Probability review (6.431x) notes.

David G. Khachatrian

September 21, 2019

1 Preamble

This was made long after having taken the course. It will likely not be exhaustive.

2 Probability models and axioms.

A *sample space* Ω is a list of possible outcomes that are mutually exclusive and collectively exhaustive. An *event* A is a subset of Ω.

Define a *probability law* $\mathbb{P} : A \to [0, 1]$ maps events to probabilities which follows the axioms of probability (e.g., countable additivity axiom).

More rigorously, a *probability law* defines a sigma algebra over Ω (i.e., $(\Omega, \{A\})$ is a measurable space).

Countable Additivity Axiom If $A_1, A_2, \cdots$ is a *countable sequence* of disjoint events, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Note: This only makes sense for countable events. For continuous sample spaces, any event A with probability $P(A) > 0$ cannot be enumerated; it must be "inherently" continuous. This is another consequence of the difference between "countable" and "uncountable" infinities.

3 Unit 2: Conditioning and independence

3.1 Conditioning and Bayes' rule

Probability of B conditioned on A is the probability that the event B occurs, given that we know that A has occurred. It is denoted $P(B \mid A)$.

We can write the probability that both A and B occur in two ways:

$$P(A \mid B)P(B) = P(B \mid A)P(A) = P(A \cap B)$$

Total Probability Theorem We can describe the probability of an event B by "conditioning out" the probability. We consider the probabilities that a different event A occurs, then the probability that B occurs given that A occurred:

$$P(B) = \sum_i P(B \mid A_i)P(A_i)$$

We can use the Total Probability Theorem to write *Bayes' rule*, which allows us to update our prior *beliefs* about the world, B, after viewing *events or evidence*, E:

$$P(B \mid E) = \frac{P(E \mid B)P(B)}{P(E)}$$
$$= \frac{P(E \mid B)P(B)}{\sum_i P(E \mid B_i)P(B_i)}$$

3.2 Independence

Two *events* A and B are considered *independent* iff $P(A \cap B) = P(A)P(B)$. The idea is that even if you know that B occurred, the probability of A occurring stays the same ($P(A \mid B) = P(A)$), and vice-versa.

Events can be *conditionally independent, given* C, meaning $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$.

There is a difference between *mutual independence* and (for example) *pairwise* independence.

(Mutual) independence. When events $A_1, A_2, \cdots, A_m$ are *(mutually) independent*, then

$$P(A_i \cap A_j \cap \cdots \cap A_m) = P(A_i)P(A_j) \cdots P(A_m)$$

for *any* distinct indices $i, j, \cdots, m$. That is to say, all subsets of $\{A_1, A_2, \cdots, A_m\}$ are independent.

Note! It may not be immediately obvious whether events are mutually/pairwise independent. It doesn't hurt to check using the above definition!

4 Unit 3: Counting

(Counting isn't necessarily easy.)

Counting (or "combinatorics") can provide exact answers to probability questions, when the probability problem can be described using a *discrete uniform law*, i.e., $P(A) = \frac{\text{\# elements in } A}{\text{\# elements in } \Omega}$.

Fundamental counting principle. If there are n_1 options for a first choice, n_2 options for the second choice, etc., and all combinations of options are possible and distinguishable from one another, then total number of combinations $= \prod_i n_i$.

When all options are *not* mutually distinguishable, one should be sure to account for over/undercounting. A commonly appearing corrective factor is the *multinomial coefficient*. If there are n_1 mutually indistinguishable balls, n_2 mutually indistinguishable blocks, etc., and $\sum_i n_i = N$, then the number of distinguishable outcomes is the multinomial coefficient $\binom{N}{n_1, n_2, \ldots} = \frac{N!}{\prod_i n_i!}$.

Another potentially useful method, often useful when partitioning items into different "bins"/containers, is the *stars-and-bars approach*.

5 Random variables.

A *random variable (RV or r.v.)* X associates events in Ω to real values $x \in \mathbb{R}$. More explicitly, say $\mathcal{F}$ has the sigma-algebra/sigma-field induced by a sample space Ω. X is a function $X : \mathcal{F} \to \mathbb{R}$, X maps every element/event ω of $\mathcal{F}$ to a real number, and "$X = x$" is shorthand for the event $\{\omega : X(\omega) = x\}$.

One descriptor for random variables is its *cumulative distribution function (CDF)*, defined as $F_X(x) = Pr[X \leq x]$.

