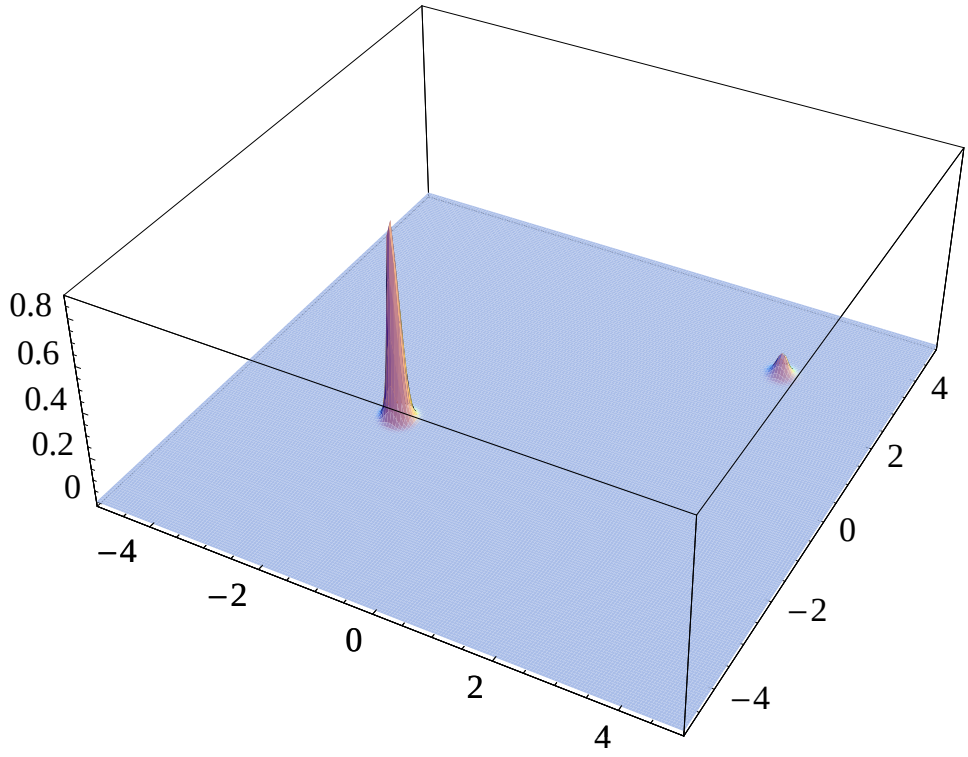


Figure 2.7: The target shape g_d is a mixture of two bivariate Gaussian shapes.

2.3.3 Metropolis-Hastings Chain

Let $\mathcal{S}$ be a finite or countable discrete state space and B be a Markov chain on $\mathcal{S}$ with transition probability matrix $q(x, y), \forall\, x, y \in \mathcal{S}$. We call B the proposal chain or base chain. Let π be a probability distribution on $\mathcal{S}$ with $\pi(x) > 0$ for every $x \in \mathcal{S}$. We are interested in sampling from π, our target distribution. The Metropolis-Hastings (M-H) algorithm produces a Markov chain $X_0, X_1, \ldots$ on $\mathcal{S}$ with stationary distribution π from the proposals of the base chain B. The base chain can propose a move to a new state y from the current state x according to its transition probability $q(x, y)$. If the proposal distribution of a base chain is independent of the current state x, i.e., $q(x, y) = q(y)$, then the base chain is said to be independent. We refer to X as the M-H chain. Armed merely with ratios of

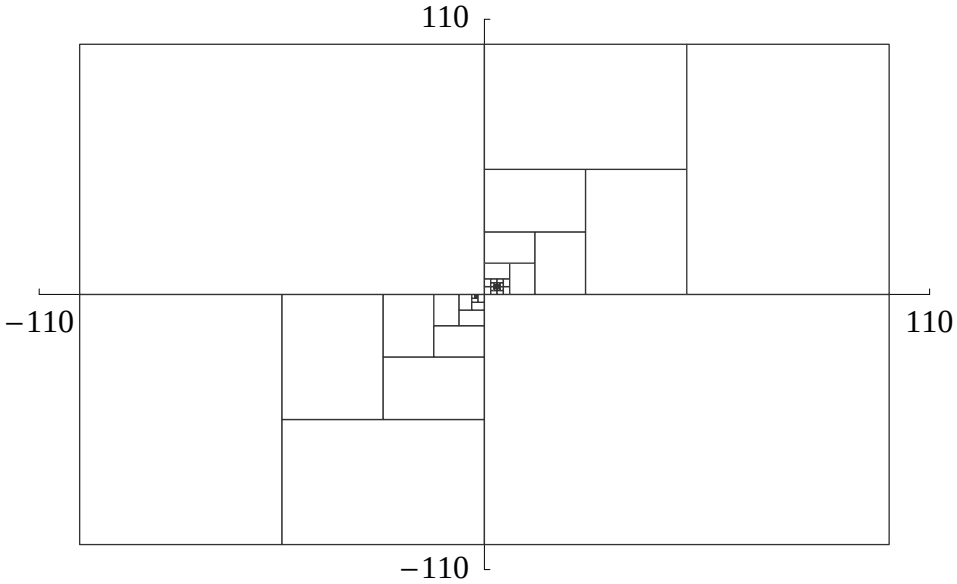


Figure 2.8: Adaptive partitioning of the domain $[-100, 100] \times [-100, 100]$ into 150 rectangles for Moore-rejection sampling from the target shape g_d.

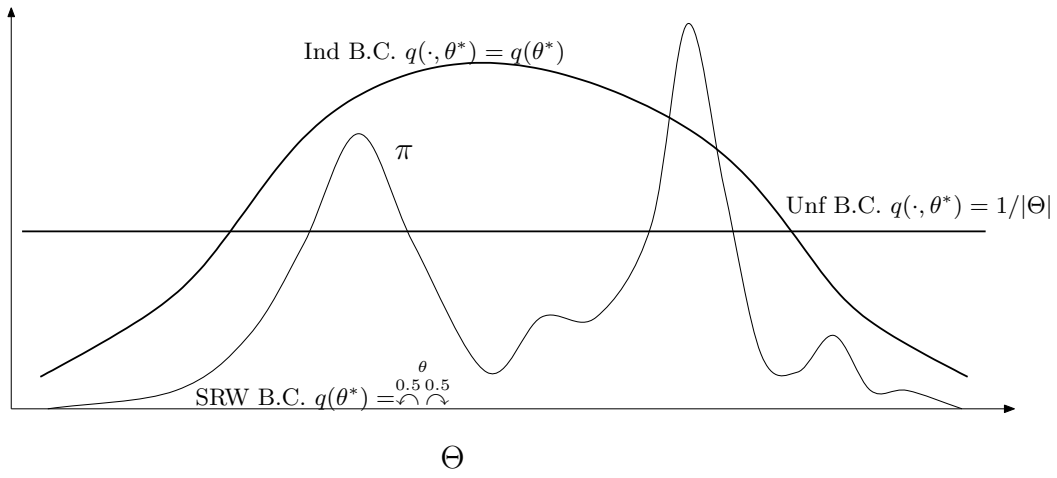


Figure 2.9: The target distribution is π. The M-H chain with independent base chain (Ind. B.C.) has the independent proposal $q(\theta^*)$. The M-H chain with uniform base chain (Unf. B.C.) has the independent proposal $1/|\boldsymbol{\Theta}|$. The M-H chain with simple random walk base chain (SRW B.C.) proposes a move to one state above or to one state below the current state with equal probability.

the form $\pi(x)/\pi(y)$ for any specified pair $(x, y) \in \mathcal{S} \times \mathcal{S}$, our objective is to draw samples from π. The following M-H algorithm produces the M-H chain on $\mathcal{S}$ with stationary distribution π.

