

<b>2</b>	<b>It's just calculus and convexity</b>	<b>1</b>
2.1	Introduction . . . . .	1
2.2	Polynomial approximations . . . . .	2
2.3	Convexity of the log MGF . . . . .	4
2.4	Conjugate convex functions . . . . .	6
2.5	Approximation of factorials . . . . .	8
2.6	Distances between probability measures . . . . .	13
2.7	Bounding sums by integrals . . . . .	14
2.8	Problems . . . . .	15
2.9	Notes . . . . .	19

Printed: 30 May 2025 at 14:31

## Chapter 2

# It's just calculus and convexity

Calculus::Calculus

SECTION 2.1 explains why you can safely just skim through this chapter on a first pass.

SECTION 2.2 Describes a simple trick using TAYLOR expansion to bound functions of the form  $2(f(x) - f(0) - xf'(0))/x^2$ . I have found the trick a very good way of avoiding masses of calculus. Two important special cases are described.

SECTION 2.3 explains why the logarithm of a moment generating function (MGF) is always convex.

SECTION 2.4 describes briefly a procedure known known in the literature by several different names, explaining how it relates to a minimization method (Chapter 3) for deriving probability tail bounds. The magical  $\psi_{\text{benn}}$  makes its first appearance.

SECTION 2.5 waxes historical about the STIRLING formula and binomial coefficients. It describes a few tricks that pop up often in the literature on high-dimensional statistics.

SECTION 2.6 is an advertisement for the convex function  $\psi_{\text{benn}}$ . It has been a hobby of mine to find hidden appeals to the helpful properties of that function buried inside complex arguments.

SECTION 2.7 explains why many important quantities are written as integrals.

## 2.1 Introduction

Calculus::S:intro

You could safely skim quickly through this chapter, just to get an idea of how much fancy modern theory depends on only a little bit of calculus with magical inequalities coming via convexity arguments, sometimes in disguise. I'll tell you in later chapters when something from here is needed, at which point you might want to read some sections from this Chapter more carefully. I prefer having the elementary—but very useful—arguments collected together in one place rather than having them buried inside complicated proofs scattered all over the place.

## 2.2 Polynomial approximations

Calculus::S:Taylor

Where would Mathematics be without approximation of functions by polynomials?

For my purposes, the integral form of the TAYLOR theorem is the most useful. Suppose  $g$  is a real-valued function defined at least on an interval  $J$  of the real line that contains both 0 and a point  $x$ . If  $g$  is twice continuously differentiable then

$$g(x) - g(0) = \int_0^1 \frac{\partial g(xt)}{\partial t} dt = x \int_0^1 \dot{g}(xt) dt \quad \text{for } x \in J$$

and

$$\begin{aligned} R_1(x, g) &= g(x) - g(0) - x\dot{g}(0) \\ &= x \int_0^1 \dot{g}(xt) - \dot{g}(0) dt \\ &= x^2 \int_0^1 \int_0^1 \{0 \leq s \leq t \leq 1\} \ddot{g}(xs) ds dt \\ &= \frac{1}{2}x^2 \Delta(x, g) \quad \text{where } \Delta(x, g) := \int_0^1 2(1-s)\ddot{g}(xs) ds \end{aligned}$$

\E@ R1.rep <1>

And so on, if higher-order derivatives exist. See Problem [1] for the analog when  $g$  has an absolutely continuous  $k$ th derivative.

**Remark.** The assumptions on  $g$  could be relaxed slightly by using some fancy facts about absolute continuity.

The behaviors of  $R_1(x, g)$  and  $\Delta(x, g)$  are controlled by  $\ddot{g}$ . For example, if  $\ddot{g}$  is non-negative then so are  $\Delta(\cdot)$  and  $R_1(\cdot)$ , and if  $\ddot{g}$  is continuous then so are  $\Delta(\cdot)$  and  $R_1(\cdot)$ . More interestingly, if  $\ddot{g}$  is a nondecreasing function then so is  $\Delta(\cdot)$ , and if  $\ddot{g}$  is a convex function then so is  $\Delta(\cdot)$ .

We can interpret the factor  $2(1-s)$  in the definition of  $\Delta$  as the density (with respect to LEBESGUE measure) of a probability measure on the interval  $(0, 1)$ , with expected value

$$\int_0^1 s 2(1-s) ds = 1/3.$$

If  $\ddot{g}$  is convex then the JENSEN inequality gives

$$\Delta(x, g) \geq \ddot{g} \left( x \int_0^1 2s(1-s) ds \right) = \ddot{g}(x/3).$$

\E@ integral.Jensen <2>

This inequality accounts for a few factors of  $1/3$  that pop up in well known inequalities.

My main interest in these calculus tricks originally came from studying two special examples, which will be important throughout this book. (Hence the funny fonts.) The two functions are

$$\begin{aligned} \text{\E@ fbb} &<3> \quad \mathfrak{f}(x) = e^x - 1 - x \quad \text{for } x \in \mathbb{R} \\ \text{\E@ hbb} &<4> \quad \mathfrak{h}(x) = (1+x) \log(1+x) - x \quad \text{for } x \geq -1. \end{aligned}$$

For  $\mathfrak{h}$  the convention  $0 \log 0 = 0$  gives  $\mathfrak{h}(-1) = 1$ . Sometimes the domain of  $\mathfrak{h}$  is extended to the whole real line by defining  $\mathfrak{h}(x) = +\infty$  when  $x < -1$ , for reasons that will be explained in the next Section.

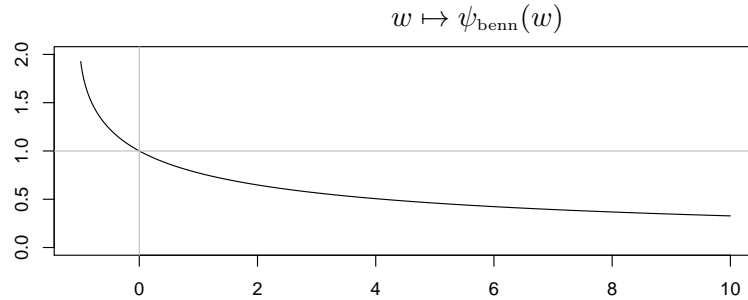
The first and second derivatives,

$$\begin{aligned} \dot{\mathfrak{f}}(x) &= e^x - 1 \quad \text{AND} \quad \ddot{\mathfrak{f}}(x) = e^x \\ \dot{\mathfrak{h}}(x) &= \log(1+x) \quad \text{AND} \quad \ddot{\mathfrak{h}}(x) = (1+x)^{-1}, \end{aligned}$$

show that both functions are non-negative and convex, with each achieving its minimum value of 0 at  $x = 0$ . From <1> we have representations

$$\begin{aligned} \text{\E@ Dele.def} &<5> \quad \mathfrak{f}(x) = \frac{1}{2}x^2 \Delta(x) \quad \text{where } \Delta(x) := \int_0^1 2(1-s)e^{xs} ds \\ \text{\E@ Bennpsi.def} &<6> \quad \mathfrak{h}(x) = \frac{1}{2}x^2 \psi_{\text{benn}}(x) \quad \text{where } \psi_{\text{benn}}(x) := \int_0^1 \frac{2(1-s)}{(1+xs)} ds \text{ for } x \geq -1. \end{aligned}$$

The function  $\Delta$  is continuous, convex, and strictly increasing. The function  $\psi_{\text{benn}}$  is continuous, convex, and strictly decreasing on  $[-1, \infty)$ .



Moreover, by Jensen's inequality,

$$\text{\E@ Benn.Bern} <7> \quad \psi_{\text{benn}}(x) \geq \ddot{\mathfrak{h}}(x/3) = (1+x/3)^{-1}.$$

Compare with the fact that  $\psi_{\text{benn}}(x)$  actually decreases like  $2x^{-1} \log(x)$  as  $x$  tends to infinity.

As you will see in Chapter 3, the function  $\psi_{\text{benn}}$  lies hidden inside a handful of very useful tail bounds. Inequality <7> will also have some surprising consequences: in Section 2.6, in the context of the PINSKER inequality; and in Chapter 8, to explain the relationship between the BENNETT and BERNSTEIN exponential inequalities.

**Remarks.** The subscript ‘benn’ stands for George Bennett, who established a now famous exponential tail bound (Bennett, 1962, equation 8b), although he did not write it using the  $\psi_{\text{benn}}$  function. This bound, and its martingale extension, are described in Chapter 8.

