CHAPTER 2

Essentials of Quantum Mechanics

We begin by building up the basic ingredients of quantum mechanics. This is not meant to be a course on quantum mechanics, and so we will proceed pragmatically and without much fanfare. We will have the luxury of working with finite-dimensional Hilbert spaces (if you do not know what this means, you will soon), since this is the setting of most present applications of quantum learning theory. Our pedagogical approach will be to revisit ordinary probability theory in a suggestive way that naturally generalizes to quantum theory. Our exposition is meant to be accessible to readers with a knowledge of linear algebra and probability theory.

1. Probability theory on vector spaces

1.1. Probability distributions and their transformations

Here we will formulate probability theory on a discrete space, with some additional linear algebraic baggage that will be useful later. If we have a set of size N we can represent a probability distribution over that set as a vector in $\mathbb{R}^N$ given by

$$\vec{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix}$$

where p_i is the probability of the ith item. We have, out of convenience, chosen an ordering on our set of items so that we can organize the probabilities into a vector, but of course this ordering is arbitrary. As usual, we require $p_i \geq 0$ for all i since probabilities cannot be negative, and also $\sum_{i=1}^N p_i = 1$ so that the probabilities are appropriately normalized. There is a natural way of packaging the normalization condition. To this end, consider the row vector

$$\vec{1}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} .$$

Then $\sum_{i=1}^N p_i = 1$ is equivalent to

$$\vec{1}^T \cdot \vec{p} = 1 ,$$

and we will use this more compact expression henceforth. It will sometimes be useful to consider the *probability simplex* Δ_N which is a subset of $\mathbb{R}^N$, where Δ_N consists of all nonnegative vectors with entries summing to one. Then we can write $\vec{p} \in \Delta_N$.

Next we consider a rudimentary version of *dynamics*. That is, what kinds of transformations on $\vec{p}$ will map it into another valid probability distribution? The

simplest kind of transformation we can imagine is a linear one, so let us examine that first. Letting M be an $N \times N$ matrix, we consider the transformation

$$\vec{p}' = M \cdot \vec{p},$$

so that $\vec{p}'$ is the new probability distribution after the transformation. But what conditions do we need to put on M such that $\vec{p}'$ is a bona fide probability distribution for all initial distributions $\vec{p}$? Well, we need for all entries of $\vec{p}'$ to be nonnegative, and for $\vec{1}^T \cdot \vec{p}' = 1$. To ensure the first property, suppose that $\vec{p}$ is all zeroes except for the jth entry which equals one. (That is, we would sample the jth object with probability 1 and never sample anything else.) To introduce some other notation, let $\vec{e}_j$ be vector which is all zeroes except for the jth entry which equals one. Then we have

$$\vec{p}' = M \cdot \vec{e}_j = \begin{bmatrix} M_{1j} \\ M_{2j} \\ \vdots \\ M_{Nj} \end{bmatrix} .$$

In order for all entries of $\vec{p}^{'}$ to be nonnegative, we evidently require $M_{ij} \geq 0$ for all j, and i fixed. Varying over i as well, we find the requirement that $M_{ij} \geq 0$ for all i, j, and so M must be a matrix with nonnegative entries. Since we also demand that $\vec{1}^T \cdot \vec{p}' = 1$, we find the condition

$$\vec{1}^T \cdot \vec{p}' = \vec{1}^T \cdot M \cdot \vec{e}_j = \vec{1}^T \cdot \begin{bmatrix} M_{1j} \\ M_{2j} \\ \vdots \\ M_{Nj} \end{bmatrix} = 1 .$$

That is, the jth column of M must sum up to one. Since this must hold for every column, we find the condition

$$\vec{1}^T \cdot M = \vec{1}^T . \tag{2}$$

Thus a nonnegative matrix satisfying (2) will send probability vectors to probability vectors. We honor this finding with a definition:

Definition 5 (Markov matrix)**.** *Let M be an $N \times N$ matrix. We say that M is a* ***Markov matrix*** *if $M_{ij} \geq 0$ for all i, j, and $\vec{1}^T \cdot M = \vec{1}^T$. Then M maps probability vectors to probability vectors.*

A few comments are in order. In many treatments of Markov matrices, there is a different convention in which M is taken to act on probability distributions 'to the left', which would give the transpose our definition above. Our conventions here are chosen to align with those of quantum mechanics, as we will see later on.