For discrete random variables (the range of X is countable), the *probability mass function (PMF)*, defined as $p_X(x) = Pr[X = x]$, fully characterizes the random variable.

For continuous random variables, we define a *probability density function (PDF)* f_X based on the CDF: $f_X(x) = \frac{dF_X}{dx}(x)$. The following intuition can be helpful: "$f_X(x)dx \approx Pr[X \in (x, x + dx)]$". (Remember: for continuous random variables, $Pr[X = x] = 0$ for any value x; only continuous ranges have nonzero probability).

One can define *functions of random variables*, which are themselves random variables. As an example, if $g(x) = x^2$, then if X is a random variable, $g(X) = X^2$ is a function of a random variable.

The *expectation of a random variable* is denoted $E[\cdot]$. For a discrete random variable X, it can be written $E[X] = \sum_{x \in X} Pr[X = x]x = \sum_{x \in X} p_X(x)x$. For continuous RVs, we have $E[X] = \int_{x \in X} f_X(x)xdx$.

Note that both $\sum$ and $\int$ are linear operators, so expectations are as well, i.e., $E[aX + b] = aE[X] + b$.

The law of the unconscious statistician is incredibly useful for calculating expectations. It states for X and $Y = g(X)$, $E[Y] = \sum_{y \in Y} p_Y(y)y = \sum_{x \in X} p_X(x)g(x)$. This is not immediately implied; it works because of a bijection one can make between values/probabilities in X and values/probabilities in $Y = g(X)$, as well as the linearity of expectations. (True to its name, the law is often invoked without anyone ever giving it much thought.)

For example, we can write the expectations of *functions* of random variables $g(X)$ as $E[g(X)] = \sum_{x \in X} p_X(x)g(x)$.

The *mean* of a random variable X is simply $E[X]$. The *variance* of a random variable X is defined as $var(X) := E[(X - E[X])^2]$. This can be rewritten as $var(X) = E[X^2] - (E[X])^2$.

A random variable conditioned on an event A (assuming $P(A) > 0$) has a new probability distribution $p_{X|A}(x) = Pr[X = x \mid A]$ and can be written $p_{X|A}(x) = \begin{cases} 0, x \notin A \\ p_X(x)/P(A), x \in A \end{cases}$

Similarly, we can write for a RV X conditioned on another RV Y that $p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}, p_Y(y) > 0$.

The *total expectation theorem* is essentially another form of the law of total probability. Given mutually exclusive and collectively exhaustive events $\{A_i\}$, we have: $E[X] = \sum_i P(A_i)E[X \mid A_i]$.

A *joint probability distribution* over random variable sets $\{X, Y, Z\}$ is written $Pr[X = x, Y = y, Z = z] = p_{X,Y,Z}(x, y, z)$. They are independent if $p_{X,Y,Z}(x, y, z) = p_X(x)p_Y(y)p_Z(z)$. If they are independent, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ and $var(X + Y) = var(X) + var(Y)$.

One can create *marginalized probability distributions* to get the dependence on just a subset of RVs. We do this by integrating over the RVs to be marginalized. In a sense, we're using the law of total probability: $Pr[X = x] = \sum_i Pr[X = x \mid Y = y_i]Pr[Y = y_i] = \sum_i Pr[X = x, Y = y_i]$.

We can "mix" PDFs and PMFs in Bayes' rule: use f for continuous r.v.'s and p for discrete r.v.'s.

6 Indicator Variables and problem solving techniques.

A very powerful method to solve many problems is to use *indicator variables* to represent events, or vice-versa. An indicator variable for an event A is a random variable X_A such that $X_A = \begin{cases} 1, A \text{ happens} \\ 0, A \text{ doesn't happen} \end{cases}$.

Note that $E[X_A] = P(A)$. Also, we can describe the complement of A as $1 - X_A$.
Note also that we can describe $Pr(A \cap B) = E[X_A X_B]$. Using De Morgan's laws, we can show that $A \cup B = (A^C \cap B^C)^C$ and so $P(A \cup B) = E[1 - (1 - X_A)(1 - X_B)]$. And so on.

Conditioning a sequence of indicator variables on each other is another useful skill. For example, what is the expected number of dice rolls until all sides of a fair k-sided die is seen? We can write $X = X_1 + X_2 + \cdots + X_k$, where $X_i \sim Geom(\frac{k-i+1}{k})$ and compute its expectation that way.