The monotonicity and convexity of  $\Delta$  are easy to check for  $x \geq 0$  from the series expansion

$$\Delta(x) = 2 \sum_{k \geq 2} x^{k-2}/k!.$$

However, even the monotonicity is less obvious when  $x$  is negative. Brute force calculation of the derivative,

$$d\Delta(x)/dx = \frac{(x-2)e^x + x + 2}{x^3/2},$$

didn’t help me much when I first encountered  $\Delta$  in the Chow and Teicher (1978, Section 11.1) book. A later edition of that book did use the shorter proof, almost as above. See also the proof by Freedman (1975, Lemma 3.1).

## 2.3 Convexity of the log MGF

Calculus::S:MGFconvexity

For a random variable  $X$ , the function  $M_X(\lambda) = \mathbb{P} \exp(\lambda X)$  for  $\lambda \in \mathbb{R}$  is called the moment generating function (MGF) for  $X$ . It is always finite at the origin because  $M_X(0) = \mathbb{P}1 = 1$ . The expectation is well defined and nonnegative for all real  $\lambda$ , although it might take the value  $+\infty$ . For example, if  $X$  has a CAUCHY distribution then  $M_X(\lambda) = +\infty$  for all nonzero  $\lambda$ ; and if  $X$  has a standard exponential distribution then

$$M_X(\lambda) = \int_0^\infty e^{\lambda x - x} dx = \begin{cases} (1 - \lambda)^{-1} & \text{for } \lambda < 1 \\ \infty & \text{otherwise} \end{cases}.$$

The function  $M_X$  inherits convexity from the exponential function. More precisely, the set

$$\text{EPI}(M_X) = \{(\lambda, t) \in \mathbb{R}^2 : t \geq M_X(\lambda)\}$$

is convex and closed as a subset of  $\mathbb{R}^2$ . The function  $M_X$  is a closed, proper ( $M_X(\lambda) > -\infty$ ), convex function, in the sense described by Rockafellar (1970, Section 7). Moreover, the restriction of the function  $M_X$  to the convex set  $\mathcal{D}_X := \{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$  is continuous and convex in the usual sense. If  $\text{int}(\mathcal{D}_X)$ , the interior of  $\mathcal{D}_X$ , is non-empty then  $M_X$  is infinitely differentiable with

\EQ MGF.derivs

<8>

$$M_X^{(k)}(\lambda) = \mathbb{P} \left( X^k e^{\lambda X} \right) \quad \text{for } k = 1, 2, \dots \text{ if } \lambda \in \text{int}(\mathcal{D}_X).$$

See Problem [3] for the (easy) proofs.

**Remark.** The differentiability can fail at the boundary of  $\mathcal{D}_X$ . For example, if  $X$  has the distribution with density  $x^{-2}\{x \geq 1\}$  with respect to Lebesgue measure then  $M_X(\lambda) < \infty$  iff  $\lambda \leq 0$ . The left-hand derivative at  $\lambda = 0$  is infinite: the ratio  $(M_X(0) - M_X(-h))/h$  tends to  $+\infty$  as  $h$  decreases to zero.

The function  $L_X := \log M_X$  inherits from  $M_X$  the lower semi-continuity and infinite differentiability on the interior of  $\mathcal{D}_X$ . More surprisingly,  $L_X$  is also a proper, closed convex function. The convexity follows from the HÖLDER inequality: for nonnegative random variables  $Y_1$  and  $Y_2$ : for positive constants  $\alpha_i$  with  $\alpha_1 + \alpha_2 = 1$ ,

$$\mathbb{P}(Y^{\alpha_1} Y_2^{\alpha_2}) \leq (\mathbb{P}Y)^{\alpha_1} (\mathbb{P}Y_2)^{\alpha_2}.$$

Taking  $Y_i = \exp(\lambda_i X)$  we get

$$M_X(\alpha_1 \lambda_1 + \alpha_2 \lambda_2) \leq M_X(\lambda_1)^{\alpha_1} M_X(\lambda_2)^{\alpha_2}.$$

By taking logs we get

$$L_X(\alpha_1 \lambda_1 + \alpha_2 \lambda_2) \leq \alpha_1 L_X(\lambda_1) + \alpha_2 L_X(\lambda_2).$$

**Remark.** If you worry about things like  $\log(6\infty) = \log(6) + \log(\infty)$  you could instead note that the restriction of  $L_X$  to  $\mathcal{D}_X$  is convex in the usual sense, with  $L_X(\lambda) = \infty$  outside  $\mathcal{D}_X$ , which is equivalent to convexity of  $\text{EPI}(L_X)$  as a subset of  $\mathbb{R}^2$ .

Formula <8> leads to interesting expressions for the first two derivatives of  $L_X$  on  $\text{int}(\mathcal{D}_X)$ :

$$\begin{aligned}\dot{L}(\lambda) &= \frac{\dot{M}_X(\lambda)}{M_X(\lambda)} = \mathbb{P}X \frac{e^{\lambda X}}{M_X(\lambda)} \\ \ddot{L}(\lambda) &= \frac{\ddot{M}_X(\lambda)}{M_X(\lambda)} - \left( \frac{\dot{M}_X(\lambda)}{M_X(\lambda)} \right)^2 = \mathbb{P}X^2 \frac{e^{\lambda X}}{M_X(\lambda)} - \left( \mathbb{P}X \frac{e^{\lambda X}}{M_X(\lambda)} \right)^2.\end{aligned}$$

If we define a probability measure  $\mathbb{P}_\lambda$  by its density,  $d\mathbb{P}_\lambda/d\mathbb{P} = e^{\lambda X}/M_X(\lambda)$ , then

$$\boxed{\text{E@ } L_X.\text{derivs}} \quad <9> \quad \dot{L}(\lambda) = \mathbb{P}_\lambda X \quad \text{AND} \quad \ddot{L}(\lambda) = \mathbb{P}_\lambda (X - \mathbb{P}_\lambda X)^2 = \text{var}_\lambda(X) \geq 0,$$

with strict inequality except in the degenerate case where  $P$  concentrates at a single point. Thus  $L$  is strictly convex except in the degenerate case. In particular, if  $0 \in \text{int}(\mathcal{D}_X)$  then  $\dot{L}(0) = \mathbb{P}X$ , so that Taylor expansion gives

$$\boxed{\text{E@ } L.\text{Taylor}} \quad <10> \quad L(\lambda) = \lambda \mathbb{P}X + \frac{1}{2} \lambda^2 \text{var}_{\lambda^*}(X) \quad \text{with } \lambda^* \text{ lying between } 0 \text{ and } \lambda.$$

I was quite surprised when I first learned how often <9> lurks behind useful facts. Problem [7] presents an example, which I learned from one of my colleagues only last year.

**Remark.** Bounds on the  $\text{var}_{\lambda^*}(X)$ , for some hard to calculate  $\lambda^*$  that depends on  $\lambda$ , lead to bounds on  $L$  and tail bounds for  $X - \mathbb{P}X$ . The proof of the Hoeffding inequality in Chapter 7 uses the representation <10> to control the MGF of a bounded random variable.

## 2.4 Conjugate convex functions

Calculus::S:Fenchel

Many probability bounds involve a minimization problem of the form

for each  $w \in \mathbb{R}$  minimize over  $\lambda \in J$  the function  $f(\lambda) - \lambda w$ ,

where  $J$  is a nonempty convex subset of the real line, and  $f : J \rightarrow \mathbb{R}$ . Of course a convex  $J$  is some sort of interval, such as  $\mathbb{R}$  itself or  $(-\infty, 3]$  or  $[-7, 3)$ , and so on. The role of  $J$  can be ignored if we extend  $f$  to a function taking values in  $\mathbb{R} \cup \{\infty\}$  by defining  $f(\lambda) = +\infty$  for  $\lambda \notin J$ . While we are at it, we should allow for the possibility that there might be no minimizing value by defining

\E@ conjugate <11>

$$g(w) = \inf_{\lambda \in \mathbb{R}} (f(\lambda) - \lambda w) \quad \text{for } w \in \mathbb{R}.$$

Equivalently,

$$-g(w) = \sup_{\lambda \in \mathbb{R}} (\lambda w - f(\lambda)) \quad \text{for } w \in \mathbb{R}.$$

The function  $f^* = -g$  is called the **convex conjugate** of  $f$ . It is also called the FENCHEL-LEGENDRE transformation. The name ‘conjugate’ was applied by Fenchel (1949). See also Kiselman (2019), which includes a translation of a letter from Fenchel that explains the connection with earlier work of Legendre and others.

