Taste the Rainbow?

This morning I took out a little fun-size packet of Skittles, and found to my surprise that of the 16 skittles inside not a single one was green. (Skittles come in five flavors - green, yellow, orange, red, purple - and we're going to assume that each skittle is i.i.d. assigned one of these with uniform probability. Incidentally, this story is 100% true.)

This surprised me, so I wondered – what is the probability of getting such a packet, where some flavor is missing? (I assumed that all packets have 16 skittles.) Well, for any given flavor (say, green), the probability that a skittle is not that flavor is 4/5, and there are 16 in a packet, so

$$\mathbb{P}[\text{packet contains no green}] = (4/5)^{16}$$

But I'm not interested in just "no green" – I want to know what the probability of missing *any* flavor is. This is upper-bounded by using the Union Bound over the 5 flavors, giving

$$\mathbb{P}[\text{packet is missing a flavor}] \leq 5 \cdot (4/5)^{16}$$

This is actually a fairly close bound, because it's only due to the possibility that *two* flavors might be missing which makes it a bound and not an equality. But missing two flavors is phenomenally unlikely – and from Problem 2 on the midterm we know that

$$\mathbb{P}[\text{packet is missing a flavor}] \geq 5 \cdot (4/5)^{16} - \binom{5}{2}(3/5)^{16}$$

We can then give both upper- and lower-bounds:

$$0.14 \leq 5 \cdot (4/5)^{16} - \binom{5}{2}(3/5)^{16} \leq \mathbb{P}[\text{packet is missing a flavor}] \leq 5 \cdot (4/5)^{16} \approx 0.14$$

This is really surprising! This means that if everything is uniform and independent, roughly *one out of every seven* packs is missing a flavor. Incidentally, the probability of there being a missing-flavor packet out of *five* random packets is

$$\mathbb{P}[\text{at least one is missing a flavor}] = 1 - \mathbb{P}[\text{no packet is missing a flavor}] \geq 1 - (0.86)^5 \approx 0.53$$

This means you have a *slightly better than* 1/2 *chance* of getting such a pack in a group of five.

I feel like there's a fortune in bet winnings just waiting here.

Characteristic Functions

First things first – make sure you are comfortable with (a) complex numbers in general, and (b) especially with expressions of the form e^{it}, notably the Euler formula

$$e^{it} = \cos(t) + i\,\sin(t) \quad \text{(note that this has L2-norm of 1)}$$

(and its extension $e^{it+s} = e^s(\cos(t) + i\,\sin(t))$).

Limitations of the MGF, and how to get around them

The MGF is a very useful tool, but it has the notable limitation of sometimes not existing. For instance, consider the *Cauchy distribution*:

Definition 0.1. The Cauchy distribution of location μ and scale γ is the continuous distribution on $\mathbb{R}$ with PDF

$$f_X(x) = \frac{1}{\pi\gamma\Big(1 + (\frac{x-x_0}{\gamma})^2\Big)}$$

This happens to have CDF of the form

$$F_X(x) = \frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right) + \frac{1}{2}$$

This is often called *pathological* because its expectation is not defined. Furthermore, the MGF is defined *nowhere* (except at $s = 0$) – we can show this by simply attempting to compute

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx = \int_{-\infty}^{\infty} e^{sx}\frac{1}{\pi\gamma\Big(1 + (\frac{x-x_0}{\gamma})^2\Big)}\,dx$$

For any $s \neq 0$, we have the following for sufficiently big positive x or big negative x:

$$e^{sx} > \pi\gamma\Big(1 + (\frac{x-x_0}{\gamma})^2\Big)$$

This immediately implies that the integral is infinite because it is > 1 on infinitely large measure.

So if we can't use the MGF on Cauchy, what can we do? Use e^{itX} instead of e^{sX} – the expression e^{itX} is always of L2-norm 1 because itX has no real part. We therefore define:

Definition 0.2. The *characteristic function* of a real-valued random variable X is a function $\phi_X : \mathbb{R} \to \mathbb{C}$ given by

$$\phi_X(t) := \mathbb{E}[e^{it}]$$

Because it has L2-norm of 1 everywhere, both the real and imaginary components of e^{itX} are absolutely bounded by 1 – and therefore by the Bounded Convergence Theorem, the expectation exists and is finite. Even more, we know that $\phi_X(t)$ is always within the unit circle around 0 in the complex plane.

Why is the characteristic function useful?

If you've seen *Fourier analysis*, you might recognize the characteristic function as being super similar to the Fourier transform (but without the -2π constant term in the exponent). Furthermore, we'll use without proof here the following facts (Yury will probably cover them sometime):

> **Proposition 0.1.** X, Y have the same distribution $\iff$ $\phi_X = \phi_Y$ everywhere.