When there is more than one source of randomness (e.g. a randomly selected number of (random) dice rolls), the *law of iterated expectations* (mentioned before) is often useful. $E_X[X] = E_Y[E_X[X \mid Y]]$. For the variance, a slightly more complicated formula (the *law of total variance*) is needed: $Var(X) = Var(E_X[X \mid Y]) + E[Var(X \mid Y)]$.

We can describe the *variance of a sum of r.v.'s* as follows:

$$Var(X_1 + X_2 + \cdots + X_n) = \sum_i Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$$

Here, the *covariance between two r.v.'s* is described as:

$$Cov(X, Y) := E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

The *correlation between two r.v's* is a scaled version of the covariance:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

(The *standard deviation of* X is denoted $\sigma_X := \sqrt{Var(X)}$.)

When r.v.'s are independent, they are uncorrelated (converse **not** necessarily true), which implies $Cov(X, Y) = 0$.

6.1 Derived distributions.

The PDF/CDF of $Y = g(X)$ can be described by the PDF/CDF of X by considering:

$$\begin{aligned} F_Y(y) &= Pr[Y \leq y] \\ &= Pr[g(X) \leq y] \\ &= Pr[X \leq g^{-1}(y)] \\ &= F_X(g^{-1}(y)) \end{aligned}$$

For monotonic random variables where $Y = g(X)$, we can simplify the calculation to:

$$\begin{aligned} f_X(x)dx &= f_Y(y)dy \\ f_X(x) &= f_Y(y)\frac{dy}{dx} \\ &= f_Y(g(x))\frac{dg(x)}{dx} \quad (y = g(x)) \end{aligned}$$

(Fun Fact: The justification and form is quite reminiscent to those for the Legendre transformation.)

For two independent r.v's X and Y, we can describe the probability distribution of $Z = X + Y$ via a *convolution*: $p_Z(z) = \sum_{x \in X} p_X(x)p_Y(z - x)$. (Use integration instead of summation for continuous r.v.'s.)

An important consequence of the nature of convolutions: the sum of independent Normal distributions is also Normal.

7 Bayesian Inference

Bayesian inference treats an unknown Θ *as a random variable* (and *not* as a fixed parameter of unknown value). This means Θ has an associated *prior distribution* p_Θ. We observe values of X (so $X = x$), and we update our prior beliefs (via Bayes' rule) to get $p_{\Theta|X}(\cdot \mid X = x)$. Bayes' rule would be:

$$\begin{aligned} p_{\Theta|X}(\theta \mid x) &= \frac{p_{X|\Theta}(x \mid \theta) p_\Theta(\theta)}{p_X(x)} \\ &= \frac{p_{X|\Theta}(x \mid \theta) p_\Theta(\theta)}{\sum_j p_{X|\Theta}(x \mid \theta_j) p_\Theta(\theta_j)} \end{aligned}$$

The full output is a posterior $p_{\Theta|X}$, but often we want point estimates. This could be:

- the *maximum a posteriori probability (MAP)* estimate ($\max_\theta p_{\Theta|X}(\theta \mid x)$, or
- the *conditional expectation* $E_\Theta[\Theta \mid X = x]$; this minimizes conditional mean squared error (MSE), i.e., provides the least-mean-squares (LMS) solution. Mathematically, $\hat{\Theta}^* = \text{argmin}_{\hat{\Theta}} E[(\hat{\Theta} - \Theta)^2 \mid X = x] = E_\Theta[\Theta \mid X = x]$.
- the *expectation*$E_\Theta[\Theta] = E_X[E_\Theta[\Theta \mid X]]$, which minimizes the overall MSE $\hat{\theta}^* = \text{argmin}_{\hat{\theta}} E[(\Theta - \hat{\theta})^2]$. (This would probably be best used if you have a strong understanding of X's distribution and aren't too concerned with the particular realizations you observed.)