**Remark.** Note that for any function  $\ell : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}$ ,

$$\ell(w) + f(\lambda) \geq w\lambda \quad \text{for all } w, \lambda \in \mathbb{R}$$

if and only if  $\ell(w) \geq f^*(w)$  for all  $w$ .

The calculation of the infimum in <11> is easier when  $f$  itself is convex. It is then usually easy to crank out  $g$  by brute force calculus by finding where the derivative is positive or negative or zero. In Chapter 3 you will encounter several such cases where  $f(\lambda) = \log \mathbb{P}e^{\lambda X}$  for a random variable  $X$ .

It is also convenient that  $g$ , as an infimum of linear functions of  $w$ , always turns out to be concave and  $f^*$  is always convex. The convexity can also be seen from consideration of the **epigraph**:

$$\begin{aligned} \text{epi}(f^*) &= \{(t, w) \in \mathbb{R}^2 : t \geq f^*(w)\} \\ &= \cap_{\lambda \in \mathbb{R}} \{(t, w) \in \mathbb{R}^2 : t - \lambda w \geq -f(\lambda)\}, \end{aligned}$$

an intersection of closed half-spaces. The epigraph is closed as a (nonempty) subset of  $\mathbb{R}^2$ , not just convex. That makes  $f^*$  a closed, proper convex function in the sense used by Rockafellar (1970, Section 7).

**Remark.** For a proper, convex function with  $J = \{\lambda : f(\lambda) < \infty\}$  the closedness property just involves the values of  $f$  at the endpoints of  $J$ , if any. This subtlety need not concern us.

The general theory is interesting because  $f^{**} = (f^*)^*$  happens to be the largest closed, convex function for which  $f(w) \geq f^{**}(w)$  for all  $w$  (see [Rockafellar \(1970, pages 103-4\)](#)). In particular, if  $f$  itself is closed and convex then  $f^{**} = f$ . Such a property plays a central role in the study of convex duality ([Rockafellar, 1974](#)).

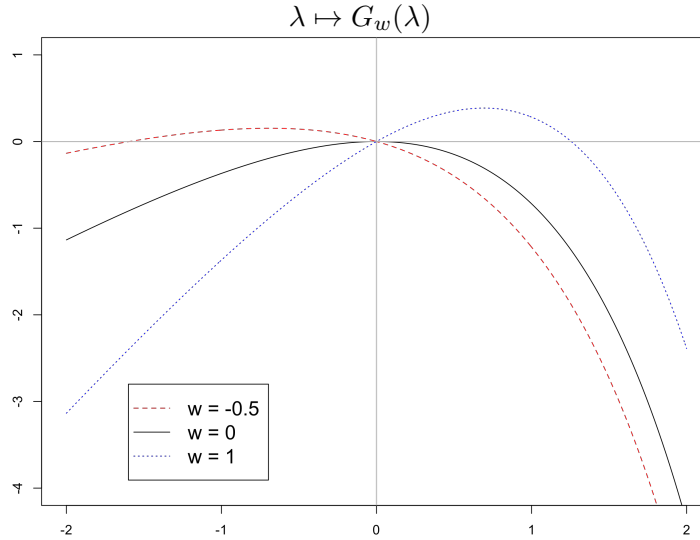
**Remark.** Unfortunately these wonderful duality facts about  $f^{**}$  don't seem to help much for the material in this book. I should perhaps look more carefully at the pages of [Boucheron et al. \(2000\)](#) pointed at by index entries of the form “duality ...”.

Calculus::exp <12>

**Example.** Consider the convex function  $\mathbb{f}(\lambda) = e^\lambda - 1 - \lambda$ , for  $\lambda \in \mathbb{R}$ , from Section 2.2. It is convex on the real line with derivative 0 at the origin. For each  $w \in \mathbb{R}$  the concave function

$$G_w(\lambda) = \lambda w - \mathbb{f}(\lambda) = \lambda(w + 1) - e^\lambda + 1$$

has  $\lambda$ -derivative  $G'_w(\lambda) = (1 + w) - e^\lambda$ .



If  $w > -1$  the maximum of  $G_w$  is achieved when  $G'_w(\lambda) = 0$ , that is, at  $\lambda = \log(1 + w)$ . For  $w = -1$  we have  $G_{-1}(\lambda) = 1 - e^\lambda$ , which approaches its supremum of 1 as  $\lambda \rightarrow \infty$ . For  $w < -1$ , we have  $G_w(\lambda) \uparrow +\infty$  as  $\lambda \downarrow -\infty$ . In summary,

\E@ fbb.star <13>

$$\mathbb{f}^*(w) = \begin{cases} (1 + w) \log(1 + w) - w & \text{if } w > -1; \text{ achieved at } \lambda = \log(1 + w) \\ 1 & \text{if } w = -1; \text{ approached as } \lambda \rightarrow \infty \\ +\infty & \text{if } w < -1; \text{ approached as } \lambda \rightarrow \infty \end{cases}$$

That is,  $\mathbb{f}^* = \mathbb{h}$ , as a map from  $\mathbb{R}$  to  $(-\infty, +\infty]$ .

The general duality theory asserts that

$$\mathbb{f}(\lambda) = \mathbb{h}^*(\lambda) = \sup_{w \in \mathbb{R}} w\lambda - \mathbb{h}(w) = \sup_{w \geq -1} w\lambda - \mathbb{h}(w).$$



Let me check.

For  $\lambda \in \mathbb{R}$  the supremum defining  $\mathfrak{h}^*(\lambda)$  need only run over  $w \geq -1$  because  $\lambda w - \mathfrak{h}(w) = -\infty$  for  $w < -1$ .

The derivative of the concave function  $w \mapsto \lambda w - \mathfrak{h}(w)$  is  $\lambda - \log(1+w)$ , which tends to  $+\infty$  as  $w \rightarrow -1$  and  $-\infty$  as  $w \rightarrow \infty$ . For the maximizing  $w$  we have  $\lambda = \log(1+w)$ , which gives

$$\mathfrak{h}^*(\lambda) = \lambda(e^\lambda - 1) - \lambda e^\lambda + (e^\lambda - 1) = e^\lambda - 1 - \lambda.$$

That is,  $\mathfrak{h}^* = \mathfrak{f}$ , as asserted by the general theory.

## 2.5 Approximation of factorials

Calculus::S:factorial

Abraham de Moivre (1756, page 244), before deriving his famous approximation to the Binomial distribution, made a comment about a constant  $B$  (which we now know is equal to  $\sqrt{2\pi}$ ), which he had approximated by adding up some terms of a slowly converging infinite series. He went on:

...seeing at the same time that what I had done answered my purpose tolerably well, I desisted from proceeding farther till my worthy and learned Friend Mr. *James Stirling*, who had applied himself after me to that inquiry, found that the Quantity  $B$  did denote the Square-root of the Circumference of a Circle whose Radius is Unity ...

Yes, it was the Stirling of the formula that is now often cited in the exquisitely sharp form: for  $n \in \mathbb{N}$ ,

$$\boxed{\backslash \mathbb{E} \text{ Stirling}} \quad \langle 14 \rangle \quad n! = C n^{n+1/2} e^{-n+r_n} \quad \text{where } b_n := \frac{1}{12n+1} < r_n < \frac{1}{12n} =: a_n,$$

where  $C$  is a universal constant. In fact, a simple CLT argument (see below) shows that  $C$  equals  $\sqrt{2\pi}$ .

**Remark.** In my opinion, the main value of the inequality  $b_n < r_n$  comes from the reassurance it gives that the upper bound is astoundingly sharp: the difference  $a_n - b_n$  between the upper and lower bounds for  $\log n!$  is of order  $O(n^{-2})$ . If instead we used the bounds  $0 < r_n < a_n$  the corresponding difference would be of order  $O(n^{-1})$ , still very respectable.