(Note: it is possible for the characteristic functions of different random variables to agree on an interval containing 0, but somehow disagree elsewhere. However, I don't know any examples and they won't be discussed here.)

> **Theorem 0.1 (Levy's Continuity Theorem).** If $X_1, X_2, \dots$ and X are random variables, and $\phi_{X_n} \to \phi_X$ (pointwise) everywhere, then $X_1, X_2, \dots \to X$ *in distribution.*

This makes it a very powerful tool for this sort of thing.

We'll also use the following, which can be proved in the same manner as for MGFs:

> **Proposition 0.2.** The characteristic function satisfies the following properties:
>
> - If a, b are real numbers, $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.
> - If X, Y are independent random variables, $\phi_{X+Y}(t) = \phi_X(t)\,\phi_Y(t)$.

Proof. For the first, we just write

$$\phi_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = \mathbb{E}[e^{itb}\, e^{it(aX)}] = e^{itb}\mathbb{E}[e^{i(at)X}] = e^{itb}\phi_X(at)$$

For the second, we use the fact that X, Y independent $\implies$ e^{itX}, e^{itY} independent. Then:

$$\phi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}\, e^{itY}] = \mathbb{E}[e^{itX}]\,\mathbb{E}[e^{itY}] = \phi_X(t)\,\phi_Y(t)$$

concluding the proof. $\square$

Some quick problems using the CF

> **Problem 0.1.** Prove that the sum of two Cauchy's is also Cauchy.

The CF of the Cauchy distribution $f_X(x) = \frac{1}{\pi\gamma\left(1+\left(\frac{x-x_0}{\gamma}\right)^2\right)}$ happens to be $\phi_X(t) = e^{itx_0-\gamma|t|}$. This is quite difficult to actually compute without complex analysis tools, but we'll use it. The rest is simple: let X, Y have parameters x_0, γ_X and y_0, γ_Y. Then

$$\phi_{X+Y}(t) = \phi_X(t)\,\phi_Y(t) = e^{itx_0-\gamma_X|t|}\, e^{ity_0-\gamma_Y|t|} = e^{it(x_0+y_0)-(\gamma_X+\gamma_Y)|t|}$$

which is also the CF of a Cauchy (with parameters $x_0 + y_0$ and $\gamma_X + \gamma_Y$).

Problem 0.2. Use characteristic functions to show that average of n i.i.d. Ber(p) converges to a constant (equal to the probability p) as $n \to \infty$.

We consider $X_k \sim \text{Ber}(p)$ (i.i.d.), and $S_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. The CF of X_k is

$$\phi_{X_k}(t) = \mathbb{E}[e^{itX_k}] = (1-p) + pe^{it}$$

Furthermore, adding independent random variables multiplies CFs (same as MGFs), giving

$$\phi_{S_n}(t) = \phi_{\sum_{k=1}^{n} X_k}(t/n) = \big((1-p) + pe^{it/n}\big)^n = \big(1 + p(e^{it/n} - 1)\big)^n$$

Note that as $n \to \infty$, we have $it/n \to 0$ – so we'll take the first-order Taylor expansion at 0:

$$e^{it/n} = 1 + it/n + O(n^{-2}) \implies (e^{it/n} - 1) = it/n + O(n^{-2})$$

(Why the first-order? Because the $O(n^{-2})$ term is too small to affect the result in the limit, even with the outer power-of-n.) This gives

$$\lim_{n\to\infty} \big(1 + p(e^{it/n} - 1)\big)^n = \lim_{n\to\infty} \big(1 + (itp)/n\big)^n = e^{itp}$$

But we can easily recognize that e^{itp} is just the CF of the distribution which returns p with probability 1. Therefore, the S_n's converge (in distribution) to that distribution.

Problem-solving about the MGF

Problem 0.3. Suppose that we know that

$$\limsup_{x\to\infty} \frac{\log\big(\mathbb{P}[X > x]\big)}{x} = -t < 0$$

We want to show that the MGF $M_X(s) < \infty$ for all $s \in [0, t)$.

