Constructive probabilistic semantics with non-spatial locales

Benjamin Sherman
MIT CSAIL
sherman@csail.mit.edu

Jared Tramontano
MIT CSAIL
tramo@mit.edu

Michael Carbin
MIT CSAIL
mcarbin@csail.mit.edu

1 Introduction

Ideally, a probabilistic programming language should admit a computable semantics, but languages often provide operators that denote uncomputable functions. While the use of these uncomputable operators may result in uncomputable programs, a programmer can productively use these operators and still produce a computable program.

For instance, consider the TrueSkill model of Fig. 1, in which three players with compete against each other in a round robin, with their performance randomly varying around their baseline skill. The TrueSkill program returns a random Boolean value (of type $\mathcal{R}(\mathbb{B})$, where $\mathbb{B}$ is the space of boolean values, true and false) that indicates what observations were made: Alice beat Bob in game 1, Bob beat Cyd in game 2, and Alice beat Cyd in game 3. Used in this program are a normal distribution function $\mathcal{N} : \mathbb{R} \times \{x : \mathbb{R} \mid x > 0\} \to \mathcal{R}(\mathbb{R})$ that takes mean and standard deviation parameters as arguments, as well as a Boolean-valued comparison function $(>) : \mathbb{R} \times \mathbb{R} \to \mathbb{B}$. TrueSkill is computable, meaning that the probability the return value is true can be computed to any finite precision within finite time. But its component computations are not computable: the comparison function $(>)$ is not even continuous[1] [2]! TrueSkill's computability, therefore, cannot be verified compositionally.

To better understand why $(>)$ is innocuous in TrueSkill, consider the simpler examples in Fig. 2. There, $\text{pos}_{\mathcal{N}}$ is computable (it's essentially the normal CDF), whereas pos_{δ} is uncomputable (which follows from uncomputability of comparison of a real number with 0). In this paper, we use locale theory to define a probabilistic programming semantics that admits computable programs (such as TrueSkill), rejects uncomputable programs (such as pos_{δ}), and for which computability is compositional.

The key idea is that though comparison with 0, $(\cdot > 0) : \mathbb{R} \to \mathbb{B}$, is not continuous, it is continuous when restricted to the domain $\{x : \mathbb{R} \mid x \neq 0\}$. Because the output of a normal distribution as used in $\text{pos}_{\mathcal{N}}$ lies in this domain with probability 1, we can type the operations as $\mathcal{N} : \mathbb{R} \times \{x : \mathbb{R} \mid x > 0\} \to \mathcal{R}(\{x : \mathbb{R} \mid x \neq 0\})$ and $(\cdot > 0) : \{x : \mathbb{R} \mid x \neq 0\} \to \mathbb{B}$, such that both functions are continuous.

The subspace $\{x : \mathbb{R} \mid x \neq 0\}$ is just one of many subspaces of $\mathbb{R}$ that has probability 1 under a normal distribution. In particular, for any $z : \mathbb{R}$, there is the subspace $\{x : \mathbb{R} \mid x \neq z\}$ excluding that point. It is equally valid to declare that $\mathcal{N}$ should output to any of its probability-1 subspaces, or better yet, it should output a value that is in *all* of these subspaces at once, i.e., their intersection. This intersection must be empty in point-based theories such as measure theory and classical topology, but not in *locale theory*, where the intersection has exactly the desired structure (though it still has no global points). Such a locale without an analogue in classical topology is called *non-spatial*.

[1] All computable functions are continuous.
[2] Since $\mathbb{R} \times \mathbb{R}$ is connected and $\mathbb{B}$ is discrete, there are no nontrivial continuous maps from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{B}$.