**Proof (of <14>).** The formula can be rewritten as

$$\boxed{\backslash \mathbb{E} \text{ sandwich}} \quad \langle 15 \rangle \quad b_n < r_n = d_n - \log C < a_n \quad \text{or} \quad d_n - a_n < \log C < d_n - b_n$$

$$\text{where } d_n := \log \left( n! e^n / n^{n+1/2} \right) = \log(n!) - (n + \tfrac{1}{2}) \log n + n.$$

This assertion will follow from the inequalities

$$\boxed{\backslash \mathbb{E} \text{ diffs}} \quad \langle 16 \rangle \quad \frac{\theta}{(n+\theta)(n+13\theta)} = b_n - b_{n+1} < d_n - d_{n+1} < a_n - a_{n+1} = \frac{\theta}{n(n+1)},$$

where  $\theta := 1/12$ . Indeed, these inequalities show that  $\{d_n - a_n : n \in \mathbb{N}\}$  is an increasing sequence and  $\{d_n - b_n : n \in \mathbb{N}\}$  is a decreasing sequence with  $d_n - a_n < d_n - b_n$ , which ensures existence of a finite constant  $c$  for which  $d_n - a_n \uparrow c$  and  $d_n - b_n \downarrow c$ . Inequality <15> will hold with  $C = e^c$ .

The proof of <16> is elementary and quite elegant. First write everything in terms of  $T_n = t_n^{-1} = 2n + 1$ . By direct calculation,

$$\begin{aligned} d_n - d_{n+1} &= (n + 1/2) \log \left( \frac{n+1}{n} \right) - 1 = \frac{1}{2} T_n \log \left( \frac{T_n + 1}{T_n - 1} \right) - 1 \\ &= \frac{1}{2t_n} \log \left( \frac{1+t_n}{1-t_n} \right) - 1 \\ &= \frac{t_n^2}{3} + \frac{t_n^4}{5} + \frac{t_n^6}{6} + \dots \quad \text{by TAYLOR expansion} \\ &< \frac{t_n^2}{3} + \frac{t_n^4}{3} + \frac{t_n^6}{3} + \dots = \frac{t_n^2}{3} (1 - t_n^2)^{-1} = \frac{1}{3} (T_n^2 - 1)^{-1} \\ &= \frac{\theta}{n(n+1)} = a_n - a_{n+1}. \end{aligned}$$

Similarly,

$$d_n - d_{n+1} > \frac{t_n^2}{3} = \frac{1}{3T_n^2} = \frac{\theta}{n^2 + n + 1/4} > b_n - b_{n+1},$$

the final inequality coming from

$$\begin{aligned} (n^2 + 14n\theta + 13\theta^2) - (n^2 + 12n\theta + 36\theta^2) &= 2n\theta - 23\theta^2 \\ &\geq \theta(2 - 23/12) = \theta^2. \end{aligned}$$

It remains only to show that  $C$  equals  $\sqrt{2\pi}$ . Here we can make use of a charming connection between  $C$  and  $p_n(i) := \mathbb{P}\{X_n = i\} = e^{-n} n^i / i!$  for  $X_n \sim \text{POISSON}(n)$ . The maximum of  $p_n(i)$  is achieved at  $i = n$ , with

$$\sqrt{n} p_n(n) = e^{-n} n^{n+1/2} / n! = e^{-r_n} / C \rightarrow 1/C \quad \text{as } n \rightarrow \infty.$$

Thus  $C = \sqrt{2\pi}$  iff  $\sqrt{n} p_n(n) \rightarrow 1/\sqrt{2\pi}$ . By the central limit theorem we have  $Z_n := n^{-1/2}(X_n - n) \rightsquigarrow \gamma := N(0, 1)$ , which suggests that

$$p_n(n) = \mathbb{P}\{n < X_n \leq n+1\} = \mathbb{P}\{0 < Z_n \leq n^{-1/2}\} \approx \gamma(0, n^{-1/2}] \approx \frac{1}{\sqrt{2\pi n}}.$$

The usual CLT does not quite justify all this handwaving but the basic idea works if we replace the  $n^{-1/2}$  by a small positive number  $\delta$ . With  $k := k_n(\delta) := \lfloor \delta\sqrt{n} \rfloor$  we then have

$$\begin{aligned} \mathbb{P}\{0 < Z_n \leq \delta\} &= \sum_{i=1}^k p_n(n+i) \rightarrow \gamma(0, \delta] \quad \text{as } n \rightarrow \infty, \text{ by CLT,} \\ &\text{and } \gamma(0, \delta]/\delta \rightarrow (2\pi)^{-1/2} \quad \text{as } \delta \rightarrow 0. \end{aligned}$$

For  $1 \leq i \leq k$  we have

$$\frac{p_n(n+i)}{p_n(n)} = \frac{n^i n!}{(n+i)!} > \prod_{i=1}^k (1+i/n)^{-1} \geq (1+k/n)^{-k} \quad \text{for } k = \lfloor \delta \sqrt{n} \rfloor.$$

Thus

$$1 \geq \sum_{i=1}^k \frac{p_n(n+i)}{k p_n(n)} \geq \exp(-k \log(1+k/n)) \rightarrow e^{-\delta^2} \quad \text{as } n \rightarrow \infty,$$

which rearranges to

$$(\delta + o(1)) \sqrt{n} p_n(n) \geq \gamma(0, \delta] + o(1) \geq (\delta + o(1)) \sqrt{n} p_n(n) (e^{-\delta^2} + o(1)),$$

with the  $o(1)$  converging to zero as  $n \rightarrow \infty$ . In the limit we have

$$e^{\delta^2} \gamma(0, \delta] / \delta \geq \limsup_n \sqrt{n} p_n(n) \geq \liminf_n \sqrt{n} p_n(n) \geq \gamma(0, \delta] / \delta$$

□ for each  $\delta > 0$ . It follows that  $\sqrt{n} p_n(n) \rightarrow 1/\sqrt{2\pi}$ , as hoped for.

The factorial function can be extended to a continuous domain by means of the **gamma function** defined by

$$\boxed{\backslash \mathbb{E} \text{ Gdef}} \quad \langle 17 \rangle \quad \Gamma(x) := \int_0^\infty r^{x-1} e^{-r} dr \quad \text{for } x > 0.$$

It has the following properties.

- (i)  $\log \Gamma$  is convex (by either of the methods from Section 2.3).
- (ii)  $\Gamma(x+1) = x\Gamma(x)$  for  $x > 0$  (integration-by-parts).
- (iii)  $\Gamma(n+1) = n!$  for  $n \in \mathbb{N}$  (from (ii) and  $\Gamma(1) = 1$ ).
- (iv)  $\inf_{x>0} \Gamma(x) \approx 0.89$ , achieving its minimum near 1.46 (according to **R**).
- (v)  $\Gamma(x) = \sqrt{2\pi} \exp((x-1/2) \log x - x + R(x))$  with  $0 < R(x) < (12x)^{-1}$  for all  $x > 0$ . See Problem [6].

Approximation (v) will be most helpful in Chapter 15.

**Example.** A lot of asymptotic statistical theory involves the binomial coefficients:

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \quad \text{for } i = 0, 1, \dots, n.$$

To avoid an overabundance of parentheses, for  $0 < i < n$  define  $\eta = i/n = 1 - \gamma$  and  $\epsilon_{i,n} = r_i + r_{n-i} - r_n$ . Then

$$\begin{aligned} \binom{n}{i} &= \frac{n!}{i!(n-i)!} = \frac{n^{n+1/2} \exp(\epsilon_{i,n})}{\sqrt{2\pi} (n\eta)^{i+1/2} (n\gamma)^{n-i+1/2}} \quad \text{by } \langle 14 \rangle \\ &= \frac{1}{\sqrt{2\pi n\eta\gamma}} \exp \left( -n \left[ \eta \log \frac{1}{\eta} + \gamma \log \frac{1}{\gamma} \right] + \epsilon_{i,n} \right). \end{aligned}$$

The  $\epsilon_{i,n}$  is usually ignored if both  $i$  and  $n-i$  are large. The term within square brackets in the exponent is called the **entropy** of the  $\text{BER}(\eta)$  distribution, a quantity much loved by information theorists (Cover and Thomas, 2012, Example 2.1.1). When multiplied by  $n$  it gives the entropy of the  $\text{BIN}(n, \eta)$  distribution.

