Multi-armed Bandits

1 Thompson sampling

Thompson sampling is a Bayesian approach to the multi-armed bandit problem, in contrast to the UCB algorithm, which takes a frequentist approach. Historically, it was the first algorithm proposed for this problem, and possibly the first formulation of the problem; see Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, 1933. However, it is only recently that a mathematical analysis of its regret was obtained, and that will be our main topic in this chapter.

In keeping with Thompson's original paper, we will restrict ourselves to Bernoulli bandits, though the algorithm can be easily extended to other distributions. The main idea behind the algorithm is very simple. Consider the case of two Bernoulli arms, with unknown mean rewards μ_1 and μ_2. The Bayesian approach to unknown parameters is to start with some prior belief about them, encoded as a probability distribution, and to update this belief in the light of evidence / data. Thus, we propose prior distributions π_1 and π_2 for μ_1 and μ_2, and compute posterior distributions based on observed rewards, using Bayes' formula. We then use these posterior distributions to decide which arm to play. How exactly should we do this? Thompson's suggestion was to take independent samples from the posterior distributions for each arm, and then play that arm whose sample value was largest. As more samples are collected, the posterior distribution concentrates more strongly around the true mean, so the best arm is more likely to be played. Nevertheless, the posterior is sufficiently spread out (if the prior is chosen properly) to ensure that each arm gets played infinitely often. The problem is to analyse how good a trade-off between exploration and exploitation is achieved by this method.

We will begin with some preliminaries about Bayesian inference. It turns out that Beta distributions are particularly convenient for working with data generated from a Bernoulli distribution, so we begin with an introduction to this family of distributions.

2 Preliminaries: Gamma and Beta distributions

A random variable X is said to have a Gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$, denoted $X \sim \text{Gamma}(\alpha, \lambda)$, if it is non-negative and has density

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0.$$

Here, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ is the constant needed to ensure that the density integrates to 1. Notice that if $\alpha \leq 0$ or $\lambda \leq 0$, then the integral of the function displayed above blows up, and it cannot be normalised to be a probability density function. Also note that if $\alpha = 1$, then X has the density of an $\text{Exp}(\lambda)$ distribution. It can be shown that, if $\alpha \in \mathbb{N}$, then the Gamma distribution with shape parameter α and scale parameter λ is the same as that of the sum of α iid $\text{Exp}(\lambda)$ random variables. If the scale parameter λ is equal to 1, we may suppress it from the notation and simply write $\text{Gamma}(\alpha)$ to denote a Gamma distribution with (explicit) shape parameter α and (implicit) scale parameter 1.

A Beta distribution is supported on the interval $[0, 1]$, and is parametrised by two positive real numbers. We say that a random variable X has a $\text{Beta}(\alpha, \beta)$ distribution, written $X \sim \text{Beta}(\alpha, \beta)$, if it takes values in $[0, 1]$, and has the density

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad x \in [0, 1].$$

The Beta distribution is closely connected with the Gamma distribution, as described in the following lemma.

Lemma 1 *Let X and Y be independent Gamma random variables, with shape parameters α and β respectively, and common scale parameter, which we take to be 1. (The value of the scale parameter doesn't matter, so long*

as it is the same for X and Y). Then,

$$V = \frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta).$$

Remark. The random variable V is obtained as a function, or transformation, of the random variables X and Y. We need a formula for the density of random variables obtained by such transformations. If you are not familiar with this material, then see the appendix at the end of this chapter for a review.

Proof. The function $(X, Y) \mapsto V = \frac{X}{X+Y}$ maps $\mathbb{R}^2$ to $\mathbb{R}$, and so we cannot directly use the formula for densities of random variables. We need to define an auxiliary random variable, so that we can construct a suitable function. One obvious choice is to set $W = X$. This leads to consider $(V, W) = g(X, Y)$ where the function $g : \mathbb{R}^2 \to \mathbb{R}^2$ is defined as

$$g(x, y) = \Big(\frac{x}{x+y}, x\Big).$$

It is easy to see that g is injective (one-to-one), but not surjective on $\mathbb{R}^2$. (Why?) Nevertheless, as it is injective, it is invertible on the set $Im(g)$, the image of $\mathbb{R}^2$ under g. To compute its inverse, not that given $(v, w) = g(x, y)$, we have $x = w$ and $(x + y)v = x$, i.e., $(w + y)v = w$, so that $y = (w/v) - 1$. In other words,