PPS'18, January 9, 2018, Los Angeles, CA, USA
2018. ACM ISBN …$15.00
https://doi.org/

$$
\begin{aligned}
&\text{alice\_baseline} \leftarrow \mathcal{N}(0, 20);\\
&\text{bob\_baseline} \leftarrow \mathcal{N}(0, 20);\\
&\text{cyd\_baseline} \leftarrow \mathcal{N}(0, 20);\\
&\text{alice}_1 \leftarrow \mathcal{N}(\text{alice\_baseline}, 1);\\
&\text{bob}_1 \leftarrow \mathcal{N}(\text{bob\_baseline}, 1);\\
&\text{bob}_2 \leftarrow \mathcal{N}(\text{bob\_baseline}, 1);\\
&\text{cyd}_2 \leftarrow \mathcal{N}(\text{cyd\_baseline}, 1);\\
&\text{alice}_3 \leftarrow \mathcal{N}(\text{alice\_baseline}, 1);\\
&\text{cyd}_3 \leftarrow \mathcal{N}(\text{cyd\_baseline}, 1);\\
&\text{return}(\text{alice}_1 > \text{bob}_1 \;\&\&\; \text{bob}_2 > \text{cyd}_2 \;\&\&\; \text{alice}_3 > \text{cyd}_3)
\end{aligned}
$$

Figure 1. A probabilistic program of type $\mathcal{R}(\mathbb{B})$ representing a TrueSkill model.

$$
\begin{aligned}
&\text{pos}_{\mathcal{N}}(x) \triangleq y \leftarrow \mathcal{N}(x, 1); && \text{pos}_{\delta}(x) \triangleq y \leftarrow \delta(x);\\
&\qquad \text{return}(y > 0) && \qquad \text{return}(y > 0)
\end{aligned}
$$

Figure 2. Two probabilistic programs of type $\mathbb{R} \to \mathcal{R}(\mathbb{B})$. $\text{pos}_{\mathcal{N}}$ is computable but not pos_{δ}. δ represents the Dirac delta.

Nontermination. This technique can similarly be used to handle functions that terminate almost surely but not always without resorting to explicit partiality (which would come at the cost of significant complication to the mathematical interpretation of programs as probability distributions, as well as to the computability of their properties). For instance, a developer may want to define the geometric distribution with probability 1/2, $\text{geometric} : \mathcal{R}(\mathbb{N})$ and have access to the coinflips : $\mathcal{R}(\text{Stream}(\mathbb{B}))$ distribution, which provides an infinite stream of random independent Boolean values. Intuitively, they can define the geometric distribution by flipping coins until it lands on heads, and counting the number of coinflips it took to get the heads. That is, they wish to define a deterministic function $\text{wait\_for\_heads} : \text{Stream}(\mathbb{B}) \to \mathbb{N}$ that computes the first true value in a sequence of booleans, and apply this to the coinflips distribution.

Unfortunately, wait_for_heads is partial, as it is undefined on the input sequence that is always false, even though this sequence occurs with probability 0 under coinflips. But the same solution applies: if we consider coinflips to return a value in the intersection of all its probability-1 subspaces, then wait_for_heads will be total on that domain.

Contributions. Locale theory enables us to understand these example programs as total and continuous (where semantic frameworks based on classical topology or measure theory would not). We propose locale theory, and particularly non-spatial sublocales,

as a semantic framework for probabilistic programming. The advantages of this framework are:

- **Totality**: There is no need to use partiality to understand programs that almost surely terminate.
- **Continuity**: Programs that are only almost surely continuous in classical topology are now continuous.
- **Computability**: Constructivity makes the semantics computable in a strong sense: the probability that a program returning a random $\mathbb{B}$ returns true can be evaluated to any precision in finite time (unlike in measure-theoretic semantics).
- **Disintegration**: Non-spatial locales afford a notion of disintegration that provides stronger uniqueness guarantees than in measure theory and stronger existence guarantees than for continuous disintegrations in classical topology.

