Notes on Markov Kernels

Prakash Panangaden Sherry Shanshan Ruan

June 30, 2014

1 Prelude: binary relations

The whole point of this note is that Markov kernels are a natural probabilistic generalization of the notion of binary relations. We first introduce some basic definitions and algebra of binary relations.

Definition A binary relation $R \subseteq X \times Y$ is a set of (x, y) pairs where $x \in X$ and $y \in Y$, and we write xRy to indicate $(x, y) \in R$

We can compose two binary relations or take the converse of a binary relation. The formal definitions are given as follows:

Definition Given $R \subseteq X \times Y$ and $S \subseteq Y \times Z$, $S \circ R$ is defined as

$$x(S \circ R)z \qquad \text{if there exists } y \in Y \text{ such that } xRy \text{ and } ySz$$

Definition Given $R \subseteq X \times Y$, $R^c \subseteq Y \times X$ is defined as

$$yR^cx \qquad \text{if } xRy$$

2 Binary relations on power sets

We then proceed by introducing binary relations from power sets, which can be regarded as the "discrete analogue" of probabilistic relations. Hence, this can facilitate the presentation of probabilistic relations later. Binary relations from power set are defined as follows:

Definition Let X, Y be sets and $f : X \to \mathcal{P}(Y)$, then binary relation $F \subseteq X \times Y$ is defined as

$$xFy \qquad \text{if } y \in f(x)$$

Given a function mapping from one set to another set, it is often useful to manipulate it to obtain some mappings defined on power sets. We have the following two special functions:

Definition Let X, Y be sets and f a function from X to Y, then function $\mathcal{P}(f) : \mathcal{P}(X) \to \mathcal{P}(Y)$ is defined as

$$\mathcal{P}(f)(A) = \{f(x) \mid x \in A\} \qquad \text{for any } A \subseteq X$$

and function $f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X)$ is defined as

$$f^{-1}(B) = \{x \mid f(x) \in B\} \qquad \text{for any } B \subseteq Y$$

Note that $\mathcal{P}(f)$ and f^{-1} are not inverse functions of each other because they may not be bijections. But we can roughly think of them as inverses. Subsequently, we introduce two functions which are essential to composing binary relations. One maps a set to its power set, and the other maps a power set to the set itself.

Definition For any set X, function $\{\cdot\}_X : X \to \mathcal{P}(X)$ is defined as

$$\{\cdot\}_X(x) = \{x\} \qquad \text{for all } x \in X$$

Definition For any set X, function $U : \mathcal{P}(\mathcal{P}(X)) \to \mathcal{P}(X)$ is defined as

$$U(\{A_i \mid i \in I\}) = \bigcup_{i \in I} A_i \qquad \text{for any } A_i \in X$$

Now we are ready to compose two functions which map a set to its power set.

Definition Given $f : X \to \mathcal{P}(Y)$ and $g : Y \to \mathcal{P}(Z)$ together with corresponding binary relations F and G, we define the composition function $g \# f : X \to \mathcal{P}(Z)$ (thereby $G \# F$) as follows:

$$g \# f = U \circ \mathcal{P}(g) \circ f : X \to \mathcal{P}(Z)$$

To justify the composition is well-defined, we first check the consistency of types:

$$\begin{aligned} \mathcal{P}(g) : &\quad \mathcal{P}(Y) \to \mathcal{P}(\mathcal{P}(Z)) \\ \mathcal{P}(g) \circ f : &\quad X \to \mathcal{P}(\mathcal{P}(Z)) \\ U \circ \mathcal{P}(g) \circ f : &\quad X \to \mathcal{P}(Z) \end{aligned}$$

Furthermore, we prove the associativity of compositions holds.

Proof. Let $f : X \to \mathcal{P}(Y)$, $g : Y \to \mathcal{P}(Z)$, $h : Z \to \mathcal{P}(W)$ be three arbitrary functions. Then $h \# (g \# f) = h \# (U \circ \mathcal{P}(g) \circ f) = U \circ \mathcal{P}(h) \circ (U \circ \mathcal{P}(g) \circ f)$. By the associativity of $\circ$, this is equal to $U \circ (\mathcal{P}(h) \circ U \circ \mathcal{P}(g)) \circ f$. Then we want to establish the equality between $\mathcal{P}(h) \circ U \circ \mathcal{P}(g)$ and $\mathcal{P}(U \circ \mathcal{P}(h) \circ g)$.