1. INITIALIZE $X_0 = x$ for some $x \in \mathcal{S}$ AND $n = 0$

2. PROPOSE candidate state $y \sim q(x, \cdot)$ and DRAW $u \sim Unif[0, 1]$

3. SET $n = n + 1$ AND the next state :

$$X_{n+1} = \begin{cases} y & \text{if } u \leq \min\{\, \mathbf{A}(x, y), 1 \,\} \\ x & \text{otherwise} \end{cases}$$

4. GOTO Step 2 UNTIL n equals the desired number of samples,

where the acceptance ratio is defined by

$$\mathbf{A}(x, y) = \frac{\pi(y)\ q(y, x)}{\pi(x)\ q(x, y)}.$$

Let p/N_p be a density that is dominated by the Lebesgue measure on a compact domain $\mathbb{I}\boldsymbol{\Theta} \Subset \mathbb{R}$, where $N_p := \int_{\boldsymbol{\Theta}} p(\theta)\, \partial\theta$. Consider the finite partition of $\boldsymbol{\Theta}$ at resolution $r := \max_i d(\Theta^{(i)})$

$$\mathfrak{P}_r := \{\, \Theta_r^{(1)}, \Theta_r^{(2)}, ..., \Theta_r^{(|\mathfrak{P}_r|)} \,\}.$$

We want to construct a M-H chain on the finite discrete state space

$$\mathcal{S}_{\mathfrak{P}_r} := \{1, 2, \ldots, |\mathfrak{P}_r|\}.$$

with stationary distribution

$$\pi_r(i) := d(\Theta_r^{(i)}) \cdot p(m(\Theta_r^{(i)})) \simeq \int_{\boldsymbol{\Theta}_\mathbf{r}^{(\mathbf{i})}} p(\theta)\, \partial\theta, \quad \text{where, } \Theta_r^{(i)} \in \mathfrak{P}_r,\ i = 1, 2, \ldots, |\mathfrak{P}_r|.$$

Thus, π_r can be thought of as a resolution specific discretization of N_p with a simple approximation. Since $\pi_r(i)$ must be easy to compute for any i, we produce it by the product of the diameter of $\Theta_r^{(i)}$ and the value of p at the midpoint of $\Theta_r^{(i)}$. The rationale for the discretization is four-fold. First, we can painlessly get rigorous results by tapping into the literature. Second, in practice, one saves the samples from the M-H chain only to count the number of visits to each bin in some appropriate binning of $\mathbf{\Theta}$ and further applies some smoothing transformation of these raw counts to reproduce p/N_p. Third, the mesh r will be so fine with typical values for $|\mathfrak{P}_r|$ so large that the problem is genuinely challenging akin to sampling from a large discrete state space $\mathcal{S}_{\mathfrak{P}_r}$. Fourth, π_r on $\mathbf{\Theta}$ will have a small enough r so that π_r can be used to sample from p/N_p. We refer to r as the target resolution, π_r as the target distribution, p as the target function, and p/N_p as the target density.

The subject of this section is the behavior of M-H chains with stationary distribution π_r on $\mathcal{S}_{\mathfrak{P}_r}$ using *enclosures* of ratios of the form $P(\Theta)/P(\mathit{\Theta})$, where $\Theta, \mathit{\Theta} \in \mathbb{I}\mathbf{\Theta}$ and P is the well-defined interval extension of p.

2.3.4 Moore-Metropolis-Hastings Chain

Now suppose that our target function $p(\theta) : \mathbf{\Theta} \to \mathbb{R}$ has a well-defined interval extension $P(\Theta) : \mathbb{I}\mathbf{\Theta} \to \mathbb{IR}$, then by Theorem 4

$$\begin{aligned} \Theta \subseteq \mathit{\Theta} \subseteq \mathbf{\Theta} \quad &\Longrightarrow \quad P(\Theta) \subseteq P(\mathit{\Theta}) \\ \Theta \subseteq \mathbf{\Theta} \quad &\Longrightarrow \quad Rng(p; \Theta) =: p(\Theta) \subseteq P(\Theta). \end{aligned}$$

Thus we are not only armed with the ratios

$$p(\theta)/p(\vartheta), \forall\, (\theta, \vartheta) \in \mathbf{\Theta} \times \mathbf{\Theta},$$

but also have at our disposal the enclosures of ratios of intervals of the form

$$P(\Theta)/P(\Theta), \forall\, (\Theta, \Theta) \in \mathbb{I}\mathbf{\Theta} \times \mathbb{I}\mathbf{\Theta}.$$

The ratios of $p(m(\Theta^{(i)}))/p(m(\Theta^{(j)}))$ yield

$$\frac{\pi_r(i)}{\pi_r(j)} \;=\; \frac{d(\Theta_r^{(i)})\cdot p(m(\Theta_r^{(i)}))}{d(\Theta_r^{(j)})\cdot p(m(\Theta_r^{(j)}))}$$

where, $\Theta_r^{(i)}, \Theta_r^{(j)} \in \mathfrak{P}_r$ and $i, j \in \mathcal{S}_{\mathfrak{P}_r} := \{1, \ldots, |\mathfrak{P}_r|\}$. We need the above ratio to drive the M-H chain on $\mathcal{S}_{\mathfrak{P}_r}$ with stationary distribution π_r.

For an arbitrary partition $\mathfrak{P}_s$ of $\mathbf{\Theta}$ at resolution s, the interval ratios

$$\frac{P(\Theta_s^{(i)})}{P(\Theta_s^{(j)})}, \quad \text{where } \Theta_s^{(i)}, \Theta_s^{(j)} \in \mathfrak{P}_s \text{ and } i, j \in \mathcal{S}_{\mathfrak{P}_s} := \{1, \ldots, |\mathfrak{P}_s|\}$$

yield the enclosures of interest at resolution s

$$\frac{\Pi_s(i)}{\Pi_s(j)} \;=\; \frac{d(\Theta_s^{(i)})\cdot P(\Theta_s^{(i)})}{d(\Theta_s^{(j)})\cdot P(\Theta_s^{(j)})} \;\supseteq\; \frac{d(\Theta_s^{(i)})\cdot p(m(\Theta_s^{(i)}))}{d(\Theta_s^{(j)})\cdot p(m(\Theta_s^{(j)}))} \;=\; \frac{\pi_s(i)}{\pi_s(j)}$$