$$g^{-1}(v, w) = \Big(w, \frac{1-v}{v} w\Big).$$

Next we calculate the Jacobian matrix,

$$J_g(x, y) = \begin{pmatrix} \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \\ 1 & 0 \end{pmatrix},$$

so that

$$|det J_g(x, y)| = \frac{x}{(x+y)^2} = \frac{v^2}{w}.$$

Since X and Y are independent, their joint distribution can be written as

$$f_{X,Y}(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta} x^{\alpha-1} e^{-x} y^{\beta-1} e^{-y}.$$

Thus, using the formula for transformations of random variables, we obtain that the joint density of V and W is given by

$$\begin{aligned} f_{V,W}(v,w) &= \frac{w}{v^2}\frac{1}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1}e^{-w}\Big(\frac{w(1-v)}{v}\Big)^{\beta-1}e^{-w(1-v)/v} \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}\Big(\frac{w}{v}\Big)^{\alpha+\beta-1} v^{\alpha-2}(1-v)^{\beta-1}e^{-w/v}. \end{aligned}$$

We can now obtain the marginal density of V, our random variable of interest, by integrating out the joint density over the auxiliary random variable W. In other words,

$$\begin{aligned} f_V(v) &= \int_{w=0}^{\infty} f_{V,W}(v,w)dw = \frac{v^{\alpha-2}(1-v)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}\int_{w=0}^{\infty}\Big(\frac{w}{v}\Big)^{\alpha+\beta-1}e^{-w/v}dw \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1}(1-v)^{\beta-1}\int_{z=0}^{\infty}\frac{1}{\Gamma(\alpha+\beta)}z^{\alpha+\beta-1}e^{-z}dz, \end{aligned}$$

where we have made the change of variables $z = w/v$ to get the last equality. Now, the last integral above is equal to 1, because it is the integral of the density of a Gamma($\alpha+\beta$) random variable (or by the definition of $\Gamma(\alpha+\beta)$). Notice also that $V = X/(X+Y) \in [0,1]$ by definition, as X and Y are non-negative random variables. Hence, the density of V is supported on the interval $[0,1]$, and is zero outside this interval. We recognise the expression for the density as that of a Beta(α,β) random variable. This concludes the proof of the lemma. □

Remarks. We now describe some properties of the Beta distribution. Observe that it is supported on $[0,1]$ and that, if $\alpha = 1$ and $\beta = 1$, then the density is a constant. Thus, Beta$(1,1)$ is the uniform distribution on $[0,1]$. If $\alpha < 1$, then the density tends to infinity as x tends to zero; likewise, if $\beta < 1$, then the density tends to infinity as x tends to 1. If $\alpha \geq 1$ and $\beta \geq 1$, then it is easy to verify by differentiating the density (or its logarithm) that its maximum value is attained at $x = \frac{\alpha-1}{\alpha+\beta-2}$; if at least one of α and β is strictly larger than 1, then this is the unique maximiser, i.e., the mode of the Beta distribution. If α and β are both very large, then it can be shown that the Beta distribution concentrates around its mode.

We are now going to use the Beta distribution, which is supported on $[0,1]$, as the prior distribution for the parameter of a Bernoulli random variable. The reason for our interest in the Beta distribution will become apparent when we compute the posterior distribution. Let $X_1, X_2, \ldots$ be an iid sequence of Bernoulli random variables with unknown mean μ. Fix $\alpha, \beta > 0$,

and let $\pi_0 \sim \text{Beta}(\alpha, \beta)$ be our prior distribution for the parameter μ. What can we say about the posterior distribution of μ given the first n elements of this sequence? Denote by $S(n) = \sum_{k=1}^{n} X_k$ the total number of 1s (or successes) seen in the first n Bernoulli samples, and by π_n the posterior distribution based on n samples. We now compute π_n. The likehood function of a sequence with $S(n)$ successes in n trials, as a function of the success probability θ, is given by

$$L(X_1, X_2, \ldots, X_n; \theta) = \theta^{S(n)}(1-\theta)^{n-S(n)}.$$

Notice that the likelihood function only depends on the number of successes and failures, and not on the order in which they occur. This is always the case with iid random variables, where the likelihood function only depends on the empirical distribution, i.e., the frequency of each possible outcome, and not on the order in which they occured.