7.1 Linear models with Normal Noise

Say that at time $t = 1, 2, \cdots$ we observe X_t, which we model as coming from signals $a_{tj}\Theta_j$ (a's known) and noise W_t, all independent and added linearly, i.e.,

$$X_t = \sum_j a_{tj}\Theta_j + W_t$$

Further suppose all the $W_t \sim N(0, \sigma_t^2), \Theta_j \sim N(x_0, \sigma_0^2)$ are independent and Normally distributed. Then we end up with a likelihood function $L(\theta_j \mid x_t) = c(x) \exp(-quadratic(\theta))$. If you'd like point estimates, you can perform the following steps:

- MAP: Find peak of $L(\theta \mid x)$. Find and compare values of solutions for $\frac{dL}{d\theta}(\theta) = 0$.
- LMS: Find $E_\Theta[\Theta \mid X = x]$, i.e., calculate the mean of the posterior distribution.

For the above problem, you'll find that the answer is

$$\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = E[\Theta \mid X = x] = \frac{\sum_{t=0}^n \frac{x_t}{\sigma_t^2}}{\sum_{t=0}^n \frac{1}{\sigma_t^2}}$$

where "x_0, σ_0^2" refer to the mean and variance of the prior distribution of Θ. (Intuitively, "our prior distribution acts as if it were 'the first observation'".)

7.2 Least-Mean-Squares and Linear LMS

We can rewrite the mean-squared-error $E[(\Theta - \hat{\theta})^2]$ as $Var(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2 = Var(\Theta) + Bias(\Theta, \hat{\theta})^2$. Even if you have a model with no bias, *on expectation* you will have MSE of $var(\Theta)$.

If we're minimizing conditional MSE $E[(\Theta - \hat{\Theta})^2 \mid X]$, the best estimator is $\hat{\Theta} = E[\Theta \mid X]$. Calling the error $\epsilon = \hat{\Theta} - \Theta$, we have $E[\epsilon \mid X = x] = 0$, $Cov(\epsilon, \hat{\Theta}) = 0$, $var(\Theta) = var(\hat{\Theta}) + var(\epsilon)$.

If we limit ourselves to estimators $\hat{\Theta} = aX + b$ (for parameters we calculate based on the data, a and b), we get

$$a = \frac{Cov(\Theta, X)}{var(X)} = \rho\frac{\sigma_\Theta}{\sigma_X}, b = E[\Theta] - aE[X]$$

We can show that $E[(\hat{\Theta_L} - \Theta)^2] = (1 - \rho^2)var(\Theta)$.

7.3 Improper, Uninformative, and Conjugate Priors

(From "Fundamentals of Statistics" course.)

Note that your prior distribution needn't be integrable as long as the posterior distribution is integrable. Such priors are called *improper priors*.

7.3.1 Jeffreys prior: a "covariant" prior.

Keep in mind that a *change of variables* from p to q implies changing f_P to f_Q such that the following holds:

$$f_P(p)dp = f_Q(q)dq$$

Let's say you'd like to use the machinery of Bayesian inference and have a model M with unknown parameters (as RVs) Θ, but have no reason to believe one value of Θ is more likely than the other. What we mean to say is that the prior should have no effect on the posterior distribution, that only the observations should affect the shape, that the prior doesn't give some parameters an "unfair advantage" in the posterior.

This is trickier to pin down than you might expect. One way of making this more concrete is the following: We want a prior $JP_\Theta(\theta)$ so that we can take the following two routes:

1. From $(JP_\Theta, L_{X|\Theta})$, calculate posterior under θ, $\pi_{\Theta|X}$, then change variables to $\pi^{(1)}_{\Phi|X}$.
2. From $(JP_\Theta, L_{X|\Theta})$, change variables to $(JP_\Phi, L_{X|\Phi})$ then calculate posterior under ϕ, $\pi^{(2)}_{\Phi|X}$.

And in both cases, the posteriors are equal, i.e. $\pi^{(1)}_{\Phi|X} = \pi^{(2)}_{\Phi|X}$. This suggests that it doesn't matter what functional form of the parameter we're studying – "equivalent" (bijective) values have equivalent probabilities before *and* after observing the data, i.e. $\pi_\Theta(\theta)d\theta = \pi_\Phi(\phi)d\phi$ and also $\pi_{\Theta|X}(\theta \mid x)d\theta = \pi_{\Phi|X}(\phi \mid x)d\phi$ for any bijective function $g : \theta \to \phi$ where $\phi = g(\theta)$.