2 Locale-theoretic probabilistic semantics

A *locale* A is a lattice $\mathcal{O}(A)$ of *opens* that has finitary meets and infinitary (arbitrary/small) joins (i.e., a frame). A *continuous map* $f : A \to_c B$ is a map of opens $f^{-1} : \mathcal{O}(B) \to \mathcal{O}(A)$ that preserves finitary meets and infinitary joins. This defines a category **Loc**. **Loc** has a terminal object $*$, and a (global) *point* of a locale A is a continuous map $* \to_c A$. Locales may be specified by giving a set S of *basic opens* and an inductive family of *covering axioms* of the form $a \leq \bigvee_{i:I} b_i$ on those basic opens $a, b_i : S$. Such locales are called *inductively generated formal spaces* [2], which form a full subcategory $\mathbf{Loc_i}$ of **Loc** that is cartesian monoidal even in a predicative setting[3]. A strong monad $\mathcal{R}$ on **Loc** maps any locale to the locale of its probability distributions [9, 10], and restricts to a strong monad on $\mathbf{Loc_i}$. Thus $\mathbf{Loc_i}$ and $\mathcal{R}$ have sufficient structure to give constructive and predicative categorical semantics for probabilistic programming languages.

Sublocales of a given inductively generated formal space are formed by adding additional covering axioms. For instance, for given an open $U : \mathcal{O}(A)$, adding the axiom $\top \leq U$ forms the open sublocale corresponding to U, while adding the axiom $U \leq \bot$ forms the closed sublocale corresponding to the set-theoretic complement of U [4]. The intersection of a family of sublocales is produced by adding all of the covering axioms from each sublocale in the family.

Random locales. Given a probability distribution $\mu : * \to_c \mathcal{R}(A)$ (where $*$ is the terminal object), one can form the locale $\mathsf{Ran}(\mu)$ that is the intersection of its probability-1 open sublocales [7][5], with the inclusion $\iota_\mu : \mathsf{Ran}(\mu) \to_c A$. $\mathsf{Ran}(\mu)$ is inductively generated when A is. There is a distribution $\mu_\restriction : * \to_c \mathcal{R}(\mathsf{Ran}(\mu))$ such that $\iota_{\mu *}\mu_\restriction = \mu$, where $\cdot_*$ denotes the pushforward, i.e., mapping ι_μ over $\mu_\restriction$. If the locale A is *fitted*, i.e., every sublocale is the intersection of all the open sublocales containing it, then the sublocale $\mathsf{Ran}(\mu)$ is in fact the smallest measure-1 sublocale. Both $\mathbb{R}$ and $\mathsf{Stream}(\mathbb{B})$ are fitted.

A probability distribution $\mu : * \to_c \mathcal{R}(A)$ is called *random* if whenever $\mu(U) = 1$, then $\top \leq U$ (we also call A random w.r.t. μ) ([7] definition 6.4). Clearly, any $\mu_\restriction : * \to_c \mathcal{R}(\mathsf{Ran}(\mu))$ is random. A distribution $\mu : * \to_c \mathcal{R}(A)$ is random iff A is homeomorphic to $\mathsf{Ran}(\mu)$.

Totality and continuity. Random sublocales give a canonical way to turn almost-surely terminating functions into total ones and almost-everywhere continuous functions into continuous ones. In the following theorems, suppose we have a probability distribution $\mu : * \to_c \mathcal{R}(A)$ on a fitted locale A.

Partiality can be modeled with a monad $\cdot_\bot$ on **Loc** [8, Ch. 10]. An open inclusion $\mathsf{up} : B \to_c B_\bot$ embeds the original locale, and in addition there is a closed point $\bot : * \to_c B_\bot$ that represents nontermination[6].

Theorem 2.1 (Totality). *Given a map $f : A \to_c B_\bot$ such that $f_*\mu(\mathsf{up}(B)) = 1$ (where $f_*\mu$ denotes the pushforward of μ by f), there is a map $g : \mathsf{Ran}(\mu) \to_c B$ such that $f \circ \iota_\mu = \mathsf{up} \circ g$, and thus $f_*\mu = (\mathsf{up} \circ g)_*\mu_\restriction$.*

Since locale theory is "point-free," it does not provide a natural setting for describing (potentially discontinuous) functions on sets. Accordingly, we define those *almost everywhere continuous* functions from A to B with respect to $\mu : * \to_c \mathcal{R}(A)$ as continuous maps $f : A' \to_c B$ on some sublocale A' of A satisfying $\mu(A') = 1$ [7][8].

