A Lambda-Calculus Foundation for Universal Probabilistic Programming

Johannes Borgström* Ugo Dal Lago† Andrew D. Gordon‡
Marcin Szymczak§

January 21, 2016

Abstract

We develop the operational semantics of a probabilistic λ-calculus with continuous distributions, as a foundation for universal probabilistic programming languages such as Church, Anglican, and Venture. Our first contribution is to adapt the classic operational semantics of λ-calculus to the continuous case, via creating a measure space on terms and defining step-indexed approximations. We prove equivalence of big-step and small-step formulations of this *distributional semantics*. Our second contribution is to formalize the implementation technique of trace MCMC for our calculus and to show correctness. A key step is defining a *sampling semantics* of a term as a function from a trace of random samples to a value, and showing that the distribution induced by integrating over all traces equals the distributional semantics. Another step is defining sufficient conditions for the distribution induced by trace MCMC to converge to the distributional semantics. To the best of our knowledge, this is the first rigorous correctness proof for trace MCMC for a higher-order functional language.

1 Introduction

Church [10] introduced *universal probabilistic programming*, the idea of writing probabilistic models for machine learning in a Turing-complete functional programming language. Church, and its descendants Venture [15], Anglican [21], and Web Church [9] are dialects of Scheme. Another example of universal probabilistic programming is webppl [8], a probabilistic interpretation of JavaScript.

For example, the following program defines a probability p at random, defines a function to flip a coin with bias p, conditions the model on single observations of 0 and 1, and returns p. We use uniform() to sample a probability from the uniform distribution on the unit interval.

```
let p = uniform()
let flip() = uniform() < p
if (flip() = 0) and (flip() = 1)
then p
else fail
```

1.1 Semantics for Universal Probabilistic Programming

The first problem we address in this work is to provide an operational semantics for universal probabilistic programming languages. Our example illustrates the common situation in machine

*Uppsala University
†University of Bologna & INRIA
‡Microsoft Research and University of Edinburgh
§University of Edinburgh

learning that models are based on continuous distributions (such as uniform), but previous works on operational semantics for probabilistic λ-calculi are based on discrete distributions.

We introduce a call-by-value λ-calculus with primitives for random draws from various continuous distributions, and exceptions to represent conditioning. We describe a metric space of λ-terms and let $\mathscr{D}$ range over *distributions*, that is, sub-probability Borel measures on terms of the λ-calculus. We define step-indexed operational semantics, in both small-step ($M \to_n \mathscr{D}$) and big-step ($M \Downarrow_n \mathscr{D}$) styles, which we prove equivalent, and enable us to define the *distributional semantics* $\mathscr{D} = [\![M]\!]$ for each closed term M.

1.2 Semantics and Correctness of Trace MCMC

The original implementation of Church introduced the implementation technique *trace MCMC* [10]. A closed λ-term M has a *trace* s if there is a run of M making a finite sequence of random choices s, which yields a result $G(s)$ functionally dependent on s. In our example, each trace has form $s = [p, q_1, q_2]$ where p is the bias of the coin, and each binary flip b_i is true if and only if $q_i < p$. We give a deterministic *sampling semantics* for our calculus, conditional on an explicit trace s of random draws, that produces an explicit weight w for each trace. We formulate the sampling semantics in big-step style ($M \Downarrow^s_w G$) and small-step style ($(M, w, s) \to (M', w', s')$) and prove equivalence. Moreover, we prove that the value distributions induced by the sampling and distributional semantics are indeed the same.

We consider trace MCMC as an instance of the Metropolis-Hastings (MH) algorithm. Given a closed term M, trace MCMC generates a Markov chain of traces, with a stationary distribution on traces that induces a distribution over values corresponding to the semantics of M. The algorithm is parametric in a target distribution $\mathscr{D}$ and a proposal kernel Q, a function that maps a trace s of M to a probability distribution over traces, used to choose the next trace in the Markov chain.

We formalize the algorithm rigorously, defining the target distribution on program traces and the proposal kernel as Lebesgue integrals of corresponding density functions. We prove these functions measurable with respect to the σ-algebra on program traces—a step usually omitted in similar developments.

We exploit the equivalence results for the different semantics to show that, subject to sufficient aperiodicity and irreducibility conditions on the transition kernel induced by Q and the acceptance ratio, the distribution on values induced by trace MCMC on M converges to the distributional semantics $\mathscr{D} = [\![M]\!]$.

1.3 Structure of the Paper

Section 2 recalls some standard definitions from measure theory.

Section 3 defines the syntax of our probabilistic λ-calculus with draws from continuous distributions. We include an exception mechanism to represent conditioning.

Section 4 defines our step-indexed operational semantics, in both small-step ($M \to_n \mathscr{D}$) and big-step ($M \Downarrow_n \mathscr{D}$) styles, which by Theorem 1 are equivalent, and define $[\![M]\!]$, a distribution over values, for each closed term M.

Section 5 makes an important auxiliary construction, a measure space on the set of program traces, that is, sequences of reals of arbitrary length.

Section 6 defines *sampling-based* operational semantics for our calculus. The semantics is based on the explicit consumption of a program trace s of random draws and production of an explicit weight w for each outcome. We formulate the operational semantics in two standard styles and show equivalence: big-step semantics, $M \Downarrow^s_w G$, and small-step semantics, $(M, w, s) \to (M', w', s')$. Theorem 2 establishes a precise equivalence between the sampling-based semantics and the distributional semantics of Section 4.

Section 7 formalizes trace MCMC for our calculus, in the spirit of Hur and others. Theorem 3 shows equivalence between the distribution computed by the algorithm and the semantics of the previous sections. Hence, Theorem 3 is the first correctness theorem for trace MCMC for a λ-calculus.

Section 8 describes related work and Section 9 concludes.
Appendix A collects various proofs of measurability.

1.4 Example of soft conditioning

This is not yet in Church style.

```
let m=Gaussian(0,1) in condition (Gaussian(m,1)) 4.5 m
```

Derivable variation:

```
condition m o y =
  let d = |m-o|  // d >=0
  if B(e^-d) then y else fail
```

2 Preliminaries

We begin by recapitulating some standard definitions for sub-probability distributions and kernels over metric spaces.

A σ-algebra is a set Σ that contains $\varnothing$, and is closed under complement and countable union (and hence is closed under countable intersection). Let the σ-algebra *generated* by S, written $\sigma(S)$, be the set $\sigma(S)$ that is the least σ-algebra over $\cup S$ that is a superset of S. In other words, $\sigma(S)$ is the least set such that:

1. we have $S \subseteq \sigma(S)$ and $\varnothing \in \sigma(S)$; and
2. $((\cup S) \setminus A) \in \sigma(S)$ if $A \in \sigma(S)$; and
3. $\cup_{i \in \mathbb{N}} A_i \in \sigma(S)$ if each $A_i \in \sigma(S)$.

An equivalent definition is that $\sigma(S) \triangleq \bigcap\{\Sigma \mid S \subseteq \Sigma \text{ and } \Sigma \text{ is a } \sigma\text{-algebra}\}$. We write $\mathbb{R}_+$ for $[0, \infty]$ and $\mathbb{R}_{[0,1]}$ for the interval $[0, 1]$. A *metric space* is a set X with a symmetric *distance function* $\delta : X \times X \to \mathbb{R}_+$ that satisfies the triangle inequality $\delta(x, z) \leq \delta(x, y) + \delta(y, z)$ and the axiom $\delta(x, x) = 0$. We write $\mathbf{B}(x, r) \triangleq \{y \mid \delta(x, y) < r\}$ for the open ball around x of radius r. We equip $\mathbb{R}_+$ and $\mathbb{R}_{[0,1]}$ with the standard metric $\delta(x, y) = |x - y|$, and products of metric spaces with the Manhattan metric (e.g., $\delta((x_1, x_2), (y_1, y_2)) = \delta(x_1, y_1) + \delta(x_2, y_2)$).

The *Borel σ-algebra* on a metric space (X, δ) is $\mathcal{B}(X, \delta) \triangleq \sigma(\{\mathbf{B}(x, r) \mid x \in X \wedge r > 0\})$. We often omit the arguments to $\mathcal{B}$ when they are clear from the context.

A *measurable space* is a pair (X, Σ) where X is a set of possible outcomes, and $\Sigma \subseteq \mathcal{P}(X)$ is a σ-algebra. As an example, we consider the extended positive real numbers $\mathbb{R}_+$ equipped with the Borel σ-algebra $\mathcal{R}$ (with respect to the standard metric) is the set $\sigma(\{(a, b) \mid a, b \geq 0\})$, which is the smallest σ-algebra containing all open (and closed) intervals. We can create finite products of measure spaces by iterating the construction $(X, \Sigma) \times (X', \Sigma') = (X \times X', \sigma(A \times B \mid A \in \Sigma \wedge B \in \Sigma'))$.

If (X, Σ) and (X', Σ') are measurable spaces, then the function $f : X \to X'$ is *measurable* if and only if for all $A \in \Sigma'$, $f^{-1}(A) \in \Sigma$, where the *inverse image* $f^{-1} : \mathcal{P}(X') \to \mathcal{P}(X)$ is given by $f^{-1}(A) \triangleq \{x \in X \mid f(x) \in A\}$.

A *measure* μ on (X, Σ) is a function from Σ to $\mathbb{R}_+$, that is (1) zero on the empty set, that is, $\mu(\varnothing) = 0$, and (2) countably additive, that is, $\mu(\cup_i A_i) = \Sigma_i \mu(A_i)$ if $A_1, A_2, \ldots$ are pair-wise disjoint. The measure μ is called a *probability measure* if $\mu(X) = 1$, and a *sub-probability measure* if $\mu(X) \leq 1$. For any element x of X, the Dirac measure $\delta(x)$ is defined as follows:

$$\delta(M)(A) = \begin{cases} 1 & \text{if } x \in A; \\ 0 & \text{otherwise} \end{cases}$$

A *measure space* is a triple $\mathscr{M} = (X, \Sigma, \mu)$ where μ is a measure on the measurable space (X, Σ). Given a measurable function $f : X \to \mathbb{R}_+$, the *integral* of f over $\mathscr{M}$ can be defined following Lebesgue's theory and denoted as either of

$$\int f \, d\mu = \int f(x) \, \mu(dx) \in \mathbb{R}_+.$$

If $A \in \Sigma$ we write 1_A for the function that is 1 on A and 0 elsewhere. We then write

$$\int_A f \, d\mu \triangleq \int f(x) \cdot 1_A(x) \, \mu(dx).$$

We equip some measurable spaces (X, Σ) with a *stock measure* μ. When f is measurable $f : X \to \mathbb{R}_+$ we write $\int f(s)\, ds$ (or shorter, $\int f$) for $\int f \, d\mu$. In particular, we let the stock measure on $(\mathbb{R}^n, \mathcal{B})$ be the Lebesgue measure.

If (X, Σ) and (Y, Σ') are measurable spaces, then a function $Q : X \times \Sigma' \to \mathbb{R}_{[0,1]}$ is called a *(subprobability) kernel* (from (X, Σ) to (Y, Σ')) if

1. for every $x \in X$, $Q(x, \cdot)$ is a subprobability measure on (Y, Σ'); and
2. for every $A \in \Sigma'$, $Q(\cdot, A)$ is a non-negative measurable function $X \to \mathbb{R}_{[0,1]}$.

Q is said to be a *probability kernel* if $Q(x, Y) = 1$ for all $x \in X$. When Q is a kernel, note that $\int f(y)\, Q(x, dy)$ denotes the integral of f with respect to the measure $Q(x, \cdot)$.

Kernels can be composed in the following ways: If Q_1 is a kernel from (X_1, Σ_1) to (X_2, Σ_2) and Q_2 is a kernel from (X_2, Σ_2) to (X_3, Σ_3), then $Q_2 \circ Q_1 : (x, A) \mapsto \int Q_2(y, A)\, Q_1(x, dy)$ is a kernel from (X_1, Σ_1) to (X_3, Σ_3). Moreover, if Q_1 is a kernel from (X_1, Σ_1) to (X_2, Σ_2) and Q_2 is a kernel from (X_1', Σ_1') to (X_2', Σ_2'), then $Q_1 \times Q_2 : ((x, y), (A \times B)) \mapsto Q_1(x, A) \cdot Q_2(y, B)$ uniquely extends to a kernel from $(X_1, \Sigma_1) \times (X_1', \Sigma_1')$ to $(X_2, \Sigma_2) \times (X_2', \Sigma_2')$.

The function f is a *density* of a measure μ (with respect to the measure ν) if $\mu(A) = \int_A f \, d\nu$ for all measurable A. Similarly, if $Q : X \times \Sigma' \to \mathbb{R}_+$ is a kernel then $q : X \times (Y) \to \mathbb{R}_+$ is said to be a density of Q if $Q(v, A) = \int_A q(v, y)\, \nu(dy)$ for all $A \in \Sigma$ and $v \in X$.

3 Syntax

We represent scalar data as real numbers $c \in \mathbb{R}$. We use 0 and 1 to represent `false` and `true`, respectively.

Let $\mathcal{E}$ be a set of exception identifiers, ranged over by metavariables like α and β. We assume that $\texttt{fail}, \texttt{error} \in \mathcal{E}$. The exception `error` represents a run-time error (such as applying a constant as if it were a function) while the exception `fail` represents conditioning of the distribution defined by a term.

Let $\mathcal{I}$ be a countable set of *distribution identifiers* (or simply *distributions*). Metavariables for distributions are D, E. Each distribution identifier D has an integer *arity* $|\mathsf{D}| > 0$. Each distribution identifier D defines the density $\mathrm{pdf}_{\mathsf{D}} : \mathbb{R}^{|\mathsf{D}|} \times \mathbb{R} \to \mathbb{R}$ of a sub-probability kernel $\mu_{\mathsf{D}(\vec{c})} : \mathcal{M}^{\mathbb{R}} \to \mathbb{R}_{[0,1]}$.

Let g be a metavariable ranging over a countable set of *function identifiers* each with an integer *arity* $|g| > 0$ and with an interpretation as a total measurable function $\sigma_g : \mathbb{R}^{|g|} \to \mathbb{R}$. Examples of function identifiers include addition $+$, comparison $>$, and equality $=$. We define the *values* V and *terms* M as follows, where x ranges over a denumerable set of variables $\mathcal{X}$.

$$\begin{aligned} V &::= c \mid x \mid \lambda x.M \\ M, N &::= V \mid M\ N \mid \mathsf{D}(V_1, \ldots, V_{|\mathsf{D}|}) \mid g(V_1, \ldots, V_{|g|}) \\ &\quad \mid \texttt{if}\ V\ \texttt{then}\ M\ \texttt{else}\ L \mid \alpha \end{aligned}$$

$$
\begin{aligned}
d(x,x) &= 0\\
d(c,d) &= |c-d|\\
d(MN,LP) &= d(M,L)+d(N,P)\\
d(g(V_1,\ldots,V_n),&g(W_1,\ldots,W_n))\\
&= d(V_1,W_1)+\cdots+d(V_n,W_n)\\
d(\lambda x.M,\lambda x.N) &= d(M,N)\\
d(\mathsf{D}(V_1,\ldots,V_n),&\mathsf{D}(W_1,\ldots,W_n))\\
&= d(V_1,W_1)+\cdots+d(V_n,W_n)\\
d(\texttt{if } V \texttt{ then } M \texttt{ else } N,&\texttt{if } W \texttt{ then } L \texttt{ else } P)\\
&= d(V,W)+d(M,L)+d(N,P)\\
d(\alpha,\alpha) &= 0\\
d(M,N) &= \infty \text{ otherwise}
\end{aligned}
$$

Figure 1: Metric d on terms.

As usual, free occurrences of x inside M are bound by $\lambda x.M$. Terms are taken modulo renaming of bound variables. Substitution of all free occurrences of x by a value V in M is defined as usual, and denoted $M\{V/x\}$. This can be easily generalized to $M\{\vec{V}/\vec{x}\}$, where $\vec{x}$ is a sequence of variables and $\vec{V}$ is a sequence of values (of the same length). K is the term $\lambda x.\lambda y.x$, which takes two arguments and discards the second one. Λ denotes the set of all terms. Given a set $X \subseteq \mathcal{X}$ of variables, we indicate with $\Lambda_P(X)$ the set of all terms with free variables among those in X, and with $C\Lambda$ the set $\Lambda_P(\varnothing)$ of *closed* terms. We define $\mathcal{M}$ to be the set of Borel-measurable sets of terms equipped with the recursively defined metric d in Figure **??**.

Given a measurable subset of terms A, $\mathcal{M}^A$ is the *restriction* of $\mathcal{M}$ to elements *in* A, *i.e.*, $\mathcal{M}^A = \{B \cap A \mid B \in \mathcal{M}\}$. Any pair in the form $\mathscr{M}_A = (A, \mathcal{M}^A)$ is by construction a measurable space. Any non-negative measure μ on $\mathcal{M}^A$ turns it into a measure space $\mathscr{M}_A^\mu = (A, \mathcal{M}^A, \mu)$.

A term M is said to be *skeleton* iff no real number occurs in M, and each variable occurs *at most once* in M. The set of skeletons is SK. Any closed term M can be written as $N\{\vec{c}/\vec{x}\}$, where N is a skeleton. The set of closed terms corresponding this way to a skeleton $M \in \mathsf{SK}$ is denoted as $\mathsf{TM}(M)$. If the underlying term is a skeleton, substitution can be defined also when the substituted terms are *sets* of values rather than mere values, because variables occurs at most once; in that case, we will used the notation $M\{X/x\}$, where X is any set of values.

4 Operational Semantics

We define call-by-value evaluation. The set of all *closed values* is $\mathcal{V}$, and we write $\mathcal{V}_\lambda$ for $\mathcal{V} \setminus \mathbb{R}$. *Evaluation contexts* are defined as follows:

$$E ::= [\cdot] \;\Big|\; EM \;\Big|\; (\lambda x.M)E$$

We let $\mathcal{C}$ be the set of all closed evaluation contexts, i.e., where every occurrence of a variable x is as a subterm of $\lambda x.M$.

The term obtained by replacing the only occurrence of $[\cdot]$ in E by M is indicated as $E[M]$. If μ is a measure on terms, we let $E\{\mu\}$ be the push-forward measure $A \mapsto \mu(\{M \mid E[M] \in A\})$. *Generalized values* are elements G, H of the set $\mathcal{GV} = \mathcal{V} \cup \mathcal{E}$, *i.e.*, generalized values are either values or exceptions. *Erroneous redexes* are closed terms in one of the following three forms:

- $\mathsf{D}(V_1,\ldots,V_{|\mathsf{D}|})$ where at least one of the V_i is a λ-abstraction.
- $g(V_1,\ldots,V_{|g|})$ where at least one of the V_i is a λ-abstraction.

$$
\begin{aligned}
E[g(\vec{c}] &\xrightarrow{\text{det}} E[\sigma_g(\vec{c})] \\
E[(\lambda x.M)\ V] &\xrightarrow{\text{det}} E[M\{V/x\}] \\
E[c\ V] &\xrightarrow{\text{det}} E[\texttt{error}] \\
E[\texttt{if 1 then } M_2 \texttt{ else } M_3] &\xrightarrow{\text{det}} E[M_2] \\
E[\texttt{if 0 then } M_2 \texttt{ else } M_3] &\xrightarrow{\text{det}} E[M_3] \\
E[T] &\xrightarrow{\text{det}} E[\texttt{error}] \\
E[\alpha] &\xrightarrow{\text{det}} \alpha \quad \text{if } E \text{ is not } [\cdot]
\end{aligned}
$$

Figure 2: Deterministic Reduction.