The appropriate prior distribution may not be immediately obvious. It turns out one should use a form of the *Fisher information* $I_{n=1}(\theta)$, specifically

$$JP_\Theta(\theta) = \sqrt{\det(I_1(\theta))}$$

to achieve this property. $J(\theta)$ is called the *Jeffreys prior*. What's perhaps most important about this choice of prior is that the reparameterization also works on the *posterior distribution* $\pi(\cdot \mid X) = \pi(\cdot)L(X \mid \cdot)$, i.e.,

$$\pi(\theta \mid X)d\theta = \pi(\phi \mid X)d\phi$$

The Fisher information can be written as

$$I_{n=1}(\theta) = Var_X\left(\boldsymbol{\nabla}_\theta\left(\ln(L_{n=1}(\theta; X))\right)\right) = -E_X\left[\mathbf{H}_\theta\left(\ln(L_{n=1}(\theta; X))\right)\right]$$

(H is the Hessian, or "second-derivative matrix".)

7.3.2 Conjugate priors

Some priors "play nicely" with certain sorts of data. By this we mean the prior and posterior distributions remain in the same family, which can help keep things computationally tractable. As one example, if we are interested in $X_i \sim Ber(p)$, if we chose the *Beta distribution* $Beta(a, b)$ as our prior for p, the posterior distribution for p given the data would end up being $Beta(a + \sum_i X_i, b + (n - \sum_i X_i))$, which is still a Beta distribution. This is considered a *conjugate prior* for the parameter in question. Different models for X_i suggest different conjugate priors.

8 Classical Statistics

8.1 Inequalities

Markov inequality For any $X \geq 0$ and $a > 0$, $Pr[X \geq a] \leq \frac{E[X]}{a}$.

Proof: Set $Y = a(if X \geq a); 0o/w$. $E[Y] = aPr[X \geq a] \leq E[X]$.

Chebyshev inequality Assume X has finite mean μ and variance σ^2. Then $Pr[|X - \mu| \geq c] \leq \frac{\sigma^2}{c^2}$.

Proof: Consequence of Markov inequality: $Pr[(X - \mu)^2 \geq c^2] \leq \frac{E[(X-\mu)^2]}{c^2}$.

Hoeffding's inequality Let X_i be *bounded and (mutually) independent*. Then the tail bound is exponential. In a particular case, for $X_i = 1, -1$ with equal probability, $Pr[\bar{X}_n \geq a] \leq e^{-na^2/2}$.

Proof:

$$Pr[\sum_i X_i \geq a] = Pr[e^{t\sum_i X_i} \geq e^{ta}] \leq \frac{E[e^{t\sum_i X_i}}{e^{ta}}$$

for all $t > 0$. Minimize w.r.t t.

(Note: This applies even for small n. It is more conservative than the Central Limit Theorem, but the r.v.'s must be *bounded* as well as independent.)

Weak Law of Large Numbers For all $\epsilon > 0$, $Pr[|\bar{X}_n - \mu| \geq \epsilon] \rightarrow^{(p)}_{n\rightarrow\infty} 0$.

Proof: Apply Chebyshev inequality to $\bar{X}_n$. (Less stringent conditions can also be used, but more work involved.)

8.2 Central Limit Theorem

Consider iid random samples X_i with mean μ and variance σ^2. We have:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \rightarrow^{(d)}_{n\rightarrow\infty} N(0, 1)$$

In terms of sums (and a "moving" convergence), we can write

$$\sum_i X_i \rightarrow^{(d)}_{n\rightarrow\infty} N(n\mu, n\sigma^2)$$

In practice, the distribution is often roughly Normal at or above $n \approx 30$ (symmetry and unimodality among the r.v.'s probability distribution helps).

When applying the CLT to discrete r.v's, it may help to use the "half-correction". Let $S_n = \sum_i X_i$. Since $Pr[S_n \leq k] = Pr[S_n \leq (k+1)]$, you can use the Normal approximations corresponding to $Pr[Z_n \leq (k+1/2)]$ and often get more accurate approximations.

(We defer further discussion of CLT and classical statistics to the "Fundamentals of Statistics" review notes.)

9 Bernoulli and Poisson Processes

Both are examples of *stochastic processes*. These can be viewed in a number of ways:

1. a sequence of r.v.'s indexed by "time": $X_1, X_2, \cdots$
2. a probability distribution over a sample space (a set of infinitely long sequences)

A useful way to determine the distributions of some r.v.'s of interest can be to *determine the CDF first* and then calculate the PDF/determine the PMF from it.