The Moore-Metropolis-Hastings (M-M-H) algorithm is the M-H algorithm described earlier with a given transition probability matrix $q(i, j), \forall\, i, j \in \mathcal{S}_{\mathfrak{P}_r}$ for the base chain B, where the matrix q is obtained from interval methods. Consider a family of such interval-based transition probability matrices $\mathcal{Q}$ corresponding to a family of base chains. Each member q of the family $\mathcal{Q}$ invokes a M-M-H Markov chain on $\mathbf{\Theta}$ with stationary distribution π^r. We will compare the rates of convergence of different members of the M-M-H family of chains next. Since $\mathbf{\Theta}$ is compact we will assume that $\mathbf{\Theta} = [0, 1]$ without loss of generality. Also, to

simplify the notation and the analysis in the sequel we take a uniform partition $\mathfrak{R}_r$ over $[0,1]$ of size $R = |\mathfrak{R}_r|$ and mesh $r = 1/2^R$ for the target distribution π_r

$$\begin{aligned}\mathfrak{R}_r &= \{\,\Theta_r^{(1)}, \Theta_r^{(2)}, \ldots, \Theta_r^{(R)}\,\} = \{\,[\tfrac{0}{2^R}, \tfrac{1}{2^R}], [\tfrac{1}{2^R}, \tfrac{2}{2^R}], \ldots, [\tfrac{2^R-1}{2^R}, \tfrac{2^R}{2^R}]\,\}, \text{ where} \\ \Theta_r(i) &= [\tfrac{i-1}{2^R}, \tfrac{i}{2^R}],\ \forall\, i \in \mathcal{S}_{\mathfrak{R}_r} = \{1, 2, \ldots, R\}\end{aligned}$$

Thus the ratio of the π_r's at our target resolution r is

$$\frac{\pi_r(i)}{\pi_r(j)} = \frac{d(\Theta_r^{(i)}) \cdot p(m(\Theta_r^{(i)}))}{d(\Theta_r^{(j)}) \cdot p(m(\Theta_r^{(j)}))} = \frac{\frac{1}{2^R} p(\frac{i}{2^R} - \frac{1}{2^{R+1}})}{\frac{1}{2^R} p(\frac{j}{2^R} - \frac{1}{2^{R+1}})},\ \forall\, i, j \in \mathcal{S}_{\mathfrak{R}_r}$$

and that of the Π_s's at any uniform partition $\mathfrak{R}_s$ over $[0,1]$ of size $S = |\mathfrak{R}_s|$ and mesh $s = 1/2^S$ is

$$\frac{\Pi_s(i)}{\Pi_s(j)} = \frac{d(\Theta_s^{(i)}) \cdot P(\Theta_s^{(i)})}{d(\Theta_s^{(j)}) \cdot P(\Theta_s^{(j)})} = \frac{\frac{1}{2^S} P([\frac{i-1}{2^S}, \frac{i}{2^S}])}{\frac{1}{2^S} P([\frac{j-1}{2^S}, \frac{j}{2^S}])} \supseteq \frac{\frac{1}{2^S} p([\frac{i-1}{2^S}, \frac{i}{2^S}])}{\frac{1}{2^S} p([\frac{j-1}{2^S}, \frac{j}{2^S}])},\ \forall\, i, j \in \mathcal{S}_{\mathfrak{R}_s}$$

Typically $s > r$ and we refer to s as the probing resolution. In general, given a probing resolution $s > r$ and an arbitrary partition $\mathfrak{P}_s$ of $\mathbf{\Theta}$, we can enclose $p(\Theta_s^{(i)})$

$$p(\Theta_s^{(i)}) \subseteq P(\Theta_s^{(i)}) := [\underline{P}(\Theta_s^{(i)}), \overline{P}(\Theta_s^{(i)})],\ \forall\, i \in \mathcal{S}_{\mathfrak{P}_s}, \Theta^{(i)} \in \mathfrak{P}_s.$$

Note that for each $i \in \mathcal{S}_{\mathfrak{R}_r}$ there exists $\Theta_s^{(K_{\mathfrak{P}_s}(i))} \in \mathfrak{P}_s$, where

$$K_{\mathfrak{P}_s}(i) : \mathcal{S}_{\mathfrak{R}_r} \to \mathcal{S}_{\mathfrak{P}_s},\ \text{such that, } \Theta_s^{(K_{\mathfrak{P}_s}(i))} \supseteq \Theta_r^{(i)} \in \mathfrak{R}_r,$$

because $\mathfrak{R}_r$ is a refinement of $\mathfrak{P}_s$ that makes $K_{\mathfrak{P}_s}$ a map. Thus there are $|K_{\mathfrak{P}_s}^{-1}(j)|$ elements of $\mathfrak{R}_r$ contained in $\Theta_s^{(j)} \in \mathfrak{P}_s$. For example, when $s \geq r$ or equivalently when $S \leq R$ and $\mathfrak{P}_s = \mathfrak{R}_s$, $K_{\mathfrak{R}_s}(i)$ has a simple expression

$$K_{\mathfrak{R}_s}(i) = \lceil \frac{i}{2^{R-S}} \rceil : \mathfrak{R}_r \to \mathfrak{R}_s,$$

where, $\lceil x \rceil := \min\{n : (n \geq x) \wedge (n \in \mathbb{N})\}$ is the natural ceiling function.

An Independent Family of Base Chains

Recall that the proposal distribution of the M-H algorithm with an independent base chain is $q(i,j) = q(j)$. Thus, the acceptance probability in Step 3 of the M-H algorithm becomes:

$$\mathbf{A}(i,j) = \frac{\pi_r(j)\ q(j,i)}{\pi_r(i)\ q(i,j)} = \frac{\pi_r(j)\ q(i)}{\pi_r(i)\ q(j)}, \quad \forall\, i,j \in \mathcal{S}_{\mathfrak{R}_r}$$

Now we need a mechanism that generates the transition probabilities for the independent base chains on $\mathcal{S}_{\mathfrak{R}_r}$

$$q(i),\ \forall\, i \in \mathcal{S}_{\mathfrak{R}_r}.$$

We will use the enclosure of p at the probing resolution s under the coarser partition $\mathfrak{P}_s$ to obtain $\mathfrak{P}_s$–specific transition probabilities $q^*_{\mathfrak{P}_s}(i)$ for a base chain B on $\mathcal{S}_{\mathfrak{R}_r}$ as follows:

$$q^*_{\mathfrak{P}_s}(i) = M^{(K_{\mathfrak{P}_s}(i))}_{\mathfrak{P}_s}(i)\ \frac{p^*(\Theta^{(K_{\mathfrak{P}_s}(i))}_s))(d(\Theta^{(K_{\mathfrak{P}_s}(i))}_s))}{V_{\mathfrak{P}_s}}$$

where for each $K_{\mathfrak{P}_s}(i) \in \mathcal{S}_{\mathfrak{P}_s}$ and the corresponding set $K^{-1}_{\mathfrak{P}_s}(K_{\mathfrak{P}_s}(i)) \subseteq \mathcal{S}_{\mathfrak{R}_r}$,

$$M^{(K_{\mathfrak{P}_s}(i))}_{\mathfrak{P}_s}(i) \text{ is a p.m.f. on } K^{-1}_{\mathfrak{P}_s}(K_{\mathfrak{P}_s}(i)),\ \text{such that, } M^{(K_{\mathfrak{P}_s}(i))}_{\mathfrak{P}_s}(i) > 0,\ \forall\, i,$$

p^*'s are a set of fixed values such that

$$p^*(\Theta^{(j)}_s) \in P(\Theta^{(j)}_s),\ \forall\, j \in \mathcal{S}_{\mathfrak{P}_s},\ \text{and } \Theta^{(j)} \in \mathfrak{P}_s,$$

and V_s is the normalizing constant

$$V_{\mathfrak{P}_s} = \sum_{j=1}^{|\mathcal{S}_{\mathfrak{P}_s}|} p^*(\Theta^{(j)}_s) \cdot d(\Theta^{(j)}_s).$$

