Semantics of Probabilistic Programming I

Xin Zhang

Peking University

Most of the content is from "Semantics of Probabilistic Programming: A Gentle Introduction" by Fredrik Dahlqvist, Alexandra Silva, and Dexter Kozen

Recap: Problem and Motivation

- Evaluate P(Z|X) and related expectations

- Problem with exact methods
  - Curse of dimensionality

  - P(Z|X) has a complex form making expectations analytically intractable

Recap: Variational Inference

- Functional: a function that maps a function to a value

$$\mathrm{H}[p] = \int p(x) \ln p(x) \, \mathrm{d}x$$

- Variational method: find a input function that maximizes the functional

- Variational inference: find a distribution q(z) to approximate p(Z|X) so a functional is maximized

Recap: Variational Inference

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

Between p(Z|X) and q(Z)

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\ \mathrm{KL}(q\|p) &= -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}\end{aligned}$$

If q can be any distribution, then variational inference is precise.
But in practice, it cannot

Is the following statement right?

- Probability $p(Z,X)$ is usually easier to evaluate compared to $P(Z|X)$.

Recap: Sampling Methods

- Stochastic methods

- Also called Monte Carlo methods

$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \quad \Rightarrow \quad \hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)})$$

$\mathbf{z_1}, \ldots, \mathbf{z_l}$ are samples from p

Recap: Sampling Methods

- Transformation method: $CDF^{-1}(uniform(0,1))$

- Rejection sampling
  - A proposal distribution $q(z)$
  - Choose k, such that $k*q(z) >= p(z)$, for any x
  - Sampling process:
    - Sample z_0 from $q(z)$
    - Sample h from $uniform(0, k*q(z_0))$
    - If $h > p(z_0)$, discard it; otherwise, keep it

Is the following statement correct?

- All primitive distributions can be constructed using the transformation method.

Is the following statement right?

- In rejection sampling, the probability whether a sample is accepted does not depend on the proposal distribution

Is the following statement correct?

- The efficiency of importance sampling depends on the choice of the proposal distribution

Recap: Sampling Methods

- Importance sampling
  - Used to evaluate f(z) where z is from p(z)

$$E(f) = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \approx \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^l)}{q(z^l)}f(z^l)$$

  - How to get real samples: create a new discrete distribution using the above samples and set their probabilities using the importance weights

Recap: Sampling Methods

- Markov Chain Monte Carlo
  - A sampling method that works with a large family of distributions and high dimensions

- Workflow
  - Start with some sample z_0
  - Suppose the current sample is z^τ . Draw next sample z^* from $q(z \mid z^\tau)$
  - Decide whether to accept z^*as the next state based some criteria. If accepted, $z^{\tau+1} = z^*$. Otherwise, $z^{\tau+1} = z^\tau$
  - Samples form a Markov chain

Recap: Sampling Methods

	Metropolis	Metropolis-Hasting
Constraints on the proposal distribution	Symmetric	None
Accepting probability	$\min(1, \frac{p(z')}{p(z)})$	$\min(1, \frac{p(z')q(z'\|z)}{p(z)q(z\|z')})$

Recap: Why MCMC works?

- Markov chain: $$p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}) = p(\mathbf{z}^{(m+1)}|\mathbf{z}^{(m)}).$$
- Stationary distribution of a Markov chain: each step in the chain does not change the distribution.
- Detailed balance: $$p^\star(\mathbf{z})T(\mathbf{z}, \mathbf{z}') = p^\star(\mathbf{z}')T(\mathbf{z}', \mathbf{z})$$
  - $p^*(\boldsymbol{z})$ is a stationary distribution
- A *ergodic* Markov chain converges to the same distribution regardless the initial distribution
  - The system does not return to the same state at fixed intervals
  - The expected number of steps for returning to the same state is finite

Is the following statement right?

- The samples drawn using MCMC are independent

Is the following statement right?

- A Markov chain can have more than one stationary distribution

Use MCMC to solve the problem below

- Super optimization
  - There is a straight-line program
  - A set of test cases are given
  - The program can be modified by deleting a statement, inserting a statement from the initial program at a given place
  - Optimize the program by using the above operations

This Class

- The lecture is heavy in math. It is OK if you only get a sense of it. We won't focus on it in exams
- Semantics of probabilistic programming
- Measure theory

Motivations

- In order to reason about properties of a program, we need formal tools

- Example questions
  - Is the postcondition satisfied?
  - Does this program halt on all inputs?
  - Does it always halt in polynomial time?