For many purposes a somewhat cruder upper bound for  $\binom{n}{i}$  suffices. For example, if  $1 \leq i \leq d \leq n$ ,

$$\binom{n}{i} = \frac{n(n-1)\dots(n-i+1)}{i!} < \left(\frac{n}{d}\right)^i \frac{d^i}{i!} < \left(\frac{n}{d}\right)^d \frac{d^i}{i!},$$

implying

$$\sum_{0 \leq i \leq d} \binom{n}{i} \leq 1 + \sum_{1 \leq i \leq d} \left(\frac{n}{d}\right)^d \frac{d^i}{i!} < \left(\frac{en}{d}\right)^d.$$

This inequality plays a central role in the ultra-famous theory initiated by Vapnik & Červonenkis. See Section 14.2.

<19> **Example.** If  $X \sim \text{BIN}(n, p)$  and  $q = 1 - p$  then, again writing  $\eta = 1 - \gamma$  for  $i/n$ , we have

$$\begin{aligned} \mathbb{P}\{X = i\} &= \binom{n}{i} p^i q^{n-i} = \binom{n}{i} \exp(n\eta \log p + n\gamma \log q) \\ &= \frac{1}{\sqrt{2\pi n\eta\gamma}} \exp\left(-n\left[\eta \log \frac{p}{\eta} + \gamma \log \frac{q}{\gamma}\right] + \epsilon_{i,n}\right). \end{aligned}$$

The quantity

$$D_{KL}(\eta \parallel p) := \eta \log(\eta/p) + \gamma \log(\gamma/q), \quad \text{for } q = 1 - p \text{ and } \gamma = 1 - \eta,$$

is called the KULLBACK-LEIBLER distance (or KL-divergence) between the  $\text{BER}(\eta)$  and  $\text{BER}(p)$  distributions. Multiplication by  $n$  gives the KL-divergence between the  $\text{BIN}(n, \eta)$  and the  $\text{BIN}(n, p)$  distributions.

The KL-divergence is also closely related to the function  $\psi_{\text{benn}}$ , defined (in <6>) by  $\frac{1}{2}t^2\psi_{\text{benn}}(t) = \mathbb{h}(t) = (1+t)\log(1+t) - t$ . If  $i = np + x$  then  $\eta = p(1 + x/(np))$  and  $\gamma = q(1 - x/(nq))$  then  $D(\text{BIN}(n, \eta) \parallel \text{BIN}(n, p))$  equals

$$\begin{aligned} &np \left(1 + \frac{x}{np}\right) \log \left(1 + \frac{x}{np}\right) + nq \left(1 - \frac{x}{nq}\right) \log \left(1 - \frac{x}{nq}\right) \\ &= np \times \mathbb{h}\left(\frac{x}{np}\right) + nq \times \mathbb{h}\left(\frac{-x}{nq}\right) + np \left(\frac{x}{np}\right) + nq \left(\frac{-x}{nq}\right) \\ &= \frac{x^2 g(x, n, p)}{2npq} \quad \text{where } g(x, n, p) := q\psi_{\text{benn}}\left(\frac{x}{np}\right) + p\psi_{\text{benn}}\left(\frac{-x}{nq}\right). \end{aligned}$$

In summary, for  $0 < i = np + x < n$  and  $X \sim \text{BIN}(n, p)$  we have

$$\boxed{\text{\texttt{\textbackslash E@ Bin.rep}}} <20> \quad \mathbb{P}\{X = i\} = \frac{1}{\sqrt{2\pi npq}} \exp(-\mathcal{Q}(x)) \times \Lambda(x, n, p)$$

where

$$\begin{aligned}\mathcal{Q}(x) &:= np \times \mathfrak{h}\left(\frac{x}{np}\right) + nq \times \mathfrak{h}\left(\frac{-x}{nq}\right) = \frac{x^2 g(x, n, p)}{2npq} \\ g(x, n, p) &= q\psi_{\text{benn}}\left(\frac{x}{np}\right) + p\psi_{\text{benn}}\left(\frac{-x}{nq}\right) \\ \log \Lambda(x, n, p) &:= r_{np+x} + r_{nq-x} - r_n - \frac{1}{2} \log [(1 + x/(np))(1 - x/(nq))].\end{aligned}$$

The constraint  $0 < i < n$  is only needed to ensure that there are no divisions by 0 in the definition of  $\Lambda$ .

If  $\min(np, nq)$  is large and  $x/\min(np, nq)$  is small then  $\Lambda(x, n, p) \approx 1$  and  $g(x, n, p) \approx 1$ , so that

$$\mathbb{P}\{X = i\} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{x^2}{2npq}\right).$$

This result should remind you of the normal approximation to the  $\text{BIN}(n, p)$ , but not just as an approximation for tail probabilities. A more careful handling of the errors of approximation would lead to a so-called **local limit theorem** for the BINOMIAL distribution (Petrov, 1975, Chapter 7).

If you don't like great gobs of algebra you could safely skip the rest of this Example. I am just about to make a small point about a refinement of the normal approximation to the BINOMIAL, with another advertisement for the methods from Section 2.2.

The appearance of the convex function  $\psi_{\text{benn}}$  in the definition of  $g(x, n, p)$  tell us something surprising about departures from normality. By convexity,

\E@ Bennpsi.convex <21>

$$g(x, n, p) \geq \psi_{\text{benn}}\left(q\frac{x}{np} + p\frac{-x}{nq}\right) = \psi_{\text{benn}}\left(\frac{(q-p)x}{npq}\right).$$

If  $p > 1/2$  then  $g(x, n, p) > 1$  for  $x > 0$ . We get something smaller than the normal exponent  $-x^2/(2npq)$  in the right tail beyond the mean  $np$ . This effect can be attributed to the skewness,  $\mathbb{P}(X - np)^3 / \text{var}^{3/2}(X) = npq(q-p)/(npq)^{3/2}$ , which is negative if  $p > 1/2$ . To a first approximation, the  $\text{BIN}(n, p)$  is symmetric around its mean  $np$ . To a finer approximation, if  $p > 1/2$  then the probabilities are slightly smaller in the left tail than in the right tail.

By the principle of no free lunches, if skewness gives us a nicer right tail then it will probably give us an uglier left tail. That is in fact the case but it cannot be deduced directly from inequality <21>. Instead we need to identify some higher order terms in the series expansion of  $\mathcal{Q}(x)$ . From

$$\begin{aligned}\dot{\mathfrak{h}}(x) &= \log(1+x) & \ddot{\mathfrak{h}}(x) &= 1/(1+x) \\ \mathfrak{h}^{(3)}(x) &= -1/(1+x)^2 & \mathfrak{h}^{(4)}(x) &= 2/(1+x)^3\end{aligned}$$

and Problem [1] we get

$$\mathfrak{h}(x) = \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{3} \int_0^1 \frac{(1-s)^3}{(1+sx)^3} ds \quad \text{for } x > -1$$

and

$$\begin{aligned}
\mathcal{Q}(x) &= nq\mathfrak{h}(x/np) + np\mathfrak{h}(-x/nq) \\
&= np \left( \frac{x^2}{2(np)^2} - \frac{x^3}{6(np)^3} + \frac{x^4}{3(np)^4} \int_0^1 \frac{(1-s)^3}{(1+sx/np)^3} ds \right) \\
&\quad + nq \left( \frac{x^2}{2(nq)^2} + \frac{x^3}{6(nq)^3} + \frac{x^4}{3(nq)^4} \int_0^1 \frac{(1-s)^3}{(1-sx/nq)^3} ds \right) \\
&= \frac{x^2}{2npq} + \frac{x^3(q-p)}{6n^2p^2q^2} \\
&\quad + \frac{x^4}{3} \int_0^1 (1-s)^3 [(np+sx)^{-3} + (nq-sx)^{-3}] ds.
\end{aligned}$$

The pattern is clearer if we consider terms  $i = np + x$  with  $x = y\sqrt{npq}$ , that is, if we scale by the standard error  $\sigma = \sqrt{npq}$  for  $X$ :

$$\begin{aligned}
\mathcal{Q}(y\sigma) &= \frac{y^2}{2} + \frac{y^3(q-p)}{6\sqrt{npq}} \\
&\quad + \frac{y^4p^2q^2}{3n} \int_0^1 (1-s)^3 [(p+sy\sigma/n)^{-3} + (q-sy\sigma/n)^{-3}] ds.
\end{aligned}$$

The  $y^2$  term contributes the normal approximation; the  $y^3$  term contributes a skewness perturbation of order  $n^{-1/2}$ ; and the  $y^4$  term scoops up the leftovers, which are of order  $n^{-1}$ . If you like this kind of thing then you will love [Petrov \(1975, Chapter 6\)](#).  $\square$

## 2.6 Distances between probability measures

Calculus::S:divergence

Here is another example of  $\psi_{\text{benn}}$  popping up in unexpected places.