Now, it is easily seen to follow from Bayes' formula that the posterior distribution is proportional to the product of the prior and the likelihood; it is equal to this product, normalised to be a probability distribution. Thus, substituting for the Beta density of the prior, we get

$$\pi_n(\theta|X_1, \ldots, X_n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{S(n)}(1-\theta)^{n-S(n)},$$

where we have ignored constants that don't depend on $S(n)$. The constant of proportionality needs to be such as to make π_n a probability distribution on $[0, 1]$, i.e., to ensure that $\int_0^1 \pi_n(\theta)d\theta = 1$. We recognise that π_n is proportional to the density of a $\text{Beta}(\alpha + S(n), \beta + n - S(n))$ random variable. Hence, it is in fact equal to that density. In other words, the posterior π_n has a $\text{Beta}(\alpha + S(n), \beta + n - S(n))$ distribution.

Notice that the posterior distribution belongs to the same family as the prior distribution, if we shoose a Beta prior. Such a family is known as a conjugate family of priors. While the posterior distribution is well defined for arbitrary prior distributions, computing the posterior can, in general, be very computationally demanding or intractable. Conjugate priors are much more tractable. Moreover, they admit analytical solutions, which can be very helpful in gaining insight as well as in doing calculations by hand.

The following fact about Beta distributions will be useful for computations.

Lemma 2 *Let X have a Beta(α, β) distribution, with $\alpha, \beta \in \mathbb{N} + 1$ (the natural numbers starting at 1). Fix $p \in (0, 1)$ and let Y have a Bin($\alpha + \beta -$*

$1, p)$ *distribution. Then,*

$$\mathbb{P}(X > p) = \mathbb{P}(Y \leq \alpha - 1).$$

Proof. The proof makes use of the following well-known fact about Poisson processeses. Let $\{N_t, t \geq 0\}$ be a Poisson process of arbitrary intensity $\lambda > 0$. Let n be a positive integer and let t be a positive real number. Then, conditional on the event $N_t = n$, the unordered increment times on $[0, t]$ (the random times at which the Poisson process sees an increment) are mutually independent, and uniformly distributed on $[0, t]$.

Recall that if X has a Beta(α, β) distribution, then we can write $X = \frac{V}{V+W}$, where V and W are independent Gamma random variables, with shape parameters α and β respectively, and common scale parameter, say 1. If α and β are integer-valued, then, since a Gamma$(\alpha, 1)$ random variable is the sum of α iid Exp(1) random variables, we can interpret V as the time of the α^{th} increment of a unit rate Poisson process, and $V + W$ as the time of the $(\alpha + \beta)^{\text{th}}$ increment. Consequently, conditional on $V + W = \tau$, the Poisson process N_t has exactly $\alpha + \beta - 1$ increments in the interval $[0, \tau)$. By the fact noted above, the unordered times of these increments are iid, uniformly distributed on $[0, \tau)$. The event $\{X > p\}$ is the same as the event $\{V > p\tau\}$, conditional on $V + W = \tau$; as V is the time of the α^{th} increment, this says that at most $\alpha - 1$ increments occur in $[0, p\tau]$. As the increments are iid, uniform in $[0, \tau)$, and there are $\alpha+\beta-1$ in total, the number that fall within $[0, p\tau]$ has aBin$(\alpha + \beta - 1, p)$ distribution, which is the distribution of the random variable Y. Thus, the events $\{X > p\}$, $\{V > p\tau | V + W = \tau\}$ and $Y \leq \alpha - 1$ all have the same probability. This completes the proof of the lemma. □

3 Thompson sampling: Algorithm and Analysis

We are now ready to formally describe the Thompson sampling algorithm. We will restrict ourselves to the case of a two-armed bandit, with iid Bernoulli rewards on each arm, mutually independent across the arms. We denote by μ_i the mean reward on arm $i \in \{1, 2\}$, which is the same as the success probability (probability of non-zero reward) for that arm. We assume without loss of generality that $\mu_1 > \mu_2$, though the player doesn't know that. We denote the difference by $\Delta = \mu_1 - \mu_2$, and call it the arm gap.