- `if` V `then` M `else` L, where V is neither `true` nor `false`.

Erroneous redexes are ranged over by metavariables like T, R. *Deterministic reduction* is the relation $\xrightarrow{\text{det}}$ on closed terms defined in Figure 2

One-step evaluation is a relation $M \to \mathscr{D}$ between closed terms M and sub-probability measures $\mathscr{D}$ on terms, defined as follows:

$$
\begin{aligned}
E[\mathsf{D}(\vec{c})] &\to E\{\mu_{\mathsf{D}(\vec{c})}\} \\
E[M] &\to \delta(E[N]) \text{ if } M \xrightarrow{\text{det}} N
\end{aligned}
$$

Redexes are generated by the following grammar

$$
\begin{aligned}
R ::= &(\lambda x.M)V \ \Big|\ cM \ \Big|\ \mathsf{D}(\vec{c}) \ \Big|\ g(\vec{c}) \ \Big| \\
&\alpha \ \Big|\ \texttt{if true then } M \texttt{ else } N \ \Big|\ \texttt{if false then } M \texttt{ else } N \ \Big|\ T
\end{aligned}
$$

Lemma 1 *For every closed term M, either M is a generalized value or there are unique E, R such that $M = E[R]$. Moreover, if M is not a generalized value and $R = \alpha$, then E is proper, that is, $E \neq [\cdot]$.*

Proof. This is an easy induction on the structure of M. □

Reducible terms are those closed terms M such that M can be written as $E[R]$. The set of all reducible terms is denoted as $R\Lambda_P$.

Lemma 2 *For every closed term M, either M is a generalized value or there is a unique $\mathscr{D}$ such that $M \to \mathscr{D}$.*

Proof. An easy consequence of Lemma 1. □

Lemma 3 *$\to$ is a subprobability kernel.*

Proof. Lemma 2 already tells us that $\to$ can be seen as a function $\hat{\to}$ defined as follows:

$$
\hat{\to}(M, A) = \begin{cases} \mathscr{D}(A) & \text{if } M \to \mathscr{D}; \\ 0 & \text{otherwise.} \end{cases}
$$

The fact that $\hat{\to}(M, \cdot)$ is a subprobability measure is easily verified. On other hand, the fact that $\hat{\to}(\cdot, A)$ is measurable amounts to proving that $OS(A, B) = (\hat{\to}(\cdot, A))^{-1}(B)$ is a measurable set of

$$\frac{n > 0}{G \to_n \delta(G)} \qquad \frac{}{M \to_0 \mathbf{0}}$$

$$\frac{M \to \mathscr{D} \qquad \{N \to_n \mathscr{E}_N\}_{N \in \mathrm{supp}(\mathscr{D})}}{M \to_{n+1} A \mapsto \int \mathscr{E}_N(A)\, \mathscr{D}(dN)}$$

Figure 3: Step-Indexed Approximation Small-Step Semantics.

terms whenever B is a measurable set of real numbers. We will do that by showing that for every skeleton N, the set $OS(A, B) \cap \mathsf{TM}(N)$ is measurable. The thesis then follows by observing that

$$OS(A, B) = \bigcup_{N \in \mathsf{SK}} OS(A, B) \cap \mathsf{TM}(N)$$

and that SK is countable. Now, let us observe that for every skeleton N, the nature of any term L in $\mathsf{TM}(N)$ as for if being a value, or containing a deterministic redex, or containing a sampling redex, only depends on N and not on the term L. As an example, terms in $\mathsf{TM}(xy)$ are nothing but deterministic redexes (actually, all of them rewrites deterministically to $\delta(\texttt{error})$. This allows us to proceed by distinguishing three cases:

- If all terms in $\mathsf{TM}(N)$ are values, then it can be easily verified that

$$OS(A, B) \cap \mathsf{TM}(N) = \begin{cases} \mathsf{TM}(N) & \text{if } 0 \in B; \\ \varnothing & \text{if } 0 \notin B. \end{cases}$$

  Both when $0 \in B$ and when $0 \notin B$, then, $OS(A, B) \cap \mathsf{TM}(N)$ is indeed measurable.
- If all terms in $\mathsf{TM}(N)$ contain deterministic redexes, then

$$OS(A, B) \cap \mathsf{TM}(N) = \begin{cases} \to^{-1}(A) \cap \mathsf{TM}(N) & \text{if } 1 \in B \\ \varnothing & \text{if } 1 \notin B. \end{cases}$$

  Since deterministic reduction $\to$ is known to be measurable, then both when $1 \in B$ and when $1 \notin B$, the set $OS(A, B) \cap \mathsf{TM}(N)$ is measurable.
- The hardest case is when N is of the form $G[\mathsf{D}(\vec{x})]$, where G is an evaluation context. In this case, however, we can proceed by decomposing the function we want to prove measurable into three measurable functions:
  - The function $app : \mathcal{C} \times C\Lambda \to C\Lambda$, which given an evaluation context E and a term M, returns the term $E[M]$. This is proved measurable in the Appendix.
  - The function $deapp : R\Lambda_P \to \mathcal{C} \times C\Lambda$ which "splits" a term in $R\Lambda_P$ into an evaluation context E and a closed term M. This is proved measurable in the Appendix.
  - For every distribution identifier D, the function $distapp_{\mathsf{D}} : \mathbb{R}^n \to C\Lambda$ (where n is the arity of D) which, given a tuple of real numbers x, returns the term $\mathsf{D}(x)$. This function is a continuous function between two metric spaces, so measurable.
  - We know that for every distribution identifier D, there is a kernel $\mu_{\mathsf{D}} : \mathbb{R}^n \times \Sigma_{\mathbb{R}} \to \mathbb{R}_{[0,1]}$. Moreover, one can also consider the Dirac kernel on evaluation contexts, namely $I : \mathcal{C} \times \Sigma_{\mathcal{C}} \to \mathbb{R}_{[0,1]}$ where $I(E, A) = \delta(E)(A)$. Then, the product $\mu_{\mathsf{D}} \times I$ is also a kernel, so measurable.

This concludes the proof.

Given a family $\{\mathscr{D}_M\}_{M \in A}$ of A-subdistributions indexed by terms in a measurable set B, and a measurable set A of terms from A, we often write, with an abuse of notation, $\mathscr{D}_M(A)$ for the function which assigns to any term $M \in A$ the real number $\mathscr{D}_M(A)$.

The *step-indexed approximation small-step semantics* is the family of n-indexed relations $M \to_n \mathscr{D}$ between terms and distributions inductively defined as follows. Since generalised values have no transitions (there is no $\mathscr{D}$ such that $G \to \mathscr{D}$), the rules above are disjoint and so there is at most one $\mathscr{D}$ such that $M \to_n \mathscr{D}$.

$$\frac{n>0}{G \Downarrow_n \delta(G)} \qquad \frac{}{M \Downarrow_0 \mathbf{0}} \qquad \frac{n>0}{T \Downarrow_n \delta(\mathtt{error})} \qquad \frac{n>0}{\mathsf{D}(\vec{c}) \Downarrow_n \mu_{\mathsf{D}(\vec{c})}} \qquad \frac{n>0}{g(\vec{c}) \Downarrow_n \delta(\sigma_g(\vec{c}))}$$

$$\frac{M \Downarrow_n \mathscr{D}}{\mathtt{if\ true\ then}\ M\ \mathtt{else}\ N \Downarrow_{n+1} \mathscr{D}} \qquad \frac{N \Downarrow_n \mathscr{D}}{\mathtt{if\ false\ then}\ M\ \mathtt{else}\ N \Downarrow_{n+1} \mathscr{D}}$$

$$\frac{M \Downarrow_n \mathscr{D} \qquad N \Downarrow_n \mathscr{E} \qquad \{L\{V/x\} \Downarrow_n \mathscr{E}_{L,V}\}_{(\lambda x.L)\in\mathrm{supp}(\mathscr{D}),V\in\mathrm{supp}(\mathscr{E})}}{MN \Downarrow_{n+1} A \mapsto \mathscr{D}^{\mathcal{E}}(A) + \mathscr{D}(\mathbb{R}) \cdot \delta(\mathtt{error}) + \mathscr{D}(\mathcal{V}_\lambda) \cdot \mathscr{E}^{\mathcal{E}}(A) + \iint \mathscr{E}_{L,V}(A)\, \mathscr{D}^{\mathcal{V}_\lambda}(\lambda x.dL)\, \mathscr{E}^{\mathcal{V}}(dV)}$$

Figure 4: Step Indexed Approximation Big-Step Semantics.

Lemma 4 *For every $n \in \mathbb{N}$, the function $\to_n$ is a kernel.*

Proof. By induction on n:

- $\to_0$ can be seen as the function $\hat{\to}_0$ which attributes 0 to any pair (M, A). This is clearly a kernel.
- $\to_{n+1}$ can be seen as the function $\hat{\to_{n+1}}$ defined as follows:

$$\hat{\to}_{n+1}(M, A) = \begin{cases} 1 & \text{if } M \in \mathcal{GV} \text{ and } M \in A; \\ 0 & \text{if } M \in \mathcal{GV} \text{ and } M \notin A; \\ (\int \hat{\to}_n(N, A)\, \mathscr{D}(dN)) & \text{if } M \to \mathscr{D}. \end{cases}$$

The fact that $\hat{\to}_{n+1}(M, \cdot)$ is a measure for every M is clear, and can be proved by case distinction on M. On the other hand, if B is a measurable set of reals, then:

$$(\hat{\to}_{n+1}(\cdot, A))^{-1}(B) = (\hat{\to}_{n+1}(\cdot, A))^{-1}(B) \cap \mathcal{GV} \cup (\hat{\to}_{n+1}(\cdot, A))^{-1}(B) \cap (C\Lambda - \mathcal{GV}).$$

Now, the fact that $(\hat{\to}_{n+1}(\cdot, A))^{-1}(B) \cap \mathcal{GV}$ is a measurable set of terms is clear: it is $A \cap \mathcal{GV}$ if $1 \in B$ and $\varnothing$ otherwise. But how about $(\hat{\to}_{n+1}(\cdot, A))^{-1}(B) \cap (C\Lambda - \mathcal{GV})$? In that case, we just need to notice that

$$(\hat{\to}_{n+1}(\cdot, A))^{-1}(B) \cap (C\Lambda - \mathcal{GV}) = (\hat{\to}_n \circ \to)^{-1}(B)$$

where $\hat{\to}_n$ and $\to$ are kernels (the former by induction hypothesis, the latter by Lemma 3). Since kernels compose, this concludes the proof.

Lemma 5 *For every closed term M and for every $n \in \mathbb{N}$ there is a unique distribution $\mathscr{D}$ such that $M \to_n \mathscr{D}$.*

Proof. This is an easy consequence of Lemma 4.

Given a measurable set of terms A, an *A-distribution* is a sub-probability measure on A, that is, a measure $\mathscr{D} : \mathcal{M}^A \to \mathbb{R}_{[0,1]}$ such that $\mathscr{D}(A) \leq 1$. We write $\mathbf{0}$ for the zero distribution $A \mapsto 0$. Given a X-distribution $\mathscr{D}$ and a measurable subset $B \subseteq A$, $\mathscr{D}^B$ is the *restriction* of $\mathscr{D}$ to elements *in* B, *i.e.*, $\mathscr{D}^B(A) = \mathscr{D}(A \cap B)$. As a consequence, the restriction of $\mathscr{D}$ to elements *not in* A is $\mathscr{D}^{\mathcal{GV}-A}$.

A *value distribution* is a *$\mathcal{V}$-distribution*, that is, a sub-probability measure $\mathscr{D} : \mathcal{M}^{\mathcal{V}} \to \mathbb{R}_{[0,1]}$ such that $\mathscr{D}(\mathcal{V}) \leq 1$.

The *step-indexed approximation big-step semantics* $M \Downarrow_n \mathscr{D}$ is the n-indexed family of relations inductively defined by the rules in Figure 4.

Above, the rule for applications is the most complex, with the resulting distribution consisting of three exceptional terms in addition to the normal case. To better understand this rule, one can study what happens if we replace general applications with a let construct plus application of values to values. Then we would end up having the following three rules, instead of the rule for application above:

$$\frac{M \Downarrow_n \mathscr{D} \qquad \{N\{V/x\} \Downarrow_n \mathscr{E}_V\}_{V \in \mathrm{supp}(\mathscr{D})}}{\texttt{let}\ x = M\ \texttt{in}\ N \Downarrow_{n+1} A \mapsto \mathscr{D}^{\mathcal{E}}(A) + \mathscr{D}(\mathbb{R}) \cdot \delta(\texttt{error}) + \int \mathscr{E}_V(A)\, \mathscr{D}^{\mathcal{V}}(dV)}$$

$$\frac{M\{V/x\} \Downarrow_n \mathscr{E}}{(\lambda x.M)V \Downarrow_{n+1} \mathscr{E}} \qquad \frac{n > 0}{c\ V \Downarrow_n \delta(\texttt{error})}$$

The set of value distributions with the pointwise order forms a ω**CPO**, and thus any denumerable, directed set of value distributions has a least upper bound. One can define the *small-step semantics* and the *big-step semantics* as, respectively, the distributions

$$[\![M]\!]_{\Rightarrow} = \sup\{\mathscr{D} \mid M \to_n \mathscr{D}\}$$
$$[\![M]\!]_{\Downarrow} = \sup\{\mathscr{D} \mid M \Downarrow_n \mathscr{D}\}$$

Lemma 6 (Monotonicity) *If $M \to_n \mathscr{D}$, $m \geq n$ and $M \to_m \mathscr{E}$, then $\mathscr{E} \geq \mathscr{D}$.*

Lemma 7 *If $M \to_n \mathscr{D}$, $N \to_m \mathscr{E}$, and for all L and V, $L\{V/x\} \to_p \mathscr{E}_{L,V}$, then $MN \to_{n+m+p} \mathscr{F}$ such that for all A*

$$\mathscr{F}(A) \geq \mathscr{D}^{\mathcal{E}}(A) + \mathscr{D}(\mathbb{R}) \cdot \delta(\texttt{error}) + \mathscr{D}(\mathcal{V}_\lambda) \cdot \mathscr{E}^{\mathcal{E}}(A) + \iint \mathscr{E}_{L,V}(A)\, \mathscr{D}^{\mathcal{V}_\lambda}(\lambda x.dL)\, \mathscr{E}^{\mathcal{V}}(dV).$$

Proof. First of all, one can prove that if $N \to_n \mathscr{D}$ and $L\{V/x\} \to_m \mathscr{E}_V$ for all V then $(\lambda x.L)N \to_{n+m} \mathscr{F}$ where $\mathscr{F}(A) \geq \mathscr{D}^{\mathcal{E}}(A) + \int \mathscr{E}_V(A)\, \mathscr{D}^{\mathcal{V}}(dV)$ for all A. This is an induction on n.

- If $n = 0$, then $\mathscr{D}$ is necessarily the zero distribution $A \mapsto 0$. Then $\mathscr{F}(A) \geq 0 = \mathscr{D}^{\mathcal{E}}(A) + \int \mathscr{E}_V(A)\, \mathscr{D}^{\mathcal{V}}(dV))$.
- Suppose the thesis holds for n, and let's try to prove the thesis for $n+1$. We proceed by further distinguishing some subcases:
  - If N is a value W, then $\mathscr{D} = \delta(W)$, $\mathscr{D}^{\mathcal{E}}$ is the zero distribution and thus
  $$(\lambda x.L)N \to_{m+1} (A \mapsto \mathscr{D}^{\mathcal{E}}(A) + \int \mathscr{E}_V(A)\, \mathscr{D}^{\mathcal{V}}(dV)).$$
  The thesis follows by monotonicity.
  - If N is an exception α, then $\mathscr{D} = \delta(\alpha)$, and since $(\lambda x.L)\alpha \to \alpha$, we can conclude that, since $\mathscr{D}^{\mathcal{V}}$ is the zero distribution,
  $$(\lambda x.L)N \to_2 \delta(\alpha) = (A \mapsto \mathscr{D}^{\mathcal{E}}(A) + \int \mathscr{E}_V(A)\, \mathscr{D}^{\mathcal{V}}(dV)).$$
  The thesis again follows by monotonicity.
  - If N is not a generalized value, then, necessarily $\mathscr{D}(A) = \int \mathscr{G}_P(A)\, \mathscr{H}(dP)$, where $N \to \mathscr{H}$ and $P \to_n \mathscr{G}_P$ for every P. By induction hypothesis, there are measures $\mathscr{I}_P$ such that $(\lambda x.L)P \to_{n+m} \mathscr{I}_P$, and, for all A,
  $$\mathscr{I}_P(A) \geq \mathscr{G}_P^{\mathcal{E}} + \int \mathscr{E}_V(A) \mathscr{G}_P^{\mathcal{V}}(dV)$$
  Let now E be the evaluation context $(\lambda x.L)[\cdot]$. Then, it holds that $(\lambda x.L)N \to E\{\mathscr{H}\}$ and thus:
  $$(\lambda x.L)N \to_{n+m+1} (A \mapsto \int \mathscr{I}_P(A)\, (E\{\mathscr{H}\}((\lambda x.L)dP))).$$

We can now observe that:

$$\begin{aligned}\int \mathscr{I}_P(A)\,(E\{\mathscr{H}\}((\lambda x.L)dP)) &= \int \mathscr{I}_P(A)\,\mathscr{H}(dP)\\ &\geq \int \mathscr{G}^{\mathcal{E}}_P(A)\,\mathscr{H}(dP) + \iint \mathscr{E}_V(A)\,\mathscr{G}^{\mathcal{V}}_P(dV)\,\mathscr{H}(dP)\\ &= \mathscr{D}^{\mathcal{E}}(A) + \iint \mathscr{E}_V(A)\,\mathscr{G}^{\mathcal{V}}_P(dV)\,\mathscr{H}(dP)\\ &= \mathscr{D}^{\mathcal{E}}(A) + \int \mathscr{E}_V(A)\,\mathscr{D}^{\mathcal{V}}(dV).\end{aligned}$$

Then one can prove the statement of the lemma, again by induction on n, following the same strategy as above.

Lemma 8 *If $M \Downarrow_n \mathscr{D}$, then there are $m, \mathscr{E}$ such that $M \to_m \mathscr{E}$ and $\mathscr{E} \geq \mathscr{D}$.*

Proof. By induction on n, exploiting Lemma 7, we can prove that $M \to_{3^n} \mathscr{E}$ where $\mathscr{E} \geq \mathscr{D}$. The only interesting case is the one in which M is an application.

Lemma 9 *If $MN \to_{n+1} \mathscr{D}$, then $M \to_n \mathscr{E}$, $N \to_n \mathscr{F}$ and for all P and V, $P\{V/x\} \to_n \mathscr{G}_{P,V}$ such that for all A,*

$$\mathscr{D}(A) \leq \mathscr{E}^{\mathcal{E}}(A) + \mathscr{E}(\mathbb{R}) \cdot \delta(\texttt{error}) + \mathscr{E}(\mathcal{V}_\lambda) \cdot \mathscr{F}^{\mathcal{E}}(A) + \iint \mathscr{G}_{P,V}(A)\,\mathscr{E}^{\mathcal{V}_\lambda}(\lambda x.dP)\,\mathscr{F}^{\mathcal{V}}(dV)$$

Proof. By induction on n.