Let $\mathcal{Q}$ be the family of independent transition probabilities, each of whose members are obtained from the interval extension P of p on a corresponding coarse partition $\mathfrak{P}_s$. Each member of $\mathcal{Q}$ induces a M-H chain on $\mathcal{S}_{\mathfrak{R}_r}$ by providing the

transition probabilities for an independent base chain. We refer to this M-H chain driven by an interval-informed independent base chain as an independent M-M-H chain. Note that each $q^*_{\mathfrak{P}_s} \in \mathcal{Q}$ is specified by three objects: (1) the coarse partition $\mathfrak{P}_s$, (2) the mass functions $M^{(K_{\mathfrak{P}_s}(i))}_{\mathfrak{P}_s}$ and (3) the choice of p^*'s. Thus, our family $\mathcal{Q}$ is too large to parametrically hunt for some optimal sequence of base chains. We consider a smaller parameterized family in $\mathcal{Q}$ by restricting (1) the coarse partition to the uniform $\mathfrak{R}_s$, (2) the mass functions to the uniform family of mass functions $U^{(K_{\mathfrak{R}_s}(i))}_{\mathfrak{R}_s}(i) = 1/|K^{-1}_{\mathfrak{R}_s}(K_{\mathfrak{R}_s}(i))| = \frac{1}{2^{R-S}}$, $\forall\, i$, $\because\ K_{\mathfrak{R}_s}(i) = \lceil \frac{i}{2^{R-S}} \rceil$, and (3) the choice of $p^*(\Theta^{(j)}_s)$'s to $\overline{P}(\Theta^{(j)}_s)$, $\forall\, j \in \mathcal{S}_{\mathfrak{R}_s}$. Therefore, for a given $s > r$ we have a base chain $\overline{B}_s$ on $\mathcal{S}_{\mathfrak{R}_r}$ with transition probabilities $\overline{q}_s(i)$ invoked by $\mathfrak{R}_s$, $U^{(\lceil \frac{i}{2^{R-S}} \rceil)}_{\mathfrak{R}_s}(i)$, and $\overline{P}(\Theta^{(j)}_s)$. Now, consider the s parametrized family of these base chains $\overline{B}_s$, with $s = 1/2^S$ and $S \in \{0, 1, \ldots, R\}$ with transition probabilities

$$\begin{aligned} \overline{q}_s(i) = & \quad \frac{1}{|K^{-1}_{\mathfrak{R}_s}(K_{\mathfrak{R}_s}(i))|} \frac{\overline{P}(\Theta^{(K_{\mathfrak{R}_s}(i))}_s) d(\Theta^{(K_{\mathfrak{R}_s}(i))}_s)}{V_s} \\ = & \quad \frac{1}{2^{R-S}} \frac{\frac{1}{2^S} \cdot \overline{P}(\Theta^{(\lceil \frac{i}{2^{R-S}} \rceil)}_s)}{\sum_{j=1}^{2^S} \frac{1}{2^S} \cdot \overline{P}(\Theta^{(j)}_s)} = \frac{1}{2^{R-S}} \frac{\overline{P}(\Theta^{(\lceil \frac{i}{2^{R-S}} \rceil)}_s)}{\sum_{j=1}^{2^S} \overline{P}(\Theta^{(j)}_s)} \end{aligned} \tag{2.4}$$

We refer to the family of M-H chains driven by members of the above family of base chains as independent M-M-H Max chains or simply (Max chains). Analogously, when $\overline{P}(\Theta^{(\lceil \frac{i}{2^{R-S}} \rceil)}_s)$ in the above expression for $\overline{q}_s(i)$ is replaced with $\underline{P}((\Theta^{(\lceil \frac{i}{2^{R-S}} \rceil)}_s))$ or with $m(P((\Theta^{(\lceil \frac{i}{2^{R-S}} \rceil)}_s)))$ we obtain the family of transition probabilities $\underline{q}_s(i)$'s or $\mathring{q}_s(i)$'s, that in turn drive the family of independent M-M-H Min chains or the family of independent M-M-H Mid chains, respectively.

Thus, the acceptance probability in Step 3 of the M-H algorithm for the max chain, $\forall\, i, j \in \mathcal{S}_{\mathfrak{R}_r}$, simplifies to

$$\mathbf{A}(i,j) = \frac{\pi_r(j)\ q(i)}{\pi_r(i)\ q(j)} = \frac{p(\frac{j}{2^R} - \frac{1}{2^{R+1}})\ \overline{P}(\Theta^{(\lceil \frac{i}{2^{R-S}} \rceil)}_s)}{p(\frac{i}{2^R} - \frac{1}{2^{R+1}})\ \overline{P}(\Theta^{(\lceil \frac{j}{2^{R-S}} \rceil)}_s)}.$$

We finally have the necessary ingredients to sample from independent M-M-H chains with stationary distribution π_r. Next we focus on understanding the behavior of these chains in terms of their efficiency.

Theorem 9 (Liu) *The total variation distance for the M-H chain X_i started at i with independent base chain whose transition probabilities are $q(i)$, $i \in \mathcal{S}_{\mathfrak{R}_r}$ is bounded from above by*

$$4\,||(X_i)^k - \pi_r||^2 \leq \frac{\beta_q^{2k}}{\pi_r(i)},$$

where, $\beta_q = 1 - \frac{1}{W_q}$, and $W_q = \max_{1 \leq i \leq 2^R} \frac{\pi_r(i)}{q(i)}$.

Proof: See [24, 8].

Theorem 9 helps us compare the performance of several M-H chains that are targeting the same π with distinct base chains whose transition probabilities are $q_1, q_2, \ldots$. If all these independent M-H chains are started at the same state i, then the best one has the smallest W_{q_j} value.