Theorem 2.2 (Continuity). *Given an almost-everywhere continuous function f from A to B defined on some sublocale A' of A satisfying $\mu(A') = 1$, there is a map $g : \mathsf{Ran}(\mu) \to_c B$ such that $f \circ \iota = g$, and thus $f_*\mu_{|A'} = g_*\mu_\restriction$, where $\iota : \mathsf{Ran}(\mu) \to_c A'$ is the inclusion of $\mathsf{Ran}(\mu)$ in A'.*

Computability. Given a point $x : * \to_c A$ of a locale A and an open cover $\top \leq \bigvee_{i:I} U_i$ of A, there (constructively) exists some $i^* : I$ such that x lies in U_{i^*}. This is the fundamental computational content of the locale-theoretic semantics[9]. For instance, for every $\varepsilon > 0$ there is a cover of $\mathbb{R}$ by open balls with rational centers and radius ε, allowing approximation of any real number to within ε. Since there is a continuous map from $\mathcal{R}(\mathbb{B})$ to $\mathbb{R}$ indicating the probability the Boolean is true, it is possible to compute the probability that a random Boolean value is true (and hence, the probability that any decidable predicate holds).

Locales A without points, such as $\mathsf{Ran}(\mathcal{N})$ (where $\mathcal{N}$ is any normal distribution — their random sublocales are the same), still have

[3] **Loc** is not closed, so it can only interpret first-order languages. Presumably, to be able to interpret higher-order functions, it could be embedded in a larger category that is cartesian closed, such as the gros topos of sheaves over the site generated by the open cover topology [3, 4].

[4] Since for any open V we have $\bot \leq V \leq \top$, these two axioms imply, respectively, that $U = \top$ and $U = \bot$.

[5] Simpson [7] for the most part uses σ-locales rather than locales, whose lattice of opens has countable joins rather than arbitrary joins. Contrary to locales, much theory related to σ-locales requires appealing to classical mathematics. Classically, every measurable space determines a σ-locale that is usually not a locale, and measurable functions determine σ-continuous maps of σ-locales. For strongly Lindelöf spaces such as $\mathbb{R}$ and $\mathsf{Stream}(\mathbb{B})$, the two notions coincide. We choose to work with locales rather than σ-locales because they can be formulated constructively and hence have stronger computational content.

[6] A brief description of the partiality construction: if A is an inductively generated formal space, so is $A_\bot$. Its basic opens are those of A together with a new basic open that represents the new top element. All covering axioms of A are added as corresponding axioms to $A_\bot$, but the new top open has no covering axioms.

[7] Simpson [7] defines outer measure for sublocales of fitted locales.

[8] Functions on sets (not necessarily continuous) can be represented in locale theory via the adjunction $\mathsf{Discrete} \dashv \mathsf{Pt}$ between **Set** and **Loc**. One might be tempted to call a function $f : \mathsf{Pt}(A) \to \mathsf{Pt}(B)$ almost everywhere continuous if there is a sublocale $\iota_{A'} : A' \to_c A$ such that $\mu(A') = 1$ and there is some $g : A' \to_c B$ such that $f \circ \mathsf{Pt}(\iota_{A'}) = \mathsf{Pt}(g)$. Any function that is almost everywhere continuous in this sense corresponds to one that is almost everywhere continuous in the other sense,

[9] For σ-locales, the analogous statement fails to hold constructively for σ-locales of interest. For instance, classically, every measurable space defines a corresponding σ-locale. In particular, the Borel σ-algebra on $\mathbb{R}$ classically gives a σ-locale with the cover $\top \leq \{x : \mathbb{R} \mid x \leq 0\} \vee \{x : \mathbb{R} \mid x > 0\}$ but the ability to determine which subset holds of an arbitrary real number implies the Limited Principle of Omniscience (LPO), which is a "constructive taboo."