Motivations

- In order to reason about properties of a program, we need formal tools

- Example questions
  - What is the probability that the postcondition is satisfied?
  - What is the probability that this program halts on all inputs?
  - What is the probability that it halts in polynomial time?

Motivations

- When designing a language, rigorous semantics is needed to guarantee its correctness

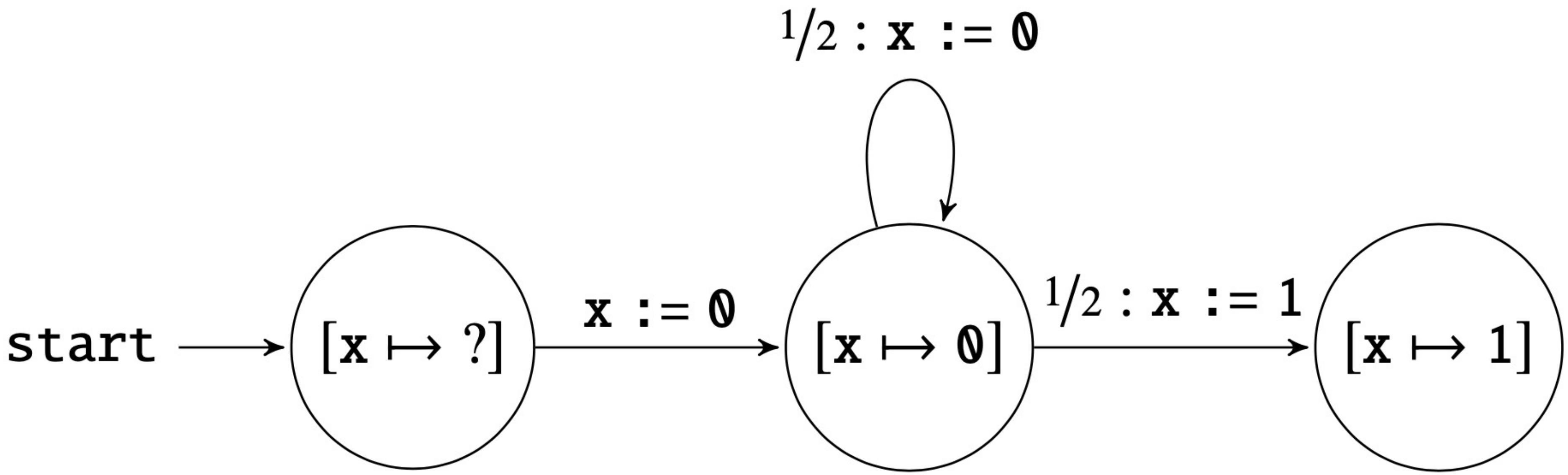
- Example that didn't have rigorous semantics: Javascript
  - https://javascriptwtf.com

Examples

We can decompose the semantics of a program into semantics of statements

```
x :=0
while x == 0 do
        x:=coin()
```

What is the probability that It runs through n iterations?
What is the expected number of iterations?
What is the probability that the program halts?



Examples

```
main{
        u:=0;
        v:=0;
        step(u,v);
        while u!=0 || v!=0 do
                step(u,v)
}

step(u,v){
        x:=coin();
        y:=coin();
        u:=u+(x-y);
        v:=v+(x+y-1)
}
```

What is the probability that the program halts?

The program is a two-dimensional random walk. According to probability theory, the probability that it returns to the origin is 1.

By relating to concepts in probabilities, we can simplify the reasoning

Examples

```
i:=0;
n:=0;
while i<1e9 do
        x:=rand();
        y:=rand();
        if (x*x+y*y) < 1 then n:=n+1;
        i:=i+1
i:=4*n/1e9;
```

What does this program compute?

How to reason about it?

Measure Theory
The mathematical foundation of probabilities and integration

Uniform(0,1) is called *Lebesgue measure*

Measure Theory

- Measures: generalization of concepts like length, area, or volume

- We will talk about
  - What is a measurable space
  - Measures on measurable spaces
  - Rich structures of spaces of measures

Measure Example: Length

- What subsets of R can meaningfully be assigned a length?

- What properties should be the length function l satisfy?