Theorem 10 *Let p be a target function on $[0,1]$ with a well-defined interval extension P. Let π_r be the target distribution on the discrete state space $\mathcal{S}_{\mathfrak{R}_r}$ that approximates the integral of p over each set in the uniform partition $\mathfrak{R}_r$ of size 2^R and mesh $r = 1/2^R$. Consider the independent M-M-H max chain driven by an independent base chain with transition probabilities $\overline{q}_s(i)$, which are informed by the interval enclosures of p over sets in the uniform partition $\mathfrak{R}_s$ of size 2^S and mesh $s = 1/2^S$ and $S \leq R$. Then*

$$W_{\overline{q}_s} = \max_{1 \leq i \leq 2^R} \frac{\pi_r(i)}{\overline{q}_s(i)} \leq \frac{2^{R-S} \sum_{j=1}^{2^S} \overline{P}(\Theta_s^{(j)})}{\sum_{j=1}^{2^R} p(\frac{j}{2^R} - \frac{1}{2^{R+1}})}$$

Furthermore, if the target function $p \in \mathfrak{E}_{\mathfrak{L}}$, the Lipschitz class of elementary functions, and $R \to \infty$, then

$$W_{\overline{q}_s} \leq 1 + \mathcal{O}\left(\frac{1}{2^S}\right)$$

Proof: Let us first simplify $\frac{\pi_r(i)}{\overline{q}_s(i)}$ for any $i \in \mathcal{S}_{\mathfrak{R}_r}$

$$\begin{aligned}\frac{\pi_r(i)}{\overline{q}_s(i)} &= \frac{\frac{p(\frac{i}{2^R}-\frac{1}{2^{R+1}})}{\sum_{j=1}^{2^R} p(\frac{j}{2^R}-\frac{1}{2^{R+1}})}}{\frac{1}{2^{R-S}}\frac{\overline{P}(\Theta_s^{(\lceil \frac{i}{2^{R-S}} \rceil)})}{\sum_{j=1}^{2^S}\overline{P}(\Theta_s^{(j)})}} \\ &= \frac{2^{R-S}\ \sum_{j=1}^{2^S}\overline{P}(\Theta_s^{(j)})}{\sum_{j=1}^{2^R} p(\frac{j}{2^R}-\frac{1}{2^{R+1}})}\frac{p(\frac{i}{2^R}-\frac{1}{2^{R+1}})}{\overline{P}(\Theta_s^{(\lceil \frac{i}{2^{R-S}} \rceil)})}\end{aligned}$$

Therefore,

$$\left(\frac{\pi_r(1)}{\overline{q}_s(1)},\ldots,\frac{\pi_r(2^R)}{\overline{q}_s(2^R)}\right) = \frac{2^{R-S}\ \sum_{j=1}^{2^S}\overline{P}(\Theta_s^{(j)})}{\sum_{j=1}^{2^R} p(\frac{j}{2^R}-\frac{1}{2^{R+1}})}\left(\frac{p(\frac{1}{2^R}-\frac{1}{2^{R+1}})}{\overline{P}(\Theta_s^{(\lceil \frac{1}{2^{R-S}} \rceil)})},\ldots,\frac{p(\frac{2^R}{2^R}-\frac{1}{2^{R+1}})}{\overline{P}(\Theta_s^{(\lceil \frac{2^R}{2^{R-S}} \rceil)})}\right)$$

Now Recall that

$$\begin{aligned}\tfrac{i}{2^R}-\tfrac{1}{2^{R+1}} = m([\ \tfrac{i-1}{2^R}, \tfrac{i}{2^R}\ ]) &\in [\ \tfrac{i-1}{2^R}, \tfrac{i}{2^R}\ ] \\ &\subseteq [\ \tfrac{(\lceil \frac{i}{2^{R-S}} \rceil)-1}{2^S}\ ,\ \tfrac{(\lceil \frac{i}{2^{R-S}} \rceil)}{2^S}\ ] = \Theta_s^{(\lceil \frac{i}{2^{R-S}} \rceil)}\end{aligned}$$

By the inclusion property of the well-defined interval extension P of p,

$$\begin{aligned}&\tfrac{i}{2^R}-\tfrac{1}{2^{R+1}} \in \Theta_s^{(\lceil \frac{i}{2^{R-S}} \rceil)} \\ \Longrightarrow\quad &\underline{P}(\Theta_s^{(\lceil \frac{i}{2^{R-S}} \rceil)}) \leq p(\tfrac{i}{2^R}-\tfrac{1}{2^{R+1}}) \leq \overline{P}(\Theta_s^{(\lceil \frac{i}{2^{R-S}} \rceil)}) \\ \Longrightarrow\quad &\left(\frac{p(\frac{1}{2^R}-\frac{1}{2^{R+1}})}{\overline{P}(\Theta_s^{(\lceil \frac{1}{2^{R-S}} \rceil)})}\right) \leq 1,\ \forall\, i \in \mathcal{S}_{\mathfrak{R}_r}.\end{aligned}$$

Therefore,

$$W_{\overline{q}_s} = \max\left(\frac{\pi_r(1)}{\overline{q}_s(1)},\ldots,\frac{\pi_r(2^R)}{\overline{q}_s(2^R)}\right) \leq \frac{2^{R-S}\ \sum_{j=1}^{2^S}\overline{P}(\Theta_s^{(j)})}{\sum_{j=1}^{2^R} p(\frac{j}{2^R}-\frac{1}{2^{R+1}})}.$$

To prove the second statement observe that as $R \to \infty$ and $|\mathcal{S}_{\mathfrak{R}_r}|$ approaches a countable state space

$$\frac{2^{R-S}\ \sum_{j=1}^{2^S} \overline{P}(\Theta_s^{(j)})}{\sum_{j=1}^{2^R} p(\frac{j}{2^R} - \frac{1}{2^{R+1}})} = \frac{2^{-S}\ \sum_{j=1}^{2^S} \overline{P}(\Theta_s^{(j)})}{2^{-R}\ \sum_{j=1}^{2^R} p(\frac{j}{2^R} - \frac{1}{2^{R+1}})} \to \frac{2^{-S}\ \sum_{j=1}^{2^S} \overline{P}(\Theta_s^{(j)})}{\int_{\mathbf{\Theta}} p(\theta)\partial\theta}.$$

Then the second statement of the Theorem is a consequence of applying Theorem 5 as done in the proof of Theorem 8 $\square$.

The above Theorem allows the target function to be any Lipschitz function in $\mathfrak{E}_{\mathfrak{L}}$. In particular multi-modal targets are fair game. Thus, the number of modes of a multi-modal Lipschitz p is irrelevant to the order of convergence of $W_{\overline{q}_s}$ toward 1 from above. In this respect, the independent M-M-H max chain is robust to multi-modality of the Lipschitz target.

Examples of Beta shapes

Example 4 *We carefully study the simple target* $p(\theta) = \theta^\alpha(1-\theta)^\beta$

Theorem 11 *Targeting* $p = \theta^\alpha(1-\theta)^\beta$ *with uniform base chain is worse than targeting with max base chain informed by interval arithmetic.*

Proof: For a uniform partition $\mathfrak{R}_r$ of $\mathbf{\Theta} = [0,1]$ with resolution $r = 1/2^R$, size 2^R and mesh $1/2^R$

$$\mathfrak{R}_r = \{\, \Theta_r^{(1)}, \Theta_r^{(2)}, \ldots, \Theta_r^{(2^R)} \,\} = \{\, [\frac{0}{2^R}, \frac{1}{2^R}], [\frac{1}{2^R}, \frac{2}{2^R}], \ldots, [\frac{2^R-1}{2^R}, \frac{2^R}{2^R}] \,\},$$

where, $\Theta_r^{(i)} := [(i-1)/2^R, i/2^R]$. Let π_r be the target distribution and $\overline{q}_s$ given by Equation 2.4 be the transition probabilities of the base chain informed by enclosures of p at the probing resolution s. Let $q_u = 1/2^R$ be the transition probabilities of the uniform base chain. If $W_{\overline{q}_s} = \max_i\ \pi_r(i)/\overline{q}_s(i)$ and $W_u = \max_i\ \pi_r(i)/q_u(i)$