- If $n = 0$, then $\mathscr{D}$ is the zero distribution, and so are $\mathscr{E}, \mathscr{F}$ and all $\mathscr{G}_{P,V}$.
- Suppose the thesis holds for every natural number smaller than n and prove it for n. Let us distinguish a few cases, and examine the most relevant ones:
  - If M is an abstraction $\lambda x.L$ and N is a value W, then $M \to \delta(L\{W/x\})$ and $L\{W/x\} \to_n \mathscr{D}$. We can then observe that

$$\begin{aligned}\mathscr{E} &= \delta(\lambda x.L)\\ \mathscr{F} &= \delta(W)\\ \mathscr{G}_{P,V} &= \mathscr{D} \text{ whenever } P = L \text{ and } V = W\end{aligned}$$

    Just observe that

$$\mathscr{D}(A) = \iint \mathscr{G}_{P,V}(A)\,\mathscr{E}(\lambda x.dP)\,\mathscr{F}(dV)$$

    and that $\mathscr{E}^{\mathcal{E}} = \mathscr{E}^{\mathbb{R}} = \mathscr{F}^{\mathcal{E}} = \mathbf{0}$.
  - If none of M and N are values, then $M \to \mathscr{L}$ and thus $MN \to E\{\mathscr{L}\}$ where $E = [\cdot]N$. Moreover, $LN \to_n \mathscr{H}_L$, where

$$\mathscr{D}(A) = \int \mathscr{H}_L(A)\,E\{\mathscr{L}\}((dL)N) = \int \mathscr{H}_L(A)\,\mathscr{L}(dL)$$

    We apply the induction hypothesis (and monotonicity) to each of the $LN \to_n \mathscr{H}_L$, and we obtain that $L \to_{n-1} \mathscr{I}_L$, $N \to_n \mathscr{F}$ and $P\{V/x\} \to_n \mathscr{G}_{P,V}$, where

$$\begin{aligned}\mathscr{H}_L(A) \leq \mathscr{I}^{\mathcal{E}}_L(A) &+ \mathscr{I}_L(\mathbb{R}) \cdot \delta(\texttt{error}) + \mathscr{I}_L(\mathcal{V}_\lambda) \cdot \mathscr{F}^{\mathcal{E}}(A)\\ &+ \iint \mathscr{G}_{P,V}(A)\,\mathscr{I}^{\mathcal{V}_\lambda}_L(\lambda x.dP)\,\mathscr{F}^{\mathcal{V}}(dV)\end{aligned}$$

    Now let $\mathscr{E}$ be the measure

$$A \mapsto \int \mathscr{I}_L(A)\,\mathscr{L}(dL).$$

Clearly, $M \to_n \mathscr{E}$. Moreover,

$$\begin{aligned}
\mathscr{D}(A) &= \int \mathscr{H}_L(A)\,\mathscr{L}(dL) \\
&\leq \int \mathscr{I}_L^{\mathcal{E}}(A)\,\mathscr{L}(dL) + \int \mathscr{I}_L(\mathbb{R}) \cdot \delta(\texttt{error})\,\mathscr{L}(dL) \\
&\quad + \int \mathscr{I}_L(\mathcal{V}_\lambda) \cdot \mathscr{F}^{\mathcal{E}}(A)\,\mathscr{L}(dL) \\
&\quad + \iiint \mathscr{G}_{P,V}(A)\,\mathscr{I}_L^{\mathcal{V}_\lambda}(\lambda x.dP)\,\mathscr{F}^{\mathcal{V}}(dV)\,\mathscr{L}(dL) \\
&= \mathscr{E}^{\mathcal{E}}(A) + \mathscr{E}(\mathbb{R}) \cdot \delta(\texttt{error}) + \mathscr{E}(\mathcal{V}_\lambda) \cdot \mathscr{F}^{\mathcal{E}}(A) \\
&\quad + \iint \mathscr{G}_{P,V}(A)\,\mathscr{E}^{\mathcal{V}_\lambda}(\lambda x.dP)\,\mathscr{F}^{\mathcal{V}}(dV)
\end{aligned}$$

Lemma 10 *If $M \to_n \mathscr{D}$, then there is $\mathscr{E}$ such that $M \Downarrow_n \mathscr{E}$ and $\mathscr{E} \geq \mathscr{D}$.*

Theorem 1 *For every term M, $[\![M]\!]_{\Rightarrow} = [\![M]\!]_{\Downarrow}$.*

Proof. This is a consequence of Lemma 8 and Lemma 10.

In the following, the value distribution $[\![M]\!]$ stands for either $[\![M]\!]_{\Rightarrow}$ or $[\![M]\!]_{\Downarrow}$.

5 Measure Space on Program Traces

Let a *program trace* s be a finite sequence of reals $s = [c_1, \ldots, c_n]$ of arbitrary length. In this section, we construct a measure space on the set $\mathbb{S}$ of program traces: (1) we define a measurable space $(\mathbb{S}, \mathcal{S})$ and (2) we equip it with a stock measure μ to obtain our measure space $(\mathbb{S}, \mathcal{S}, \mu)$.

5.1 The Measurable Space of Program Traces

To define the semantics of a program as a measure on the space of random choices, we first need to define a measurable space of program traces. Since a program trace is a sequence of random real variables of an arbitrary length, the set of all program traces is $\mathbb{S} = \biguplus_{n\in\mathbb{N}} \mathbb{R}^n$. Now, let us define the σ-algebra $\mathcal{S}$ on $\mathbb{S}$ as follows: let $\mathcal{R}^n$ be the Borel σ-algebra on $\mathbb{R}^n$. Consider the class of sets $\mathcal{S}$ of the form:

$$A = \bigbiguplus_{n\in\mathbb{N}} H_n$$

where $H_n \in \mathcal{R}^n$ for all n. Then $\mathcal{S}$ is a σ-algebra.

Lemma 11 *$\mathcal{S}$ is a σ-algebra on $\mathbb{S}$.*

Proof. We have $\mathbb{S} = \biguplus_{n\in\mathbb{N}} \mathbb{R}^n$ and $\mathbb{R}^n \in \mathcal{R}^n$ for all n, so $\mathbb{S} \in \mathcal{S}$.

If A is defined as above, then $\mathbb{S} - A = \biguplus_{n\in\mathbb{N}}(\mathbb{R}^n - H_n)$, where $\mathbb{R}^n - H_n \in \mathcal{R}^n$ for all n, so $\mathbb{S} - A \in \mathcal{S}$.

For closure under countable union, take $A_i = \biguplus_{n\in\mathbb{N}} H_{in}$ for all $i \in \mathbb{N}$, where $H_{in} \in \mathcal{R}^n$ for all i, n. Then $\bigcup_{i\in\mathbb{N}} A_i = \bigcup_{i\in\mathbb{N}} \biguplus_{n\in\mathbb{N}} H_{in} = \biguplus_{n\in\mathbb{N}}(\bigcup_{i\in\mathbb{N}} H_{in}) \in \mathcal{S}$, because $\bigcup_{i\in\mathbb{N}} H_{in} \in \mathcal{R}^n$.

Thus, $\mathcal{S}$ is a σ-algebra on E.

Since $\mathcal{S}$ is a σ-algebra on $\mathbb{S}$ it follows that $(\mathbb{S}, \mathcal{S})$ is a measurable space.

5.2 Stock measure on program traces

Since each distribution $\mathscr{D}$ has a density, the probability of each random value (and thus of each trace of random values) is zero. Instead, we define the trace and transition probabilities in terms of densities, with respect to the stock measure μ on $(\mathbb{S}, \mathcal{S})$ defined below.

$$\mu\left(\biguplus_{n\in\mathbb{N}} H_n\right) = \sum_{n=1}^{\infty} \lambda_n(H_n)$$

where λ_n is the Lebesgue measure on $\mathbb{R}^n$.

Lemma 12 *μ is a measure on $(\mathbb{S}, \mathcal{S})$.*

Proof. We check the three properties:

1. Since for all $n \in \mathbb{N}$ and $H_n \in \mathcal{R}^n$, we have $\lambda_n(H_n) \in [0, \infty]$, obviously $\mu(\biguplus_{n\in\mathbb{N}} H_n) = \sum_{n=1}^{\infty} \lambda_n(H_n) \in [0, \infty]$
2. If $H = \biguplus_{n\in\mathbb{N}} H_n = \varnothing$, then $H_n = \varnothing$ for all n, so $\mu(H) = \sum_{n=1}^{\infty} \lambda_n(\varnothing) = 0$.
3. Countable additivity: if $H_1 = \biguplus_{n\in\mathbb{N}} H_{1n}, H_2 = \biguplus_{n\in\mathbb{N}} H_{2n}, \ldots$ is a sequence of disjoint sets in $\mathcal{S}$, then:

$$\begin{aligned}
\mu(\biguplus_{m=1}^{\infty} H_m) &= \mu(\biguplus_{m=1}^{\infty}\biguplus_{n=1}^{\infty} H_{mn}) \\
&= \mu(\biguplus_{n=1}^{\infty}\biguplus_{m=1}^{\infty} H_{mn}) \\
&= \sum_{n=1}^{\infty} \lambda_n(\biguplus_{m=1}^{\infty} H_{mn}) \\
&= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \lambda_n(H_{mn}) \\
&= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \lambda_n(H_{mn}) \\
&= \sum_{m=1}^{\infty} \mu(H_m)
\end{aligned}$$

   where the equality $\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \lambda_n(H_{mn}) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} \lambda_n(H_{mn})$ follows from Tonelli's theorem for series.

A measure μ on (X, Σ) is σ-finite if $X = \bigcup_i A_i$ for some countable (finite or infinite) sequence of sets $A_i \in \Sigma$ such that $\mu(A_i) < \infty$. If μ is a σ-finite measure on (X, Σ), the measure space (X, Σ, μ) is also called σ-finite. σ-finite measure spaces behave better with respect to integration than those who are not.

In the following, let $[a, b]^n = \{(x_1, \ldots, x_n) \mid x_i \in [a, b] \quad \forall i \in 1..n\}$.

Lemma 13 *The measure μ on $(\mathbb{S}, \mathcal{S})$ is σ-finite.*

Proof. For every $n \in \mathbb{N}$, we have that $\mathbb{R}^n = \bigcup_{k\in\mathbb{N}} [-k, k]^n$. Hence, $\mathbb{S} = \biguplus_{n\in\mathbb{N}} \mathbb{R}^n = \biguplus_{n\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} [-k, k]^n$ is a countable union of sets in $\mathcal{S}$ of the form $[-k, k]^n$. Finally, for all $k, n \in \mathbb{N}$ we have $\mu([-k, k]^n) = \lambda_n([-k, k]^n) = (2k)^n < \infty$.

It follows that $(\mathbb{S}, \mathcal{S}, \mu)$ is a σ-finite measure space.

$$\frac{G \in \mathcal{GV}}{G \Downarrow_1^{[]} G}(\textsc{Eval Val}) \qquad \frac{w = \mathrm{pdf}_{\mathsf{D}}(\vec{c}) \qquad c, w > 0}{\mathsf{D}(\vec{c}) \Downarrow_w^{[c]} c}(\textsc{Eval Random}) \qquad \frac{}{g(\vec{c}) \Downarrow_1^{[]} \sigma_g(\vec{c})}(\textsc{Eval Prim})$$

$$\frac{M \Downarrow_{w_1}^{s_1} \lambda x.P \qquad N \Downarrow_{w_2}^{s_2} V \qquad P[V/x] \Downarrow_{w_3}^{s_3} G}{M\ N \Downarrow_{w_1 w_2 w_3}^{s_1 @ s_2 @ s_3} G}(\textsc{Eval Appl}) \qquad \frac{M \Downarrow_w^s \alpha}{M\ N \Downarrow_w^s \alpha}(\textsc{Eval Appl Raise1})$$

$$\frac{M \Downarrow_w^s c}{M\ N \Downarrow_w^s \texttt{error}}(\textsc{EvalApplRaise2}) \qquad \frac{M \Downarrow_{w_1}^{s_1} \lambda x.P \qquad N \Downarrow_{w_2}^{s_2} \alpha}{M\ N \Downarrow_{w_1 w_2}^{s_1 @ s_2} \alpha}(\textsc{Eval Appl Raise2})$$

$$\frac{M \Downarrow_w^s G}{\texttt{if true then } M \texttt{ else } N \Downarrow_w^s G}(\textsc{Eval If True}) \qquad \frac{N \Downarrow_w^s G}{\texttt{if false then } M \texttt{ else } N \Downarrow_w^s G}(\textsc{Eval If False})$$

Figure 5: Sampling-Based Big Step Semantics

6 Sampling-based semantics

6.1 Evaluation rules for computing the density on traces

Lemma 14 *If $M \xrightarrow{det} M'$ and $M \xrightarrow{det} M''$ then $M' = M''$.*

Proof. Since $M \xrightarrow{\mathrm{det}} M'$ implies that M is not a generalized value, Lemma 1 states that $M = E[R]$ for some unique E, R. If $R = \alpha$, then E is proper and $E[R]$ can only reduce to α. Otherwise, it follows immediately by inspection of the reduction rules that $E[R] \xrightarrow{\mathrm{det}} E[N]$ for some N, which is uniquely determined by the redex R.

Let us define composition of contexts $E \circ E'$ inductively as:

$$\begin{aligned} [\,] \circ E' &\triangleq E' \\ (E\ M) \circ E' &\triangleq (E \circ E')\ M \\ ((\lambda x.M)\ E) \circ E' &\triangleq (\lambda x.M)\ (E \circ E') \end{aligned}$$

Lemma 15 $(E \circ E')[M] = E[E'[M]]$.

Proof. By induction on the structure of E.

Lemma 16 *For any E and M such that $M \neq E'[\alpha]$, $M \xrightarrow{det} M'$ if and only if $E[M] \xrightarrow{det} E[M']$.*

Proof. Standard, using Lemmas 1 and 14.

$\Rightarrow$: Since $M \xrightarrow{\mathrm{det}} M'$, M is not a generalized value. By Lemma 1, $M = E'[R]$ for some E', R. By assumption, $R \neq \alpha$, so by inspection of the reduction rules $E'[R] \xrightarrow{\mathrm{det}} E'[N]$ for some N. By Lemma 14, $E'[N] = M'$. By Lemma 15, $E[M] = (E \circ E')[R]$ and $E[M'] = (E \circ E')[N]$. Since the context in the reduction rules is arbitrary, we can replace E' with $(E \circ E')$ in $E'[R] \xrightarrow{\mathrm{det}} E'[N]$, obtaining $(E \circ E')[R] \xrightarrow{\mathrm{det}} (E \circ E')[N]$, which implies $E[M] \xrightarrow{\mathrm{det}} E[M']$.

$\Leftarrow$: As before, we have $M = E'[R]$, so $E[M] = (E \circ E')[R]$. Since $R \neq \alpha$, $(E \circ E')[R] \xrightarrow{\mathrm{det}} (E \circ E')[N]$, where $(E \circ E')[N] = E[M']$ by Lemma 14.

By replacing the context $(E \circ E')$ with E', we get $E'[R] \xrightarrow{\mathrm{det}} E'[N]$, which implies $M \xrightarrow{\mathrm{det}} M'$.

Lemma 17 *If $(M, w, s) \to (M', w', s')$ and $(M, w, s) \to (M'', w'', s')$, then $M' = M''$ and $w' = w''$.*

$$\frac{w' = \mathrm{pdf}_{\mathsf{D}}(\vec{c}, c) \qquad w' > 0}{(E[\mathsf{D}(\vec{c})], w, s) \to (E[c], ww', s@[c])}(\textsc{Red Pure})$$

$$\frac{M \to N}{(M, w, s) \to (N, w, s)}(\textsc{Red Random})$$

Figure 6: Small-step sampling-based operational semantics

Proof. By inversion, using Lemmas 1 and 14. If $M \neq E[\mathsf{D}(\vec{c})]$, then both $(M, w, s) \to (M', w', s')$ and $(M, w, s) \to (M'', w'', s')$ must have been derived with (Red Pure), so $w'' = w' = w$, $M \xrightarrow{\det} M'$ and $M \xrightarrow{\det} M''$. By Lemma 14, $M' = M''$.

If $M = E[\mathsf{D}(\vec{c})]$, then $(M, w, s) \to (M', w', s')$ and $(M, w, s) \to (M'', w'', s')$ must have been derived with (Red Random). Hence, $s' = s@[c]$ for some c, so $w' = w'' = w\,\mathrm{pdf}_{\mathsf{D}}(\vec{c}, c)$ and $M' = M'' = E[c]$.

We let *multi-step reduction* be the inductively defined relation $(M, w, s) \Rightarrow (M', w', s')$ if and only if $(M, w, s) = (M', w', s')$ or $(M, w, s) \Rightarrow (M'', w'', s'') \to (M', w', s')$ for some M'', w'', s''.

Lemma 18 *If $(M, w, s) \Rightarrow (M', w', s')$, then $s' = s@s''$ for some s''.*

Proof. By induction on the derivation of $(M, w, s) \Rightarrow (M', w', s')$.

Lemma 19 *If $(M, w, s) \Rightarrow (G', w', s')$ and $(M, w, s) \Rightarrow (G'', w'', s')$, then $G' = G''$ and $w' = w''$.*

Proof. By induction on the derivation of $(M, w, s) \Rightarrow (G', w', s')$, with appeal to Lemmas 17 and 18.

- Base case: $(M, w, s) = (G', w', s')$. Generalized values do not reduce, so $G'' = G$ and $w'' = w$.
- Induction case: $(M, w, s) \to (\hat{M}, \hat{w}, \hat{s}) \Rightarrow (G', w', s')$. Since $M \neq G''$, we also have $(M, w, s) \to (M^*, w^*, s^*) \Rightarrow (G'', w'', s')$.
  By inspection of the reduction rules, either $\hat{s} = s^* = s$ or $\hat{s} = s@[c]$ and $s^* = s@[d]$ for some c, d. In the latter case, by Lemma 18, $s' = s@[c]@s'' = s@[d]@s'''$ for some s'', s''' ,which implies $c = d$. Thus, $\hat{s} = s^*$.
  By Lemma 17, $\hat{M} = M^*$ and $\hat{w} = w^*$. The induction hypothesis implies $G'' = G'$ and $w'' = w'$, as required.

Lemma 20 *For any E and M such that $M \neq E'[\alpha]$, $(M, w, s) \to (M', w', s')$ if and only if $(E[M], w, s) \to (E[M'], w', s')$*

Proof. By inversion of $\to$, using Lemma 16. If $M \neq E[\mathsf{D}(\vec{c})]$, then $(E[M], w, s) \to (E[M'], w', s')$ was derived with (Red Pure), so the result follows by Lemma 16.

If $M = E[\mathsf{D}(\vec{c})]$, then $(E[M], w, s) \to (E[M'], w', s')$ was derived with (Red Random), so $w' = ww''$ and $s' = s@[c]$, where $w'' = \mathrm{pdf}_{\mathsf{D}}(\vec{c}, c)$. From the assumptions $w'' = \mathrm{pdf}_{\mathsf{D}}(\vec{c}, c)$ and $w'' > 0$, we can derive $(E[M], w, s) \to (E[M'], ww'', s@[c])$ by (Red Random) for any E.