Modern statistical theory makes extensive use of various measures of ‘distance’ between probability measures  $P$  and  $Q$  defined on the same sigma-field. To keep the following discussion simple I’ll assume that both probabilities are defined on the same finite set  $\mathcal{X}$  with  $Q\{x\} > 0$  for each  $x \in \mathcal{X}$ . Define

$$\rho(x) = \frac{P\{x\}}{Q\{x\}} \quad \text{for } x \in \mathcal{X}.$$

For each real-valued function  $g$  on  $\mathcal{X}$  think of its  $Q$  expectation as the linear functional

$$Qg = Qg(x) := \sum_{x \in \mathcal{X}} Q\{x\}g(x).$$

**Remark.** The quantity  $\rho(x)$  is sometimes called the *likelihood ratio*. If you desire greater generality, feel free to extend the arguments to general sigma-fields with  $P$  absolutely continuous with respect to  $Q$  with density (RADON-NIKODYM derivative)  $\rho = dP/dQ$ . You will also need to worry about integrability. I would even forgive you for adopting the hideous notation  $\mathbb{E}_Q g$  for  $Qg$ , but the even more hideous  $\mathbb{E}_{x \sim Q} g(x)$  is beyond the pale, in my humble opinion.

For each convex function  $f$  on  $\mathbb{R}^+$  for which  $f(1) = 0$ , the  $f$ -divergence is defined as  $D_f(P \parallel Q) := Qf(\rho(x))$ . Three particularly popular special cases correspond to the functions  $f(x) = |x - 1|$  (the  $L^1$  distance), and  $f(x) = x \log x$  (the KULLBACK-LEIBLER divergence) and  $f(x) = (x - 1)^2$  (the  $\chi^2$ -divergence):

$$\begin{aligned}\|P - Q\|_1 &:= Q|\rho(x) - 1| = \sum_x |P\{x\} - Q\{x\}| \\ D_{KL}(P \parallel Q) &:= Q\rho(x) \log(\rho(x)) = \sum_x P\{x\} \log(P\{x\}/Q\{x\}) \\ D_{\chi^2}(P \parallel Q) &:= Q(\rho(x) - 1)^2 = \sum_x \frac{(P\{x\} - Q\{x\})^2}{Q\{x\}}\end{aligned}$$

Of these, only the  $L^1$  distance is a metric.

The relationship between the three divergences becomes clearer when they are written as integrals involving  $\delta(x) = \rho(x) - 1$ :

$$\begin{aligned}\|P - Q\|_1 &:= Q|\delta| \\ D_{KL}(P \parallel Q) &:= Q\mathfrak{h}(\delta) = \frac{1}{2}Q(\delta^2\psi_{\text{benn}}(\delta)) \\ D_{\chi^2}(P \parallel Q) &:= Q\delta^2\end{aligned}$$

As before,  $\mathfrak{h}(x) = (1 + x) \log(1 + x) - x$ . The extra  $-x$  does not change the integral because  $Q\delta = Q\rho - 1 = 1 - 1$ .

The divergences are related by the inequalities

$$D_{\chi^2}(P \parallel Q) \geq D_{KL}(P \parallel Q) \geq \frac{1}{2} \|P - Q\|_1^2.$$

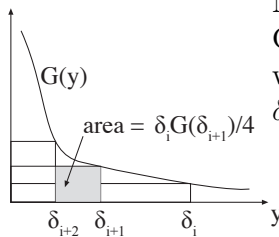
I don't know if anyone has claimed parentage for the first inequality, which follows from the fact that  $\psi_{\text{benn}}(\delta) \leq 2$ . The second inequality, which is often called the PINSKER inequality, follows from CAUCHY-SCHWARZ:

$$\begin{aligned}\frac{1}{2} (Q|\delta|)^2 &= \frac{1}{2} Q \left( \frac{|\delta|}{\sqrt{1 + \delta/3}} \sqrt{1 + \delta/3} \right)^2 \\ &\leq \frac{1}{2} Q \left( \frac{\delta^2}{1 + \delta/3} \right) Q(1 + \delta/3) \quad \text{by CAUCHY-SCHWARZ} \\ &\leq \frac{1}{2} Q(\delta^2\psi_{\text{benn}}(\delta)) \quad \text{by } \langle 7 \rangle \text{ and } Q\delta = 0.\end{aligned}$$

I learned of this neat argument from [Kemperman \(1969\)](#).

## 2.7 Bounding sums by integrals

calculus::S:sum.to.integral



Many approximation schemes (particularly those that will be discussed in Chapter 10) bound useful quantities by sums of the form  $\sum_{i=0}^k \delta_i G(\delta_{i+1})$ , with  $G(\cdot)$  a decreasing function on  $\mathbb{R}^+$ . If the  $\delta_i$ 's decrease geometrically, say  $\delta_i = \delta_0/2^i$ , then the sum can be bounded above by an integral,

$$\sum_{i=0}^k \delta_i G(\delta_{i+1}) \leq 4 \int_{\delta_{k+2}}^{\delta_1} G(r) dr \leq 4 \int_0^{\delta_1} G(r) dr.$$

If  $G$  is integrable the replacement of the lower terminal by 0 costs little.

In the early empirical process literature, some authors chose the  $\delta_i$ 's to make the  $G(\delta_i)$ 's increase geometrically, for similar reasons. (Use horizontal slices.)

In delicate situations, sometimes the  $\{\delta_i\}$  sequence needs to be chosen in a more cunning way. See Section [Dudley \(1973, p 74\)](#), for example.

## 2.8 Problems

Calculus::S:Problems

Calculus::P:Taylor.Rk

- [1] Suppose  $g$  is a real-valued function, defined at least on an interval  $J$  of the real line that contains both 0 and a point  $x$ , that is  $(k+1)$ -times continuously differentiable. For  $x$  in the interior of  $J$  show that

$$\begin{aligned} R_k(x, g) &:= g(x) - \left[ g(0) + x\dot{g}(0) + \cdots + x^k g^{(k)}(0)/(k!) \right] \\ &= x^{k+1} \int \cdots \int \{0 \leq t_{k+1} \leq t_k \leq \cdots \leq t_1 \leq 1\} g^{(k)}(t_{k+1}x) dt_{k+1} \cdots dt_1 \\ &= x^{k+1} \int_0^1 \frac{(1-s)^k}{k!} g^{(k+1)}(sx) ds. \end{aligned}$$

Calculus::P:eix

- [2] By splitting into real and imaginary parts, show that

$$e^{ix} - \left( 1 + ix + (ix)^2/2! + \cdots + (ix)^k/k! \right) = x^{k+1} \int_0^1 \frac{(1-s)^k}{k!} i^{k+1} e^{isx} ds.$$

for all real  $x$ . Deduce that

$$\left| e^{ix} - \left( 1 + ix + (ix)^2/2! + \cdots + (ix)^k/k! \right) \right| \leq \frac{|x|^{k+1}}{(k+1)!},$$

an inequality that is useful for arguments involving FOURIER transforms.

Calculus::P:smooth.MGF

- [3] Suppose  $M_X(\lambda) = \mathbb{P}e^{\lambda X}$  for  $\lambda \in \mathbb{R}$  and  $\mathcal{D}_X = \{\lambda \in \mathbb{R} : M_X(\lambda) < \infty\}$ .
- (i) Show that  $M_X(\lambda)$  is lower semi-continuous at each point of  $\mathbb{R}$ . Equivalently,  $\text{EPI}(M_X)$  is a closed subset of  $\mathbb{R}^2$ . Hint: If  $\lambda_n \rightarrow \lambda \in \mathbb{R}$ , the FATOU lemma tells you something about  $\liminf_n \mathbb{P}e^{\lambda_n X}$ . Argue similarly if  $(t_n, \lambda_n) \in \text{EPI}(M_X)$  and  $(t_n, \lambda_n) \rightarrow (t, \lambda) \in \mathbb{R}^2$ . (As [Rockafellar \(1970, page 52\)](#) argued, lower semi-continuity of  $M_X$  is equivalent to  $\text{EPI}(M_X)$  being closed as a subset of  $\mathbb{R}^2$ .)
- (ii) Show that the restriction of  $M_X$  to  $\mathcal{D}_X$  is continuous as a function on  $\mathcal{D}_X$ . Hint: Suppose  $[\lambda, \gamma] \subset \mathcal{D}_X$  and  $\lambda_n = \alpha_n \lambda + (1 - \alpha_n) \gamma$ . Convexity of  $M_X$  gives

$$M_X(\lambda_n) \leq \alpha_n M_X(\lambda) + (1 - \alpha_n) M_X(\gamma). \quad \text{for } 0 \leq \alpha_n \leq 1.$$

Take the limsup as  $\alpha_n$  tends to 0 to deduce that  $\limsup M_X(\lambda_n) \leq M_X(\lambda)$ , which together with lower semi-continuity proves continuity from the right at  $\lambda$ . Argue similarly if  $\alpha_n \rightarrow 1$  for continuity from the left at  $\gamma$ .