We immediately notice that Markov matrices behave nicely under composition. Specifically, we have the useful lemma:

Lemma 6 (Composition of Markov matrices)**.** *If $M_1, M_2, ..., M_k$ are Markov matrices, then $M_k \cdots M_2 \cdot M_1$ is also a Markov matrix.*

The proof of this useful fact follows by a short calculation using the definition (which you should do if you have not thought it through before). The upshot of

this lemma is that we can consider transformations like

$$\vec{p}\,' = M_k \cdots M_2 \cdot M_1 \cdot \vec{p}$$

as instantiating a type of 'circuit', with depth k. That is, we could say the words: starting with $\vec{p}$ we apply M_1 followed by M_2 followed by M_3 and so on, and then finally apply M_k.

Before moving on to increasing levels of sophistication, we consider a simple example:

Example 1 (Bernoulli coin, $N = 2$). We now specialize to a two–outcome space and fix the ordering so that the first coordinate is outcome 0 ("success") and the second is outcome 1 ("failure"). A Bernoulli distribution with success probability θ is therefore represented by

$$\vec{p}_\theta = \begin{bmatrix} \Pr[0] \\ \Pr[1] \end{bmatrix} = \begin{bmatrix} \theta \\ 1-\theta \end{bmatrix}, \qquad \theta \in [0,1].$$

Consider the *bit–flip* dynamics with flip probability $\varepsilon \in [0,1]$,

$$M_\varepsilon = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix},$$

whose entries are nonnegative and whose columns each sum to 1, so M_ε is a Markov matrix in our sense. Acting on $\vec{p}$ produces

$$\vec{p}_{\theta'} = M_\varepsilon\, \vec{p}_\theta = \begin{bmatrix} (1-\varepsilon)\theta + \varepsilon(1-\theta) \\ \varepsilon\theta + (1-\varepsilon)(1-\theta) \end{bmatrix} \quad \Longrightarrow \quad \theta' = (1-2\varepsilon)\,\theta + \varepsilon,$$

where $\theta' = \Pr'[0]$ is the new success probability.

Some immediate checks help build intuition. When $\varepsilon = 0$ the map is the identity; when $\varepsilon = 1$ it deterministically flips $0 \leftrightarrow 1$; and when $\varepsilon = \frac{1}{2}$ it sends every input to the uniform distribution $\vec{p}_{1/2} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$ in one step. For any $0 < \varepsilon < 1$, the unique fixed point solves $\theta' = \theta$ and is $\theta_* = \frac{1}{2}$. (To see this, simply solve $\theta_* = (1-2\varepsilon)\theta_* + \varepsilon$ for θ_*). Iterating M_ε a total of k times yields exponential mixing toward the fixed point θ_* at rate $|1-2\varepsilon|$:

$$\theta^{(k)} = \big(1-2\varepsilon\big)^k \Big(\theta^{(0)} - \frac{1}{2}\Big) + \frac{1}{2}\,.$$

Finally, the family M_ε of Markov matrices is closed under composition (illustrating the lemma above): a short calculation shows

$$M_\eta\, M_\varepsilon = M_{\varepsilon+\eta-2\varepsilon\eta}\,,$$

and in particular $M_\varepsilon^k = M_{\varepsilon_{\text{eff}}}$ with

$$\varepsilon_{\text{eff}} = \frac{1-(1-2\varepsilon)^k}{2}.$$

This two–state example already displays dynamics, fixed points, and circuit composition within the linear–algebraic language we have been developing.

Moving on, it is useful to recount a few features of probability distributions. If

we have k probability distributions $\vec{p}_1, ..., \vec{p}_k$, then we can form a new probability distribution by forming a convex combination

$$\vec{p}' = \sum_{j=1}^{k} r_j \, \vec{p}_j \tag{3}$$

where $r_j \geq 0$ and $\sum_{j=1}^{k} r_j = 1$. To see this, notice that $\vec{p}'$ has nonnegative entries and that $\vec{1}^T \cdot \vec{p}' = \sum_{j=1}^{k} r_j \, (\vec{1}^T \cdot \vec{p}_j) = \sum_{j=1}^{k} r_j = 1$. We can interpret $r_1, ..., r_k$ as a probability distribution over k items in its own right, and say of (3) that we have a probabilistic mixture of k probability distributions wherein we sample from $\vec{p}_j$ with probability r_j. That is, $r_1, ..., r_k$ is a probability distribution over probability distributions. (You can use this 'meta' statement to impress your friends, if you like.) To make this concrete, consider the following example:

Example 2 (Sampling two coins, $N = 2$). Suppose we have two Bernoulli coins, represented by the probability vectors $\vec{p}_{1/2}$ and $\vec{p}_{1/3}$, respectively. The first one gives heads with probability $1/2$ and tails with probability $1/2$, and the second gives heads with probability $1/3$ and tails with probability $2/3$. Now suppose I have both coins in my pocket in such a way that when I reach in, I grab the first coin with probability $1/4$ and the second coin with probability $3/4$. Then if I reach in and grab a coin and toss it, what is the probability that I would output heads? This is described by the convex combination

$$\frac{1}{4} \, \vec{p}_{1/2} + \frac{3}{4} \, \vec{p}_{1/3} = \begin{bmatrix} 3/8 \\ 5/8 \end{bmatrix},$$

and so evidently the probability of heads is $3/8$.

So far we have only considered *linear* transformations on $\vec{p}$ that map it into another probability distribution. What if we consider nonlinear transformations? One example would be the nonlinear transformation

$$T(\vec{p}) = \begin{bmatrix} \frac{p_1^2}{\sum_{i=1}^{N} p_i^2} \\ \frac{p_2^2}{\sum_{i=1}^{N} p_i^2} \\ \vdots \\ \frac{p_N^2}{\sum_{i=1}^{N} p_i^2} \end{bmatrix}.$$

Another example would be a Bayesian update. There are clearly a vast infinitude of other possibilities as well. Among this infinitude of transformations there is a natural class that interfaces well with convex combinations of probability distributions. In particular, suppose we mandate that T satisfies

$$T\left(\sum_{j=1}^{k} r_j \, \vec{p}_j \right) = \sum_{j=1}^{k} r_j \, T(\vec{p}_j) \tag{4}$$

for any $\vec{p}_1, ..., \vec{p}_k$ and any valid $r_1, ..., r_k$. In words, we are requiring that a transformation of a probabilistic mixture is a probabilistic mixture of transformations (and specifically, the same transformation). Such T's satisfy a nice structure theorem:

Theorem 7 (Mixture-preserving transformations are Markov matrices)**.** *Suppose that $T : \Delta_N \to \Delta_N$ is a mixture-preserving transformation, namely that (4) is satisfied. Then there exists a Markov matrix M such that $T(\vec{p}) = M \cdot \vec{p}$ for all $\vec{p}$.*

PROOF. Write $\vec{p} = \sum_{j=1}^{N} p_j \, \vec{e}_j$. Using the mixture-preserving property of T, we have

$$T(\vec{p}) = T\left(\sum_{j=1}^{N} p_j \, \vec{e}_j\right) = \sum_{j=1}^{N} p_j \, T(\vec{e}_j) \, .$$

Let M be the matrix whose jth column is $T(\vec{e}_j)$. Then $T(\vec{p}) = M \cdot \vec{p}$. Each column $T(\vec{e}_j)$ is a probability vector, so $M_{ij} \geq 0$ and $\vec{1}^T \cdot M = \vec{1}^T$. Thus M is a Markov matrix, as claimed. □

Mixture-preserving transformations are natural from a physical point of view. Imagine a preparation device that, with probabilities $r_1, ..., r_k$, produces one of the distributions $\vec{p}_1, ..., \vec{p}_k$ by consulting some randomly tossed coins you do not get to see. If dynamics could distinguish whether this randomization happened "before" or "after" the transformation, then the timing of the unseen coin flips would be observable from the output statistics alone. Requiring that they not be observable is exactly the statement of (4).

Two simple consequences are worth keeping in mind. First, the admissible dynamics are closed under randomized control: if with probability r_j you implement a Markov matrix M_j, then the overall map is

$$M' = \sum_{j=1}^{k} r_j \, M_j \, ,$$

which is again a Markov matrix since $\vec{1}^T \cdot M' = \sum_{j=1}^{k} r_j \, (\vec{1}^T \cdot M_j) = \vec{1}^T$ and all entries are nonnegative. Second, if one further insists that deterministic states are carried to deterministic states, so that $\vec{e}_j$ never acquires additional randomness, then each column $T(\vec{e}_j)$ must itself be a basis vector. Equivalently, M has exactly one 1 (and zeros elsewhere) in each column. Such matrices are sometimes called *deterministic* or *functional* Markov matrices. If in addition the mapping $j \mapsto i(j)$ is injective (no two distinct columns point to the same basis vector), then M is a permutation matrix.