Lemma 21 *For any E, if $(M, w, s) \Rightarrow (M', w', s')$ and $M' \neq \alpha$ then we have $(E[M], w, s) \Rightarrow (E[M'], w', s')$.*

Proof. By induction on the number of steps in the derivation of $(M, w, s) \Rightarrow (M', w', s')$, with appeal to Lemma 20. Since $M' \neq \alpha$, no expression in the derivation chain (other than the last one) can be of the form $E'[\alpha]$.

Lemma 22 *For any E, if $(M, w, s) \Rightarrow (\alpha, w', s')$ then*
$(E[M], w, s) \Rightarrow (\alpha, w', s')$.

Proof. By induction on the number of steps in the derivation, using Lemmas 20 and 21. If $(M, w, s) \Rightarrow (\alpha, w', s')$ was derived in 0 steps, then $M = \alpha$, $w' = w'$ and $s' = s$. By (Red Pure), $(E[\alpha], w, s) \Rightarrow (\alpha, w, s)$, as required.

If $(M, w, s) \Rightarrow (\alpha, w', s')$ was derived in 1 or more steps, then there exist $\hat{M}$, $\hat{w}$, $\hat{s}$ such that $(M, w, s) \Rightarrow (\hat{M}, \hat{w}, \hat{s}) \rightarrow (\alpha, w', s')$, where $\hat{M} \notin \mathcal{GV}$. Because $\alpha \neq E'[c]$ for any E', c, $(\hat{M}, \hat{w}, \hat{s}) \rightarrow (\alpha, w', s')$ must have been derived with (Red Pure), which implies $\hat{w} = w'$ and $\hat{s} = s'$.

By Lemma 21, $(E[M], w, s) \Rightarrow (E[\hat{M}], w', s')$.

If $\hat{M} = E'[\beta]$ for some E', β, then $\beta = \alpha$ and by (Red Pure), $((E \circ E')[\alpha], w', s') \rightarrow (\alpha, w', s')$. Thus, $(E[M], w, s) \Rightarrow (\alpha, w', s')$, as required.

If $\hat{M} \neq E'[\beta]$, then by Lemma 20, $(E[\hat{M}], w', s') \rightarrow (E[\alpha], w', s')$. By (Red Pure), $(E[\alpha], w', s') \rightarrow (\alpha, w', s')$. Thus, $(E[M], w, s) \Rightarrow (\alpha, w', s')$.

Lemma 23 *If $M \xrightarrow{det} M'$ and $M' \Downarrow_s^w G$, then $M \Downarrow_s^w G$.*

Proof. By induction on the structure of M, using Lemma 16. Note that $M \xrightarrow{\text{det}} M'$ implies that M is not a generalized value.

- If $M = g(\vec{c})$ or $M = c\ V$ or $M = T$ or $M = E[\alpha]$ for some E, then M reduces to a generalized value in 1 step, so the result holds trivially (by one of the evaluation rules).
- Case $M = \texttt{if}\ 1\ \texttt{then}\ M_2\ \texttt{else}\ M_3$: We have $\texttt{if}\ 1\ \texttt{then}\ M_2\ \texttt{else}\ M_3 \xrightarrow{\text{det}} M_2$. By assumption, $M_2 \Downarrow_s^w G$. Thus, the desired result holds by (Eval If True).
- Case $M = \texttt{if}\ 0\ \texttt{then}\ M_2\ \texttt{else}\ M_3$: analogous to the previous case.
- Case $M = (\lambda x.N_1)\ V$: We have $(\lambda x.N_1)\ V \xrightarrow{\text{det}} N_1\{V/x\}$. Since $(\lambda x.N_1)$ and V are already values and $N_1\{V/x\} \Downarrow_s^w G$ by assumption, (Eval Appl) yields $(\lambda x.N_1)\ V \Downarrow_s^w G$.
- Case $M = (\lambda x.N_1)\ N_2$, $N_2 \notin \mathcal{GV}$, $N_2 \neq E[\alpha]$. We have $(\lambda x.N_1)\ N_2 \xrightarrow{\text{det}} (\lambda x.N_1)\ N_2'$ for some N_2', so by Lemma 16, $N_2 \xrightarrow{\text{det}} N_2'$. By assumption $(\lambda x.N_1)\ N_2' \Downarrow_s^w G$.
  - If $(\lambda x.N_1)\ N_2' \Downarrow_s^w G$ was derived with (Eval Appl), then $N_2' \Downarrow_{s_1}^{w_1} V$ and $(\lambda x.N_1)\ V \Downarrow_{s_2}^{w_2} G$, where $w = w_1 w_2$ and $s = s_1 @ s_2$. By induction hypothesis, $N_2 \Downarrow_{s_1}^{w_1} V$, so (Eval Appl) gives $(\lambda x.N_1)\ N_2 \Downarrow_s^w G$, as required.
  - If $(\lambda x.N_1)\ N_2' \Downarrow_s^w G$ was derived with (Eval Appl Raise3), then $G = \alpha$ and $N_2' \Downarrow_{s_1}^{w_1} \alpha$. By induction hypothesis, $N_2 \Downarrow_{s_1}^{w_1} \alpha$, so by (Eval Appl Raise3), $(\lambda x.N_1)\ N_2 \Downarrow_s^w \alpha$
- Case $M = N_1\ N_2$, $N_1 \notin \mathcal{GV}$, $N_1 \neq E[\alpha]$. We have $N_1\ N_2 \xrightarrow{\text{det}} N_1'\ N_2$ for some N_1', which implies $N_1 \xrightarrow{\text{det}} N_1'$ by Lemma 16. By assumption $N_1'\ N_2 \Downarrow_s^w G$.
  - If $N_1'\ N_2 \Downarrow_s^w G$ was derived with (Eval Appl), then $N_1' \Downarrow_{s_1}^{w_1} (\lambda x.N_1'')$, $N_2 \Downarrow_{s_2}^{w_2} V$ and $(\lambda x.N_1'')\ V \Downarrow_{s_3}^{w_3} G$, where $w = w_1 w_2 w_3$ and $s = s_1 @ s_2 @ s_3$. By induction hypothesis, $N_1 \Downarrow_{s_1}^{w_1} (\lambda x.N_1'')$, so (Eval Appl) gives $N_1\ N_2 \Downarrow_s^w G$, as required.
  - If $N_1'\ N_2 \Downarrow_s^w G$ was derived with (Eval Appl Raise1), then $G = \alpha$ and $N_1' \Downarrow_s^w \alpha$. By induction hypothesis, $N_1 \Downarrow_{s_1}^{w_1} \alpha$, so by (Eval Appl Raise1), $N_1\ N_2 \Downarrow_s^w \alpha$
  - If $N_1'\ N_2 \Downarrow_s^w G$ was derived with (Eval Appl Raise3), then $N_1' \Downarrow_{s_1}^{w_1} (\lambda x.N_1'')$ and $N_2 \Downarrow_{s_2}^{w_2} \alpha$, where $w = w_1 w_2$ and $s = s_1 @ s_2$. By induction hypothesis, $N_1 \Downarrow_{s_1}^{w_1} (\lambda x.N_1'')$, so (Eval Appl Raise3) gives $N_1\ N_2 \Downarrow_s^w \alpha$, as required.
  - If $N_1'\ N_2 \Downarrow_s^w G$ was derived with (Eval Appl Raise1), then $G = \texttt{error}$ and $N_1' \Downarrow_s^w c$. By induction hypothesis, $N_1 \Downarrow_{s_1}^{w_1} c$, so by (Eval Appl Raise1), $N_1\ N_2 \Downarrow_s^w \texttt{error}$.

Lemma 24 *If $(E[\mathsf{D}(c)], 1, []) \rightarrow (M', w_1, s_1)$ and $M' \Downarrow_{w_2}^{s_2} G$, then $M \Downarrow_{w_1 w_2}^{s_1 @ s_2} G$.*

Proof. Since $(E[\mathsf{D}(c)], 1, []) \rightarrow (M', w_1, s_1)$ must have been derived with (Red Random), we have $s_1 = [c]$ for some c such that $M' = E[c]$ and $w_1 = \text{pdf}_{\mathsf{D}}(\vec{c}, c)$. Thus, we have $(E[\mathsf{D}(c)], 1, []) \rightarrow (E[c], \text{pdf}_{\mathsf{D}}(\vec{c}, c), [c])$ The result follows by induction on the structure of E.

- Base case: $E = [\ ]$: the result follows directly from (Eval Random).
- Case $E = (\lambda x.N)\ E'$: By applying lemma 20 twice (in different directions), we can deduce $(E'[\mathsf{D}(\vec{c})], 1, []) \rightarrow (E'[c], w_1, s_1)$

- If $(\lambda x.N)\ E'[c] \Downarrow^{s_2}_{w_2} G$ was derived with (Eval Appl), then $E'[c] \Downarrow^{s_2'}_{w_2'} V$ for some w_2', s_2', V' and $N[V/x] \Downarrow^{s_2''}_{w_2''} G$ for some w_2'', s_2'', G, where $w_2 = w_2'w_2''$ and $s_2 = s_2'@s_2''$. By induction hypothesis, $E'[D(\vec{c})] \Downarrow^{s_1@s_2'}_{w_1w_2'} V$. Hence, by (Eval Appl), $E[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} G$.
- If $(\lambda x.N)\ E'[c] \Downarrow^{s_2}_{w_2} G$ was derived with (Eval Appl Raise3), then $E'[\vec{c}] \Downarrow^{s_2}_{w_2} \alpha$.By induction hypothesis, $E'[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} \alpha$. Hence, by (Eval Appl Raise3), $E[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} G$.

- Case $E = E'\ N$: Again, we have $(E'[\mathsf{D}(\vec{c})], 1, []) \to (E'[c], w_1, s_1)$.
  - If $E'[c]\ N \Downarrow^{s_2}_{w_2} G$ was derived with (Eval Appl), then $E'[c] \Downarrow^{s_2'}_{w_2'} (\lambda x.N')$ for some w_2', s_2', $\lambda x.N'$, $N \Downarrow^{s_2''}_{w_2''} V$ for some w_2'', s_2'', V and $N'[V/x] \Downarrow^{s_2'''}_{w_2'''} G$ for some w_2''', s_2''', G, where $w_2 = w_2'w_2''w_2'''$ and $s_2 = s_2'@s_2''@s_2'''$. By induction hypothesis, $E'[D(\vec{c})] \Downarrow^{s_1@s_2'}_{w_1w_2'} \lambda x.N'$. Hence, by (Eval Appl), $E[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} G$.
  - If $E'[c]\ N \Downarrow^{s_2}_{w_2} G$ was derived with (Eval Appl Raise1), then $E'[c] \Downarrow^{s_2}_{w_2} \alpha$ for some α. By induction hypothesis, $E'[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} \alpha$. Hence, by (Eval Appl Raise1), $E[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} G$.
  - If $E'[c]\ N \Downarrow^{s_2}_{w_2} G$ was derived with (Eval Appl Raise3), then $E'[c] \Downarrow^{s_2'}_{w_2'} (\lambda x.N')$ for some w_2', s_2', $\lambda x.N'$, and $N \Downarrow^{s_2''}_{w_2''} \alpha$ for some w_2'', s_2'', α, where $w_2 = w_2'w_2''$ and $s_2 = s_2'@s_2''$. By induction hypothesis, $E'[D(\vec{c})] \Downarrow^{s_1@s_2'}_{w_1w_2'} \lambda x.N'$. Hence, by (Eval Appl Raise3), $E[D(\vec{c})] \Downarrow^{s_1@s_2}_{w_1w_2} G$.

Lemma 25 *If $(M, 1, []) \to (M', w_1, s_1)$ and $M' \Downarrow^{s_2}_{w_2} G$, then $M \Downarrow^{s_1@s_2}_{w_1w_2} G$.*

Proof. By inversion of $\to$, using Lemmas 23 and 24.

If $(M, 1, []) \to (M', w_1, s_1)$ was derived with (Red Pure), then $M \xrightarrow{\text{det}} M'$, $w_1 = 1$ and $s_1 = []$, so the result follows directly from Lemma 23.

If $(M, 1, []) \to (M', w_1, s_1)$ was derived with (Red Random), then $M = E[\mathsf{D}(\vec{c})]$, so the result follows by Lemma 24.

Lemma 26 *If $(M, w, s) \Rightarrow (M', w', s')$ and $w > 0$, then $w' > 0$.*

Proof. By induction on the number of steps in the derivation.
- If $(M, w, s) \Rightarrow (M', w', s')$ was derived in 0 steps, then $w' = w$, so $w' > 0$.
- If $(M, w, s) \Rightarrow (M', w', s')$ was derived in 1 or more steps, then $(M, w, s) \Rightarrow (M^*, w^*, s^*) \to (M', w', s')$. By induction hypothesis, $w* > 0$.
  If $(M^*, w^*, s^*) \to (M', w', s')$ was derived with (Red Pure), then $w' = w^* > 0$.
  If $(M^*, w^*, s^*) \to (M', w', s')$ was derived with (Red Random), then $w' = w^*w''$ for some $w'' > 0$, so $w' > 0$.

Lemma 27 *For any w_0, s_0,*

1. *If $(M, 1, []) \to^n (M', w, s)$, then $(M, w_0, s_0) \to^n (M', w_0w, s_0@s)$.*
2. *If $(M, w_0, s_0) \to^n (M', w, s)$, then $(M, 1, []) \to^n (M', w_1, s_1)$ such that $w = w_0 \cdot w_1$ and $s = s_0@s_1$.*

Proof. By induction on n.

1. - If $n = 0$, then $M' = M$, $w = 1$ and $s = []$. Thus, $(M, w_0, s_0) \to^0 (M', w_0w, s_0@s)$.
   - If $(M, 1, []) \to^{n+1} (M', w, s)$, then $(M, 1, []) \to^n (M^*, w^*, s^*) \to (M', w, s)$ for some M^*, w^*, s^*. By induction hypothesis, $(M, w_0, s_0) \to^n (M^*, w_0w^*, s_0@s^*)$.
     - If $(M^*, w^*, s^*) \to (M', w, s)$ was derived with (Red Pure), then $w = w^*$, $s = s^*$ and $M^* \xrightarrow{\text{det}} M'$, so by (Red Pure), $(M^*, w_0w^*, s_0@s^*) \to (M', w_0w^*, s_0@s^*)$. Hence, $(M, w_0, s_0) \to^{n+1} (M', w_0w, s_0@s)$.

- If $(M^*, w^*, s^*) \to (M', w, s)$ was derived with (Red Random), then $M^* = E[\mathsf{D}(\vec{c})]$, $M' = E[c]$, $w = w^* \operatorname{pdf}_{\mathsf{D}}(\vec{c}, c)$ (which implies $w_0 w = w_0 w^* \operatorname{pdf}_{\mathsf{D}}(\vec{c}, c)$), $\operatorname{pdf}_{\mathsf{D}}(\vec{c}, c) > 0$ and $s = s^*@[c]$. By (Red Random), $(M^*, w_0 w^*, s_0@s^*) \to (M', w_0 w, s_0@s)$. Therefore, $(M, w_0, s_0) \to^{n+1} (M', w_0 w, s_0@s)$.

2. - If $n = 0$, then $(M, w_0, s_0) = (M', w, s_0@s)$, so, since $w_0 > 0$, obviously $(M, 1, []) \to^0 (M', w/w_0, s)$.
   - If $(M, w_0, s_0) \to^{n+1} (M', w_0 w, s_0@s)$, then $(M, w_0, s_0) \to^n (M^*, w^*, s_0@s^*) \to (M', w, s_0@s)$ for some M^*, w^*, s^*. By induction hypothesis, $(M, 1, []) \to^n (M^*, w^*/w_0, s^*)$.
     - If $(M^*, w^*, s_0@s^*) \to (M', w, s_0@s)$ was derived with (Red Pure), then $w = w^*$ (which implies $w/w_0 = w^*/w_0$), $s = s^*$ and $M^* \xrightarrow{\text{det}} M'$. By (Red Pure), $(M^*, w^*/w_0, s^*) \to (M', w/w_0, s)$. Hence, $(M, 1, []) \to^{n+1} (M', w/w_0, s)$.
     - If $(M^*, w^*, s_0@s^*) \to (M', w, s_0@s)$ was derived with (Red Random), then $M^* = E[\mathsf{D}(\vec{c})]$, $M' = E[c]$, $w = w^* \operatorname{pdf}_{\mathsf{D}}(\vec{c}, c)$ (which implies $w/w_0 = (w^*/w_0) \operatorname{pdf}_{\mathsf{D}}(\vec{c}, c)$), $\operatorname{pdf}_{\mathsf{D}}(\vec{c}, c) > 0$ and $s = s^*@[c]$. By (Red Random), $(M^*, w^*/w_0, s^*) \to (M', w/w_0, s)$, so $(M, 1, []) \to^{n+1} (M', w/w_0, s)$, as required.

Lemma 28 *For any E, if $E[\alpha] \Downarrow_w^s G$, then $s = []$ and $w = 1$.*

Proof. By induction on the derivation of $E[\alpha] \Downarrow_w^s G$.

The small-step and the big-step sampling semantics both compute the same traces with the same weights.

Proposition 1 *$M \Downarrow_w^s G$ if and only if $(M, 1, []) \Rightarrow (G, w, s)$.*

Proof.

$\Rightarrow$: By induction on the derivation of $M \Downarrow_w^s G$.

- Case: $\dfrac{\text{(Eval Val)}\quad V \in \mathcal{V}}{V \Downarrow_1^{[]} V}$

  Here, $M = V$, $w = 1$ and $s = []$. so (M, w_0, s_0) reduces to (V, w_0, s_0) in 0 steps by the small-step semantics.

- Case: $\dfrac{\text{(Eval Random)}\quad w = \operatorname{pdf}_{\mathsf{D}}(\vec{c}, v)\quad w > 0}{\mathsf{D}(\vec{c}) \Downarrow_w^{[v]} v}$

  By (Red Random) (taking $E = [\ ]$), $(\mathsf{D}(\vec{c}), 1, []) \to (v, w, [v])$, so $(M, 1, [])$ reduces to (v, w, s) in 1 step.

- Case: $\dfrac{\text{(Eval Prim)}}{g(\vec{c}) \Downarrow_1^{[]} \sigma_g(\vec{c})}$

  By (Red Pure) (taking $E = [\ ]$), $(g(\vec{c}), 1, []) \to (\sigma_g(\vec{c}), 1, [])$, so $(M, 1, [])$ reduces to $(\sigma_g(\vec{c}), 1, [])$ in 1 step.

- Case: $\dfrac{\text{(Eval Appl)}\quad M \Downarrow_{w_1}^{s_1} \lambda x.M'\quad N \Downarrow_{w_2}^{s_2} V\quad M'[V/x] \Downarrow_{w_3}^{s_3} G}{M\ N \Downarrow_{w_1 w_2 w_3}^{s_1@s_2@s_3} G}$

  By induction hypothesis, $(M, 1, []) \Rightarrow (\lambda x.M', w_1, s_1)$, $(N, 1, []) \Rightarrow (V, w_2, s_2)$ and $(M'[V/x], 1, []) \Rightarrow (G, w_3, s_3)$. By Lemma 27, $(N, w_1, s_1) \Rightarrow (V, w_1 w_2, s_1@s_2)$ and $(M'[V/x], w_1 w_2, s_1@s_2) \Rightarrow (G, w_1 w_2 w_3, s_1@s_2@s_3)$

By Lemma 21 (for $E = [\ ]\ N$), $(M\ N, 1, []) \Rightarrow ((\lambda x.M')\ N, w_1, s_1)$
By Lemma 21 again (for $E = (\lambda x.M')\ [\ ]$), $((\lambda x.M')\ N, w_1, s_1) \Rightarrow ((\lambda x.M')\ V, w_1 w_2, s_1 \ @\ s_2)$.
By (Red Pure), $((\lambda x.M')\ V, w_1 w_2, s_1 \ @\ s_2) \to (M'[V/x], w_1 w_2, s_1 @ s_2)$.
Thus, the desired result follows by transitivity of $\Rightarrow$.