Thompson Sampling Algorithm: Two Bernoulli arms

1. Start with independent Beta$(1, 1)$ priors for each of the parameters μ_1 and μ_2.
2. In each time step t, independently sample μ_1 and μ_2 from the current posterior distributions for these parameters. Call the samples $\theta_1(t)$ and $\theta_2(t)$.
3. Play the arm with the higher sample value, i.e., play arm 1 if $\theta_1(t) > \theta_2(t)$ and arm 2 otherwise.
4. Update the posterior for the parameter of the arm played, based on the reward obtained from playing it.
5. Move to the next time step.

The extension of the algorithm to the case of more than two Bernoulli arms is self-evident. If there are K arms, with Bernoulli paramaters $\mu_1, \ldots, \mu_K$, then we start with independent Beta(1,1) priors for each of these parameters. In each time step, we sample $\theta_1, \ldots, \theta_K$ from the current posteriors for $\mu_1, \ldots, \mu_K$. We play the arm whose index corresponds to the largest of the θ_i; there is zero probability of a tie. Based on the reward obtained, we update the posterior distribution for the arm which was played. The posteriors for all other arms are unchanged

Thompson sampling can be similarly extended to reward distributions other than Bernoulli. However, the ease of implementing it depends on how easy it is to sample from the posterior distribution. Conjugate priors are available for some commonly used distributions. For other distributions, any of a wide range of techniques from computational Bayesian statistics can be used, with varying amounts of accuracy and computational cost. Bounds on regret are not known in complete generality.

We now return to the case of two Bernoulli arms. The analysis of Thompson sampling is rather intricate and we do not study it in full in this course. Instead, we shall analyse a simplified, genie-assisted version of the Thompson sampling algorithm that we now present. The simplified analysis contains many of the key ideas needed to fully analyse the actual algorithm.

Genie-Assisted Thompson Sampling: Two Bernoulli arms

1. Suppose a genie gives you the value of μ_1 but does not tell you that arm 1 is better.
2. Start with a Beta$(1, 1)$ prior for μ_2, the only unknown parameter.
3. In each time step t, sample μ_2 from its current posterior distribution. Call the sample $\theta_2(t)$.
4. Play arm 1 if $\mu_1 > \theta_2(t)$ and arm 2 otherwise.
5. Update the posterior for arm 2 when it is played, based on the reward obtained from playing it.
6. Move to the next time step.

Notice that we do not need to maintain a prior for μ_1 as its true value is known (if we trust the genie, as we assume we do) and there is no uncertainty about it. We are now ready to state our main result about the regret incurred by the genie-assisted Thompson sampling algorithm described above. Let $\mathcal{R}(T)$ denote the regret up to time T, and recall that $\Delta = \mu_1 - \mu_2$ is the gap in mean rewards between the two arms.

Theorem 1 *Consider genie-assisted Thompson sampling as described above, applied to a multi-armed bandit with two Bernoulli arms. Its regret is bounded as follows:*

$$\mathcal{R}(T) \leq \frac{2 \log T}{\Delta} + 3\Delta.$$

We need to define some notation before stating these lemmas. We denote by $N_i(t)$ the number of times that arm i has been played in the first t rounds or time steps, and by $S_i(t)$ the number of successes observed in these plays; we might also denote the number of successes by $S_{i,N_i(t)}$ which makes the number of attempts more explicit in the notation.

Notice that if arm i is played in time step t, and its posterior distribution at the beginning of this time step is Beta$(\alpha_i(t), \beta_i(t)))$, then the posterior distribution at the end of the time step is Beta$(\alpha_i(t)+X_i(t), \beta_i(t)+1-X_i(t))$, where $X_i(t) \in \{0, 1\}$ is the reward obtained on playing arm i in time step t. The posterior distribution of tan arm that was not played doesn't change.