computational content indirectly, as their derived locales $\mathcal{R}(A)$ or $\mathcal{P}^+_\Diamond(A)$, the locale of their nondeterministic values, may have points, as is the case for $\mathsf{Ran}(\mathcal{N})$. For instance, for any probability distribution $\mu : \mathcal{R}(\mathsf{Ran}(\mathcal{N}))$, we can compute the probability that μ is larger than 0 by mapping a function $(\cdot > 0) : \mathsf{Ran}(\mathcal{N}) \to_c \mathbb{B}$ over it to produce a probability distribution on $\mathbb{B}$. One can view randomly sampling from a locale A as producing a point of the nondeterministic powerlocale $\mathcal{P}^+_\Diamond(A)$, and accordingly, one can for instance run deterministic computations $A \to_c B$ on these random samples to produce random samples from B, even though A has no points.

3 Conditional probability and disintegration

Statistical inference can usually be phrased as the problem of finding *disintegrations*, making the question of their existence and uniqueness important. A disintegration of a probability distribution $\mu : * \to_c \mathcal{R}(A \times B)$ is a decomposition into the distribution $\mathsf{fst}_*\mu : * \to_c \mathcal{R}(A)$ and a conditional distribution $f : A \to_c \mathcal{R}(B)$ such that $\mu = \int (\mathrm{id} \times f) d(\mathsf{fst}_*\mu)$.

Disintegrations exist (constructively) when A is discrete, but there are also instances where they do not exist, even classically. For instance, the distribution on $*_\bot \times *_\bot$ that has half its mass on the point $(\mathsf{up}(\mathsf{tt}), \bot)$ and its other half on the point $(\bot, \mathsf{up}(\mathsf{tt}))$ cannot be disintegrated.

With random sublocales, it becomes particularly interesting to consider disintegrations of the above form where $\mathsf{fst}_*\mu$ is random, or in particular a random sublocale A' of a larger locale A. Using the random sublocale A' makes both existence and uniqueness of the conditional distribution $f : A' \to_c \mathcal{R}(B)$ *easier* in comparison to A, both for the same reason: any $f : A \to_c \mathcal{R}(B)$ can be restricted to the domain A'.

Classically, disintegrations are only unique up to measure 1, which Shan and Ramsey [6] find problematic:

> The definition of disintegration allows latitude that our disintegrator does not take: When we disintegrate $\xi = \Lambda \otimes \kappa$, the output κ is unique only almost everywhere–κx may return an arbitrary measure at, for example, any finite set of x's. But our disintegrator never invents an arbitrary measure at any point. The mathematical definition of disintegration is therefore a bit too loose to describe what our disintegrator actually does. How to describe our disintegrator by a tighter class of "well-behaved disintegrations" is a question for future research. In particular, the notion of continuous disintegrations (Ackerman, Freer, and Roy 2016) is too tight, because depending on the input term, our disintegrator does not always return a continuous disintegration, even if one exists.

Locale theory's notion of continuity is an improvement on Ackerman, Freer, and Roy [1] in this regard: it is no longer too strict to require a continuous disintegration on the random sublocale, which is a much smaller sublocale than is possible classically. As an example, consider a distribution $\mu : * \to_c \mathcal{R}(\mathsf{Ran}(\mathcal{N}) \times \mathbb{B})$ which is a standard normal distribution on the $\mathsf{Ran}(\mathcal{N})$ part and the $\mathbb{B}$ part is true if the $\mathsf{Ran}(\mathcal{N})$ component is larger than 0 and false if it's smaller than 0 (since there's no mass at exactly 0, this specifies the distribution completely). Suppose we want to condition on the $\mathsf{Ran}(\mathcal{N})$ component. The conditional distribution would be considered discontinuous according to Ackerman, Freer, and Roy [1] (who would necessarily use $\mathbb{R}$ in place of $\mathsf{Ran}(\mathcal{N})$) because both $\delta(\mathsf{true})$ and $\delta(\mathsf{false})$ (where δ is the Dirac delta) are in the image of any neighborhood of 0. However, it would be continuous on $\mathsf{Ran}(\mathcal{N})$, defined by the continuous map