- Case: $\dfrac{M \Downarrow_w^s \alpha}{M\ N \Downarrow_w^s \alpha}$ (Eval Appl Raise1)

By induction hypothesis, $(M, 1, []) \Rightarrow (\alpha, w, s)$.
By Lemma 22 (with $E = [\ ]\ N$)), $(M\ N, 1, []) \Rightarrow (\alpha, w, s)$.

- Case: $\dfrac{M \Downarrow_w^s c}{M\ N \Downarrow_w^s \texttt{error}}$ (Eval Appl Raise2)

By induction hypothesis, $(M, 1, []) \Rightarrow (c, w, s)$. By Lemma 21 (with $E = [\ ]\ N$)), $(M\ N, 1, []) \Rightarrow (c\ N, w, s)$.
By (Red Pure), $(c\ N, w, s) \to (\texttt{error}, w, s)$.
Thus, $(M\ N, 1, []) \Rightarrow (\texttt{error}, w, s)$.

- Case: $\dfrac{M \Downarrow_{w_1}^{s_1} \lambda x.M' \quad N \Downarrow_{w_2}^{s_2} \alpha}{M\ N \Downarrow_{w_1 w_2}^{s_1 @ s_2} \alpha}$ (Eval Appl Raise3)

By induction hypothesis, $(M, 1, []) \Rightarrow (\lambda x.M', w_1, s_1)$, and $(N, 1, []) \Rightarrow (\alpha, w_2, s_2)$.
By Lemma 21, $(M\ N, 1, []) \Rightarrow ((\lambda x.M')\ N, w_1, s_1)$,
By Lemma 27, $(N, w_1, s_1) \Rightarrow (\alpha, w_1 w_2, s_1 @ s_2)$.
By Lemma 22, $((\lambda x.M')\ N, w_1, s_1) \Rightarrow (\alpha, w_1 w_2, s_1 @ s_2)$.
Hence, the desired result follows by transitivity of $\Rightarrow$.

- Case: $\dfrac{M_2 \Downarrow_w^s G}{\texttt{if}\ 1\ \texttt{then}\ M_2\ \texttt{else}\ M_3 \Downarrow_w^s G}$ (Eval If True)

By (Red Pure) (taking $E = [\ ]$), $(\texttt{if}\ 1\ \texttt{then}\ M_2\ \texttt{else}\ M_3, 1, []) \to (M_2, 1, [])$. By induction hypothesis, $(M_2, 1, []) \Rightarrow (G, w, s)$.
Hence $(\texttt{if}\ 1\ \texttt{then}\ M_2\ \texttt{else}\ M_3, 1, []) \Rightarrow (G, w, s)$

- Case (Eval If False): analogous to (Eval If True)

- Case: $\dfrac{}{T \Downarrow_1^{[]} \texttt{error}}$ (Eval Error)

By (Red Pure), $(T, 1, []) \to (\texttt{error}, 1, [])$.

$\Leftarrow$: By induction on the length of the derivation of $(M, 1, []) \Rightarrow (G, w, s)$.

- Base case: If $(M, 1, []) = (G, w, s)$, then $M \Downarrow_s^w G$ by (Eval Val).
- Induction step: assume $(M, 1, []) \to (M', w_1, s_1) \to^n (G, w, s)$. By Lemma 27, $(M', 1, []) \to^n (G, w_2, s_2)$ for some w_2, s_2 such that $w = w_1 w_2$ and $s = s_1 @ s_2$.
  By induction hypothesis, $M' \Downarrow_{w_2}^{s_2} G$. By Lemma 25, $M \Downarrow_{w_1 w_2}^{s_1 s_2} G$, and so $M \Downarrow_w^s G$.

Lemma 29 *If $M \Downarrow_w^s G$ and $M \Downarrow_{w'}^s G'$ then $w = w'$ and $G = G'$.*

Proof. Corollary of Proposition 1 and Lemma 19.

6.2 From Trace Semantics to Value Distributions

Recall that A stands for a measurable set of terms. $\mathbb{S}$ is the set of finite-length sample traces, which can be given the structure of a measure space, as we showed in Section 5. The first thing to do is to define the following function:

$$\mathbf{P}_M(s) = \begin{cases} w & \text{if } M \Downarrow_w^s G \text{ for some } G \\ 0 & \text{otherwise} \end{cases}$$

Lemma 30 *For any program M, $\mathbf{P}_M$ is a measurable function $\mathbb{S} \to \mathbb{R}_+$.*

Proof. See Appendix A.

The trace sub-probability measure $[\![M]\!]_{\Downarrow}^{\mathbb{S}}$ is defined by

$$[\![M]\!]_{\Downarrow}^{\mathbb{S}}(A) = \int_A \mathbf{P}_M.$$

Lemma 31 *The function $[\![M]\!]_{\Downarrow}^{\mathbb{S}}$ is a measure on $(\mathbb{S}, \mathcal{S})$.*

Proof. Since $\mathbf{P}_M$ is a non-negative $\mathcal{S}$-measurable function,
$[\![M]\!]_{\Downarrow}^{\mathbb{S}}(A) = \int_A \mathbf{P}_M(s)\mu(ds)$ is a measure definition.

The return value of a trace s is given by

$$\mathbf{O}_M(s) = \begin{cases} G & \text{if } M \Downarrow_w^s G \text{ for some } w \\ \texttt{fail} & \text{otherwise} \end{cases}$$

Lemma 32 *For each M, $\mathbf{O}_M$ is a measurable function $\mathbb{S} \to \mathcal{GV}$.*

Proof. See Appendix A.

The sampling-based semantics then induces a value distribution $[\![M]\!]_{\mathbb{S}}$ defined by :

$$[\![M]\!]_{\mathbb{S}}(A) := [\![M]\!]_{\Downarrow}^{\mathbb{S}} \mathbf{O}_M^{-1} = \int \mathbf{P}_M(s) \cdot \delta(\mathbf{O}_M(s))(A)\, ds$$

6.3 Equivalence of Sampling-Based and Distribution-Based Semantics

This section is a proof of Theorem 2.

Theorem 2 $[\![M]\!]_{\mathbb{S}} = [\![M]\!]$.

For every $n \in \mathbb{N}$, let $\mathbb{S}_n$ be the set of sample traces of length at most n. It has itself the structure of a measure space. Let us then define the value distribution $[\![M]\!]_{\mathbb{S}}^n$ as follows:

$$[\![M]\!]_{\mathbb{S}}^n(A) = \int_{\mathbb{S}_n} \mathbf{P}_M(s) \cdot \delta(\mathbf{O}_M(s))(A)\, ds$$

If $\mathrm{dom}(f) = \mathbb{S}$ we write $f|_n$ for the restriction of f to $\mathbb{S}_n$.

Lemma 33 $\int f = \sup_n \int f|_n$ *whenever f is measurable wrt. the stock measure on $\mathbb{S}$.*

Proof. Let g_n be the function that is as f on $\mathbb{S}_n$ and 0 outside. Then $\int f|_n = \int g_n$. Since the g_n are converging to f pointwise from below, we also have $\int f = \sup_n \int g|_n$ by the monotone convergence theorem.

A corollary is that $[\![M]\!]_{\mathbb{S}} = \sup_{n \in \mathbb{N}} [\![M]\!]_{\mathbb{S}}^n$.

Lemma 34 *If $M \to \mathscr{D}$, then $[\![M]\!] = A \mapsto \int [\![N]\!](A)\ \mathscr{D}(dN)$*

Proof. For every n and for every term N, let $\mathscr{E}_N^n$ be the unique value distribution such that $N \to_n \mathscr{E}_N^n$. By definition, we have that

$$\mathscr{E}_M^{n+1}(A) = \int \mathscr{E}_N^n(A)\ \mathscr{D}(dN).$$

By the monotone convergence theorem, then,

$$\begin{aligned}[\![M]\!](A) &= \sup_n \mathscr{E}_M^{n+1}(A) = \sup_n \int \mathscr{E}_N^n(A)\ \mathscr{D}(dN) \\ &= \int (\sup_n \mathscr{E}_N^n(A))\ \mathscr{D}(dN) = \int [\![N]\!](A)\ \mathscr{D}(dN).\end{aligned}$$

Lemma 35 $[\![E[\mathsf{D}(\vec{c})]]\!]_{\mathbb{S}}^{n+1}(A) = \int [\![N]\!]_{\mathbb{S}}^{n+1}(A)\ E\{\mu_{\mathsf{D}}(\vec{c})\}(dN)$.

Lemma 36 *If $M \to \mathscr{D}$, then $[\![M]\!]_{\mathbb{S}} = A \mapsto \int [\![N]\!]_{\mathbb{S}}(A)\ \mathscr{D}(dN)$*

A program M is said to *deterministically diverge* iff $(M, 1, []) \Rightarrow (N, w, s)$ implies $w = 1$, $s = []$ and N is *not* a generalized value. A program M is said to *determinsitically converge to a program* N iff $(M, 1, []) \Rightarrow (N, 1, [])$.

Lemma 37 *If M deterministically diverges, then $[\![M]\!] = [\![M]\!]_{\mathbb{S}} = \mathbf{0}$.*

Proof. One can easily prove, by induction on n, that if M deterministically diverges, then $M \to_n \mathbf{0}$:

- If $n = 0$, then $M \to_n \mathbf{0}$ by definition.
- About the inductive case, since M cannot be a generalized value, it must be that $M \to \delta(N)$ (where N deterministically diverges) and that $M \to_{n+1} \mathscr{D}$, where $N \to_n \mathscr{D}$. By induction hypothesis, $\mathscr{D} = \mathbf{0}$.

The fact that $[\![M]\!]_{\mathbb{S}} = \mathbf{0}$ is even simpler to prove, since if M deterxministically diverges, then there cannot be any s, w, V such that $M \Downarrow_w^s V$, and thus $\mathbf{P}_M(s)$ is necessarily 0.

Lemma 38 *Let M be a term that deterministically converge to a term N. Then:*

- *$\mathscr{D} \leq \mathscr{E}$ whenever $M \to_n \mathscr{D}$; and $N \to_n \mathscr{E}$;*
- *$[\![M]\!]_{\mathbb{S}}^n = [\![N]\!]_{\mathbb{S}}^n$;*
- *$[\![M]\!] = [\![N]\!]$ and $[\![M]\!]_{\mathbb{S}} = [\![N]\!]_{\mathbb{S}}$*

Proof. The first point is an induction on the structure of the proof that M deterministically converge to N. Let us consider the second and third points. Since equality is transitive, we can assume, without losing any generality, that $(M, 1, []) \to (N, 1, [])$, namely that $M \xrightarrow{\text{det}} N$. With the latter hypothesis, it is easy to realize that $M \Downarrow_s^w V$ iff $N \Downarrow_s^w V$ and that $M \to_{n+1} \mathscr{D}$ iff $N \to_n \mathscr{D}$. The thesis easily follows.

Lemma 39 *For every generalized value G, it holds that $[\![V]\!] = [\![G]\!]_{\mathbb{S}} = \delta(G)$.*

Lemma 40 *For every program M, exactly one of the following conditions holds:*

- *M deterministically diverges;*
- *There is generalized value G such that M deterministically converges to G*
- *There are $E, \mathsf{D}, c_1, \ldots, c_{|\mathsf{D}|}$ such that M deterministically converges to $E[\mathsf{D}(c_1, \ldots, c_{|\mathsf{D}|})]$.*

Proof. Easy.

Lemma 41 *If $M \to_n \mathscr{D}$, then $\mathscr{D} \leq [\![M]\!]_{\mathbb{S}}$.*

Proof. By induction on n:

- If $n = 0$, then $\mathscr{D}$ is necessarily $\mathbf{0}$, and we are done.

- About the inductive case, let's distinguish three cases depending on the three cases of Lemma 40, applied to M:
  - If M deterministically diverges, then by Lemma 37, $\mathscr{D} \leq [\![M]\!] = [\![M]\!]_{\mathbb{S}}$.
  - If M deterministically converges to a generalized value G, then by Lemma 38 and Lemma 39, it holds that
$$\mathscr{D} \leq [\![M]\!] = [\![G]\!] = \delta(G) = [\![G]\!]_{\mathbb{S}} = [\![M]\!]_{\mathbb{S}}.$$
  - If M deterministically converges to $E[\mathsf{D}(\vec{c})]$, let $\mathscr{E}$ be such that $E[\mathsf{D}(\vec{c})] \to_{n+1} \mathscr{E}$. By Lemma 38 and Lemma 35 we have, by induction hypothesis, that
$$\begin{aligned}\mathscr{D}(A) \leq \mathscr{E}(A) &= \int \mathscr{F}_N(A)\ E\{\mu_{\mathsf{D}}(\vec{c})\}(dN)\\ &\leq \int [\![N]\!]_{\mathbb{S}}(A)\ E\{\mu_{\mathsf{D}}(\vec{c})\}(dN)\\ &= [\![M]\!]_{\mathbb{S}}\end{aligned}$$
    where $N \to_n \mathscr{F}_N$.

Lemma 42 *For every $n \in \mathbb{N}$, $[\![M]\!]^n_{\mathbb{S}} \leq [\![M]\!]$.*

Proof. By induction on n:
- In the base case, then let us distinguish three cases depending on the three cases of Lemma 40, applied to M:
  - If M deterministically diverges, then by Lemma 37, $[\![M]\!]^0_{\mathbb{S}} \leq [\![M]\!]_{\mathbb{S}} = [\![M]\!]$.
  - If M deterministically converges to a generalized value G, then by Lemma 38 and Lemma 39, it holds that
$$[\![M]\!]^0_{\mathbb{S}} = [\![G]\!]^0_{\mathbb{S}} \leq [\![G]\!]_{\mathbb{S}} = \delta(G) = [\![G]\!] = [\![M]\!].$$
  - If M deterministically converges to $E[\mathsf{D}(\vec{c})]$, then $[\![M]\!]^0_{\mathbb{S}} = [\![E[\mathsf{D}(\vec{c})]]\!]^0_{\mathbb{S}} = \mathbf{0} \leq [\![M]\!]$.
- About the inductive case, let us again distinguish three cases depending on the three cases of Lemma 40, applied to M:
  - If M deterministically diverges, then by Lemma 37, $[\![M]\!]^{n+1}_{\mathbb{S}} \leq [\![M]\!]_{\mathbb{S}} = [\![M]\!]$.
  - If M deterministically converges to a generalized value G, then by Lemma 38 and Lemma 39, it holds that
$$[\![M]\!]^{n+1}_{\mathbb{S}} = [\![G]\!]^{n+1}_{\mathbb{S}} \leq [\![G]\!]_{\mathbb{S}} = \delta(G) = [\![G]\!] = [\![M]\!].$$
  - If M deterministically converges to $E[\mathsf{D}(\vec{c})]$, by Lemma 38 and Lemma 35 we have, by induction hypothesis, that
$$\begin{aligned}[\![M]\!]^{n+1}_{\mathbb{S}}(A) = [\![E[\mathsf{D}(\vec{c})]]\!]^{n+1}_{\mathbb{S}}(A) &= \int [\![N]\!]^n_{\mathbb{S}}(A)\ E\{\mu_{\mathsf{D}}(\vec{c})\}(dN)\\ &\leq \int [\![N]\!](A)\ E\{\mu_{\mathsf{D}}(\vec{c})\}(dN)\\ &= [\![M]\!].\end{aligned}$$

Restatement of Theorem 2 $[\![M]\!]_{\mathbb{S}} = [\![M]\!]$.

Proof.
$$\begin{aligned}[\![M]\!]_{\Rightarrow} &= \sup_{n\in\mathbb{N}}\{\mathscr{D} \mid M \to_n \mathscr{D}\} && \text{(by definition)}\\ &\leq [\![M]\!]_{\mathbb{S}} && \text{(by Lemma 41)}\\ &= \sup_{n\in\mathbb{N}} [\![M]\!]^n_{\mathbb{S}} && \text{(by Lemma 33)}\\ &\leq [\![M]\!] && \text{(by Lemma 42)}\\ &= [\![M]\!]_{\Rightarrow} && \text{(by Theorem 1)}\end{aligned}$$

Corollary 1 *The measure $[\![M]\!]^{\mathbb{S}}_{\Downarrow}$ is a sub-probability measure.*

7 Inference

In this section, we are interested in sampling the return values of a particular closed term $M \in \Lambda$. To avoid trivial cases, we assume that M has positive success probability and does not behave deterministically, i.e., that $[\![M]\!](\mathcal{V}) > 0$ and $[\![M]\!]^{\mathbb{S}}_{\Downarrow}(\{[]\}) = 0$. We can then let the probability distribution over the return values of M be defined as $[\![M]\!]_{\mathcal{GV}}(A) = [\![M]\!](A)/[\![M]\!](\mathcal{GV})$. We can sample from the distribution $[\![M]\!]_{\mathcal{GV}}$ using the Metropolis-Hastings (MH) algorithm [11] over the space of traces $s \in \mathbb{S}$. This algorithm yields consecutive samples from a Markov chain over $\mathbb{S}$, such that the density of the samples converges to $\mathbf{P}_M/[\![M]\!](\mathcal{GV})$. We can then apply the function $\mathbf{O}_M$ to obtain the return value of M for a given trace. The algorithm is parametric in a proposal density function $q(s,t)$. and consists of three steps:

1. Pick an initial state s such that $\mathbf{P}_M(s) \neq 0$ (e.g., by running M).
2. Draw the next state t at random with probability density $q(s,t)$.
3. Compute α as below.
$$\alpha = \min\left(1, \frac{\mathbf{P}_M(t)}{\mathbf{P}_M(s)} \cdot \frac{q(t,s)}{q(s,t)}\right)$$
   (a) With probability α, output t and continue from 2 with $s := t$.
   (b) Otherwise, output s and continue from 2 with s unchanged.

The formula used for the number α above is often called the Hastings acceptance probability. Different probabilistic programming language implementations use different choices for the density q above, based on pragmatics. The trivial choice would be to let $q(s,t) = \mathbf{P}_M(t)$ for all s, which always yields $\alpha = 1$ and so is equivalent to rejection sampling. We here define another simple density function q (based on Hur et al. [12]), giving emphasis to the conditions that it needs to satisfy in order to prove the convergence of the Markov chain given by the Metropolis-Hastings algorithm (Theorem 3).

7.1 A Metropolis-Hastings Transition Kernel

In the following, let M be a fixed program. Given a trace $s = [c_1, \ldots, c_n]$, we write $s_{i..j}$ for the trace $[c_i, \ldots, c_j]$ when $1 \leq i \leq j \leq n$. Intuitively, the following procedure describes how to obtain the proposal kernel density (q above):

1. Given a trace s of length n, let $t = [t_1, \ldots, t_n]$ where each t_i is drawn independently from a normal distribution with mean s_i and variance σ^2, and let p_i be the probability density of t_i.
2. Let $k \leq n$ be the largest number such that $(M, 1, []) \Rightarrow (M', w, t_{1..k})$. There are three cases:
   (a) If $k = n$, run $M' \Downarrow^{w'}_{t'} G$, and let $q(s, t@t') = p_1 \ldots p_n w'$.
   (b) If $k < n$ and $M' \Downarrow^{1}_{[]} G$, let $q(s, t_{1..k}) = p_1 \ldots p_k$.
   (c) Otherwise, let $q(s, t_{1..k}) = 0$ and propose the trace $[]$.