By contrast, nonlinear updates arise when you condition on a revealed outcome and then renormalize; the rule in that case depends on which outcome was announced, so it is not a single fixed map on Δ_N and does not represent closed-system dynamics. This classical discussion sets the stage for the quantum case, which we will treat soon. (There, the state space becomes the convex set of density operators, mixture-preserving maps become convex-linear "channels," and the role of Markov matrices is played by completely positive, trace-preserving maps.)

1.2. Joint distributions and tensor products

In probability theory it is essential to consider joint distributions. Here we develop the basic operations of joint distributions in a convenient and illuminating linear algebraic notation. First we require some additional tools on the linear algebra side. Specifically, we will upgrade our linear algebraic toolkit to *multi-linear*

algebra. The key operation will be the **tensor product**, which is an operation for joining two or more vector spaces.

We will proceed by motivating the tensor product informally through simple examples, and then give the abstract definition. It is worth paying close attention as the tensor product will serve as an essential piece of mathematical architecture for almost everything in quantum learning theory.

Consider two vectors $\vec{v}, \vec{w}$ in $\mathbb{R}^N$. We denote their tensor product by $\vec{v} \otimes \vec{w}$. To develop what this means, consider the example below.

Example 3. Let $\vec{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\vec{w} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$. Then their tensor product $\vec{v} \otimes \vec{w}$ is represented by

$$\vec{v} \otimes \vec{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} \\ \\ 2 \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 6 \\ 8 \end{bmatrix} .$$

In words, $\vec{w}$ gets 'sucked in' to $\vec{v}$. Now let us take the tensor product in the other order, namely $\vec{w} \otimes \vec{v}$:

$$\vec{w} \otimes \vec{v} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ \\ 4 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \\ 4 \\ 8 \end{bmatrix} .$$

From this we glean that, in general, $\vec{v} \otimes \vec{w} \neq \vec{w} \otimes \vec{v}$. Moreover, since $\vec{v} \in \mathbb{R}^2$ and $\vec{w} \in \mathbb{R}^2$, we notice that $\vec{v} \otimes \vec{w} \in \mathbb{R}^4$. To this end we write $\vec{v} \otimes \vec{w} \in \mathbb{R}^2 \otimes \mathbb{R}^2 \simeq \mathbb{R}^4$.

Example 4. Suppose $\vec{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\vec{w} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$ so that $\vec{v} \in \mathbb{R}^2$ and $\vec{w} \in \mathbb{R}^3$. Then

$$\vec{v} \otimes \vec{w} = \begin{bmatrix} 3 \\ 4 \\ 5 \\ 6 \\ 8 \\ 10 \end{bmatrix} \in \mathbb{R}^6 ,$$

and we write $\vec{v} \otimes \vec{w} \in \mathbb{R}^2 \otimes \mathbb{R}^3 \simeq \mathbb{R}^6$.

From the previous two examples we see the general rule that if $\vec{v} \in \mathbb{R}^N$ and $\vec{w} \in \mathbb{R}^M$, then $\vec{v} \otimes \vec{w} \in \mathbb{R}^N \otimes \mathbb{R}^M \simeq \mathbb{R}^{NM}$. So upon taking the tensor product of two vector spaces, the dimensions multiply. We can generalize this further by contemplating another example:

Example 5. Let $\vec{v} = \begin{bmatrix}1\\2\end{bmatrix}$, $\vec{w} = \begin{bmatrix}3\\4\end{bmatrix}$, and $\vec{u} = \begin{bmatrix}5\\6\end{bmatrix}$. Then we have

$$\vec{v}\otimes\vec{w}\otimes\vec{u} = (\vec{v}\otimes\vec{w})\otimes\vec{u} = \begin{bmatrix}3\\4\\6\\8\end{bmatrix}\otimes\begin{bmatrix}5\\6\end{bmatrix} = \begin{bmatrix}15\\18\\20\\24\\30\\36\\40\\48\end{bmatrix}$$

and $\vec{v}\otimes\vec{w}\otimes\vec{u} \in \mathbb{R}^2\otimes\mathbb{R}^2\otimes\mathbb{R}^2 \simeq \mathbb{R}^8$.

The above example indicates that

$$\mathbb{R}^{N_1}\otimes\mathbb{R}^{N_2}\otimes\cdots\otimes\mathbb{R}^{N_k} \simeq \mathbb{R}^{N_1N_2\cdots N_k}\,,$$

namely that if we take the tensor product of k vector spaces then the result is a vector space which is the product of the dimensions of the constituents.