Note that e^{sX} is actually nonnegative. This is very useful because we can now use that nice little formula of computing the expectation of a nonnegative variable using $\mathbb{P}[X > x]$:

$$\mathbb{E}[e^{sX}] = \int_0^\infty \mathbb{P}[e^{sX} > y]\, dy$$

This is good, so far, but we really want $\mathbb{P}[X > x]$ – so we'll rewrite $y = e^{sx}$. Note that because e^{sx} is (strictly) monotonically increasing, $e^{sX} > e^{sx} \iff X > x$. The transformation takes y on $(0, \infty)$ to x on $(-\infty, \infty)$, and $dy = s\, e^{sx}\, dx$, giving

$$\mathbb{E}[e^{sX}] = s \int_{-\infty}^\infty e^{sx}\, \mathbb{P}[X > x]\, dx$$

Note the intuition here **(warning - not rigorous!)**:

$$\begin{aligned} \frac{\log\big(\mathbb{P}[X > x]\big)}{x} \le -t &\implies \mathbb{P}[X > t] \le e^{-tx} \\ &\implies s \int_{-\infty}^\infty e^{sx}\, \mathbb{P}[X > x]\, dx \le s + s \int_0^\infty e^{(s-t)x}\, dx < \infty \end{aligned}$$

(taking advantage of the fact that for $x \le 0$, we have $e^{sx}\mathbb{P}[X > x] \le 1$).

How do we make this rigorous? Use an ε.

$$\limsup_{x\to\infty} \frac{\log\big(\mathbb{P}[X > x]\big)}{x} = -t$$

really means that for all $\varepsilon > 0$, we have some x_ε such that

$$\frac{\log\big(\mathbb{P}[X > x]\big)}{x} \le -t + \varepsilon \quad \text{for all } x > x_\varepsilon$$

This condition is equivalent to $\mathbb{P}[X > x] \le e^{(-t+\varepsilon)x}$ for all $x > x_\varepsilon$. Now let us fix $s \in [0, t)$ and $\varepsilon < t - s$. Now we split the integral:

$$\mathbb{E}[e^{sX}] = s \int_{-\infty}^\infty e^{sx}\, \mathbb{P}[X > x]\, dx = s \int_{-\infty}^{x_\varepsilon} e^{sx}\, \mathbb{P}[X > x]\, dx + s \int_{x_\varepsilon}^\infty e^{sx}\, \mathbb{P}[X > x]\, dx$$

The integral on the left is finite, as it decays exponentially going to $-\infty$ and is bounded above by e^{sx_ε}. The integral on the right is then upper-bounded by our result for $\mathbb{P}[X > x]$, yielding in total (for some constant C)

$$\mathbb{E}[e^{sX}] \le C + s \int_{x_\varepsilon}^\infty e^{(s-t+\varepsilon)x}\, dx < \infty$$

because, of course, we chose $\varepsilon > 0$ such that $s - t + \varepsilon < 0$.

Multivariate normal - conditional expectation

Problem 0.4. Suppose that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $\sim \mathcal{N}(0,1)$; let $X_1, \ldots, X_n$ be linear combinations of these

$$X_j = \sum_{r=1}^{n} C_{j,r} Y_r \text{ for some constants } C_{j,r}$$

What is the conditional expectation $\mathbb{E}[X_j | X_k]$?

Note that all the normals discussed here have expectation 0, which simplifies things. We have the formula (Theorem 1 in Lecture 14 notes)

$$\mathbb{E}[X_j|X_k] = \mu_{X_j} + V_{X_j X_k} V_{X_k X_k}^{-1} (X_k - \mu_{X_k}) = V_{X_j X_k} V_{X_k X_k}^{-1} X_k$$

where $V_{Z_1 Z_2} = \mathrm{Cov}(Z_1, Z_2)$. The zero means also make the covariance calculations simpler:

$$V_{X_j X_k} = \mathbb{E}[X_j X_k] \quad \text{and} \quad V_{X_k X_k} = \mathbb{E}[X_k X_k]$$

Note that if we have Y_{i_1}, Y_{i_2} (for $i_1 \neq i_2$) which are therefore independent, we get

$$\mathbb{E}[Y_{i_1} Y_{i_2}] = \mathbb{E}[Y_{i_1}]\,\mathbb{E}[Y_{i_2}] = 0 \quad \text{and} \quad \mathbb{E}[Y_i Y_i] = \mathrm{Var}(Y_i) = 1$$

(by definition since $Y_i \sim \mathcal{N}(0,1)$).

Now we note the following, and use linearity of expectation:

$$\mathbb{E}[X_j X_k] = \mathbb{E}\Big[\sum_{r,s} C_{j,r} C_{k,s} Y_r Y_s\Big] = \sum_{r,s} C_{j,r} C_{k,s} \mathbb{E}[Y_r Y_s] = \sum_r C_{j,r} C_{k,r}$$

Note that the above holds also if $j = k$. Therefore,

$$V_{X_j X_k} = \sum_r C_{j,r} C_{k,r} \quad \text{and} \quad V_{X_k X_k} = \sum_r C_{k,r}^2$$

Plugging back in, we get

$$\mathbb{E}[X_j|X_k] = V_{X_j X_k} V_{X_k X_k}^{-1} X_k = \left(\frac{\sum_r C_{j,r} C_{k,r}}{\sum_r C_{k,r}^2} \right) X_k$$