- (iii) Suppose  $\lambda$  is an interior point of  $\mathcal{D}_X$ . Choose  $\delta > 0$  so that  $[\lambda - 2\delta, \lambda + 2\delta] \subset \mathcal{D}_X$ . For  $|h| < \delta$ , establish the domination bound

$$\begin{aligned} |e^{(\lambda+h)X} - e^{\lambda X}|/|h| &\leq \left| \int_0^1 X e^{\lambda X} e^{shX} ds \right| \\ &\leq e^{\lambda X} \left( e^{\delta X} + e^{-\delta X} \right)^2 / \delta, \end{aligned}$$

which is integrable. Deduce via a dominated convergence argument that  $M'(\lambda) = \mathbb{P}(X e^{\lambda X})$ .

- (iv) Argue similarly to show that  $M_X$  is infinitely differentiable on the interior of  $\mathcal{D}_X$ , with  $M_X^{(k)}(\lambda) = \mathbb{P}(X^k e^{\lambda X})$ .

Calculus::P:crude.Stirling

- [4] For  $k \in \mathbb{N}$  prove that  $e^k \geq k^k/k! \geq 1$ . Deduce that  $(k/e)^k \leq k! \leq k^k$  and  $k/e \leq (k!)^{1/k} \leq k$ . Compare with the sharper bounds given by the STIRLING formula:  $(k!)^{1/k} = \rho_k k/e$ , where  $\rho_k = \left( \sqrt{2\pi k} e^{r_k} \right)^{1/k}$  is a decreasing function of  $k$  that converges to 1 as  $k \rightarrow \infty$ .

$k$	1	2	3	4	5	6	7	8	9	10
$\rho_k$	2.718	1.922	1.646	1.504	1.416	1.356	1.313	1.279	1.253	1.231

Calculus::P:BohrM

- [5] (The Bohr-Mollerup theorem, apparently first stated in a 1922 textbook by H. Bohr and J. Mollerup. The following derivation is based on the exposition by [Artin, 1964](#).) Suppose  $f$  is a strictly positive function on  $(0, \infty)$  for which:

- (a)  $f(x+1) = xf(x)$  for each  $x > 0$ .
- (b)  $g := \log f$  is convex.
- (c)  $f(1) = 1$ .

Prove that  $f(x) = \Gamma(x)$  for all  $x > 0$  by arguing as follows.

- (i) Note that  $f(x+n) = f(x)\rho_n(x)$  and  $\Gamma(x+n) = \Gamma(x)\rho_n(x)$  for  $x > 0$  and  $n \in \mathbb{N}$ , where  $\rho_n(x) := x(x+1)\dots(x+n-1)$ . In particular,  $f(1+n) = n! = \Gamma(1+n)$  for each  $n \in \mathbb{N}_0$ .
- (ii) Convexity of  $g$  tells us that the slope  $H(a, b) := (g(b) - g(a))/(b - a)$ , for  $0 < a < b$ , is an increasing function of each argument. In particular, for an integer  $n \geq 2$  and  $0 < x \leq 1$  we have

$$\begin{aligned} g(n) - g(n-1) &= H(n-1, n) \leq H(n, n+x) = (g(n+x) - g(n))/x \\ &\leq H(n, n+1) = g(n+1) - g(n). \end{aligned}$$

Thus

$$(n-1)^x = \left( \frac{f(n)}{f(n-1)} \right)^x \leq \frac{f(n+x)}{f(n)} = \frac{f(x)\rho_n(x)}{(n-1)!} \leq n^x,$$

implying

$$(n-1)^x (n-1)! / \rho_n(x) \leq f(x) \leq n^x (n-1)! / \rho_n(x) \quad \text{for } n \geq 2.$$

The first inequality also holds with  $n$  replaced by  $n + 1$ . Thus

$$\frac{n^x n!}{\rho_{n+1}(x)} \leq f(x) \leq \frac{n^x (n-1)!}{\rho_n(x)} = \frac{n^x n!}{\rho_{n+1}(x)} \frac{x+n}{n} \quad \text{for } n \geq 2.$$

- (iii) The same inequalities also hold with  $f(x)$  replaced by  $\Gamma(x)$ , because the analogs of (a), (b), and (c) also hold for  $\Gamma$ . Taking the ratios we then get

$$n/(x+n) \leq f(x)/\Gamma(x) \leq (x+n)/n,$$

which in the limit as  $n \rightarrow \infty$  gives  $f(x) = \Gamma(x)$  for  $0 < x \leq 1$ . The factorizations in (i) extend the equality to all  $x > 0$ .

Calculus::P:GammaStirling

- [6] The STIRLING formula,

$$n! = \sqrt{2\pi} \exp((n + 1/2) \log n - n + r_n) \quad \text{with } (12n)^{-1} < r_n < (12n)^{-1},$$

has a continuous analog: for  $x > 0$ ,

$$\Gamma(x) = \sqrt{2\pi} \exp((x - 1/2) \log x - x + R(x)) \quad \text{with } 0 < R(x) < (12x)^{-1}.$$

Verify the following assertions to establish this approximation.

- (i) Define  $h(x) := (x - 1/2) \log x - x$  and  $f(x) := e^{h(x)+R(x)}$  for an  $R(x)$  to be determined. We have  $f(x+1) = xf(x)$  iff

$$\begin{aligned} R(x) - R(x+1) &= h(x+1) - h(x) - \log x \\ &= g(x) := (x + 1/2) \log(1 + x^{-1}) - 1 \\ &= (2y)^{-1} \log\left(\frac{1+y}{1-y}\right) - 1 \quad \text{where } y := 1/(2x+1) \\ &= y^2/3 + y^4/5 + \dots \quad \text{by TAYLOR for } \log(1 \pm y) \\ &\leq \frac{1}{3}(y^2 + y^4 + \dots) = \frac{y^2}{3(1-y^2)} \\ &= (12x(x+1))^{-1} = (12x)^{-1} - (12x+12)^{-1}. \end{aligned}$$

- (ii) Define

$$R(x) := \sum_{i \in \mathbb{N}_0} g(x+i) = \frac{1}{3(2x+1)^2} + \frac{1}{5(2x+1)^4} + \dots,$$

a convergent sum of positive convex functions. Thus  $0 < R(x) \leq (12x)^{-1}$  and  $R$  is convex.

- (iii) The function  $h$  is convex because  $\ddot{h}(x) = x^{-1} + (2x^2)^{-1} > 0$ . Thus  $\log f(x) := h(x) + R(x)$  is convex. By the Bohr-Mollerup theorem (an idea due to [Artin, 1964](#)),  $Cf(x) = \Gamma(x)$  for all  $x > 0$ , for some constant  $C$ . It remains only to show that  $C = \sqrt{2\pi}$ . To that end, note that STIRLING gives

$$Cf(n+1) = \Gamma(n+1) = n! = \sqrt{2\pi} \exp((n-1/2) \log n - n + \log n + r_n)$$

which rearranges to

$$\log(C/\sqrt{2\pi}) = h(n) + \log n + r_n - h(n+1) - R(n+1) = r_n - g(n) - R(n+1).$$

That is, the constant  $\log(C/\sqrt{2\pi})$  equals  $r_n - R(n) = o(1)$  as  $n \rightarrow \infty$ . It follows that  $C = \sqrt{2\pi}$ .