then by applying Theorem 10 and operating with interval arithmetic the ratio $W_{\overline{q}_s}/W_u$ for any $\alpha, \beta > 0$

$$\begin{aligned}
\frac{W_{\overline{q}_s}}{W_u} &= \frac{2^{(R-S)} \sum_{i=1}^{2^S} \{(\frac{i}{2^S})^\alpha (1-\frac{i-1}{2^S})^\beta\}}{2^{(R-S)} 2^S \max_i \{ \frac{1}{2^R}(i-\frac{1}{2})^\alpha (1-\frac{1}{2^R}(i-\frac{1}{2})^\beta \}} \\
&\simeq \frac{2^{-S} \sum_{i=1}^{2^S} \{(\frac{i}{2^S})^\alpha (1-\frac{i-1}{2^S})^\beta\}}{(\frac{\alpha}{\alpha+\beta})^\alpha (1-\frac{\alpha}{\alpha+\beta})^\beta} \\
&\xrightarrow{O(2^{-S})} \frac{\frac{\Gamma(\alpha+1)\ \Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)}}{(\frac{\alpha}{\alpha+\beta})^\alpha (1-\frac{\alpha}{\alpha+\beta})^\beta} < 1.
\end{aligned}$$

Setting $\alpha = \beta$ in the expression for $W_{\overline{q}_s}/W_u$ allows us to visualize the phases of $W_{\overline{q}_s}/W_u$ as a function of s in Figure 2.10, where the region to the right of the red contour is where $W_{\overline{q}_s}/W_u < 1$ implying that the independent M-M-H max chain is superior to the M-H chain driven by the uniform base chain.

In fact, the above argument can be applied even if the target is multi-modal, where the tallest mode will determine W_u. We confirm the superiority of independent M-M-H chains through simulations in our next example.

Example 5 *Next we study a two parameter family of equi-weighted, symmetric mixture of two Beta densities*

$$p_{\alpha,\beta}(\theta) = \left(\theta^{\alpha-1}(1-\theta)^{\beta-1} + \theta^{\beta-1}(1-\theta)^{\alpha-1}\right), \quad \theta \in [0,1]\,. \tag{2.5}$$

We did simulations and computed the decay in total variation for six M-H chains for eight target shapes given by Equation 2.5. The results are shown in Figures 2.11 and 2.12. The six M-H chains included the M-H chain with a uniform base chain (Unf), the M-H chain with an independent base chain given by the target density itself (Tgt), the M-H chain with a simple random walk base chain (Srw), and the three independent M-M-H chains Min, Mid and Max chains. The Min,

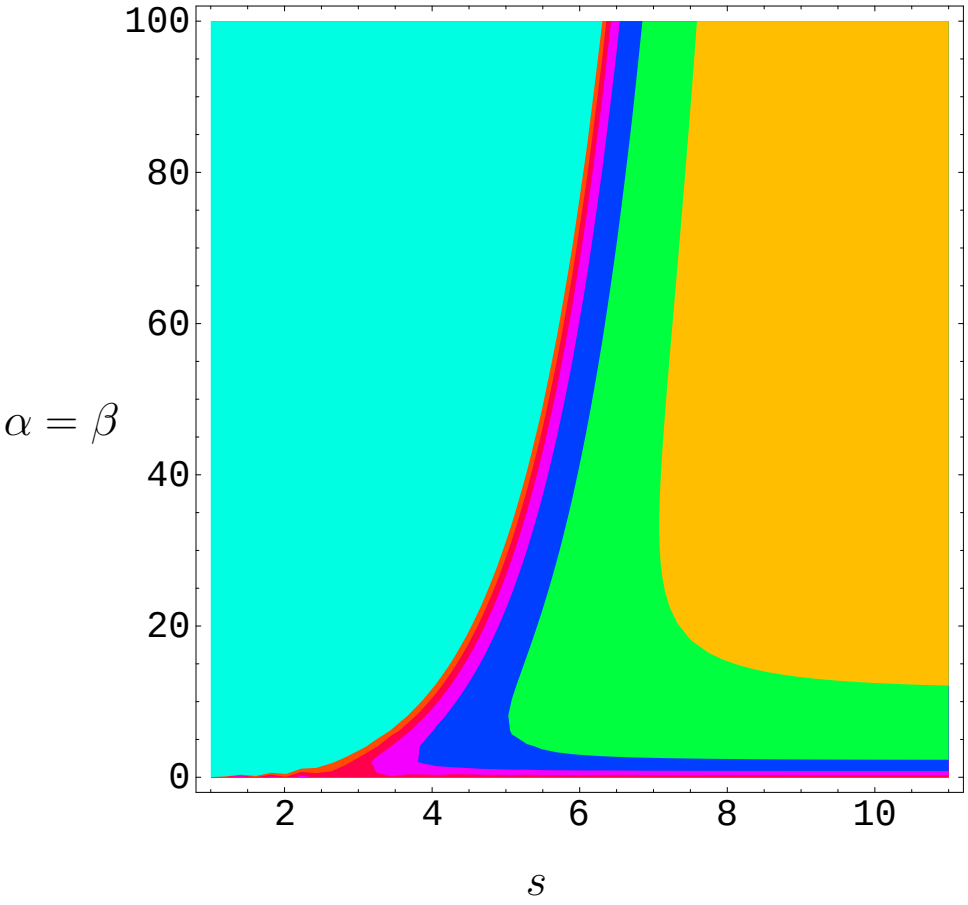
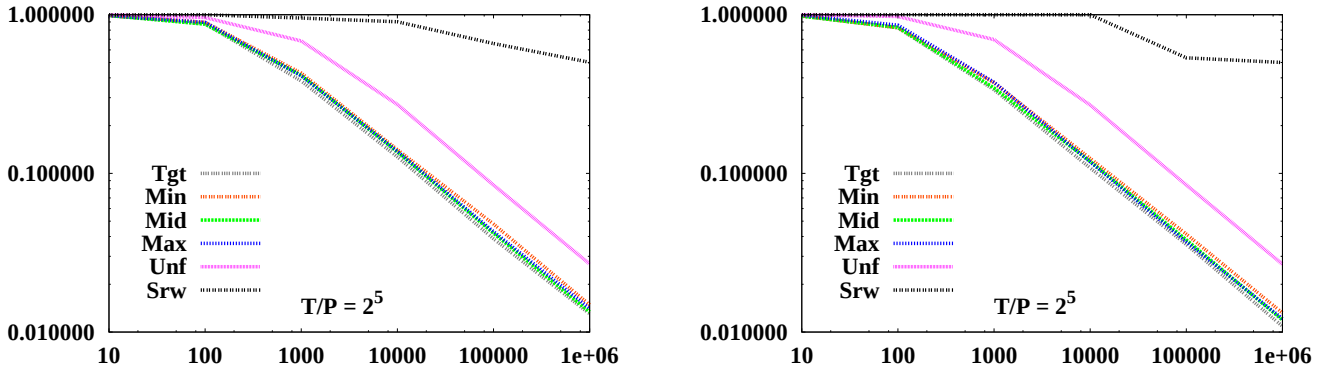


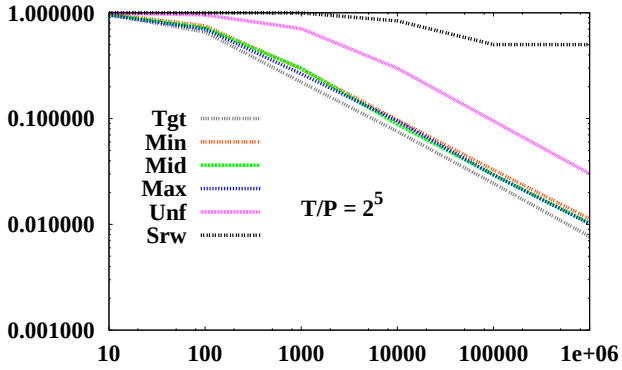
Figure 2.10: Contour of $W_{\overline{q}_s}/W_u$ as a function of s. and $\alpha = \beta$.