Measure Example: Length

$$\ell([a_1, b_1] \cup [a_2, b_2]) = \ell([a_1, b_1]) + \ell([a_2, b_2]) = (b_1 - a_1) + (b_2 - a_2).$$ $b_1 < a_2$

$$\ell\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \ell(A_i).$$ $A_i\ and\ A_j\ are\ disjoined\ .l\ is\ called\ additive$

$$\ell\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \ell(A_i).$$ $A_i\ and\ A_j\ are\ disjoined\ .The\ set\ is\ countable.$
$l\ is\ called\ countably\ additive\ or\ \sigma - additive$

$l(R) = \infty$, but we are only going to talk about finite measures

$$\ell(B \setminus A) = \ell(B) - \ell(A)$$ Domain should be closed under complementation

Measure Example: Length

- Can we extend the domain of length l to all subsets of R?
- No. Counterexample: Vitali sets
  - $V \subseteq [0,1]$, such that for each real number r, there exists exactly one number $v \in V$ such that $v - r$ is rational
  - Let $q_1, q_2, \dots$ be the rational numbers in $[-1,1]$, construct sets $V_k = V + q_k$
  - $[0,1] \subseteq \bigcup_k V_k \subseteq [-1,2]$
  - $l(V_k) = l(V)$, and there finitely many V_k
- l is called the *Lebesgue measure* on real numbers

Measurable Spaces and Measures

- (**S**, **B**) is a measurable space
  - **S** is a set
  - **B** is a σ-algebra on **S**, which is a collection of subsets of **S**
    - It contains $\emptyset$
    - Closed under complementation in **S**
    - Closed under countable union
  - The elements of **B** are called measurable sets

- If **F** is a collection of subsets of **S**, $\sigma(\boldsymbol{F})$ is the smallest σ-algebra containing **F**, or $\sigma(\mathcal{F}) \triangleq \bigcap\{\mathcal{A} \mid \mathcal{F} \subseteq \mathcal{A} \text{ and } \mathcal{A} \text{ is a } \sigma\text{-algebra}\}$. We say (S, $\sigma(\boldsymbol{F})$) is generated by **F**.

Measurable Functions

- $(\boldsymbol{S}, \boldsymbol{B_S})$ and $(\boldsymbol{T}, \boldsymbol{B_T})$ are measurable spaces. A function $f: \boldsymbol{S} \to \boldsymbol{T}$ is measurable if $f^{-1}(\boldsymbol{B}) = \{x \in S | f(x) \in B\}$ for every $\boldsymbol{B} \in \boldsymbol{B_T}$ is a measurable subset of S

Example:

$$\chi_B(s) = \begin{cases} 1, & s \in B, \\ 0, & s \notin B. \end{cases}$$

Measures: Definitions

- A signed (finite) measure on $(\boldsymbol{S}, \boldsymbol{B})$ is a countably additive map $\mu: \boldsymbol{B} \to \boldsymbol{R}$ such that $\mu(\emptyset) = 0$

- Positive signed measure: $\mu(A) \geq 0$ for all $A \in \boldsymbol{B}$

- A positive measure is a probability measure if $\mu(S) = 1$

- …is a subprobability measure if $\mu(S) \leq 1$

Measures: Definitions

- If $\mu(B) = 0$, then B is a μ-nullset

- A property is said to hold μ-almost surely (everywhere) if the sets of points on which it does not hold is contained in nullset

- In probability theory, measures are sometimes called distributions

Measures: Discrete Measures

- For $s \in S$, the Diract measure, or Diract delta, or point mass on s:

$$\delta_s(B) = \begin{cases} 1, & s \in B, \\ 0, & s \notin B. \end{cases}$$

- A measure is discrete if it is a countable weighted sum of Dirac measures
  - If the weights add up to one, then it is a discrete probability measure

- Continues measure: $\mu(\{s\}) = 0$ for all singleton sets {s} in $\boldsymbol{B}$ of $(\mathbf{S}, \boldsymbol{B})$

Measures: Pushforward Measure and Lebesgue Integration

- Given $f: (\boldsymbol{S}, \boldsymbol{B_S}) \to (\boldsymbol{T}, \boldsymbol{B_T})$ measurable an a measure μ on $\boldsymbol{B_S}$, the **pushfoward measure** $\mu(f^{-1}(B))$ on $\boldsymbol{B_T}$ is defined as

$$f_*(\mu)(B) = \mu(f^{-1}(B)), \quad B \in \mathcal{B}_T .$$

- **Lebesgue integration**: given $(\boldsymbol{S}, \boldsymbol{B})$, $\mu: \boldsymbol{B} \to \boldsymbol{R}$, $f: \boldsymbol{S} \to \boldsymbol{R}$, where $\mathrm{m} < f < M$

$$\int f \, d\mu = \lim_{n \to max} \sum_{i=0}^{n} f(s_i)\mu(B_i)$$

where $B_0, .., B_n$ is a measurable partition of $\boldsymbol{S}$, and the value of f does not vary more than $(M - m)/n$ in any B_i and $s_i \in B_i$

