A ProbProg Language Taxonomy

Ohad Kammar

36th International Conference on
Mathematical Foundations
of
Programming Semantics
(MFPS'20)

THE UNIVERSITY of EDINBURGH informatics lfcs Laboratory for Foundations of Computer Science BayesCentre

supported by: THE ROYAL SOCIETY The Alan Turing Institute Facebook Research

Probabilistic programming languages everywhere

Languages: Anglican, BLOGS, BUGS, Church, Edward, Gen, HackPPL, Hakaru, Monad-Bayes, Pyro, Stan, Turing, Venture, WebPPL, ...

Semantics: Boolean-valued models, Banach spaces, probabilistic coherence spaces, event structures, s-finite kernels ...

Communities: MFPS, LICS, POPL, NeurIPS, ICFP, PLDI, ...

How to organise ProbProg Languages?

Proposed Taxonomy:

Sampling \ Conditioning	density	distribution
graded	Stan, Pyro	Hakaru, Trace-Typed Gen
non-graded	Anglican, Monad-Bayes	?

Talk structure

ProbProg Basics

Dataset: (x, y)

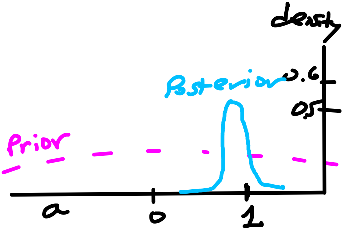
Model $\mathcal{M}$ (Bayesian Linear Regression)

$y = a \cdot x$

$$a \sim \mathcal{N}(0, 2)$$

$(1, 1.1)$	$1.1 \sim \mathcal{N}(a \times 1, \frac{1}{4})$
$(2, 1.9)$	$1.9 \sim \mathcal{N}(a \times 2, \frac{1}{4})$
$(3, 2.7)$	$2.7 \sim \mathcal{N}(a \times 3, \frac{1}{4})$

$$P_{a \sim [\![\mathcal{M}]\!]}[l \le a \le u] \propto \int_l^u dx\, e^{\frac{x^2}{2^2}} \cdot e^{4(1.1 - x)^2} \cdot e^{4(1.9 - x \cdot 2)^2} \cdot e^{4(2.7 - x \cdot 3)^2}$$



Ingredients of a ProbProg Language

([Ramsey-Shan'17] notation)

Sampling

$x \sim \mu$ in
N

Conditioning

$M \rightsquigarrow \mu$

Sequencing

let $x = M$ in N

+ other PL features (H-O functions, algebraic datatypes, state, ...
Not in this talk)

Base types (this talk)

$A, B ::=$ Syntax for spaces

$\{\ell_1 \mid \cdots \mid \ell_n\}$ finite discrete

$\mid \mathbb{N}$ Countable discrete

$\mid [a, b) \mid [a, b] \mid \cdots$ Continuous $(a, b \in [-\infty, \infty])$

$\mid A \times B$ products

$\Gamma ::= x_1 : A_1, \ldots, x_n : A_n$ Contexts

Graded/density

$P ::=$
$\quad \mathrm{Cat}\{\ell_1 | \cdots | \ell_n\}$
$\quad |\ \mathrm{Counting}$
$\quad |\ \mathrm{Lebesgue}\,[a,b]$
$\quad |\ \mathrm{Lebesgue}$

$\Omega ::= P_1, \ldots, P_n$

$\mu, \nu ::=$
$\quad \{\ell_1 : \mu_1 | \cdots | \ell_n : \mu_n\}$
$\quad |\ \mathrm{Geometric}\ \mu$
$\quad |\ \mathrm{Uniform}$
$\quad |\ \mathcal{N}(\mu_{mean}, \mu_{sdv})$

Stock measures
categorical/discrete
contable conting
bounded Lebesgue ($a, b \in (-\infty, \infty)$)
Unbounded Lebesgue

Sample spaces

built-in probability distributions

Type system

Term judgements

$\Gamma \mid \Omega \vdash M : A$

graded type system

Mostly standard, e.g.:

$$\frac{\Gamma \mid \Omega_1 \vdash M : A \qquad \Gamma, x : A \mid \Omega_2 \vdash N : B}{\Gamma \mid \Omega_1, \Omega_2 \vdash \mathrm{let}\ x = M\ \mathrm{in}\ N : B}$$

Distribution judgements

$\Gamma \mid \Omega \vdash \mu \ll \rho$

$$\frac{\Gamma \mid \Omega_1 \vdash \mathrm{mean} : \mathbb{R} \qquad \Gamma \mid \Omega_2 \vdash \mathrm{sdv} : (0, \infty)}{\Gamma \mid \Omega_1, \Omega_2 \vdash \mathcal{N}(\mathrm{mean}, \mathrm{sdv}) \ll \mathrm{Lebesgue}}$$

Typing judgements

e.g. $\underline{\text{Lebesgue}} := \mathbb{R}$
etc.