To define this kernel formally, we first give a function that partially evaluates M given a trace. Let

$$\texttt{peval}(M, s) = \begin{cases} M & \text{if } s = [] \\ M' & \text{if } (M, 1, []) \Rightarrow (M_k, w_k, s_k) \to (M', w', s) \\ & \text{for some } M_k, w_k, s_k, w' \text{ such that } s_k \neq s \\ \texttt{fail} & \text{otherwise} \end{cases}$$

For technical reasons, we need to ensure that Q is a probability kernel. We normalize $q(s,\cdot)$ by giving non-zero probability $q(s,[])$ to transitions ending in [] (which is not a completed trace of M by assumption).

Transition Density $q(s,t)$ and Kernel $Q(s,A)$ for Program M:

$q(s,t) = (\Pi_{i=1}^{k} \mathrm{pdf}_{\mathsf{Gaussian}}(s_i, \sigma^2, t_i)) \cdot \mathbf{P}_N(t_{k+1..|t|})$ if $|t| \neq 0$
where $k = \min\{|s|, |t|\}$ and $N = \mathtt{peval}(M, t_{1..k})$
$q(s,[]) = 1 - \int_A q(s,t)\,dt$ where $A = \{t \mid |t| \neq 0\}$
$Q(s,A) = \int_A q(s,t)\,dt$

We prove the following lemmas in Appendix A.

Lemma 43 `peval` *is a measurable function* $C\Lambda \times \mathbb{S} \to C\Lambda$.

Lemma 44 *For any program* M*, the transition density* $q(\cdot,\cdot) : (\mathbb{S} \times \mathbb{S}) \to \mathbb{R}_+$ *is measurable.*

Lemma 45 *The function* $Q(s,A)$ *is a probability kernel on* $(\mathbb{S}, \mathcal{S})$.

Hastings' Acceptance Probability α

$\alpha(s,t) = \min\{1, \frac{\mathbf{P}_M(t)q(t,s)}{\mathbf{P}_M(s)q(s,t)}\}$ where we let $\alpha = 0$ if $\mathbf{P}_M(t) = 0$
and otherwise $\alpha = 1$ if $\mathbf{P}_M(s)q(s,t) = 0$.

Given the proposal transition kernel Q and the acceptance ratio α, the Metropolis-Hastings algorithm yields a Markov chain over traces with the following transition probability kernel:

$$P(s,A) = \int_A \alpha(s,t)\,Q(s,dt) + \delta(s)(A) \cdot \int (1 - \alpha(s,t))\,Q(s,dt) \tag{1}$$

Define $P^n(s,A)$ to be the probability of the n-th element of the chain with transition kernel P starting at s being in A:

$$\begin{aligned} P^0(s,A) &= \delta(s)(A) \\ P^{n+1}(s,A) &= \int P(t,A)P^n(s,dt) \end{aligned}$$

If $f : X \to \mathbb{R}_+$ we let $\mathrm{supp}(f)$ be the *support* of f, that is, $\{x \in X \mid f(x) \neq 0\}$.

Lemma 46 *If* $s_0 \in \mathrm{supp}(\mathbf{P}_M)$ *then* $P^n(s_0, \mathrm{supp}(\mathbf{P}_M)) = 1$.

Proof. By induction on n. The base case holds, since $s_0 \in \mathrm{supp}(\mathbf{P}_M)$ by assumption. For the induction case, we have $P^{n+1}(s_0, \mathrm{supp}(\mathbf{P}_M)) = \int P(s, \mathrm{supp}(\mathbf{P}_M))P^n(s_0, ds)$. If $s \in \mathrm{supp}(\mathbf{P}_M)$ we have

$$\begin{aligned} P(s, \mathrm{supp}(\mathbf{P}_M)) &= \int_{\mathrm{supp}(\mathbf{P}_M)} \alpha(s,t)\,Q(s,dt) + \int (1 - \alpha(s,t))\,Q(s,dt) \\ &= \int_{\mathrm{supp}(\mathbf{P}_M)} q(s,t)\,dt + (1 - \alpha(s,[]))q(s,[]) \\ &= \int_{\mathrm{supp}(\mathbf{P}_M)} q(s,t)\,dt + (1 - \alpha(s,[]))(1 - \int_{\mathrm{supp}(\mathbf{P}_M)} q(s,t)\,dt) \\ &= 1 - \alpha(s,[])(1 - \int_{\mathrm{supp}(\mathbf{P}_M)} q(s,t)\,dt) \end{aligned}$$

where $\alpha(s, []) = 0$ since $\mathbf{P}_M([]) = 0$ by assumption. Then

$$\begin{aligned}\int P(s, \mathrm{supp}(\mathbf{P}_M))P^n(s_0, ds) &= \int_{\mathrm{supp}(\mathbf{P}_M)} P(s, \mathrm{supp}(\mathbf{P}_M))P^n(s_0, ds) \\ &= \int_{\mathrm{supp}(\mathbf{P}_M)} 1P^n(s_0, ds) \\ &= 1\end{aligned}$$

where the first and the third equality follow from the induction hypothesis.

Lemma 47 *There is* $0 \leq c < 1$ *such that* $P^n([], \mathrm{supp}(\mathbf{P}_M)) = 1 - c^n$ *and* $P^n([], \{[]\}) = c^n$.

Proof. Let $c = 1 - [\![M]\!]^{\mathbb{S}}_{\Downarrow}(\mathbb{S} \setminus \{[]\})$. By assumption $[] \notin \mathrm{supp}(\mathbf{P}_M)$ and $c < 1$, and since $[\![M]\!]^{\mathbb{S}}_{\Downarrow}$ is a sub-probability distribution we have $0 \leq c$. We proceed by induction on n. The base case is trivial. For the induction case, we have $P(s, \mathbb{S} \setminus \{[]\}) = 1$ for all $s \in \mathrm{supp}(\mathbf{P}_M)$. Finally

$$\begin{aligned}P([], \mathrm{supp}(\mathbf{P}_M)) = \int_{\mathrm{supp}(\mathbf{P}_M)} \mathbf{P}_M \\ = [\![M]\!]^{\mathbb{S}}_{\Downarrow}(\mathrm{supp}(\mathbf{P}_M)) = [\![M]\!]^{\mathbb{S}}_{\Downarrow}(\mathbb{S} \setminus \{[]\}).\end{aligned}$$

Based on Lemma 46 and 47, we below consider the Markov chain with kernel P restricted to $\mathrm{supp}(\mathbf{P}_M) \cup \{[]\}$.

7.2 Correctness of Inference

By saying that the inference algorithm is correct, we mean that as the number of steps goes to infinity, the distribution of generated samples approaches the distribution specified by the sampling-based semantics of the program.

Formally, we define $T^n(s, A) = P^n(s, \mathbf{O}_M^{-1}(A))$ as the value sample distribution at step n of the Metropolis-Hastings Markov chain. For two measures defined on the same measurable space (X, Σ), we also define the variation norm $||\mu_1 - \mu_2||$ as:

$$||\mu_1 - \mu_2|| = \sup_{A \in \Sigma} |\mu_1(A) - \mu_2(A)|$$

We want to prove the following theorem:

Theorem 3 (Correctness) *For every trace* s *with* $\mathbf{P}_M(s) \neq 0$,

$$\lim_{n \to \infty} ||T^n(s, \cdot) - [\![M]\!]_{\mathcal{GV}}|| = 0.$$

To do so, we first need to investigate convergence of P^n. It is convenient to define a measure for its target distribution.

Target Distribution π

$$\pi(A) = [\![M]\!]^{\mathbb{S}}_{\Downarrow}(A)/[\![M]\!]^{\mathbb{S}}_{\Downarrow}(\mathbb{S})$$

We use a sequence of known results for Metropolis-Hastings Markov chains [20] to prove that P^n converges to π. We say that a Markov chain transition kernel P is $\mathscr{D}$*-irreducible* if $\mathscr{D}$ is a non-zero sub-probability measure on $(\mathbb{S}, \mathcal{S})$, and for all $x \in \mathbb{S}, A \in \mathcal{S}$ there exists an integer $n > 0$ such that $\mathscr{D}(A) > 0$ implies $P^n(x, A) > 0$. We say that P is $\mathscr{D}$*-aperiodic* if there do not exist $d \geq 2$ and disjoint $B_1, \ldots, B_d$ such that $\mathscr{D}(B_1) > 0$, and $x \in B_d$ implies $P(x, B_1) = 1$, and $x \in B_i$ implies that $P(x, B_{i+1}) = 1$ for $i \in \{1, \ldots, d-1\}$.

Lemma 48 (Tierney [20], Theorem 1 and Corollary 2) *Let K be the transition kernel of a Markov chain given by the Metropolis-Hastings algorithm with target distribution $\mathscr{D}$. If K is $\mathscr{D}$-irreducible and aperiodic, then for all s, $\lim_{n\to\infty} ||K^n(s,\cdot) - \mathscr{D}|| = 0$.*

Lemma 49 (Strong Irreducibility) *If $\mathbf{P}_M(s) > 0$ and $[\![M]\!]^{\mathbb{S}}_{\Downarrow}(A) > 0$ then $P(s,A) > 0$.*

Proof. There is n such that $[\![M]\!]^{\mathbb{S}}_{\Downarrow}(A \cap \mathbb{S}_n) > 0$. Write $A|_n = A \cap \mathbb{S}_n$. For all $t \in A|_n$, $q(s,t) > 0$ by case analysis on whether $n \leq |s|$. If $n \leq |s|$, then for all $t \in A|_n$,

$$q(s,t) = \Pi_{i=1}^{n} \mathrm{pdf}_{\mathsf{Gaussian}}(s_i, \sigma^2, t_i) > 0 \qquad \text{and}$$
$$q(t,s) = (\Pi_{i=1}^{n} \mathrm{pdf}_{\mathsf{Gaussian}}(t_i, \sigma^2, s_i)) \cdot \mathbf{P}_{\mathtt{peval}(M, s_{1..n})}(s_{(n+1)}, \ldots, s_n) > 0.$$

Similarly, if $n > |s|$, then for all $t \in A|_n$,

$$q(s,t) = (\Pi_{i=1}^{|s|} \mathrm{pdf}_{\mathsf{Gaussian}}(s_i, \sigma^2, t_i)) \cdot \mathbf{P}_{\mathtt{peval}(M, t_{1..|s|})}(t_{(|s|+1)}, \ldots, t_{|s|}) > 0 \qquad \text{and}$$
$$q(t,s) = \Pi_{i=1}^{|s|} \mathrm{pdf}_{\mathsf{Gaussian}}(t_i, \sigma^2, s_i) > 0.$$

Since $\mu(A|_n) > 0$ and $\mathbf{P}_M(t) > 0$ for all $t \in A|_n$,

$$\begin{aligned} P(s,A) &\geq P(s, A|_n) \\ &\geq \int_{A|_n} \alpha(s,t)\, Q(s, dt) \\ &= \int_{A|_n} \alpha(s,t) q(s,t)\, dt \\ &= \int_{A|_n} \min\{q(s,t), \frac{\mathbf{P}_M(t) q(t,s)}{\mathbf{P}_M(s)}\}\, dt \\ &> 0. \end{aligned}$$

Corollary 2 (Irreducibility) *P as given by Equation (1) is π-irreducible.*

Lemma 50 (Aperiodicity) *P as given by Equation (1) is π-aperiodic.*

Proof. Assume that B_1, B_2 are disjoint sets such that $\pi(B_1) > 0$ and $P(s, B_2) = 1$ for all $s \in B_1$. If $s \in B_1$, Lemma 49 gives that $P(s, B_1) > 0$, so $P(s, B_2) < P(s, \mathbb{S}) = 1$, which is a contradiction. A fortiori, P is π-aperiodic.

Lemma 51 *If μ_1 and μ_2 are measures on (X_1, Σ_1) and $f : X_1 \to X_2$ is measurable Σ_1/Σ_2, then*

$$\sup_{B \in \Sigma_2} ||\mu_1 f^{-1} - \mu_2 f^{-1}|| \leq \sup_{A \in \Sigma_1} ||\mu_1 - \mu_2||$$

Proof. We have $\sup_{B\in\Sigma_2} ||\mu_1 f^{-1} - \mu_2 f^{-1}|| = \sup_{A\in\Sigma_1'} ||\mu_1 - \mu_2||$, where $\Sigma_1' = \{f^{-1}(B) | B \in \Sigma_2\}$. By measurability of f we get $\Sigma_1' \subseteq \Sigma_1$, so by monotonicity of sup we get $\sup_{A\in\Sigma_1'} ||\mu_1 - \mu_2|| \leq \sup_{A\in\Sigma_1} ||\mu_1 - \mu_2||$.

Restatement of Theorem 3 *For every trace s with $\mathbf{P}_M(s) \neq 0$,*

$$\lim_{n\to\infty} ||T_M^n(s,\cdot) - [\![M]\!]_{\mathcal{GV}}|| = 0.$$

Proof. By Corrollary 2, P is π-irreducible, and by Lemma 50, P is π-aperiodic. Lemma 48 then yields that

$$\lim_{n\to\infty} ||P^n(x,\cdot) - \pi|| = 0.$$

By definition, $T^n(s,A) = P^n(s, \mathbf{O}_M^{-1}(A))$ and $[\![M]\!]_{\mathcal{GV}}(A) = [\![M]\!](A)/[\![M]\!](\mathcal{GV})$, where $[\![M]\!]_{\mathbb{S}} = [\![M]\!]^{\mathbb{S}}_{\Downarrow}\mathbf{O}_M^{-1}$ by Theorem 2. By Lemma 51 we finally get

$$\lim_{n\to\infty} ||T^n(s,\cdot) - [\![M]\!]_{\mathcal{GV}}|| = 0.$$

8 Related Work

To the best of our knowledge, the only previous theoretical justification for trace MCMC is the recent work by Hur et al. [12], who show correctness of trace MCMC for the imperative probabilistic language R2 [16]. Their result does not apply to higher-order languages such as Church or our λ-calculus. The authors do state that the space of traces in their language is equipped with a "stock" measure, and the probabilities of program traces and transitions can be treated as densities with respect to that measure. They do not, however, show that these densities are measurable. The proof of correctness in [12] only shows that the acceptance ratio of the algorithm matches the standard formula for the MH algorithm on spaces of fixed dimensionality: the authors prove neither irreducibility nor aperiodicity of the resulting Markov chain.

Other current implementations of probabilistic languages also use MCMC inference, including Stan [7], Church [10], Venture [15], and Anglican [21]. These works do not make formal correctness claims for their implementations, instead focusing on efficiency and convergence properties of their implementations for a number of representative programs.

Wingate et al. [23] give a general program transformation for a probabilistic language to support trace MCMC, with a focus on labelling sample points in order to maximise sample reuse. Our trace semantics could easily be extended with such labelling.

Kozen [14] gives a semantics of imperative probabilistic programs as partial measurable functions from infinite random traces to final states, and prove this semantics equivalent to a domain-theoretic one. Such operational semantics are akin to the **peval** function in this paper, and appear in many works. Park et al. [18] study a similar semantics for functional programs, but "do not investigate measure-theoretic properties". Cousot and Monerau [3] generalise Kozen's operational semantics to consider probabilistic programs as measurable functions from a probability space into a semantics domain, and study several kinds of abstract interpretation in this setting. Recently, Toronto et al. [22] use a pre-image version of Kozen's semantics to obtain an efficient implementation using rejection sampling. Scibior et al. [19] define a monadic embedding of probabilistic programming in Haskell along the lines of Kozen's semantics; their paper describes various inference algorithms but has no formal correctness results.

Like Kozen's denotational semantics, our distributional semantics makes use of the partially additive structure on the category of sub-probability kernels [17] in order to treat programs that may take an unbounded number of computation steps.

While giving a fully abstract domain theory for probabilistic λ-calculi is known to be hard [13], there have been recent advances using probabilistic coherence spaces [4, 6] and game semantics [5], which in some cases are fully abstract. We don't see strong obstacles in applying any of these to our calculus, but this remains outside the scope of this contribution.

9 Conclusions

We defined a probabilistic λ-calculus with draws from continuous probability distributions, defined its semantics as distributions on terms, and proved correctness of a trace MCMC inference algorithm via a sampling semantics for the calculus.

A Proofs of Measurability

First we recap some useful results from measure theory:

Lemma 52 ([1, ex. 13.1]) *Let (Ω, Σ) and (Ω', Σ') be two measurable spaces, $T : \Omega \to \Omega'$ a function and $A_1, A_2, \ldots$ a countable collection of sets in Σ whose union is Ω. Let $\Sigma_n = \{A \mid A \subseteq A_N, A \in \Sigma\}$ be a σ-algebra in A_n and $T_n : A_n \to \Omega'$ a restriction of T to A_n. Then T is measurable Σ/Σ' if and only if T_n is measurable Σ_n/Σ' for every n.*

- A function $f : X_1 \to X_2$ between metric spaces (X_1, d_1) and (X_2, d_2) is *continuous* if for every $x \in X_1$ and $\epsilon > 0$, there exists δ such that for every $y \in X_1$, if $d_1(x, y) < \delta$, then $d_2(f(x), f(y)) < \epsilon$.
- A sequence $\{x_n\}_{n\in\mathbb{N}}$ is *Cauchy* if

$$\forall \epsilon > 0\ \exists n \in \mathbb{N}\ \forall k, l \geq n \quad d(x_k, x_l) < \epsilon$$

- A metric space (X, d) is *complete* if every Cauchy sequence $\{x_n\}$ on X has a limit in X.
- A subset A of a metric space (X, d) is *dense* if

$$\forall x \in X, \epsilon > 0\ \exists y \in A \quad d(x, y) < \epsilon$$

- A metric space is *separable* if it has a countable dense subset.

Lemma 53 ([2, Propositions 1 and 6]) *Let X_1, X_2 be complete separable metric spaces, $\mathcal{B}_1, \mathcal{B}_2$ their Borel σ-algebras, and f a function from X_1 to X_2. The function f is measurable $\mathcal{B}_1/\mathcal{B}_2$ if and only if its graph $\{(x, f(x)) \mid x \in X_1\}$ is measurable in the product space $(X_1, \mathcal{B}_1) \times (X_2, \mathcal{B}_2)$.*

Lemma 54 (Parthasarathy, 3.9) *If X_1, X_2 are metric spaces and E_1 and E_2 are subsets of respectively X_1 and X_2 and E_1 is a Borel set and f a measurable injective function from E_1 to E_2 such that $f(E_1) = E_2$, then E_2 is a Borel set.*

We can use lemma 52 to split the space $\mathcal{M}$ of expressions into subspaces of expressions of different type, and restrict functions (such as the reduction relation) to a given type of expression, to process different cases separately.

We write $\mathbf{Subst}(M, x, v)$ for $M\{V/x\}$, to emphasize the fact that substitution is a function.