Let $A \in \mathcal{P}(Y)$ be arbitrary. We want to show that $(\mathcal{P}(h) \circ U \circ \mathcal{P}(g))(A) = (\mathcal{P}(U \circ \mathcal{P}(h) \circ g))(A)$.

$$\begin{aligned} & z \in (\mathcal{P}(h) \circ U \circ \mathcal{P}(g))(A) \\ \Leftrightarrow & z \in \mathcal{P}(h)(U(\{g(x) \mid x \in A\})) \\ \Leftrightarrow & z \in \mathcal{P}(h)(\bigcup_{x \in A} g(x)) \\ \Leftrightarrow & z \in \{h(y) \mid y \in \bigcup_{x \in A} g(x)\} \\ \Leftrightarrow & z \in \{\bigcup_{y \in g(x)} h(y) \mid x \in A\} \\ \Leftrightarrow & z \in \{U(\{h(y) \mid y \in g(x)\}) \mid x \in A\} \\ \Leftrightarrow & z \in \{U(\mathcal{P}(h)(g(x))) \mid x \in A\} \\ \Leftrightarrow & z \in \{(U \circ \mathcal{P}(h) \circ g)(x) \mid x \in A\} \\ \Leftrightarrow & z \in (\mathcal{P}(U \circ \mathcal{P}(h) \circ g))(A) \end{aligned}$$

Therefore, $U \circ (\mathcal{P}(h) \circ U \circ \mathcal{P}(g)) \circ f = U \circ \mathcal{P}(U \circ \mathcal{P}(h) \circ g) \circ f = (U \circ \mathcal{P}(h) \circ g) \# f = (h \# g) \# f$. Thus we obtain $h \# (g \# f) = (h \# g) \# f$

□

Also note that this indeed defines a composition of binary relations, as justified below:

$$
\begin{aligned}
& x\,(G\#F)\,z \\
\Leftrightarrow & z \in g\#f(x) \\
\Leftrightarrow & z \in U(\mathcal{P}(g)(f(x))) \\
\Leftrightarrow & z \in U(\{g(y) \mid y \in f(x)\}) \\
\Leftrightarrow & z \in \bigcup_{y \in f(x)} g(y) \\
\Leftrightarrow & \exists\, y,\ y \in f(x) \wedge z \in g(y) \\
\Leftrightarrow & \exists\, y,\ x\,F\,y \wedge y\,G\,z
\end{aligned}
$$

Since we have established a well-defined composition operation, we can regard $\{\cdot\}_X$ as the identity (or the equivalence relation). Given any function $f : X \to \mathcal{P}(X)$, it is easy to see that $\{\cdot\}\,\#\,f = f\,\#\,\{\cdot\} = f$. We give the proof of the second equality, and then the first holds due to the commutative property. Let $x \in X$ be given, then

$$
\begin{aligned}
& (f\,\#\,\{\cdot\})(x) \\
= \ & (U \circ \mathcal{P}(g) \circ \{\cdot\})(x) \\
= \ & U(\{g(x)\}) \\
= \ & g(x)
\end{aligned}
$$

3 Probabilistic relations

Probabilistic relation can be treated as the probabilistic analogue of power sets. Note that $\mathcal{P}(X)$ can be seen as a map from $X \to \{0,1\}$, so similarly, we can define ν as a subprobability measure from Σ_X to $[0,1]$. The probabilistic analogue of (X, Σ_X) is given below:

Definition Given a set X, we define

$$\Pi X = \{\nu \mid \nu : \Sigma_X \to [0,1] \text{ and } \nu \text{ is a subprobability measure}\}$$

Definition For every $A \in \Sigma_X$, we define $P_A : \Pi X \to [0,1]$ by

$$P_A(\nu) = \nu(A)$$

Definition We define $\Sigma_{\Pi X}$ as the smallest Σ-algebra on ΠX such that $\forall A \in \Sigma_X$, P_A is measurable.

Note such a Σ-algebra exists because the Σ-algebra taken as the power set $\mathcal{P}(\Pi X)$ can guarantee that all P_A maps are measurable.