We are now ready to define tensor products abstractly, and to really appreciate what it means. Consider the following definition:

Definition 8 (Tensor product)**.** *Let V and W be real vector spaces. A **tensor product** of V and W is a vector space $V\otimes W$ together with a map*

$$\otimes : V\times W \to V\otimes W, \qquad (v,w)\mapsto v\otimes w,$$

that is bilinear in each argument, i.e. for all scalars $a,b,c\in\mathbb{R}$ and vectors $\vec{v},\vec{w},\vec{u}$,

$$(a\,\vec{v}+b\,\vec{w})\otimes\vec{u} = a\,(\vec{v}\otimes\vec{u}) + b\,(\vec{w}\otimes\vec{u}),$$
$$\vec{v}\otimes(b\,\vec{w}+c\,\vec{u}) = b\,(\vec{v}\otimes\vec{w}) + c\,(\vec{v}\otimes\vec{u}),$$

and in particular $(a\vec{v})\otimes\vec{w} = \vec{v}\otimes(a\vec{w}) = a(\vec{v}\otimes\vec{w})$. Concretely, one may construct $V\otimes W$ as the vector space spanned by formal symbols $v\otimes w$ modulo the above bilinearity relations.

To connect this with coordinates, fix bases $\{\vec{e}_i\}_{i=1}^N$ of $\mathbb{R}^N$ and $\{\vec{f}_j\}_{j=1}^M$ of $\mathbb{R}^M$. Then the NM simple tensors $\{\vec{e}_i\otimes\vec{f}_j\}_{i,j}$ form a basis of $\mathbb{R}^N\otimes\mathbb{R}^M$, and so $\dim(\mathbb{R}^N\otimes\mathbb{R}^M) = NM$. If $\vec{v} = \sum_i v_i\,\vec{e}_i$ and $\vec{w} = \sum_j w_j\,\vec{f}_j$, then

$$\vec{v}\otimes\vec{w} = \sum_{i,j} v_iw_j\,(\vec{e}_i\otimes\vec{f}_j)\,,$$

which recovers the stacking rules seen in the earlier examples and realizes the identification $\mathbb{R}^N\otimes\mathbb{R}^M \simeq \mathbb{R}^{NM}$.

Identifying $\mathbb{R}$ with the one–dimensional space spanned by 1, there are canonical isomorphisms $V\otimes\mathbb{R}\simeq V\simeq\mathbb{R}\otimes V$ given by $\vec{v}\otimes a\mapsto a\,\vec{v}$ and $a\otimes\vec{v}\mapsto a\,\vec{v}$. Hence $\mathbb{R}^N\otimes\mathbb{R}^1\simeq\mathbb{R}^N\simeq\mathbb{R}^1\otimes\mathbb{R}^N$.

Linear maps interact nicely with tensor products. If $A:\mathbb{R}^N\to\mathbb{R}^{N'}$ and $B:\mathbb{R}^M\to\mathbb{R}^{M'}$ are linear, there is a linear map $A\otimes B:\mathbb{R}^N\otimes\mathbb{R}^M\to\mathbb{R}^{N'}\otimes\mathbb{R}^{M'}$ defined by

$$(A\otimes B)(\vec{v}\otimes\vec{w}) = (A\vec{v})\otimes(B\vec{w})$$

which in matrix form is the familiar Kronecker product.

Remark 9 (Associativity of tensor products)**.** *For our purposes, it does not matter whether we first form $(V \otimes W)$ and then tensor with U from the right, or first form $(W \otimes U)$ and then tensor with V from the left. There is a canonical identification between*

$$(V \otimes W) \otimes U \quad \text{and} \quad V \otimes (W \otimes U),$$

and so we will simply write

$$V \otimes W \otimes U$$

without worrying about parentheses. This scales to many tensor factors. For a vector space V we write

$$V^{\otimes k} := \underbrace{V \otimes \cdots \otimes V}_{k \text{ copies}},$$

which has dimension $(\dim V)^k$ and a basis $\{\vec{e}_{i_1} \otimes \cdots \otimes \vec{e}_{i_k}\}$. We will use this to model multi–part systems: for example, a register of k N-ary variables naturally lives in $(\mathbb{R}^N)^{\otimes k} \simeq \mathbb{R}^{N^k}$.

As a word of caution, order still matters*. As we explained before, in general we have $\vec{v} \otimes \vec{w} \neq \vec{w} \otimes \vec{v}$. When we want to swap the order of a tensor product we will use the linear map $\mathsf{SWAP} : V \otimes W \to W \otimes V$, acting by*

$$\mathsf{SWAP} \cdot (\vec{v} \otimes \vec{w}) = \vec{w} \otimes \vec{v}\,.$$

In summary, associativity lets us ignore parentheses; SWAP lets us reorder factors when needed.