Sampling

$$\frac{\Gamma \mid \Omega_1 \vdash \mu \ll \rho \qquad \Gamma, x : \underline{\rho} \mid \Omega_2 \vdash N : A}{\Gamma \mid \Omega_1, \rho, \Omega_2 \vdash x \mathrel{\leftarrow\!\sim} \mu \text{ in } N : A}$$

Conditioning

$$\frac{\Gamma \mid \Omega_1 \vdash M : \underline{\rho} \qquad \Gamma \mid \Omega_2 \vdash \mu \ll \rho}{\Gamma \mid \Omega_1, \Omega_2 \vdash M \rightsquigarrow \mu : \mathbb{1}}$$

$:= \{\star\}$

Semantics

Syntax — denotes

- Space type A — Standard Borel space (SBS)

- Stock measure space ρ — SBS + σ-finite distribution (counting or Lebesgue)

e.g.

$$\int [\![\mathrm{Cat}\{\ell_1 | \cdots | \ell_n\}]\!]_{\mathrm{dist}} (d\ell)\, f(\ell) = \sum_{i=1}^{n} f(\ell_i)$$

$$\int [\![\mathrm{Lebesgue}]\!]_{\mathrm{dist}} dx\, f(x) = \int dx\, f(x)$$

etc.

$$[\![\Omega]\!] = \left(\prod_{\rho \in \Omega} [\![\rho]\!] ,\ \bigotimes_{\rho \in \Omega} [\![\rho]\!]_{\mathrm{dist}} \right)$$

product measure

Semantics (ctd)

Syntax

denotes

- terms

density + random variable

$$[\![\Gamma \mid \Omega \vdash M : A]\!] : [\![\Gamma]\!] \xrightarrow{([\![M]\!]_{\mathrm{den}},\ [\![M]\!]_{\mathrm{val}})} \mathbb{W}^{[\![\Omega]\!]} \times [\![A]\!]^{[\![\Omega]\!]}$$

- distributions

(Parameterised) density functions

$$[\![\Gamma \mid \Omega \vdash \mu \ll \rho]\!] : [\![\Gamma]\!] \longrightarrow \mathbb{W}^{\underline{\Omega} \times \underline{\rho}}$$

Semantics (ctd)

E.g.:

"'Bayes' Theorem"

$$[\![a \sim \mu \text{ in } N]\!]^{\gamma}_{den} := \lambda(\omega_1, a, \omega_2).$$

$$[\![\mu]\!]_{den}(\gamma; \omega_1) \times$$

$$[\![N]\!]_{den}(\gamma[x \mapsto a]; \omega_2)$$

$$[\![M \rightsquigarrow \mu]\!]^{\gamma}_{den} := \lambda(\omega_1, \omega_2).$$

$$[\![M]\!]_{den}(\gamma; \omega_1) \times [\![\mu]\!]_{den}(\gamma; \omega_2, [\![M]\!]_{val}(\gamma; \omega_1))$$

etc.

A graded monad

Rest is standard [Katsumata'14]:

$$T_\Omega S := \mathbb{W}^\Omega \times S^\Omega \cong (\mathbb{W} \times S)^\Omega$$

$(\mathbb{W}, \cdot, 1)$ monoid

Model evidence

Each model $\Gamma \mid \Omega \vdash M : A$ gives:

- A kernel

$$[\![\Gamma]\!] \xrightarrow{[\![M]\!]_{dist}} [\![A]\!]$$

$$[\![M]\!]_{dist}(\gamma, U) := \int [\![\Omega]\!] \, d\omega \; [\![M]\!]_{den}(\gamma; \omega) \times [\,[\![M]\!]_{val}(\gamma; \omega) \in U\,]$$

- model evidence: used to evaluate / diagnose fit

$$[\![M]\!]_{ev}\,\gamma := [\![M]\!]_{dist}(\gamma, [\![A]\!])$$

Foundations

W^{Ω} not a measurable space

e.g. $[0,\infty]^{\mathbb{R}}$

(Aumann's Thm)

I used quasi-Borel spaces: [Heunen, Staton, Kammar, Yang '17]

set

$$(\underline{X}, M_X)$$

subset $M_X \subseteq \underline{X}^{\mathbb{R}}$

+ closure axioms.