Measures: Absolute Continuity

- Given two measures μ and v, we say μ is absolute continuous with respect to v for all measurable sets B iff $v(B) = 0 \Longrightarrow \mu(B) = 0$
  - $\mu \ll v$

Theorem 1.1 (Radon–Nikodym) *Let μ, ν be two finite measures on a measurable space $(S, \mathcal{B})$ and assume that μ is absolutely continuous with respect to ν. Then there exists a measurable function $f : S \to \mathbb{R}$ defined uniquely up to a μ-nullset such that*

$$\mu(B) = \int_B f \, d\nu.$$

The function f is called the Radon–Nikodym derivative *of μ with respect to ν.*

Measures: More on Radon-Nikodym

- Not related to semantics, but one pillar of the probability theory

- f is called the Radon-Nikodym derivative. One example is density function

- Extends probability masses and probability measures to measures over arbitrary set

- Example: μ: $gaussain, v$: $Lebesgue\ measure\ on\ R$

Products of Measurable Spaces

- Given $(\boldsymbol{S_1}, \boldsymbol{B_1})$ and $(\boldsymbol{S_2}, \boldsymbol{B_2})$, their product is $(\boldsymbol{S_1} \times \boldsymbol{S_2}, \boldsymbol{B_1} \otimes \boldsymbol{B_2})$ where $$\boldsymbol{B_1} \otimes \boldsymbol{B_2} = \sigma(\{B_1 \times B_2 \mid B_1 \in \boldsymbol{B_1}, B_2 \in \boldsymbol{B_2}\})$$

- A measure on $(\boldsymbol{S_1} \times \boldsymbol{S_2}, \boldsymbol{B_1} \otimes \boldsymbol{B_2})$ is sometimes called a joint distribution

- A special case $(\mu_1 \otimes \mu_2)(B_1 \times B_2) \triangleq \mu_1(B_1)\mu_2(B_2).$

μ_1 and μ_2 are independent

Markov Kernels

- Given $(\boldsymbol{S}, \boldsymbol{B_S})$ and $(\boldsymbol{T}, \boldsymbol{B_T})$, $P: \boldsymbol{S} \times \boldsymbol{B_T} \rightarrow \boldsymbol{R}$ is called a Markov kernel if
  - For fixed $A \in \boldsymbol{B_T}$, the map $\lambda s. P(s, A) \rightarrow \boldsymbol{R}$ is a measurable function on $(\boldsymbol{S}, \boldsymbol{B_S})$
  - For fixed $s \in \boldsymbol{S}$, the map $\lambda A. P(s, A) \rightarrow \boldsymbol{R}$ is a probability measure on $(\boldsymbol{T}, \boldsymbol{B_T})$

- Composition of two Markov kernels
  - Given $P: S \rightarrow T$, $Q: T \rightarrow U$ $(P \,;\, Q)(s, A) = \displaystyle\int_{t \in T} P(s, dt) \cdot Q(t, A).$

- Given μ on $\boldsymbol{B_S}$, its push forward under the Markov Kernel P is

$$P_*(\mu)(B) = \int_{s \in S} P(s, B)\ \mu(ds).$$

More on Markov Kernels

- $(\boldsymbol{S}, \boldsymbol{B_S})$: x = … (x>0)
- $(\boldsymbol{T}, \boldsymbol{B_T})$: y = uniform(0,x)
- Markov kernel $P(x, \bigcup_{i=1}^{i=M}[a_i, b_i]) = \sum_{i=1}^{i=M} length([a_i, b_i] \cap [0, x])/x$

More on Markov Kernels

- $(\boldsymbol{S}, \boldsymbol{B_S})$: x = … (x>0)
- $(\boldsymbol{T}, \boldsymbol{B_T})$: y = uniform(0,x)
- $(\boldsymbol{T}, \boldsymbol{B_T})$: z = uniform(0,y)
- Composition: $(P;Q)(x,[0,z]) = \int_{y\in[0,\infty]} P(x,dy) * Q(y,[0,z])$

z < x

$$= \int_{y\in[0,x]} \frac{dy}{x} * \frac{length([0,z] \cap [0,y])}{y}$$

$$= \int_{y\in[0,z]} \frac{dy}{x} * \frac{y}{y} + \int_{y\in[z,x]} \frac{dy}{x} * \frac{z}{y} = \frac{z}{x} + \frac{z}{x}(lnx - lnz)$$