Going from the abstract back to the concrete, we have the example below:

Example 6. Suppose you are faced with this mess:

$$(a\,\vec{v} + b\,\vec{w}) \otimes (c\,\vec{s} + d\,\vec{t} + e\,\vec{u}) \otimes (f\,\vec{q} + g\,\vec{r})\,.$$

To expand it, what do you do? *Don't panic.* If you have a long list of things to do, just do them *one at a time.* Specifically in this case, use associativity to expand the bracketed terms first:

$$\begin{aligned}&\underbrace{(a\,\vec{v} + b\,\vec{w}) \otimes (c\,\vec{s} + d\,\vec{t} + e\,\vec{u})} \otimes (f\,\vec{q} + g\,\vec{r}) \\ &= (ac\,\vec{v} \otimes \vec{s} + ad\,\vec{v} \otimes \vec{t} + ae\,\vec{v} \otimes \vec{u} + bc\,\vec{w} \otimes \vec{s} + bd\,\vec{w} \otimes \vec{t} + be\,\vec{w} \otimes \vec{u}) \otimes (f\,\vec{q} + g\,\vec{r}).\end{aligned}$$

Now you can multiply through and expand the rest of the terms as

$$\begin{aligned}&acf\,\vec{v} \otimes \vec{s} \otimes \vec{q} + acg\,\vec{v} \otimes \vec{s} \otimes \vec{r} + adf\,\vec{v} \otimes \vec{t} \otimes \vec{q} + adg\,\vec{v} \otimes \vec{t} \otimes \vec{r} \\ &\quad + aef\,\vec{v} \otimes \vec{u} \otimes \vec{q} + aeg\,\vec{v} \otimes \vec{u} \otimes \vec{r} + bcf\,\vec{w} \otimes \vec{s} \otimes \vec{q} + bcg\,\vec{w} \otimes \vec{s} \otimes \vec{r} \\ &\quad + bdf\,\vec{w} \otimes \vec{t} \otimes \vec{q} + bdg\,\vec{w} \otimes \vec{t} \otimes \vec{r} + bef\,\vec{w} \otimes \vec{u} \otimes \vec{q} + beg\,\vec{w} \otimes \vec{u} \otimes \vec{r}\,,\end{aligned}$$

which is the desired expansion.

With some basic tensor product definitions at hand, we can now leverage them to discuss joint probability distributions in a slick vector space formalism.

Respecting historical tradition,[1] suppose we have two urns, where the first urn has N objects and the second urn has M objects. Suppose that the probability that we select one of the N items in the first urn is described by the probability

[1] See *Ars Conjectandi* by Jacob Bernoulli, published posthumously in 1713.

vector $\vec{p} \in \mathbb{R}^N$, and the probability that we select one of the M items in the second urn is described by the probability vector $\vec{q} \in \mathbb{R}^M$. Then if we select an item from the first urn followed by the second urn, what is the probability that we sampled item i from the first urn *and* item j from the second urn? The answer is encoded in the tensor product $\vec{p} \otimes \vec{q}$, and in particular its $(i-1)M + j$th entry:

$$[\vec{p} \otimes \vec{q}]_{(i-1)M+j} = p_i q_j \, .$$

We can extract this entry by dotting $\vec{p} \otimes \vec{q}$ against $\vec{e}_i^{\,T} \otimes \vec{e}_j^{\,T}$, namely

$$(\vec{e}_i^{\,T} \otimes \vec{e}_j^{\,T}) \cdot (\vec{p} \otimes \vec{q}) = p_i q_j \, .$$

The vector $\vec{p} \otimes \vec{q}$ is itself a probability vector living in $\Delta_{NM} \subset \mathbb{R}^{NM}$; thus it is a probability distribution on NM outcomes, as we wanted.

So far we have examined $\vec{p} \otimes \vec{q}$ which is a product distribution, assuming in our example that our sampling from each of the two urns is uncorrelated. Below we show in an example that convex combinations of tensor products can represent a correlated, joint distribution.

Example 7. Suppose the first urn has two items ($N = 2$), say a ring and a watch, and the second urn has three items ($M = 3$), say a tissue, a match, and a rubber band. The urns were prepared by the ghost of Jacob Bernoulli. We are told that with probability $1/3$ he put a ring in the first urn *and* a rubber band in the second urn, and with probability $2/3$ he put a watch in the first urn *and* a match in the second urn. Then the joint distribution over the urns is described by

$$\frac{1}{3} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \frac{2}{3} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1/3 \\ 0 \\ 2/3 \\ 0 \end{bmatrix} .$$

This distribution does not factorize into a tensor product of two individual vectors.

