Making sense of
Phase-Type and Matrix-Exponential Distributions

Simon Savine

September 2024

1 Matrix exponential

Given a square matrix M and a smooth function f (a real function of one real variable), what sense can we give to the expression $f(M)$? In other terms, what does it mean to apply a function to a matrix? One natural semantics would be to apply f coefficient-wise to all the entries of M. This is implemented, for example, in NumPy, Python's de-facto math library: the expression "np.exp(M)" returns a matrix of the same shape, whose coefficients are the exponentials of the coefficients of M.

This is not (at least not exactly) how mathematicians define $f(M)$, although, as we will see, this is not unrelated.

Denote $M = (m_{ij})$ and consider the expression M^2: it could mean (m_{ij}^2), as in NummPy's "M**2", or it could mean the matrix product MM (or "M @ M" in NumPy lingo). Mathematicians chose the second definition. Why? Because if M represents some linear map g in a given basis, then $M^2 = MM$ represents the *composition* $g \cdot g$ (apply g and apply g again to the result) in the same basis. Hence, this definition allows to consistently reason about matrices or linear maps, contrarily to the member-wise definition.

More generally, we define the expression M^n to mean $M^n = MM...M$ (multiply n times by itself in the sense of a matrix product).

Now, we can easily extend to arbitrary smooth functions, because these are combinations of power functions, by Taylor expansion:

$$f(x) = \sum_i c_i(f) x^i$$

We *define*:

$$f(M) = \sum_i c_i(f) M^i = \sum_i c_i(f) MM...M \text{ (i times)}$$

In particular, with $f = exp$ (and recalling the Taylor expansion of $exp(x) = \sum_i x^i / i!$) we have:

$$exp(M) = \sum_i c_i(exp) M^i = \sum_i M^i / i!$$

How is this useful? Matrix functions $f(M)$ defined in this manner share many properties with their "vanilla" counterparts $f(x)$, in particular in terms of differentiation, integration, etc. For example, we are going to derive the intuitive, and very useful result:

$$\frac{\partial exp(Mx)}{\partial x} = M exp(Mx) = exp(Mx) M$$

To better understand why so many properties of $f(x)$ carry over to $f(M)$, consider the special case of a diagonalisable matrix M. Then, it exists an orthonormal matrix P (whose columns are eigenvectors of M) and a diagonal matrix D (whose entries are corresponding eigenvalues) such that $M = PDP^T$.

Then:

$$\begin{aligned} M^n = MM...M = PDP^T PDP^T ...PDP^T \\ = PD(P^T P)D(P^T P)...(P^T P)DP^T = PDI_n DI_n...I_n DP^T = PD^n P^T \end{aligned}$$

where I_n is the identity matrix and $P^T P = PP^T = I_n$ since $P^T = P^{-1}$ by definition of an orthonormal matrix.

Now, pause for a minute to appreciate that for *diagonal* matrices, the matrix product and the coefficient-wise product coincide: $D^n = (d_{ij}^n)$.

It immediately follows that:

$$f(M) = \sum_i c_i(f)M^i = \sum_i c_i(f)M^i = P[\sum_i c_i(f)D^i]P^T = Pf(D)P^T = P[f(d_{ij})]P^T$$

$f(M)$ is the matrix that has the same eigenvectors as M, and eigenvalues obtained by application of f to the eigenvalues of M.

In particular,

$$exp(M) = Pexp(D)P^T = Pexp[(d_{ij})]P^T$$

Now, let us derive the differential of $exp(Mx)$ (here ' means derivative wrt x):

$$\begin{aligned} exp(Mx)' = Pexp(Dx)'P^T = Pexp(Dx)DP^T = Pexp(Dx)I_n DP^T \\ = Pexp(Dx)P^T PDP^T = [Pexp(Dx)P^T][PDP^T] = exp(Mx)M \end{aligned}$$

where we also note that, since diagonal matrices commute, $exp(Mx)' = exp(Mx)M = Mexp(Mx)$. M commutes with $exp(M)$.

This result, and similar ones, carry over to the general case where M may not be diagonalisable, with proof left as an exercise.

References:

- Wikipedia article https://en.wikipedia.org/wiki/Matrix_exponential

- A *stellar* intruduction by 3blue1brown: https://www.youtube.com/watch?v=O85OWBJ2ayo

2 Application to probability and stochastic processes

Consider now some process that finds itself in one of n different states $S_1, ..., S_n$ at every time t, with probabilities given by a vector P_t. So $P_t[1]$ is the probability of being in state S_1 at time t, $P_t[2]$ is the probability of being in state S_2, etc, and $P_t[n]$ is the probability of being in state S_n. P_0, the distribution of the initial state at time 0, is given.

The state transition matrix $R = (r_{ij})$ contains the probabilities r_{ij} to jump from state S_j to state S_i between times t and $t + dt$. Its row i contains the probabilities of jumping into state S_i from states S_1 to S_n. Its column S_j contains the probabilities of jumping from state S_j into states S_1 to S_n and sums to 1. Its diagonal contains the probabilities of staying in place. The probabilities of jumping to a different state scale with dt, so R is the form (here $n = 3$ for illustration):

$$R = \begin{pmatrix} 1-(r_{21}+r_{31})dt & r_{12}dt & r_{13}dt \\ r_{21}dt & 1-(r_{12}+r_{32})dt & r_{23}dt \\ r_{31}dt & r_{32}dt & 1-(r_{13}+r_{23})dt \end{pmatrix}$$

$$= In + \begin{pmatrix} -(r_{21}+r_{31}) & r_{12} & r_{13} \\ r_{21} & -(r_{12}+r_{32}) & r_{23} \\ r_{31} & r_{32} & -(r_{13}+r_{23}) \end{pmatrix} dt$$

where In is the identity matrix in dimension n (here, 3).