More on Markov Kernels

- $(\boldsymbol{S}, \boldsymbol{B_S})$: x = uniform(0.1, 1.1) $\mu([a, b]) = \text{length}([\text{a}, \text{b}] \cap [0.1,1.1])$
- $(\boldsymbol{T}, \boldsymbol{B_T})$: y = uniform(0,x)
- Markov kernel $P(x, \cup_{i=1}^{i=M}[a_i, b_i]) = \sum_{i=1}^{i=M} length([a_i, b_i] \cap [0, x])/x$
- μ's pushforward under P is

$$P_*(\mu)(B_T) = \int_{x \in [0.1,1.1]} B_T \cap [0, x] * \mu(dx)$$

More on Markov Kernels

- We can use Markov kernels to define the meanings of statements

- A program can be seen as a Markov kernel that links the input variable (can be a distribution) with the output distribution

Spaces of Measures

- We now talk about the structures of the spaces of measures
  - This will allow us to talk about general properties of measures

- $\boldsymbol{M(S,B)}$ or $\boldsymbol{MS}$ is the set of all finite, signed measures on a measurable set $(\boldsymbol{S},\boldsymbol{B})$

Vector Space Structure

- $\boldsymbol{MS}$ is always a real vector space

$$(\mu + \nu)(B) \triangleq \mu(B) + \nu(B) \qquad\qquad (a\mu)(B) \triangleq a\mu(B)$$

Normed Space Structure

- Every measure has a norm

$$\|\mu\| \triangleq \sup \left\{ \sum_{i=1}^{n} |\mu(B_i)| : \{B_1, \ldots, B_n\} \text{ is a finite measurable partition of } S \right\}.$$

- For positive measures, $||\mu|| = \mu(S)$

- A complete normed vector space is a Banach space

Order Structure

- Measures have a natural pointwise order: $\mu \leq v\ if\ \mu(B) \leq v(B), \forall B$

- Are two distinct probability measures comparable?

- The partial order is compatible with the vector space structure:

  - if $\mu \leq \nu$, then $\mu + \rho \leq \nu + \rho$; and
  - if $0 \leq a \in \mathbb{R}$ and $\mu \leq \nu$, then $a\mu \leq a\nu$.

- Additions and multiplications by a positive scalar are monotone

Order Structure

- The partial order defines a lattice

$$(\mu \vee \nu)(B) \triangleq \sup\{\mu(A \cap B) + \nu(A^c \cap B) \mid A \in \mathcal{B}\}$$
$$(\mu \wedge \nu)(B) \triangleq \inf\{\mu(A \cap B) + \nu(A^c \cap B) \mid A \in \mathcal{B}\}.$$

- Why do we care? It will be used to deal with loops

- $\mu^+ = \mu \vee 0, \quad \mu^{-1} = -\mu \vee 0, \mu = \mu^+ - \mu^-$, modulus $|\mu| = \mu^+ + \mu^-$

- The order is compatible with the norm: $|\mu| \leq |v| \Rightarrow \big||\mu|\big| \leq \big||v|\big|$

Order Structure

- $\boldsymbol{MS}$ is a Banach lattice:
  - A Banach space with a lattice structure that is compatible with both the linear and normed structures

- A Banach lattice is σ-order-complete if every countable order-bounded set of measures in $\boldsymbol{MS}$ has a supremum in **MS**

- Every measure space is σ-order-complete

Order Structure

- For any measure set **MS,** every countable order-bounded set of measures in $\boldsymbol{MS}$ has a supremum in **MS**

- This will help us deal with loops
  - Every iteration can be seeing as joining measures
  - The measures are bounded
  - They will converge to the supremum

Operators

- Since spaces of measures are vector spaces, we can do linear algebra
- Linear operator T: $T(x) + T(y) = T(x + y), T(ax) = aT(x)$
- We are mostly interested in operators that send probability measures to subprobability measures (conditional probabilities)

Summary

- To reason about properties and correctness of probabilistic programs, we need semantics

- To define semantics, we can
  - Decompose it into semantics of program structures
  - Link it with mathematical concepts

Summary

- Measure theory is the theory about measures (generalization of length, area, volume…)
  - Foundation of probabilities and integration

- Measurable space
- Measures: distribution, state of a program
- Markov kernels: allows us to model statements
- The space of measures on a given measure space is a σ-order-complete Banach lattice

Next Class

- Semantics of probabilistic programs
  - Operational semantics
  - Denotational semantics