As arm 1 was assumed to be better, a regret of Δ is incurred each time arm 2 is played. Hence,

$$\mathcal{R}(T) = \Delta \mathbb{E}[N_2(T)],$$

and the challenge is to bound the latter expectation. The idea behind our analysis is the following. As the number of times arm 2 is played increases, its posterior distribution concentrates increasingly sharply around the true parameter value, μ_2. We use Hoeffding's inequality to show that after sufficiently many plays of arm 2, firstly, the observed frequency of successes is very unlikely to exceed $\mu_2 + \Delta/2$, and secondly, that the posterior sample $\theta_2(t)$ is very unlikely to exceed $\mu_2 + \Delta = \mu_1$.

Fix a time horizon T, and define

$$L = \left\lceil \frac{2 \log T}{\Delta^2} \right\rceil, \quad \tau = \inf\{0 < t \leq T : N_2(t) \geq L\}.$$

As usual, we take τ to be infinite if no such time t exists. The theorem asserts that the regret up to time T is not much larger than ΔL, i.e., that arm 2 is not played too often after time τ.

As regret is incurred only when arm 2 is played, in order to bound the regret it suffices to bound the expected number of times that arm 2 is played up to time T. If $\tau = \infty$, then arm 2 has been played fewer than L times up to time T, i.e., $N_2(T) \leq L$, so the regret is bounded by $\Delta L = (2 \log T)/\Delta$, and we are done. Hence, in the following, we will restrict ourselves to the situation that $\tau \leq T$.

Lemma 3 *Suppose $\tau \leq t \leq T$. Then,*

$$\mathbb{P}\Big(\theta_2(t) \geq \mu_1\Big) \leq \frac{2}{T}.$$

Proof. By definition of τ, if $t \geq \tau$, then $N_2(t) \geq L$. We can write

$$\begin{aligned}
\mathbb{P}(\theta_2(t) \geq \mu_1) &= \mathbb{P}\Big(\theta_2(t) \geq \mu_2 + \Delta, \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{2}\Big) \\
&\quad + \mathbb{P}\Big(\theta_2(t) \geq \mu_2 + \Delta, \frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{2}\Big) \\
&\leq \mathbb{P}\Big(\theta_2(t) \geq \mu_2 + \Delta \;\Big|\; \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{2}\Big) + \mathbb{P}\Big(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{2}\Big). \qquad (1)
\end{aligned}$$

We will bound each of the last two terms. Firstly, conditional on the number of times arm 2 is played, namely $N_2(t)$, the total reward from these plays $S_2(t)$ is the sum of $N_2(t)$ independent Bern(μ_2) random variables. Hence,

using Hoeffding's inequality, we have

$$\mathbb{P}\Big(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{2} \;\Big|\; N_2(t)\Big) \le \exp\Big(-2N_2(t)\frac{\Delta^2}{4}\Big).$$

As we have assumed that $N_2(t) \ge L \ge (2\log T)/\Delta^2$, we conclude that

$$\mathbb{P}\Big(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{2}\Big) \le \exp(-\log T) = \frac{1}{T}. \tag{2}$$

Next, we note that conditional on $S_2(t)$ and $N_2(t)$, the distribution of $\theta_2(t)$ is Beta($S_2(t)+1$, $N_2(t) - S_2(t) + 1$). Consequently, by Lemma 2, we have that

$$\mathbb{P}(\theta_2(t) \ge \mu_2 + \Delta) = \mathbb{P}\big(\mathrm{Bin}(N_2(t)+1, \mu_2+\Delta) \le S_2(t)\big).$$

Applying Hoeffding's inequality to the term on the right, we see that, for $N_2(t) \ge L$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\theta_2(t) \ge \mu_2 + \Delta \;\Big|\; \frac{S_2(t)}{N_2(t)} \le \mu_2 + \frac{\Delta}{2}\Big) \\
&\le \mathbb{P}\Big(\mathrm{Bin}(N_2(t)+1, \mu_2+\Delta) \le \Big(\mu_2 + \frac{\Delta}{2}\Big)N_2(t)\Big) \\
&\le \exp\Big(-2(N_2(t)+1)\frac{\Delta^2}{4}\Big) \le \exp(-L\Delta^2/2) \le \frac{1}{T}.
\end{aligned} \tag{3}$$