Detailed definition of substitution

$$\begin{array}{l}
\mathbf{Subst}(c, x, V) \triangleq c \\
\mathbf{Subst}(x, x, V) \triangleq V \\
\mathbf{Subst}(x, y, V) \triangleq y \quad \text{if } x \neq y \\
\mathbf{Subst}(\lambda x.M, x, V) \triangleq \lambda x.M \\
\mathbf{Subst}(\lambda x.M, y, V) \triangleq \lambda x.(\mathbf{Subst}(M, y, V)) \quad \text{if } x \neq y \\
\mathbf{Subst}(M\ N, x, V) \triangleq \mathbf{Subst}(M, x, V)\ \mathbf{Subst}(N, x, V) \\
\mathbf{Subst}(\mathsf{D}(V_1, \ldots, V_{|\mathsf{D}|}), x, V) \triangleq \mathsf{D}(\mathbf{Subst}(V_1, x, V), \ldots, \mathbf{Subst}(V_{|\mathsf{D}|}, x, V)) \\
\mathbf{Subst}(g(V_1, \ldots, V_{|g|}), x, V) \triangleq g(\mathbf{Subst}(V_1, x, V), \ldots, \mathbf{Subst}(V_{|g|}, x, V)) \\
\mathbf{Subst}(\texttt{if } W \texttt{ then } M \texttt{ else } L, x, V) \triangleq \\
\quad \texttt{if } \mathbf{Subst}(W, x, V) \texttt{ then } \mathbf{Subst}(M, x, V) \texttt{ else } \mathbf{Subst}(L, x, V) \\
\mathbf{Subst}(\alpha, x, V) \triangleq \alpha
\end{array}$$

Let us define the metric on triples of arguments of **Subst** and on contexts.

$$\begin{array}{rcl}
d((M, x, V), (N, x, W)) & \triangleq & d(M, N) + d(V, W) \\
d((M, x, V), (N, y, W)) & \triangleq & \infty \quad \text{if } x \neq y
\end{array}$$

$$\begin{array}{rcl}
d([\cdot], [\cdot]) & \triangleq & 0 \\
d(EM, FN) & \triangleq & d(E, F) + d(M, N) \\
d((\lambda x.M)E, (\lambda x.N)F) & \triangleq & d(M, N) + d(E, F) \\
d(E, F) & \triangleq & \infty \quad \text{otherwise}
\end{array}$$

Lemma 55 $d(E[M], F[N]) \leq d(E, F) + d(M, N)$

Proof. If $d(E,F)=\infty$, then the result is obvious, since $d(M',N')\leq\infty$ for all M',N'.

Now let us assume $d(E,F)\neq\infty$ and prove the result by simultaneous induction on the structure on E and F:

- Case $E=F=[\cdot]$: in this case, $E[M]=M$, $F[N]=N$, and $d(E,F)=0$, so obviously $d(E[M],F[N])=d(E,F)+d(M,N)$
- Case $E=E'\ L_1$, $F=F'\ L_2$:
  We have $d(E[M],F[N])=d(E'[M]\ L_1,F'[N]\ L_2)=d(E'[M],F'[N])+d(L_1,L_2)$. By induction hypothesis, $d(E'[M],F'[N])\leq d(E',F')+d(M,N)$, so $d(E[M],F[N])\leq d(E',F')+d(M,N)+d(L_1,L_2)=d(E,F)+d(M,N)$.
- Case $E=(\lambda x.L_1)\ E'$, $F=(\lambda x.L_2)\ F'$:
  We have $d(E[M],F[N])=d((\lambda x.L_1)(E'[M]),(\lambda x.L_2)(F'[N]))=d(\lambda x.L_1,\lambda x.L_2)+d(E'[M],F'[N])$. By induction hypothesis, $d(E'[M],F'[N])\leq d(E',F')+d(M,N)$, so $d(E[M],F[N])\leq d(E',F')+d(\lambda x.L_1,\lambda x.L_2)+d(M,N)=d(E,F)+d(M,N)$.

Lemma 56 *$d(\mathbf{Subst}(M,x,V),\mathbf{Subst}(N,x,W))\leq d(M,N)+k\cdot d(V,W)$ where k is the max of the multiplicities of x in M and N*

Proof. By simultaneous induction on the structure of M and N.

Lemma 57 **Subst**(M,x,v) *is continuous, and so measurable* $(\mathcal{M}\times\mathcal{X}\times\mathcal{M})/\mathcal{M}$.

Proof. Corollary of Lemma 56.

Deterministic reduction:

$$\begin{aligned}
E[(\lambda x.M)V] &\to E[\mathbf{Subst}(M,x,V)]\\
E[g(\vec{c})] &\to E[\sigma_g(\vec{c})]\\
E[\alpha] &\to \alpha\\
E[\texttt{if true then } M \texttt{ else } N] &\to E[M]\\
E[\texttt{if false then } M \texttt{ else } N] &\to E[N]
\end{aligned}$$

Let $\mathcal{C}$ denote the set of contexts and $\mathcal{G}$ the set of primitive functions. Let:

- $P_A\triangleq\{E[(\lambda x.M)V]\mid E\in\mathcal{C},M\in\Lambda,V\in\mathcal{V}\}$
- $P_T\triangleq\{E[\texttt{if true then } M \texttt{ else } N]\mid E\in\mathcal{C},M,N\in\Lambda\}$
- $P_F\triangleq\{E[\texttt{if false then } M \texttt{ else } N]\mid E\in\mathcal{C},M,N\in\Lambda\}$
- $P_E\triangleq\{E[\alpha]\mid E\in\mathcal{C}\setminus\{[\,]\},\alpha\in\mathcal{S}\}$
- $P_G(g)\triangleq\{E[g(\vec{c})]\mid E\in\mathcal{C},\vec{c}\in\mathbb{R}^{|g|}\}$
- $P_G\triangleq\bigcup_{g\in\mathcal{G}}P_G(g)$
- $P_{det}\triangleq P_A\cup P_T\cup P_F\cup P_E\cup P_G$
- $P_{rnd}(D)\triangleq\{E[D(\vec{c})]\mid E\in\mathcal{C},\vec{c}\in\mathbb{R}^{|D|}\}$
- $P_{rnd}\triangleq\bigcup_{D\in\mathcal{D}}P_{rnd}(D)$

Lemma 58 *The sets P_A, P_T, P_F, P_E, P_G are Borel-measurable.*

Proof. All these sets are closed, so they are obviously measurable.

Lemma 59 *Every Cauchy sequence $\{M_i\}$ of terms has a limit in Λ_P.*

Proof. The sequence $\{M_i\}$ is a Cauchy sequence iff:

$$\forall\epsilon>0\ \exists n\in\mathbb{N}\ \forall k,l\geq n\quad d(M_k,M_l)\leq\epsilon$$

Fix ϵ and n. It is easy to show that if $d(M_k,M_l)\leq\infty$ for $k,l\geq n$, then all expressions $M_n,M_{n+1},\ldots$ must have the same form.

We prove the result by induction on the structure of M_n. Interesting cases:

- Case $M_n = c_n$: Here, for all $k \geq n$, $M_k = c_k$ for some $c_k \in \mathbb{R}$, so the result follows from the completeness of $\mathbb{R}$ with the standard metric $d(x, y) = |x - y|$.
- Case $M_n = x$: The only possibility is $M_k = x$ for all $k \geq n$, so the sequence trivially converges to x.
- Case $M_n = \alpha$: This implies $M_k = \alpha$ for all $k \geq n$, so the sequence converges to α.
- Case $M_n = N_n L_n$: for all $k \geq n$, $M_k = N_k L_k$ for some N_k, L_k. Then, for $k, l \geq n$, $d(N_k, N_l) \leq d(M_k, M_l) < \epsilon$ and $d(L_k, L_l) \leq d(M_k, M_l) < \epsilon$, so $\{N_k\}_{k \geq n}$ and $\{L_k\}_{k \geq n}$ are Cauchy sequences. By induction hypothesis, $\{N_k\}_{k \geq n}$ converges to some $N \in \Lambda_P$ and $\{L_k\}_{k \geq n}$ converges to some $L \in \Lambda_P$. Thus,

  $$\forall \epsilon > 0 \; \exists n' > n \in \mathbb{N} \; \forall k > n' \quad d(N_k, N) \leq \frac{\epsilon}{2} \wedge d(L_k, L) \leq \frac{\epsilon}{2}$$

  Thus:

  $$\forall \epsilon > 0 \; \exists n' > n \in \mathbb{N} \; \forall k > n' \quad d(N_k L_k, NL) \leq \epsilon$$

  and so $\{M_i\}$ converges to $NL \in \Lambda_P$.
- Other cases analogous to $M_n = N_n L_n$

Lemma 60 *The metric space (Λ_P, d) is complete.*

Proof. Corollary of Lemma 59.

Lemma 61 *The metric space (P_{det}, d) is complete.*

Proof. A closed subset of a complete metric space is complete.

Let Λ_Q be the subset of Λ_P in which all constants are rational. Then, it is easy to show that Λ_Q is countable.

Lemma 62 *Λ_Q is a dense subset of (Λ_P, d)*

Proof. We need to prove that

$$\forall M \in \Lambda_P, \epsilon > 0 \; \exists M_Q \in \Lambda_Q \quad d(M, M_q) < \epsilon$$

This can be easily shown by induction (the base case follows from the fact that $\mathbb{Q}$ is a dense subset of $\mathbb{R}$.

Lemma 63 *The metric space (Λ_P, d) is separable.*

Proof. Corollary of Lemma 62.

Lemma 64 *The metric space (P_{det}, d) is separable.*

Proof. Analogous to the proof of separability of (Λ_P, d).

We need to define a metric space on the set $\Lambda \times \mathbb{R}_+ \times \mathbb{S}$, show that it is complete separable and that the σ-algebra generated by its metric is the product σ-algebra. Let:

$$\begin{aligned} d(w, w') &\triangleq |w - w'| \\ d(s, s') &\triangleq \begin{cases} \sum_{i=1}^{|s|} |s_i - s'| & \text{if } |s| = |s'| \\ 0 & \text{otherwise} \end{cases} \\ d((M, w, s), (M', w', s')) &\triangleq d(M, M') + d(w, w') + d(s, s') \end{aligned}$$

Let $\mathcal{T} = \Lambda \times \mathbb{R}_+ \times \mathbb{S} \cup \{\perp\}$. Define:

$$\begin{aligned} d((M, w, s), \perp) &\triangleq \infty \\ d(\perp, \perp) &\triangleq 0 \end{aligned}$$

Lemma 65 *The σ-algebra on $\Lambda \times \mathbb{R}_+ \times \mathbb{S}$ generated by the metric d is $\mathcal{M} \times \mathcal{R}_+ \times \mathcal{S}$*

Lemma 66 *The metric space $(\Lambda \times \mathbb{R}_+ \times \mathbb{S}, d)$ is separable and complete.*

Lemma 67 *The deterministic reduction $M \to N$ is measurable.*

Proof. Split the space of all deterministic redexes (in contexts)into subspaces of expressions of different types (i.e. applications, deterministic function calls and if statements where condition is true/false, in arbitrary contexts). By Lemma 58, these subspaces are measurable.

By Lemma 14, the deterministic reduction $(\to)$ is a graph of some function $f_d : P_{det} \to \Lambda_P$ and it is easy to see that it is a total function on P_{det} (the reduction relation has no assumptions).

Now let f_A, f_T, f_F, f_E, f_G be f restricted to P_A, P_T, P_F, P_E, P_G, respectively. Let us show that all these restrictions are measurable:

- Case f_A:
  We have $f_A(E[(\lambda x.M)V]) = E[\mathbf{Subst}(M, x, V)]$. We will show that this function is continuous: By Lemma 55, we have $d(E[(\lambda x.M)V], F[(\lambda x.N)W]) = d(E, F) + d(M, N) + d(V, W)$ and by Lemma 56, $d(E[\mathbf{Subst}(M, x, V)], F[\mathbf{Subst}(N, x, W)]) \leq d(E, F) + d(M, N) + k \cdot d(V, W)$, where k is the maximum of the multiplicities of x in M and N.
  For any $\epsilon > 0$, take $\delta = \frac{\epsilon}{k+1}$. Then, if $d(E[(\lambda x.M)V], F[(\lambda x.N)W]) < \delta$, then

$$\begin{aligned} d(E[\mathbf{Subst}(M, x, V)], F[\mathbf{Subst}(N, x, W)]) &\leq d(E, F) + d(M, N) + k \cdot d(V, W) \\ &\leq (k+1) \cdot (d(E, F) + d(M, N) + d(V, W)) \\ &= (k+1) \cdot d(E[(\lambda x.M)V], F[(\lambda x.N)W]) \\ &< \epsilon \end{aligned}$$

  Thus, f_A is continuous, and so measurable.
- Case f_G:
  Follows from the assumption (all built-in deterministic functions are total and measurable).
- Case f_T:
  We have $f_T(E[\texttt{if true then } M \texttt{ else } N]) = E[M]$. Thus:
  $d(E[\texttt{if true then } M_1 \texttt{ else } N_1], F[\texttt{if true then } M_2 \texttt{ else } N_2]) = d(E, F) + d(M_1, M_2) + d(N_1, N_2) \geq d(E[M_1], F[M_2])$, so f_T is continuous, and so measurable
- Case f_F: analogous
- Case f_E: We have $f_E(E[\alpha]) = \alpha$. Hence $d(E[\alpha], F[\beta]) = d(E, F) + d(\alpha, \beta) \geq d(\alpha, \beta)$, so h_e is continuous, and hence measurable.
  By Lemma 52, f is measurable $\mathcal{M}^{P_{det}}/\mathcal{M}$.
  By lemmas 61, 64 and 53, $(\to) \in \mathcal{M}^{P_{det}} \times \mathcal{M}$

Lemma 68 *$\mathcal{S}$ is the Borel σ-algebra on E.*

Proof. $\mathcal{S}$ is a countable direct sum of Borel σ-algebras.

Let $\mathcal{T}_{rnd} = \{(E[D(\vec{c})], w, s@[c]) \mid E \in \mathcal{C}, D \in \mathcal{D}, \vec{c} \in \mathbb{R}^{|D|}, w \in \mathbb{R}, s \in \mathbb{S}, c \in \mathbb{R}, \mathrm{pdf}_D(\vec{c}, c) > 0\}$

Lemma 69 *$\mathcal{T}_{rnd}$ is measurable.*

Proof. For each distribution D, define a function $i_D : P_{rnd}(D) \times \mathbb{R} \times (\mathbb{S} \setminus \{[]\}) \to \mathbb{R}^{|D|} \times \mathbb{R}$ by $i_D(E[D(\vec{c})], w, s@[c]) = (\vec{c}, c)$. This function is continuous, and so measurable. Then, since for each D, pdf_D is measurable by assumption, the function $j_d = \mathrm{pdf}_D \circ i_D$ is measurable. Then, $\mathcal{T}_{rnd} = \bigcup_{D \in \mathcal{D}} j_D^{-1}((0, \infty))$, and since the set of distribution is countable, $\mathcal{T}_{rnd}$ is measurable.

Lemma 70 *The reduction relation $(M, w, s) \to (M', w', s')$ is measurable.*

Proof. Define a function $g : \mathcal{T} \to \mathcal{T}$ as:

$$\begin{aligned} g(M,w,s) &= \begin{cases} g_d(M,w,s) & \text{if}(M,w,s) \in P_{det} \times \mathbb{R} \times \mathbb{S} \\ g_r(M,w,s) & \text{if}(M,w,s) \in \mathcal{T}_{rnd} \\ g_e(M,w,s) & \text{otherwise} \end{cases} \\ g(\bot) &= \bot \end{aligned}$$

where:

$$g_d(M,w,s) \triangleq (N,w,s) \quad \text{where } M \to N$$

$$\begin{aligned} g_r &\triangleq (g_1, g_2, g_3) \\ g_1(E[\mathsf{D}(\vec{c})], w, s@[c]) &\triangleq E[c] \\ g_2(E[\mathsf{D}(\vec{c})], w, s@[c]) &\triangleq w \operatorname{pdf}_{\mathsf{D}}(\vec{c}, c), \\ g_3(E[\mathsf{D}(\vec{c})], w, s@[c]) &\triangleq s \end{aligned}$$

$$g_e(M,w,s) \triangleq \bot$$

Then:

- By Lemma 17, the function g_d is well-defined.
  The Borel σ-algebra on $\Lambda \times \mathbb{R} \times \mathbb{S}$ is generated by measurable rectangles of the form $A \times R \times S$, where $A \in \mathcal{M}$, $R \in \mathbb{R}$, $S \in \mathbb{S}$. We have $g_d^{-1}(A,R,S) = f_d^{-1}(A) \times R \times S$, where f_d is the measurable function from the proof of Lemma 67, so g_d is Borel-measurable.
- For g_1, we have $d(E[c], E'[c']) \leq d(E,E') + d(c,c') \leq d(E,E') + d(\vec{c},\vec{c'}) + d(w,w') + d(s,s') = d((E[\mathsf{D}(\vec{c})], w, s@[c]), (E'[\mathsf{D}(\vec{c'})], w', s'@[c']))$ and $d((E[\mathsf{D}(\vec{c})], w, s@[c]), (E'[\mathsf{E}(\vec{c'})], w', s'@[c'])) = \infty$ if $\mathsf{D} \neq \mathsf{E}$, so g_1 is continuous and hence Borel-measurable.
  For g_2, we have $g_2(E[\mathsf{D}(\vec{c})], w, s@[c]) = g_w(E[\mathsf{D}(\vec{c})], w, s@[c]) \times (\operatorname{pdf}_{\mathsf{D}} \circ g_c)(E[\mathsf{D}(\vec{c})], w, s@[c])$, where $g_w(E[\mathsf{D}(\vec{c})], w, s@[c]) = w$ and $g_c(E[\mathsf{D}(\vec{c})], w, s@[c]) = (\vec{c}, c)$. The continuity (and so measurability) of g_w and g_c can be easily checked (as for g_1 above). Thus, $\operatorname{pdf}_{\mathsf{D}} \circ g_c$ is a composition of measurable functions (since distributions are assumed to be measurable), and so g_2 is a pointwise product of measurable functions, so it is measurable.
  The continuity (and so measurability) of g_3 can be shown in a similar way to g_1.
  Hence, g_r is measurable.
- Since the domains of the measurable functions g_d and g_r are Borel sets, the domain of g_e, the set $D = \mathcal{T} \setminus (P_{det} \times \mathbb{R} \times \mathbb{S} \cup \mathcal{T}_{rnd})$, is also measurable. Since $g_e^{-1}(\bot) = D$, g_e is measurable.
- The constant function sending $\bot$ to $\bot$ is trivially measurable.
  By Lemma 53, the graph G of g is measurable. We have $G = G_d \cup G_d \cup G_e \cup \{(\bot,\bot)\}$, where

$$\begin{aligned} G_d &= \{((M,w,s),(N,w,s)) \mid M \to N, M \in P_{det}, N \in \Lambda, w \in \mathbb{R}_+, s \in E\} \\ G_r &= \{((E[\mathsf{D}(\vec{c})], w, s@[c]), (E[c], w \operatorname{pdf}_{\mathsf{D}}(\vec{c}, c), s)) \\ &\qquad \mid E \in \mathcal{C}, \mathsf{D} \in \mathcal{D}, \vec{c} \in \mathbb{R}^{|D|}, c \in \operatorname{supp}(\mathsf{D}(\vec{c})), w \in \mathbb{R}, s \in E\} \end{aligned}$$

and G_e is the graph of g_e. Since the image of g_e is $\{\bot\}$, we have $G \cap ((\mathcal{T} \setminus \{\bot\}) \times (\mathcal{T} \setminus \{\bot\})) = G_d \cup G_r$, which implies that $G_d \cup G_r$ is measurable.