We abstract this example in the following remark.

Remark 10 (Joint distributions and multi-index notation)**.** *Given k probability spaces represented by $\Delta_{N_i} \subset \mathbb{R}^{N_i}$ for $i = 1, ..., k$, a distribution on the joint space is represented by*

$$\Delta_{N_1 \cdots N_k} \subset \mathbb{R}^{N_1 \cdots N_k} \simeq \mathbb{R}^{N_1} \otimes \cdots \otimes \mathbb{R}^{N_k} .$$

Product (independent) distributions have the special form $\vec{p}^{\,(1)} \otimes \vec{p}^{\,(2)} \otimes \cdots \otimes \vec{p}^{\,(k)}$, and general joint distributions are convex combinations of such products. For example, if $\vec{p}_i^{\,(j)}$ represents a distribution in $\mathbb{R}^{N_j}$, then

$$\sum_{i_1, i_2, \ldots, i_k} r_{i_1 i_2 \cdots i_k} \, \vec{p}_{i_1}^{\,(1)} \otimes \vec{p}_{i_2}^{\,(2)} \otimes \cdots \otimes \vec{p}_{i_k}^{\,(k)}$$

is a joint distribution so long as $r_{i_1 i_2 \cdots i_k} \geq 0$ for all $i_1, i_2, ..., i_k$ and additionally $\sum_{i_1, i_2, \ldots, i_k} r_{i_1 i_2 \cdots i_k} = 1$. Here we have used a multi-index notation, in which we are putting subscripts on subscripts; this is to avoid notation like $\sum_{a,b,c,\ldots} r_{abc\cdots}$ which do not specify the total number of subscripts, which in our case is k. (Moreover, there are only 26 letters of the Latin alphabet.) Multi-index notation may initially

seem like gross notation, but you will soon grow accustomed to it, like generations have before you.

Joint distributions interface nicely with the $\vec{1}^T$ row vector in a number of ways. For clarity, let us write $\vec{1}_N^T$ to denote the all-ones row vector with N entries. Then we have the nice identity

$$\vec{1}_{N_1}^T \otimes \vec{1}_{N_2}^T \otimes \cdots \otimes \vec{1}_{N_k}^T = \vec{1}_{N_1 N_2 \cdots N_k}^T .$$

Thus if $\vec{p}$ is a joint distribution living in $\Delta_{N_1 N_2 \cdots N_k}$, then we have

$$(\vec{1}_{N_1}^T \otimes \vec{1}_{N_2}^T \otimes \cdots \otimes \vec{1}_{N_k}^T) \cdot \vec{p} = \vec{1}_{N_1 N_2 \cdots N_k}^T \cdot \vec{p} = 1 .$$

We can also use the all-one row vector to formulate a nice way of computing marginal distributions. To illustrate, we proceed with the example below.

Example 8. Consider a joint distribution on $\Delta_6 \subset \mathbb{R}^2 \otimes \mathbb{R}^3$. Let us denote the joint distribution by $\vec{p}_{AB}$ where A represents the first subsystem of two items, and B represents the second subsystems of three items. Then we can write $\vec{p}_{AB}$ as

$$\vec{p}_{AB} = \begin{bmatrix} p_{AB}(1,1) \\ p_{AB}(1,2) \\ p_{AB}(1,3) \\ p_{AB}(2,1) \\ p_{AB}(2,2) \\ p_{AB}(2,3) \end{bmatrix} .$$

Suppose we want to marginalize over the second probability space (the one over three items). Letting $\mathbb{1}_N$ denote the $N \times N$ identity matrix, we marvel at the linear operator $\mathbb{1}_2 \otimes \vec{1}_3^T$ which maps $\mathbb{R}^2 \otimes \mathbb{R}^3 \to \mathbb{R}^2$. We marvel at it because applying the operator to $\vec{p}_{AB}$ we find

$$(\mathbb{1}_2 \otimes \vec{1}_3^T) \cdot \vec{p}_{AB} = \begin{bmatrix} p_{AB}(1,1) + p_{AB}(1,2) + p_{AB}(1,3) \\ p_{AB}(2,1) + p_{AB}(2,2) + p_{AB}(2,3) \end{bmatrix} = \begin{bmatrix} p_A(1) \\ p_A(2) \end{bmatrix} = \vec{p}_A$$

where $\vec{p}_A$ is the marginal distribution on the first subsystem A, which has two items.