Calculus::P:condit.mean

- [7] (See [Wu and Zhou \(2021, §1.1\)](#) for an example of how the results from this Problem can be used.) Suppose  $X \sim P$  and  $Z \sim \gamma_1 := N(0, 1)$  are independent. Define  $U = \alpha X + \beta Z$  for positive constants  $\alpha$  and  $\beta$ . Define  $h(u) = \mathbb{P}(X \mid U = u)$ . Follow these steps to show that

- (a)  $h$  is an increasing function
- (b) If  $P$  is symmetric around 0 then  $h$  is an odd function.
- (i) Explain why the result is trivial if either of  $\alpha$  or  $\beta$  is zero. Deduce that, without loss of generality, we may suppose  $\alpha = \beta = 1$ .
- (ii) For suitably integrable  $f$ , show that

$$\mathbb{P}f(X, X + Z) = P^x \int_{\mathbb{R}} f(x, u) m(x, u) \phi(u) du$$

where  $m(x, u) = \phi(x - u)/\phi(u) = \exp(xu - x^2/2)$ . Deduce that: the joint distribution  $\mathbb{M}$  for  $(X, U)$  has density  $m$  with respect to  $P \otimes \gamma_1$ ; the marginal distribution (with respect to  $\gamma_1$ ) for  $U$  is  $q(u) = P^x m(x, u)$ ; and the conditional density (with respect to  $P$ ) for  $X$  given  $U = u$  is  $m(x \mid u) = m(x, u)/q(u)$ .

- (iii) Assuming the usual conditions for differentiating inside  $P$  integrals, deduce that  $u \mapsto \log q(u)$  is convex, so that its derivative,  $q'(u)/q(u)$ , is an increasing function.
- (iv) Use the fact that  $\partial m(x, u) = xm(x, u)$  to show that

$$h(u) = P^x xm(x, u)/q(u) = P^x \frac{\partial m(x, u)}{\partial u} / q(u) = q'(u)/q(u),$$

an increasing function.

- (v) Note that  $m(-x, -u) = m(x, u)$  for all  $x$  and  $u$ . For symmetric  $P$  deduce that

$$q(-u) = P^x m(x, -u) = P^x m(-x, -u) = q(u)$$

and

$$q(-u)h(-u) = q(u)P^x xm(x, -u) = q(u)P^x(-x)m(-x, -u) = -q(u)h(u),$$

which implies that  $h$  is an odd function.

## 2.9 Notes

Calculus::S:notes

See [Stigler \(1986, pages 70-76\)](#) for an illuminating discussion of the interaction between de Moivre and Stirling. I learned the STIRLING formula, as in <14>, from [Feller \(1968, page 52\)](#). He cited [Robbins \(1955\)](#), who commented that the only novelty in his proof was in the derivation of the inequality  $b_n < r_n$ ; he acknowledged that the standard parts of his argument were taken from [Darmois \(1928, pages 315-317\)](#). Robbins also commented that the “editor has pointed out that the inequalities ... permit the following brief proof”, which was a “modification of that attributed to Cesàro by A. Fisher, *Mathematical theory of probabilities*, New York, 1936, pp. 93-95.” (The 1955 Mathematical Monthly listed F. A. Ficken as the editor of the section containing Robbins’s note.) The Cesàro book was published in 1884. Even though I cannot read Italian, I feel the proof on page 270 is very similar to the one presented by Feller. As they say, there is nothing new under the sun.

[Feller \(1968, footnote to page 53\)](#) cited [Feller \(1967\)](#) for the identification of the limit of  $\{e^{d_n}\}$  as  $\sqrt{2\pi}$ . In another footnote [Feller \(1968, Section VII.2\)](#) recovered the  $\sqrt{2\pi}$  from a local limit theorem for the  $\text{BIN}(n, 1/2)$ . The argument in Section 2.5 for determining  $C$  is essentially that of [de Moivre \(1756\)](#), applied to the POISSON instead of the BINOMIAL. See also [Pitman \(1993, Section 2.3\)](#) for a beautiful exposition of de Moivre’s method. At the end of Section 3.6 of the same textbook, Pitman noted that de Moivre’s method also works with the HYPERGEOMETRIC distribution. It can be also be applied to other distributions  $P$  on the integers for which the ratios  $P\{i+1\}/P\{i\}$  have a tractable form. Similar ideas also play a role in the the study of log-concave discrete distributions and even in the approximations developed by Chen and Stein ([Barbour, Holst, and Janson, 1992, Section 9.2](#)).

Initially I included Section 2.4 mostly because I feared I had missed some subtle point made by [Boucheron, Lugosi, and Massart \(2013, Sections 2.2\)](#), who seemed to be suggesting that the FENCHEL-LEGENDRE transform plays a role far more important than just as a name for a minimization procedure. To me it didn’t seem very helpful just to know that a particular minimization of a convex function was an example of a technique named after famous mathematicians. Eventually my opinion changed when it became clearer to me that FENCHEL-LEGENDRE conjugates were lurking in the background for many useful ideas. See, for example, the discussion of the YOUNG inequality in Section 5.5. The general theory also extends easily to functions on  $\mathbb{R}^d$ , where it plays an important role in the study of large deviations for sums of independent random vectors—see [Dembo and Zeitouni \(1998, Chapter 2\)](#).

There is currently an archive of Fenchel’s papers at

<http://web.math.ku.dk/arkivet/fenchel/wfenpapr.htm>

There is also a most informative note (“Werner Fenchel, a pioneer in convexity theory and a migrant scientist” by Christer Oscar Kiselman) about Fenchel and duality at

<http://www.cb.uu.se/~kiselman/bibliography.html>

The note is listed on the web site as having appeared in *Normat. Nordisk matematisk tidskrift* 61, No. 2-4, 133-152, but I have only seen the online version.

## References

- [Artin1964Gamma](#) Artin, E. (1964). *The Gamma Function*. Holt, Rinehart and Winston. (Translated from the 1931 German original.).
- [BarbourHolstJanson92](#) Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson Approximation*. Oxford University Press.
- [Bennett62jasa](#) Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* 57, 33–45.
- [BoucheronLugosimassart2000ras](#) Boucheron, S., G. Lugosi, and P. Massart (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms* 16, 277–292.
- [BLM2013Concentration](#) Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- [ChowTeicher78book](#) Chow, Y. S. and H. Teicher (1978). *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer.
- [CoverThomas2012](#) Cover, T. M. and J. A. Thomas (2012). *Elements of Information Theory*. John Wiley & Sons.
- [Darmois1928](#) Darmois, G. (1928). *Statistique Mathématique*. Doin & cie.
- [deMoivre1756](#) de Moivre, A. (1756). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*. New York: Chelsea. Third edition (fuller, clearer, and more correct than the former), reprinted in 1967. First edition 1718.
- [DemboZeitouni98](#) Dembo, A. and O. Zeitouni (1998). *Large Deviations Techniques and Applications* (2 ed.). Springer-Verlag.
- [Dudley73gauss](#) Dudley, R. M. (1973). Sample functions of the Gaussian process. *Annals of Probability* 1, 66–103.
- [Feller1967AMM](#) Feller, W. (1967). A direct proof of Stirling’s formula. *The American Mathematical Monthly* 74(10), 1223–1225.
- [Feller1](#) Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.
- [Fenchel1949](#) Fenchel, W. (1949). On conjugate convex functions. *Canadian Journal of Mathematics* 1, 73–77.

[Freedman1975AnnProb](#)

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability* 3(1), 100–118.

[kemperman69](#)

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*. Springer-Verlag. Lecture Notes in Mathematics, 89, pages 126–169.

[Kiselman2018](#)

Kiselman, C. O. (2019). Werner Fenchel, a pioneer in convexity theory and a migrant scientist. *Normat: Nordisk matematisk tidskrift* 61(2–4), 133–152.

[Petrov1975](#)

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer-Verlag. English translation from 1972 Russian edition.

[Pitman1993Prob](#)

Pitman, J. (1993). *Probability*. Springer.

[Robbins1955AMM](#)

Robbins, H. (1955). A remark on Stirling’s formula. *The American Mathematical Monthly* 62(1), 26–29.

[Rockafellar70book](#)

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press.

[Rockafellar1974conjugate](#)

Rockafellar, R. T. (1974). *Conjugate Duality and Optimization*, Volume 16 of *Regional Conference Series in Applied Mathematics*. Siam.

[Stigler86book](#)

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, Massachusetts: Harvard University Press.

[WuZhou2021MSL](#)

Wu, Y. and H. H. Zhou (2021). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in  $O(\sqrt{n})$  iterations. *Mathematical Statistics and Learning* 4(3), 143–220.