9.1 Bernoulli process

A Bernoulli process with parameter p (Bernoulli(p)) can be characterized by the following: for any i, $X_i \sim Ber(p)$ and X_{i+1} are independent of $i, X_1, \cdots, X_i$. This suggests

- the *fresh start* property. If you begin observing the process at time T, you will know the same amount about future events as someone who started observing at time 0.
- The *interarrival time* between successes follows a $Geom(1-p)$.

Time until the k'th success can be calculated as:

$$Pr[k-1 \text{ arrivals in time } t-1] \times Pr[\text{arrival at time } t] = \binom{t-1}{k-1} p^{k-1}(1-p)^{t-k} \times p$$

The above can be called a *Pascal distribution*.

Splitting a Bernoulli process into two streams using a biased coin $Ber(q)$ will split $Bernoulli(p)$ into $Bernoulli(pq)$ and $Bernoulli(p(1-q))$. These split streams are *not* independent of each other, since in this discrete setting, if Stream 1 has an arrival, Stream 2 cannot.

9.2 Poisson Process

We discuss a Poisson process with arrival rate λ (PoisProc(λ)) based on the following assumptions:

- number of arrivals in disjoint time intervals are independent.
- Let $P(k, \tau)$ be the probability of k arrivals in interval of duration τ. For very small δ,

$$P(k, \delta) \approx \begin{cases} 1 - \lambda\delta, k = 0 \\ \lambda\delta, k = 1 \\ 0, k > 1 \end{cases}$$

The idea is that second-order values of order $O(\delta^2)$ are so small as to be negligible.

As a consequence of the above, the Poisson Process also has the "fresh-start" property.
These assumptions also mean that $p(k, \tau)$ has the PMF of $Pois(\lambda\tau)$.

To determine time of k'th arrival, we can say:

$$\begin{aligned} f_k(t)\delta &\approx Pr[k\text{'th arrival at time } t] \\ &\approx Pr[(k-1) \text{ arrivals in the interval just before time } t] \times Pr[\text{arrival in time } (t, t+\delta)] \\ &\approx \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} \times \lambda\delta \end{aligned}$$

$f_k(t)$ above described the PDF for an *Erlang distribution of order k*. Note that $Erlang(\lambda, k = 1) = Exp(\lambda)$.

If you merge PoisProc(λ_1) and PoisProc(λ_2), you simply have a PoisProc($\lambda_1 + \lambda_2$).

Splitting a Poisson Process creates two new Poisson processes. Interestingly (due to the curiosities of making time intervals infinitesimally small), these two new processes *are independent*.

10 Markov chains

Markov chains are another stochastic process. A fundamental assumption is the *Markov property*:

$$Pr[X_{n+1} = j \mid X_1, X_2, \cdots, X_n] = Pr[X_{n+1} - j \mid X_n]$$

That is to say, all relevant information about the history is contained in the current state.

One can distinguish between *transient* states and *recurrent* states. A state i is recurrent if "starting from i, whatever path you take, there will be a path that lets you return to i". States that don't satisfy this property are transient. One can have "islands/groups" of states where the groups are isolated from one another, but each state is well-connected within the group; these are called "recurrent classes" of the Markov chain.

It can also be useful to note whether the states in a recurrent class are *periodic (of order $d > 1$)*. This implies that one can separate/color the states so that every time, the state moves from one color to the next in the sequence.

10.1 Matrix description and long-term behavior.

We can describe discrete-time, finite-state Markov chains as a *transition probability matrix* M, where $M_{i,j}$ describes the probability of going from State i to State j. (Each row of M must sum to 1.)

A key recursion for Markov chains is that

$$r_{i,j}(n) = \sum_k r_{i,k}(n-1)p_{k,j}$$

.

We can use this to determine long-term behavior. We assume that there's a "long-run frequency of visiting State j" π_j, that is, $r_{i,j}(n) \to_{n\to\infty} \pi_j$. We can then solve the *balance equations*

$$\pi_j = \sum_k \pi_k p_{k,j}$$

The interpretation of π_j as the long-term visitation frequency is very useful for answer such questions as "How often does one visit State j from State i?" (Would be $\pi_i p_{i,j}$.)

For questions like "expected time to absorption by an absorbing state j, starting from state i", "expected time to pass state j, starting from state i", and "expected recurrence time for state i": you can expect to create recurrence relations based on the transition probabilities and the analogous quantities for all state $k \neq i$, then solving them.