Now, define a continuous and measurable function $h((M,w,s),(M',w',s')) = ((M,w,s'),(M',w',s))$. We have $h^{-1}(G_d \cup G_r) = (\to)$, which implies that $\to$ is measurable.

Lemma 71 *The reduction relation $(M,w,s) \to^k (M',w',s')$ is measurable for all $k \geq 0$.*

Proof. The k-fold composition of the function g from the proof of Lemma 70 with itself, g^k, is measurable. By Lemma 53, its graph G_k is also measurable. By taking the intersection of G_k and

the measurable set $(\mathcal{T}\setminus\{\bot\})\times(\mathcal{T}\setminus\{\bot\})$, and swapping the initial and final trace (which preserves measurability, as shown in the proof of Lemma 70), we obtain $(\to^k)$. Thus, $(\to^k)$ is measurable.

Lemma 72 *The reduction relation $(M,w,s)\Rightarrow(M',w',s')$ is measurable.*

Proof. By Lemma 71, $\to^k$ is measurable for all k. The relation $\Rightarrow$ is measurable, since it is a countable union of measurable sets.

To avoid any confusion, let us denote the restricted relation $(M,1,[])\Rightarrow(G,w,s)$ by $\rightsquigarrow$.

Lemma 73 *The reduction relation $(M,1,[])\rightsquigarrow(G,w,s)$ is measurable.*

Proof. We have $\rightsquigarrow\;=\;\Rightarrow\cap((\Lambda\times\{1\}\times\{[]\})\times(\mathcal{GV}\times\mathbb{R}\times\mathbb{S}))$, so $\rightsquigarrow$ is an intersection of two measurable sets.

Let $\mathcal{T}_{red}=\{(M,s)\mid(M,1,[])\Rightarrow(G,w,s),G\in\mathcal{GV},w\in\mathbb{R}\}$

Lemma 74 *$\mathcal{T}_{red}$ is Borel-measurable.*

Proof. Let g be as defined in the proof of Lemma 70.

For each $k\geq 0$, let

$$\begin{aligned}\underline{G^k} &= (g^k)^{-1}(\mathcal{GV}\times\mathbb{R}\times\{[]\})\cap(\Lambda\times\{1\}\times\mathbb{S})\\ &= \{(M,1,s)\mid(M,1,[])\to^k(G,w,s),w\in\mathbb{R},G\in\mathcal{GV}\}\end{aligned}$$

Since g is measurable, and composition of measurable functions is measurable, every $\underline{G^k}$ is measurable.

Then:

$$\bigcup_k\underline{G^k}=\{(M,1,s)\mid(M,1,[])\Rightarrow(G,w,s),w\in\mathbb{R},G\in\mathcal{GV}\}$$

Now, define a continuous, measurable bijection $i:\Lambda\times\mathbb{S}\to\Lambda\times\{1\}\times\mathbb{S}$ as $i(M,s)=(M,1,s)$. Then

$$\begin{aligned}i^{-1}(\bigcup_k\underline{G^k}) &= \{(M,s)\mid(M,1,[])\Rightarrow(G,w,s),w\in\mathbb{R},G\in\mathcal{GV}\}\\ &= \mathcal{T}_{red}\end{aligned}$$

Hence, $\mathcal{T}_{red}$ is measurable.

Restatement of Lemma 30 and Lemma 32 *For any program M, $\mathbf{P}_M$ is a measurable function $\mathbb{S}\to\mathbb{R}_+$ and $\mathbf{O}_M$ is a measurable function $\mathbb{S}\to\mathcal{GV}$.*

Proof. For each M, the relation $\mathbf{Run}_M(s,(G,w))$, defined as $\mathbf{Run}_M(s,(G,w))$ iff $(M,1,[])\Rightarrow(G,w,s)$, is measurable (by taking the intersection of $\rightsquigarrow$ with a rectangular set whose first component is $\{M\}$, and rearranging the components continuously). Note that $\mathbf{Run}_M(s,(G,w))$ iff $M\Downarrow^s_w G$, by Proposition 1.

By Lemma 19, $\mathbf{Run}_M$ is the graph of a partial function from traces to pairs of a generalized value and a real number. The total function $f:\mathbb{S}\to\mathcal{GV}\times\mathbb{R}_+$ defined by

$$f(s)=\begin{cases}\mathbf{Run}_M(s) & \text{if defined}\\ (\mathtt{fail},0) & \text{otherwise}\end{cases}$$

is measurable, since every function from one complete, separable metric space to another, with a Borel-measurable graph is Borel-measurable (Lemma 53). Note that the graph of the second case is the set $(\mathbb{S}\setminus\mathcal{T}^M_{red})\times\{(\mathtt{fail},0)\}$, where $\mathcal{T}^M_{red}=\{s\mid(M,s)\in\mathcal{T}_{red}\}$. The set $\mathcal{T}^M_{red}$ is measurable (Lemma 18.1(i) from Billingsley), so the second graph is measurable. Since $f(s)=(\mathbf{P}_M(s),\mathbf{O}_M(s))$, the two functions $\mathbf{P}_M$ and $\mathbf{O}_M$ are measurable.

We need the following results to show that `peval` function is deterministic:

Lemma 75 *If $(M, w, s) \Rightarrow (M', w', s)$ and $(M, w, s) \Rightarrow (M'', w'', s)$ and there is no N, such that $M' \to N$ or $M'' \to N$, then $M'' = M'$ and $w'' = w'$.*

Proof. By induction on the number of steps in the derivation of $(M, w, s) \Rightarrow (M', w', s)$:
- If $(M, w, s) \Rightarrow (M', w', s)$ was derived in 0 steps, then $M' = M$ and $w' = w$. Because there is no N such that $M \to N$ and only (Red Pure) can reduce an expression without changing the trace, $(M, w, s) \Rightarrow (M'', w'', s)$ also must have been derived in 0 steps, so $M'' = M = M'$ and $w'' = w = w'$.
- If $(M, w, s) \Rightarrow (M', w', s)$ was derived in 1 or more steps, then $(M, w, s) \to (M^*, w^*, s) \Rightarrow (M', w', s)$ and, since M reduces deterministically, $(M, w, s) \to (\hat{M}, \hat{w}, s) \Rightarrow (M'', w'', s)$.
  By Lemma 17, $\hat{M} = M^*$ and $\hat{w} = w^*$. Hence,by induction hypothesis, $M'' = M'$ and $w'' = w'$.

Lemma 76 *If $(M, 1, []) \Rightarrow (M', w, s)$ and $(M, 1, []) \Rightarrow (M'', w', s)$ and there is no N, such that $M' \to N$ or $M'' \to N$, then $M' = M''$ and $w = w'$.*

Proof. By induction on the length of s:
- If $s = []$, then the result follows by Lemma 75
- If $s = s'@[c]$, then because only one element can be added to a trace in one step, there must be some M^*, w^*, $\hat{M}$, $\hat{w}$ such that $(M, 1, []) \Rightarrow (M^*, w^*, s') \to (M^{**}, w^{**}, s) \Rightarrow (M', w, s)$ and $(M, 1, []) \Rightarrow (\hat{M}, \hat{w}, s') \to (\hat{M}', \hat{w}', s) \Rightarrow (M'', w', s)$.
  Since only the (Red Random) rule can add an element to the trace, we have $M^* = E[\mathsf{D}(\vec{c})]$ and $\hat{M} = E[\mathsf{D}(\vec{c})]$, which implies that M^* and $\hat{M}$ do not reduce deterministically.
  Thus, by induction hypothesis, $\hat{M} = M^*$ and $\hat{w} = w^*$, and by Lemma 17, $\hat{M}' = M^{**}$ and $\hat{w}' = w^{**}$. Then, Lemma 75 gives $M'' = M'$ and $w' = w$, as required.

Lemma 77 *$\Lambda \times \mathbb{S}^+$ with Manhattan product metric is complete and separable.*

Let $M \Rightarrow N$ be the reflexive and transitive closure or $M \to N$.

Lemma 78 *If $M \Rightarrow N$ and $M \Rightarrow N'$ and $N, N' \notin P_{det}$, then $N = N'$*

Proof. By induction on the number of steps in derivation of $M \Rightarrow N$, with appeal to lemma 17.

Lemma 79 *The relation $M \Rightarrow N$ is measurable.*

Proof. Similar to the proof of measurability of $(M, w, s) \Rightarrow (M', w', s')$.

Let us write $M \Longrightarrow N$ if $M \Rightarrow N$ and $N \notin P_{det}$

Lemma 80 *The relation $M \Longrightarrow N$ is measurable.*

Proof. Similar to the proof of measurability of $(M, 1, []) \Rightarrow (G, w, s)$.

Let $\mathcal{P}^*_{det} = \{M \mid M \Longrightarrow N\}$

Lemma 81 *$\mathcal{P}^*_{det}$ is measurable.*

Proof. Similar to the proof of Lemma 74.

Restatement of Lemma 43 `peval` *is a measurable function $(\Lambda \times \mathbb{S}) \to \Lambda$.*

Proof. We can represent `peval` as

$$\texttt{peval}(M, s) = (p \circ F')(M, s)$$

where

$$\begin{aligned} p &: \mathcal{T} \to \Lambda \\ p(M, w, s) &= M \\ p(\bot) &= \texttt{fail} \end{aligned}$$

$$\begin{aligned} F &: \Lambda \times \mathbb{R} \times \mathbb{S} \to \mathcal{T} \\ F(M, w, s) &= (g'_r \circ g'_d)^{|s|}(M, w, s) \end{aligned}$$

$$\begin{aligned} e &: \Lambda \times \mathbb{S} \to \Lambda \times \mathbb{R} \times \mathbb{S} \\ e(M, s) &= (M, 1, s) \end{aligned}$$

$$\begin{aligned} F' &: \Lambda \times \mathbb{S} \to \mathcal{T} \\ F'(M, s) &= (F \circ e)(M, s) \end{aligned}$$

$$\begin{aligned} h'_d &: \Lambda \to \Lambda \cup \{\bot\} \\ h'_d(M) &= \begin{cases} N & \text{if } M \Longrightarrow N \\ \bot & \text{otherwise} \end{cases} \end{aligned}$$

$$\begin{aligned} h''_d &: P^*_{det} \to \Lambda \\ h''_d(M) &= N \quad \text{where } M \Longrightarrow N \end{aligned}$$

$$\begin{aligned} g'_d &: \mathcal{T} \to \mathcal{T} \\ g'_d(M, w, s) &= \begin{cases} (h''_d(M), w, s) & \text{if } M \in P^*_{det} \\ \bot & \text{otherwise} \end{cases} \\ g'_d(\bot) &= \bot \end{aligned}$$

$$\begin{aligned} g'_r &: \mathcal{T} \to \mathcal{T} \\ g'_r(M, w, s) &= \begin{cases} g_r(M, w, s) & \text{if } (M, w, s) \in \mathcal{T}_{rnd} \\ \bot & \text{otherwise} \end{cases} \\ g'_r(\bot) &= \bot \end{aligned}$$

The graph of h'_d is $\Longrightarrow \cup\, (\Lambda \setminus P^*_{det}) \times \{\bot\}$, which is measurable. Thus, h'_d is measurable by Lemma 53, and so h''_d, being a restriction of h'_d to the measurable set P^*_{det}, is measurable by Lemma 52.

For every $A \in \Lambda$, $R \in \mathbb{R}$ and $S \in \mathbb{S}$, we have $g'^{-1}_d(A \times R \times S) = (h''^{-1}_d(A)) \times R \times S$. which is measurable since h''_d is measurable. Meanwhile, $g'^{-1}_d(\bot) = \bot \cup (\Lambda \setminus P^*_{det}) \times \mathbb{R} \times \mathbb{S}$, which is also measurable. Hence, the preimage of g'_d is measurable for all elements of the set generating $\mathcal{T}$, so g'_d is measurable.

The measurability of g'_r is obvious, since we have already shown that $\mathcal{T}_{rnd}$ is measurable and that g_r is measurable. Because composition of measurable functions is measurable, we can

conclude that $(g'_r \circ g'_d)$ is measurable, and that $(g'_r \circ g'_d)^n$ is measurable for all n. Thus, for every nonnegative integer n, the restriction of F to triples in which the trace s has length n is measurable. Because each such set of arguments is closed (and so Borel-measurable), Lemma 52 says that F is measurable.

The function e is continuous, and so measurable.

F' is a composition of two measurable functions, and so measurable.

The function p is clearly measurable, as it is continuous on $\Lambda \times \mathbb{R} \times \mathbb{S}$ and constant on the one-element set $\{\bot\}$.

Hence, `peval` is measurable as a composition of two measurable functions.

Restatement of Lemma 44 *For any program M, the transition density $q(\cdot,\cdot) : (\mathbb{S} \times \mathbb{S}) \to \mathbb{R}_+$ is measurable.*

Proof. Outline:

1. q restricted to $\mathbb{S} \times (\mathbb{S} \setminus \{[]\})$ is measurable, since it is composed from $\mathrm{pdf}_{\mathsf{Gaussian}}$ (which is continuous) and `peval`(which is measurable).

2. $\int_{\mathbb{S}\setminus\{[]\}} q(s, dt) \leq 1$ for all s, by the same property for $\mathrm{pdf}_{\mathsf{Gaussian}}$ and $[\![M]\!]^{\mathbb{S}}_{\Downarrow}$.

Restatement of Lemma 45 *The function Q is a probability kernel on $(\mathbb{S}, \mathcal{S})$.*

Proof. We need to verify the two properties of probability kernels:

1. For every $s \in \mathbb{S}$, $Q(s,\cdot)$ is a probability measure on $\mathbb{S}$. Since for every $s \in \mathbb{S}$, $q(s,\cdot)$ is non-negative measurable $\mathcal{S}$ (by [1, Theorem 18.1]), $Q(s,B) = \int_B q(s,y)\mu(dy)$ (as a function of B) is a well-defined measure for all $s \in \mathbb{S}$. Finally, $Q(s,\mathbb{S}) = Q(s,[]) + Q(s,\mathbb{S} \setminus \{[]\}) = 1$.

2. For every $B \in \mathcal{S}$, $Q(\cdot,B)$ is a non-negative measurable function on $\mathbb{S}$: Since $(\mathbb{S}, \mathcal{S}, \mu)$ is a σ-finite measure space, $q(\cdot,\cdot)$ is non-negative and measurable $\mathcal{S} \times \mathcal{S}$ and $Q(s,B) = \int_B q(s,y)\mu(ds)$, this follows from [1, Theorem 18.3]. □

References

[1] P. Billingsley. *Probability and Measure*. Wiley-Interscience, third edition, 1995.

[2] J. J. Buckley. Graphs of measurable functions. *Proceedings of the American Mathematical Society*, 44(1):78–80, 1974.

[3] P. Cousot and M. Monerau. Probabilistic abstract interpretation. In *Proeedings of ESOP 2012*, volume 7211 of *LNCS*, pages 166–190. Springer, 2012.

[4] V. Danos and T. Ehrhard. Probabilistic coherence spaces as a model of higher-order probabilistic computation. *Information and Computation*, 209(6):966–991, 2011.

[5] V. Danos and R. Harmer. Probabilistic game semantics. *ACM Transactions on Computational Logic*, 3(3):359–382, 2002.

[6] T. Ehrhard, C. Tasson, and M. Pagani. Probabilistic coherence spaces are fully abstract for probabilistic PCF. In *Proceedings of POPL 2014*, pages 309–320, 2014.

[7] A. Gelman, D. Lee, and J. Guo. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015. doi: 10.3102/1076998615606113. URL `http://jeb.sagepub.com/content/40/5/530.abstract`.

[8] N. D. Goodman and A. Stuhlmüller. The design and implementation of probabilistic programming languages. `http://dippl.org`, 2014.

[9] N. D. Goodman and J. B. Tenenbaum. Probabilistic models of cognition. http://probmods.org, 2014.

[10] N. D. Goodman, V. K. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. In D. A. McAllester and P. Myllymäki, editors, *Proceedings of UAI 2008*, pages 220–229. AUAI Press.

[11] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. doi: 10.1093/biomet/57.1.97.

[12] C. Hur, A. V. Nori, S. K. Rajamani, S. Samuel, and D. Vijaykeerthy. Implementing a correct sampler for imperative probabilistic programs. In *Proc. FSTTCS '15: Foundations of Software Technology and Theoretical Computer Science*, 2015. To appear.

[13] C. Jones and G. D. Plotkin. A probabilistic powerdomain of evaluations. In *Proceedings of LICS 1989*, pages 186–195, 1989.

[14] D. Kozen. Semantics of probabilistic programs. In *20th Annual Symposium on Foundations of Computer Science, San Juan, Puerto Rico, 29-31 October 1979*, pages 101–114. IEEE Computer Society, 1979. doi: 10.1109/SFCS.1979.38. URL http://dx.doi.org/10.1109/SFCS.1979.38.

[15] V. K. Mansinghka, D. Selsam, and Y. N. Perov. Venture: a higher-order probabilistic programming platform with programmable inference. *CoRR*, abs/1404.0099, 2014. URL http://arxiv.org/abs/1404.0099.

[16] A. V. Nori, C. Hur, S. K. Rajamani, and S. Samuel. R2: an efficient MCMC sampler for probabilistic programs. In C. E. Brodley and P. Stone, editors, *Proceedings of AAAI 2014.*, pages 2476–2482. AAAI Press, 2014.

[17] P. Panangaden. The category of Markov kernels. *ENTCS*, 22:171–187, 1999. In proceedings of PROBMIV 1998.

[18] S. Park, F. Pfenning, and S. Thrun. A probabilistic language based upon sampling functions. In J. Palsberg and M. Abadi, editors, *Proceedings of the 32nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL 2005, Long Beach, California, USA, January 12-14, 2005*, pages 171–182. ACM, 2005.

[19] A. Scibior, Z. Ghahramani, and A. D. Gordon. Practical probabilistic programming with monads. In B. Lippmeier, editor, *Proceedings of the 8th ACM SIGPLAN Symposium on Haskell, Haskell 2015, Vancouver, BC, Canada, September 3-4, 2015*, pages 165–176. ACM, 2015. doi: 10.1145/2804302.2804317. URL http://doi.acm.org/10.1145/2804302.2804317.

[20] L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994. doi: 10.1214/aos/1176325750. URL http://dx.doi.org/10.1214/aos/1176325750.

[21] D. Tolpin, J. van de Meent, and F. Wood. Probabilistic programming in Anglican. In A. Bifet, M. May, B. Zadrozny, R. Gavaldà, D. Pedreschi, F. Bonchi, J. S. Cardoso, and M. Spiliopoulou, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*, volume 9286 of *Lecture Notes in Computer Science*, pages 308–311. Springer, 2015. doi: 10.1007/978-3-319-23461-8_36. URL http://dx.doi.org/10.1007/978-3-319-23461-8_36.

[22] N. Toronto, J. McCarthy, and D. V. Horn. Running probabilistic programs backwards. In J. Vitek, editor, *Proceedings of ESOP 2015*, volume 9032 of *LNCS*, pages 53–79. Springer, 2015.

[23] D. Wingate, A. Stuhlmueller, and N. D. Goodman. Lightweight implementations of probabilistic programming languages via transformational compilation. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, page 131, 2011.