Now, let us derive the derivative dP_t/dt of state probabilities. Let us denote X_t the (random) state at time t.

$$\begin{aligned} P_{t+dt}[i] &= Pr(X_{t+dt} = S_i) && \text{by definition of } P_{t+dt} \\ &= \sum_j Pr(X_{t+dt} = S_i | X_t = S_j) Pr(X_t = S_j) && \text{by the law of total probabilities} \\ &= \sum_j r_{ij} P_t[j] && \text{by definition of } r_{ij} \end{aligned}$$

Hence: $P_{t+dt} = RP_t$. And it follows that:

$$\begin{aligned} \frac{dP_t}{dt} &= \frac{P_{t+dt} - P_t}{dt} \\ &= \frac{RP_t - P_t}{dt} \\ &= \frac{\left[In + \begin{pmatrix} -(r_{21}+r_{31}) & r_{12} & r_{13} \\ r_{21} & -(r_{12}+r_{32}) & r_{23} \\ r_{31} & r_{32} & -(r_{13}+r_{23}) \end{pmatrix} dt \right] P_t - P_t}{dt} \\ &= \begin{pmatrix} -(r_{21}+r_{31}) & r_{12} & r_{13} \\ r_{21} & -(r_{12}+r_{32}) & r_{23} \\ r_{31} & r_{32} & -(r_{13}+r_{23}) \end{pmatrix} P_t \end{aligned}$$

Let us call this matrix Q:

$$Q = \begin{pmatrix} -(r_{21}+r_{31}) & r_{12} & r_{13} \\ r_{21} & -(r_{12}+r_{32}) & r_{23} \\ r_{31} & r_{32} & -(r_{13}+r_{23}) \end{pmatrix} = \frac{R - I_n}{dt}$$

Note that since the columns of R sum to 1, those of Q sum to 0.

If these were numbers ($n = 1$), this would read as a textbook differential equation $f' = qf$, with well-known solution $f(t) = exp(qt)f(0)$. The same applies to matrices, as we have seen in the previous section. Hence:

$$P_t = exp(Qt)P_0$$

where Qt is the matrix Q scaled by time t, and $exp(Qt)$ is its matrix exponential.

Voila. Term-t state probabilities are computed with matrix exponentials.

We derived the distribution of state probabilities at all times t, and it involves matrix exponentials, but this is not (yet) the phase-type/matrix-exponential distribution.

For that, we need an additional bit of logic and consider state S_n as absorbing. This simply means that there is no escape from it. The last column of the transition matrix R is $(0, 0, ..., 0, 1)^T$: $r_{in} = 0$ for $i < n$ and $r_{nn} = 1$. If you are in state S_n, you have zero probability to jump out of it, you stay in state n with probability 1.

Now what is the probability distribution of τ, the first time you hit the absorbed state?

First, notice this:

$$CDF_\tau(t) = Pr(\tau \leq t) = P_t[n]$$

You have been absorbed on or before t if and only if you are in the absorbed state at t.

Finally: $P_t[n] = (0, 0, ..., 0, 1)P_t = \alpha P_t$ where α is the row vector of all zeroes except its last entry 1, and putting it all together:

$$CDF_\tau(t) = \alpha P_t = \alpha exp(Qt)P_0$$

And we can easily compute the density of τ by differentiation:

$$PDF_\tau(t) = \frac{CDF_\tau(t)}{\partial t} = \alpha exp(Qt)QP_0 = \alpha exp(Qt)Q_0$$

where $Q_0 = QP_0$. This distribution is called "phase-type" distribution, and we can easily compute its mean, variance, etc. (left as exercise).

Exercise: prove that this density integrates to 1.

Now what are "matrix-exponential" distributions? Note that in the definition above, there are constraints on the parameters: P_0 must be a vector of probabilities, that is, non-negative entries summing to 1. And R must be a transition matrix, with non-negative entries and columns summing to 1.

Suppose that we decide to release those constraints, and simply reuse the expression of the density $PDF_\tau(t) = \alpha exp(Qt)Q_0$ with arbitrary α, Q and Q_0. Then it is no longer called a "phase-type" distribution but a "matrix-exponential" distribution. It loses the physical interpretation of the absorption time distribution of a state transition process, and it may well not be a probability distribution at all (we may end-up with negative densities and/or densities not integrating to 1), but those things have interesting mathematical and computational properties, and are being researched presently for this reason. But this is a story for another day.

References:

- Wikipedia article https://en.wikipedia.org/wiki/Phase-type_distribution

- Andras Horvath, Marco Scarpa, Miklos Telek, Phase Type and Matrix Exponential distributions in stochastic modelling: https://webspn.hit.bme.hu/~telek/cikkek/horv16h.pdf