Mid, as well as the Max chains at a probing resolution of merely 32 bins (T/P) are almost as efficient as the Tgt chain for all eight targets. All three M-M-H chains are superior to the Srw chain because the latter gets stuck in one of the modes and also superior to the Unf chain because several proposals of the Unf chain are never accepted. The problem of slow mixing for the Unf chain is particularly acute when the target is strongly peaked for larger values of the parameters.

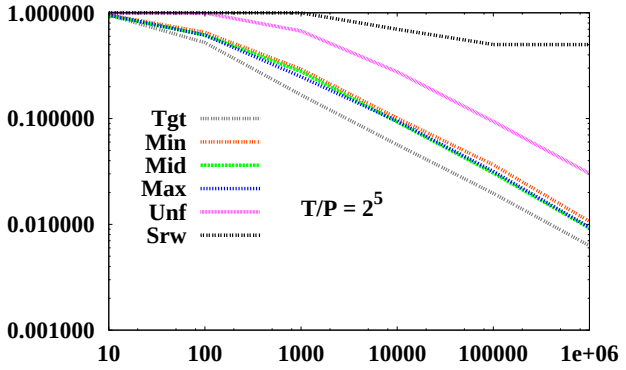


(a)

(b)



(c)



(d)

Figure 2.11: The decay in total variation for six M-H chains as function of the number of runs is shown for four targets given by Equation 2.5. The (α, β) values for the four targets are $(5, 35)$, $(10, 70)$, $(50, 350)$, and $(100, 700)$, respectively.

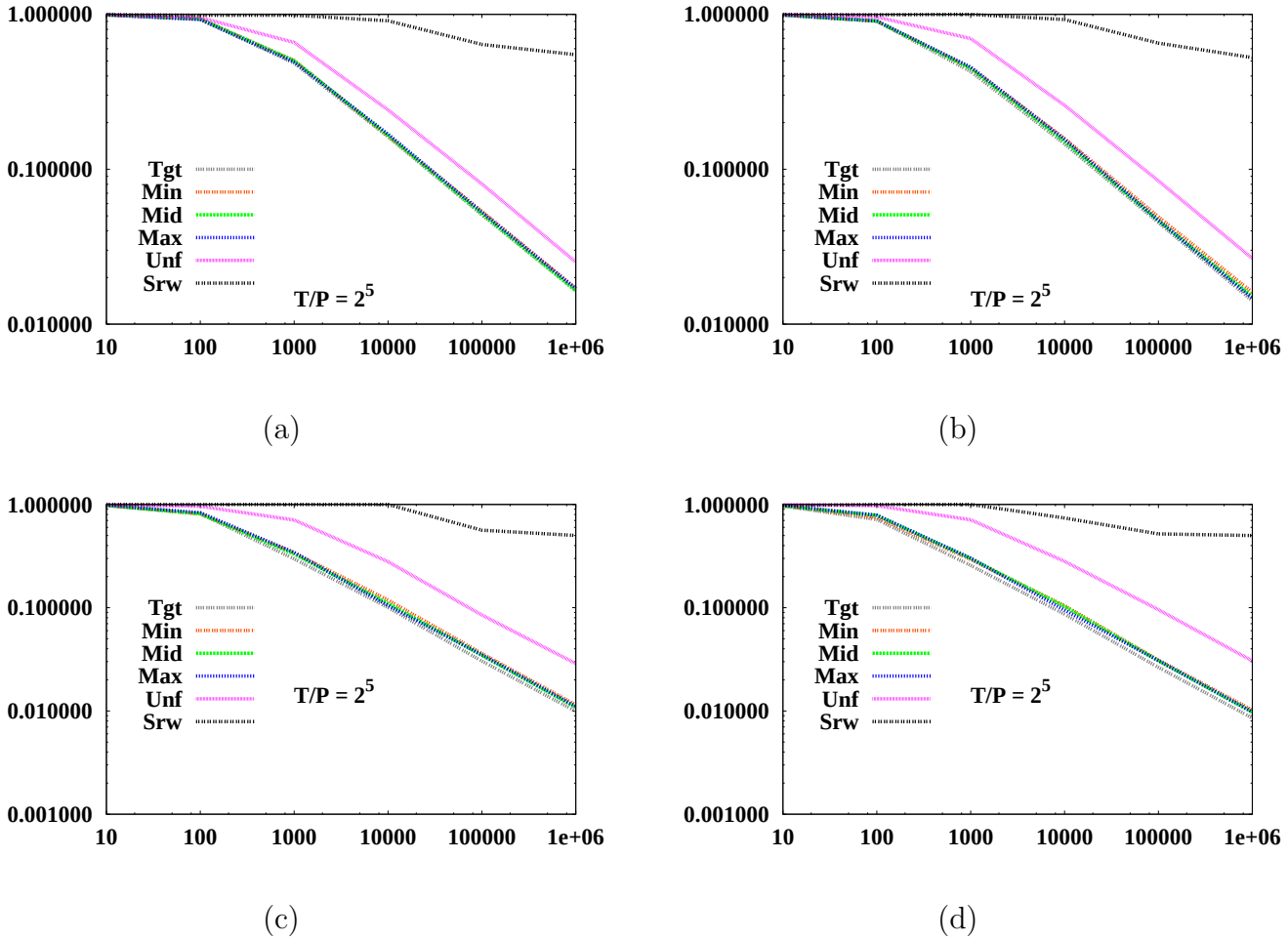


Figure 2.12: The decay in total variation for six M-H chains as function of the number of runs is shown for four targets given by Equation 2.5. The (α, β) values for the four targets are $(5, 15)$, $(10, 30)$, $(50, 150)$, and $(100, 300)$, respectively.

Discussion

When $M_{\mathfrak{P}_s}(i) = 1/|K^{-1}_{\mathfrak{P}_s}(K_{\mathfrak{P}_s}(i))|$ in the first expression for $q_s(i)$ in Equation 2.4, the probability mass in the second ratio is uniformly distributed among all the states at the target resolution r that are contained inside $\Theta_s^{(K_{\mathfrak{P}_s}(i))}$. In general one need not uniformly spread the mass. For instance, if it is known that the enclosure of the derivative over $\Theta_s^{(i)}$ does not contain zero, then one may spread the probability mass in a linear fashion over states at the target resolution by using this information.