The insight in the above example generalizes in the following way.

Remark 11 (Marginalizing any subset of subsystems)**.** *Let $\vec{p} \in \Delta_{N_1 \cdots N_k}$ be a joint distribution on k subsystems with sizes $N_1, ..., N_k$. For any subset $S \subseteq \{1, ..., k\}$, define the linear "marginalization" map*

$$\mathcal{M}_S := \bigotimes_{j=1}^{k} K_j = K_1 \otimes K_2 \otimes \cdots \otimes K_k , \qquad K_j = \begin{cases} \mathbb{1}_{N_j} & \text{if } j \in S \\ \vec{1}_{N_j}^T & \text{if } j \notin S \end{cases} ,$$

and so $\mathcal{M}_S : \mathbb{R}^{N_1 \cdots N_k} \to \mathbb{R}^{Prod_{j \in S} N_j}$. Then $\mathcal{M}_S \cdot \vec{p}$ is the marginal over the subsystems indexed by S.

To summarize, we have recast ordinary probability theory (on discrete probability spaces) in a linear-algebraic language, which has motivated us to develop the fundamentals of multi-linear algebra and tensor products. This mathematical technology certainly illuminates aspects of multi-linearity lurking in ordinary probability theory. But our true motivation was to set up probability theory in such a way as to make (finite-dimensional) quantum mechanics appear as a natural generalization, using many of the same ingredients. In this next section when

we introduce quantum mechanics, we will relentlessly capitalize on parallels with probability theory, but also take care to point out where such parallels break down.

2. Quantum theory in finite dimensions

We begin with a very brief history of quantum theory. Circa 1900 Max Planck studied blackbody radiation, and solved an inadequacy in the extant equations by stipulating that energy is quantized in units of his eponymous constant. Then in 1905, Einstein suggests that light itself is quantized as "photons", providing an explanation for the photoelectric effect. In the ensuing decade, Bohr makes a first pass at quantum theory (the so-called 'old' quantum theory), and correctly predicts the spectral lines of hydrogen. This first pass at quantum theory only goes so far, and a second pass is made in the 1920's. In 1924, de Broglie postulates that a particle with momentum p has 'wavelength' $\lambda = h/p$, which is soon confirmed by electron diffraction experiments. Thereafter, Heisenberg, Born, and Jordan developed matrix mechanics in 1925 (although they did not yet understand the connection to de Broglie). In 1926, Schrödinger leveraged de Broglie's insight to develop wave mechanics, and that same year showed the equivalence with matrix mechanics. That year as well, Born gave a 'probabilistic' interpretation of quantum mechanics which clarified its connections to measurable quantities in experiments. In 1927, Heisenberg wrote down his famous uncertainty principle. Most of the abstract mathematical foundations of quantum mechanics were consolidated by Dirac and von Neumann in the early 1930's, and Einstein-Podolsky-Rosen as well as Schrödinger highlighted the importance of entanglement in 1935. The year after in 1936, Birkoff and von Neumann investigated how quantum mechanics leads to a new form of logical reasoning that goes beyond classical Boolean logic; in hindsight this may be regarded as the first hint of the possibility of quantum computing (although it was not understood as such at the time).

Having completed our brief historical diegesis, we now turn to presenting the axioms of quantum mechanics. There are various ways of 'motivating' the axioms of quantum mechanics, although at some level they were *guessed* by very clever people and experimentally confirmed by very clever people (sometimes in the opposite order). We will, however, give some intuition. But first, a word of caution. When someone asks for a motivation for quantum mechanics in terms of classical mechanics, this is philosophically backwards; it would be like asking for a derivation of special relativity starting from Newton's equations. Indeed, just as special relativity reduces to Newtonian physics in a certain regime of validity, so too does quantum mechanics reduce to classical mechanics in a certain regime of validity. Nonetheless, we will proceed with an idiosyncratic way of 'guessing' some of the axioms of quantum mechanics starting from classical intuitions.

2.1. Mechanics on ℓ^p spaces: from classical to quantum

Let us begin by contemplating the salient mathematical structures undergirding the dynamics of probability distributions discussed above. For this, it is useful to have the following definition:

Definition 12 (Normed vector space)**.** *Let V be a vector space over a field K; we will consider either $V = \mathbb{R}^N$ (with $K = \mathbb{R}$), or $V = \mathbb{C}^N$ (with $K = \mathbb{C}$). A* ***normed***