E.g. S measurable space $\Rightarrow M_S := \mathrm{Meas}(\mathbb{R}, S)$

Foundations (ctd)

Qbs as a category:

$f : X \to Y$ is function $\underline{\underline{X}} \xrightarrow{f} \underline{\underline{Y}}$

s.t. $\forall \alpha \in M_X$.

$\mathbb{R} \xrightarrow{\alpha} \underline{\underline{X}} \xrightarrow{f} \underline{\underline{Y}}$

$\in M_Y$

Thm [HSKY'17]:

For SBSs S, T:

$$\mathbf{Qbs}(S, T) = \mathbf{Meas}(S, T)$$

So Qbs is a conservative extension ~~of~~ SBS.

Distributions

[Ścibior et al. '18]

For distributions, a monad $\mathrm{Dist} : \mathbf{Qbs} \to \mathbf{Qbs}$

S SBS, $\mathrm{Dist}\, S$ = s-finite measures on S

$M_{\mathrm{Dist}\, S}$ = s-finite kernels $\mathbb{R} \rightsquigarrow S$

[Staton'17, Kallenberg'17]

k s-finite = σ-affine combination of prob. kernels:

$$k = \sum_{n=0}^{\infty} w_n \cdot k_n \qquad w_n \in [0, \infty]$$

Programming with Ω explicitly is tedious

Prefer:

$$\Gamma \vdash M : A \qquad \Gamma \vdash \mu \ll \mathrm{Dist}\ P$$

and semantically:

$$[\![\Gamma]\!] \xrightarrow{[\![M]\!]_{dist}} \mathrm{Dist}\ [\![A]\!] \qquad [\![\Gamma]\!] \xrightarrow{[\![\mu]\!]_{dist}} W^{[\![\underline{P}]\!]}$$

[Staton'17]

We can now include any primitive probability distributions for sampling

Conditioning still requires density:

$$\frac{\Gamma \vdash \mu \ll P \qquad \Gamma \vdash M : \underline{P}}{\Gamma \vdash M \leadsto \mu : \mathbb{1}}$$

Trade-off

- Inference <u>benefits</u> from graded information (e.g. Stan's HMC, ADVI)
- Programming is easier without grading

We can:

- pay the price (either way)
- use <u>static analysis</u> to compile non-graded → graded (e.g., SlicStan [Gorinova, Gordon, Sutton '19] TraceTypes [Lew et al. '20])

Graded/distributions

(ongoing!)

$\mathrm{Dist}_{\ll} \rho$ = distributions with density w.r.t. $[\![\rho]\!]$

So: $[\![\Gamma \mid \Omega \vdash M : A]\!] : [\![\Gamma]\!] \xrightarrow{([\![M]\!]_{lat},\ [\![M]\!]_{rel})} \mathrm{Dist}_{\ll}\left([\![\Omega]\!] \times [\![A]\!]\right)^{[\![\Omega]\!]}$

$[\![\Gamma \mid \Omega \vdash \mu \ll \rho]\!] : [\![\Gamma]\!] \xrightarrow{([\![\mu]\!]_{lat},\ [\![\mu]\!]_{rel})} \mathrm{Dist}_{\ll}\left([\![\Omega]\!] \times \mathrm{Dist}_{\ll}[\![\rho]\!]\right)^{[\![\Omega]\!]}$

$[\![M \rightsquigarrow \mu]\!]^{r} := \varphi \circ [\![M]\!]^{r}_{lat}$ where

$\varphi : [\![\Omega]\!] \to \mathbb{W}$ and

$$\begin{tikzcd} {[\![\Omega]\!]} \arrow[r, bend left, "\alpha"] \arrow[r, phantom, "\text{disintegrate w.r.t. } \Omega"] & {[\![\rho]\!]} \arrow[l, bend left, "\alpha^{+\Omega}"] \end{tikzcd}$$

$\varphi := \dfrac{d\, \alpha_{*}^{+\Omega}[\![\mu]\!]^{r}_{dist}}{d[\![\Omega]\!]}$

Bridges

(Ongoing!)

- build on/relate to
  - Hakaru [Ramsey-Shan '17, Narayanan '19]
  - Ong-Mattinson [unpublished]
  - Bayesian inversion [Dahlquist et al. '16–'18, '20]

- Foundational hazards: disintegrating σ-finite measures [c.f. Vákár-Ong '18]
- non-graded

Summary

- Proposed Taxonomy:

Sampling \ Conditioning	density	distribution
graded	Stan, Pyro	Hakaru, Trace-Typed Gen
non-graded	Anglican, Monad-Bayes	?

- quasi-Borel spaces as a convenient meta-theory