Substituting (2) and (3) in (1), we conclude that if $t \ge \tau$, i.e., $N_2(t) \ge L$, then

$$\mathbb{P}(\theta_2(t) \ge \mu_1) \le \frac{2}{T},$$

as claimed in the statement of the lemma. □

Proof of Theorem 1. In order to prove the theorem, we need to bound the number of times that arm 2 is played after time τ, which is the same as the number of times that $\theta_2(t)$, the sample from the posterior for μ_2 at time t, exceeds μ_1, the true mean reward on arm 1, given to us by the genie. Notice that, by definition of τ, arm 2 is played at most L times by time τ, and exactly L times if τ is finite. Hence, we obtain from Lemma 3 that

$$\begin{aligned}
\mathcal{R}(T) &\le \Delta L + \Delta \sum_{t=\tau+1}^{T} \mathbb{P}(\theta_2(t) \ge \mu_1) \\
&\le \Delta\Big(\frac{2\log T}{\Delta^2} + 1\Big) + \frac{2\Delta}{T}(T - \tau - 1) \le \frac{2\log T}{\Delta} + 3\Delta.
\end{aligned}$$

This completes the proof of the theorem. □

A Transformation of random variables

Example: Consider the probability space $\Omega = \{1, \ldots, 6\}$, $\mathcal{F}$ = all subsets of Ω, with probabilities $P(\omega) = 1/6$ for all $\omega \in \Omega$.
(a) On this space, define the random variable $X(\omega) = \omega$. Then the pmf of X is $\{1/6, \ldots, 1/6\}$ on the set $\{1, \ldots, 6\}$. Suppose $Y = X^2$. Then what is the pmf of Y?
(b) On the same space, suppose that X is defined instead as $X(\omega) = \omega - 2$, and that again $Y = X^2$. What are the pmfs of X and Y?

The idea can be extended to continuous random variables, but there is one subtlety involved.

Example: Suppose X is Uniform($[0, 1]$) and $Y = 2X$. What are the cdf and pdf of Y? We first compute the cdf. It is obvious that $F_Y(y) = 0$ for $y < 0$. Also,

$$P(Y \leq y) = P(2X \leq y) = P(X \leq y/2) = y/2 \text{ for } y \in [0, 2).$$

Finally, $F_Y(y) = 1$ for $y \geq 2$. Differentiating the above cdf, we get $f_Y(y) = 1/2$ for $y \in (0, 1)$ and $f_Y(y) = 0$ otherwise.

Could we have guessed this? Intuitively, for an infinitesimal dy,

$$P(Y \in (y, y + dy)) = P(2X \in (y, y + dy)) = P\Big(X \in \Big(\frac{y}{2}, \frac{y}{2} + dy2\Big)\Big),$$

so that

$$f_Y(y)dy = f_X\Big(\frac{y}{2}\Big)\frac{1}{2}dy,$$

which gives the same answer. This intuition can be extended.

Let X be a random variable, g be a differentiable and strictly monotone function, and let $Y = g(X)$. Then, by the same reasoning as above,

$$f_Y(y)dy = f_X(x)dx,$$

where $y = g(x)$. How are dy and dx related? We want $y + dy = g(x + dx)$, so we must have $dy = g'(x)dx$. We are almost there, except that the sign of $g'(x)$ doesn't matter. (It may be the interval $(x - dx, x)$ that gets mapped to $(y, y + dy)$.) So, we have

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y)|}, \tag{4}$$

where the inverse g^{-1} of the function g is well-defined by the assumption that g is strictly monotone. (The domain of g^{-1} is the range of g.)

What if g isn't monotone? Then the equation $y = g(x)$ may have many solutions for x, and we have to add up the probability contributions from all of them. If there are only countably many solutions, then (4) changes to

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x) \frac{1}{|g'(x)|}. \tag{5}$$

The same idea extends to joint distributions. Suppose $X_1, \ldots, X_n$ are random variables on the same sample space and $(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$ for some differentiable function $g : \mathbb{R}^n \to \mathbb{R}^n$. Then, using boldface to denote vectors,

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}:g(\mathbf{x})=\mathbf{y}} f_{\mathbf{X}}(\mathbf{x}) \frac{1}{|\det(J_g(\mathbf{x}))|}. \tag{6}$$

Here, $\det(J_g(\mathbf{x}))$ denotes the determinant of the Jacobian matrix

$$J_g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_1}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n}(\mathbf{x}) & \cdots & \frac{\partial g_n}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$