$$f : \mathsf{Ran}(\mathcal{N}) \to_c \mathcal{R}(\mathbb{B})$$
$$f(x) \triangleq \delta(x > 0),$$

where $(\cdot > 0) : \mathsf{Ran}(\mathcal{N}) \to_c \mathbb{B}$ *is* continuous due to its smaller domain.

It is desirable to be able to extend the domain of a conditional distribution from a random sublocale to some larger sublocale in a unique manner. For instance, in the previous example, the conditional distribution can be extended in a unique way from $\mathsf{Ran}(\mathcal{N})$ to $\{x : \mathbb{R} \mid x \neq 0\}$. This follows from several facts:

- Given $f : S \to_c B$, where S is a sublocale of A, then continuous extensions of f to A are unique if S is dense in A and B is Hausdorff.
- If a locale X is *regular*, meaning that every open U of X is a union of opens O such that $\top \leq U \vee \neg O$, then $\mathcal{R}(X)$ is Hausdorff.
- $\mathbb{B}$ is regular.
- $\mathsf{Ran}(\mathcal{N})$ is dense in $\{x : \mathbb{R} \mid x \neq 0\}$.

4 Related Work

Jones [5] gives semantics of probabilistic programs with valuations on CPOs rather than locales in general. The probability that a program terminates is a lower real number, which is only lower semi-computable. Faissole and Spitters [3] propose using valuations in the context of synthetic topology to give meaning to probabilistic programs. Similarly, properties are only lower-semicomputable in this framework. Synthetic topology is only capable of representing spatial locales, so most random locales of interest cannot be represented. Their technique of constructing structures in the internal language of a topos allows their notions to be naturally extended to a higher-order setting. We imagine that our framework could similarly be extended to a higher-order setting by using the gros topos of sheaves over formal spaces with the open cover topology.

Simpson [7] elucidates how a valuation on locales can be restricted to a smallest sublocale of measure 1, but does not consider continuous maps on them.

References

[1] Nathanael L Ackerman, Cameron E Freer, and Daniel M Roy. 2016. On computability and disintegration. *Mathematical Structures in Computer Science* (2016), 1–28.

[2] Thierry Coquand, Giovanni Sambin, Jan Smith, and Silvio Valentini. 2003. Inductively generated formal topologies. *Annals of Pure and Applied Logic* 124, 1 (2003), 71–106.

[3] Florian Faissole and Bas Spitters. 2017. Synthetic topology in homotopy type theory for probabilistic programming. In *Probabilistic Programming Semantics*.

[4] Michael P Fourman. 1984. Continuous Truth I: Non-constructive Objects. *Studies in Logic and the Foundations of Mathematics* 112 (1984), 161–180.

[5] Claire Jones. 1990. Probabilistic non-determinism. (1990).

[6] Chung-chieh Shan and Norman Ramsey. 2017. Exact Bayesian inference by symbolic disintegration. In *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*. ACM, 130–144.

[7] Alex Simpson. 2012. Measure, randomness and sublocales. *Annals of Pure and Applied Logic* 163, 11 (2012), 1642–1659.

[8] Steven Vickers. 1989. *Topology via logic*. Cambridge University Press.

[9] Steven Vickers. 2008. A localic theory of lower and upper integrals. *Mathematical Logic Quarterly* 54, 1 (2008), 109–123.

[10] Steven Vickers. 2011. A monad of valuation locales. *Preprint at http://www.cs.bham.ac.uk/~sjv/Riesz.pdf* (2011).