Recall uncurrying is the technique of transforming a higher-order function that returns a new function as output into a function that takes a tuple of arguments. Therefore, given $f : X \to \mathcal{P}(Y)$, i.e. $f : X \to (Y \to \{0,1\})$, we can uncurry it to $f : X \times Y \to \{0,1\}$. Now we can define a *probabilistic relation* to be a measurable function $h : X \to \Pi Y$, or $(X \times \Sigma_Y) \to [0,1]$ by uncurrying.

Definition A probabilistic relation $h : X \times \Sigma_Y \to [0,1]$ is a *Markov kernel* if

(1) $\forall B \in \Sigma_Y, \lambda x.h(x,B)$ is a measurable function

(2) $\forall x \in X, \lambda B.h(x,B)$ is a measurable function

It is under the auspices of $\mathcal{P}(f)$, U, and $\{\cdot\}$ that we successfully composed binary relations on power sets. Thus, in order to compose probabilistic relations, we need to describe their probabilistic counterparts first.

Definition Let X and Y be two sets and f a measurable function from (X, Σ_X) to (Y, Σ_Y), we define $\Pi f : \Pi X \to \Pi Y$ as

$$(\Pi f)(\nu)(B) = \nu(f^{-1}(B)) \qquad \text{for any } \nu \in \Pi X, B \in \Sigma_Y$$

Hence, similar to $\mathcal{P}(f) : \mathcal{P}(X) \to \mathcal{P}(Y)$, we have $\Pi f : \Pi X \to \Pi Y$ as a probabilistic counterpart. The followings are analogous to $\{\cdot\}_X : X \to \mathcal{P}(X)$ and $U : \mathcal{P}(\mathcal{P}(X)) \to \mathcal{P}(X)$ respectively.

Definition Given (X, Σ_X) we define $\eta_X : X \to \Pi X$ by

$$\eta_X(x)(A) = \delta_x = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases} \qquad \text{for } A \in \Sigma_X$$

Definition We define $\xi : \Pi^2 X \to \Pi X$ as

$$\xi(\Omega)(A) = \int_{\Pi X} P_A d_\Omega \qquad \text{for } \Omega \in \Pi^2 X \text{ and } A \in \Sigma_X$$

Note that P_A is a measurable function from ΠX to $[0, 1]$ and Ω is a measure on $\Sigma_{\Pi X}$.

Having acquired these auxiliary functions, we are finally able to define the composition of two probabilistic relations:

Definition Let $f : X \to \Pi Y$ and $g : Y \to \Pi Z$, we define $g \# f : X \to \Pi Z$ as

$$(g \# f)(x, C) = (\xi \circ \Pi g \circ f)(x, C) = \int_Y g(y, C) f(x, dy) \qquad \text{for } x \in X \text{ and } C \in \Sigma_Z$$

Note that $g(y, C)$ is a measurable function from Y to $[0, 1]$, and $f(x, dy)$ is a measure defined on Σ_Y

We need to justify that this is well-defined (the second equality in the above definition).

First we prove a preliminary proposition useful in a variety of contexts. This is called the "change of variables" formula.

Proposition 3.1. *Suppose that (X, Σ_X) and (Y, Σ_Y) are measurable spaces and $f : X \to Y$ is a measurable function. Suppose that $g : Y \to [0, 1]$ is a measurable function and that $\tau \in \Pi X$; so that $\Pi(f)(\tau)$ is a measure on Y (i.e. is in ΠY). Then*

$$\int_Y g \, \mathrm{d}\Pi(f)(\tau) = \int_X g \circ f \mathrm{d}\tau.$$

Proof. We prove this for the special case of g being a characteristic function first. Let B be a measurable subset of Y and consider the special case with $g = \chi_B$. Then the lhs reduces to $\Pi(f)(\tau)(B) = \tau(f^{-1}(B))$. To evaluate the rhs note that $\chi_B \circ f = \chi_{f^{-1}(B)}$. Thus the rhs is $\tau(f^{-1}(B))$, which is the same as the lhs. Since the integral is linear the claimed equality holds for all simple functions. By the monotone convergence theorem it then holds for all measurable functions. □