The M-M-H algorithm with independent base chains informed by interval arithmetic approaches the optimal M-H algorithm with independent base chain given by the target π itself at a rate that improves no slower than linearly with the mesh. The performance of the M-H algorithms with 'simple random walk' base chains or Uniform independent base chains, in terms of the total variation distance to the target is significantly worse. Using Jun Liu's spectral decompositions available for the M-H algorithms with independent base chains [24], we have theoretically confirmed the above observations for the Beta target.

We have restricted ourselves to the independent family of M-H chains primarily because we can tap into the analytical results available for such chains. It is possible to extend interval methods to obtain more general M-H chains with base chains that are dependent on the current state. It becomes difficult to analytically study the behavior of more general Moore-Metropolis-Hastings chains although simulation studies confirm their superiority when compared to M-H chains that do not use information from interval methods.

We only focussed on the interval extension of the simplest samplers. This allowed us to obtain some rigorous results. Other interesting and straight-forward

applications of interval methods include importance sampling as well as sequential Monte Carlo methods. There is a clear need to compare the performance of more general randomized enclosure algorithms with more sophisticated randomized floating-point algorithms, such as simulated tempering and umbrella-sampling, in terms of efficiently sampling from complicated targets. It is of interest to apply efficient interval-based M-H algorithms to integrate over or sample from $\mathbf{\Theta} \subset \mathbb{R}^n$, which typically parameterizes Markov models in molecular evolution or branch lengths of trees in phylo-population genetics. The obvious improvements on the theoretical front include higher-order stochastic enclosure algorithms that employ automatic differentiation to produce tighter bounds on the integrals via Taylor expansions.

BIBLIOGRAPHY

[1] G Alefeld and J Herzberger. *An introduction to interval computations*. Academic press, 1983.

[2] M Berz. Forward algorithms for high orders and many variables with application to beam physics. In A Griewank and G Corliss, editors, *Automatic differentiation of algorithms: theory, implementation and applications*, pages 147–156. SIAM, 1991.

[3] LJ Billera, S Holmes, and K Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Math.*, 2004.

[4] D Böhning, E Dietz, R Schaub, P Schlattmann, and B Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Stat. Math.* , 46:373–388, 1994.

[5] WM Brown, EM Prager, A Wang, and AC Wilson. Mitochondrial DNA sequences of primates, tempo and mode of evolution. *Journal of Molecular Evolution*, 18:225–239, 1982.

[6] B Chor. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Mol. Biol. Evol.*, 17:1529–1541, 2000.

[7] A Cuyt, B Verdonk, S Becuwe, and P Kuterna. A remarkable example of catastrophic cancellation unraveled. *Computing*, 66:309–320, 2001.

[8] P. Diaconis and L. Saloff-Coste. What do we know about the metropolis algorithm ? *Jnl. Comp. Sys. Sci.*, 57:20–36, 1998.

[9] E Dogen and K Tanahashi. *Moon in a dew drop: Writings of Zen master Dogen*. Ferrar, Straus and Giroux, 1995.

[10] RT Durrett. *Probability: Theory and Examples*. Duxbury, 1996.

[11] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[12] R Hammer, M Hocks, U Kulisch, and D Ratz. *C++ toolbox for verified computing: basic numerical problems*. Springer-Verlag, 1995.

[13] E Hansen. Global optimization using interval analysis - the multi-dimensional case. *Numerische Mathematik*, 34:247–270, 1980.

[14] E Hansen. *Global optimization using interval analysis*. Marcel Dekker, 1992.

[15] E Hansen and S Sengupta. Bounding solutions of systems of equations using interval analysis. *BIT*, 21:203–211, 1981.

[16] W Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[17] D Hume. An enquiry concerning human understanding: Section VI - OF PROBABILITY. In CW Eliot, editor, *The Harvard classics:English philosophers of the seventeenth and eighteenth centuries, 1910 edition*, volume 37. The Collier Press, 1777.

[18] Information management and technology division. Patriot missile defense: Software problem led to system failure at Dhahran, Saudi Arabia. Report GAO/IMTEC-92-26, United States, General accounting office, 1992.

[19] JL Lions, Chairman of the inquiry board. Ariane 5, flight 501 failure: Report by the inquiry board. Inquiry, European Space Agency, 1996.

[20] TH Jukes and C Cantor. Evolution of protein molecules. In HN Munro, editor, *Mammalian Protein Metabolism*, pages 21–32. New York Academic Press, 1969.

[21] DJ Kalupahana. *Mūlamadhyamakakārikā of Nāgārjuna: The philosophy of the middle way*. Motilal Banarsidass, 1991.

[22] U Kulisch. Advanced arithmetic for the digital computer, interval arithmetic revisited. In U Kulisch, R Lohner, and A Facius, editors, *Perspectives on encolsure methods*, pages 50–70. Springer-Verlag, 2001.

[23] U Kulisch, R Lohner, and A Facius, editors. *Perspectives on encolsure methods*. Springer-Verlag, 2001.

[24] J Liu. Metropolised independent sampling and comparisons to rejection sampling and importance sampling. *Statist. and Comput.* , 6:113–119, 1995.

[25] E Loh and GW Walster. Rump's example revisited. *Reliable Computing*, 8:245–248, 2002.

[26] N Madras. *Lecture notes on Monte Carlo methods*. American Mathematical Society, 2002.

[27] G Mayer. Result verification for eigenvectors and eigenvalues. In J Herzberger, editor, *Topics in validated computations*, volume 5 of *Studies in computational mathematics*, pages 209–276. North-Holland, New York, 1994.

[28] G McLachlan and D Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[29] N Metropolis, A Rosenbluth, M Rosenbluth, A Teller, and E Teller. Equations of state calculations by fast computing machines. *Jnl. Chem. Phys.* , 21:1087–1092, 1953.

[30] RE Moore. *Interval analysis*. Prentice-Hall, 1967.

[31] RE Moore. *Methods and Applications of Interval analysis*. SIAM, Philadelphia, Pennsylvania, 1979.

[32] A Neumaier. *Interval methods for systems of equations*. Cambridge university press, 1990.

[33] JV Neumann. Various techniques used in connection with random digits. In *John Von Neumann, Collected Works*, volume V. Oxford University Press, 1963.

[34] IEEE Task P754. *ANSI/IEEE 754-1985, Standard for Binary Floating-Point Arithmetic*. IEEE, New York, 1985.

[35] WH Press, SA Teukolsky, WT Vetterling, and BP Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 1992.

[36] LB Rall. *Automatic differentiation, techniques and applications*, volume 120 of *Springer lecture notes in computer science*. Springer-Verlag, 1981.

[37] D Ratz. *Automatische ergebnisverifikation bei globalen optimierungsproblemen*. pHd dissertation, Universitat Karlsruhe, Karlsruhe, Germany, 1992.

[38] R Sainudiin and R Yoshida. Applications of interval methods to phylogenetic trees. In L Pachter and B Sturmfels, editors, *Algebraic statistics for computational biology*. Cambridge University Press, 2005.

[39] C Semple and M Steel. *Phylogenetics*. Oxford University Press, 2003.

[40] W Tucker. Auto-validating numerical methods. Lecture notes, Uppsala University, 2004.

[41] L Wittgenstein. *Philosophical Investigations*. The Macmillan Company, 1965.

[42] Z Yang. Complexity of the simplest phylogenetic estimation problem. *Proceedings Royal Soc. London B Biol. Sci.*, 267:109–119, 2000.