Now we can complete the verification of the formula for composing Markov kernels. Suppose $x \in X$ and $D \subseteq Z$, then

$$\begin{aligned}
&(\xi \circ \Pi g \circ f)(x, D) \\
=&\xi(\Pi g(f(x))(D)) \\
=&\int_{\Pi Z} P_D \, \mathrm{d}(\Pi g(f(x))) \\
=&\int_Y (P_D \circ g) \, \mathrm{d}f(x) \\
=&\int_Y g(y)(D) \\
=&\int_Y g(y, D) f(x, dy)
\end{aligned}$$

We have used the change of variables formula to covert the integral over ΠZ to an integral over Y in the penultimate step.

As in previous section, we also verify the associativity of compositions:

Proof. Let $f : X \to \Pi Y$, $g : Y \to \Pi Z$, $h : Z \to \Pi W$ be three arbitrary functions. Suppose $x \in X$ and $D \subseteq W$.

Then $((h \# g) \# f)(x, D) = \displaystyle\int_Y \left(\int_Z h(z, D) g(y, dz) \right) f(x, dy)$

Since $h(z, D)$ is non-negative, by the Simple Approximation Theorem, $h(z, D)$ can be written as the limit of a monotone sequence of simple functions h_n, i.e. $h = \lim_{n\to\infty} h_n = \lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} \mathcal{X}_{C_{n,i}}$. Thus,

$$
\begin{aligned}
&((h \# g) \# f)(x, D) \\
&= \int_Y \left(\int_Z h(z, D) g(y, dz) \right) f(x, dy) && \text{by definition} \\
&= \int_Y \left(\int_Z \lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} \mathcal{X}_{C_{n,i}} g(y, dz) \right) f(x, dy) && \text{by substitution} \\
&= \int_Y \left(\lim_{n\to\infty} \int_Z \sum_{i=1}^{m_n} a_{n,i} \mathcal{X}_{C_{n,i}} g(y, dz) \right) f(x, dy) && \text{by monotone convergence theorem} \\
&= \int_Y \left(\lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} \int_Z \mathcal{X}_{C_{n,i}} g(y, dz) \right) f(x, dy) && \text{by linearity} \\
&= \int_Y \left(\lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} g(y, C_{n,i}) \right) f(x, dy) && \text{by evaluation of the inner integral} \\
&= \lim_{n\to\infty} \int_Y \left(\sum_{i=1}^{m_n} a_{n,i} g(y, C_{n,i}) \right) f(x, dy) && \text{by monotone convergence theorem} \\
&= \lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} \int_Y g(y, C_{n,i}) f(x, dy) && \text{by linearity} \\
&= \lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} \int_Z \mathcal{X}_{C_{n,i}} \left(\int_Y g(y, dz) f(x, dy) \right) && \text{by definition of characteristic functions} \\
&= \lim_{n\to\infty} \int_Z \left(\sum_{i=1}^{m_n} a_{n,i} \mathcal{X}_{C_{n,i}} \right) \left(\int_Y g(y, dz) f(x, dy) \right) && \text{by linearity} \\
&= \int_Z \left(\lim_{n\to\infty} \sum_{i=1}^{m_n} a_{n,i} \mathcal{X}_{C_{n,i}} \right) \left(\int_Y g(y, dz) f(x, dy) \right) && \text{by monotone convergence theorem} \\
&= \int_Z h(z, D) \left(\int_Y g(y, dz) f(x, dy) \right) && \text{by substitution} \\
&= (h \# (g \# f))(x, D) && \text{by definition}
\end{aligned}
$$

□

In the discrete case, the composition operation coincides with the operation of matrix multiplication:

$$(g \# f)(x, z) = \sum_{y \in Y} f(x, y) g(y, z)$$

4 Application

We can use probabilistic relations to describe conditional probability in LMPs. For instance, given a *labeled Markov process* $(S, \Sigma_S, \mathcal{L}, \tau_a : S \times \Sigma_S \to [0, 1],\ \forall a \in \mathcal{L})$, then

$$\tau_a : S \times \Sigma_S \to [0, 1]$$

can be interpreted as: for any $x \in S$,$B \subseteq \Sigma_S$, $\tau_a(x, B)$ is the conditional probability that the system ends up in B given it starts in x