APMA 1740 Homework 1

Milan Capoor

31 Jan 2025

For 1740 and 2610:

1. **Convex functions of one variable.** Loosely speaking, a convex function is bowl shaped. Formally speaking, a convex function satisfies

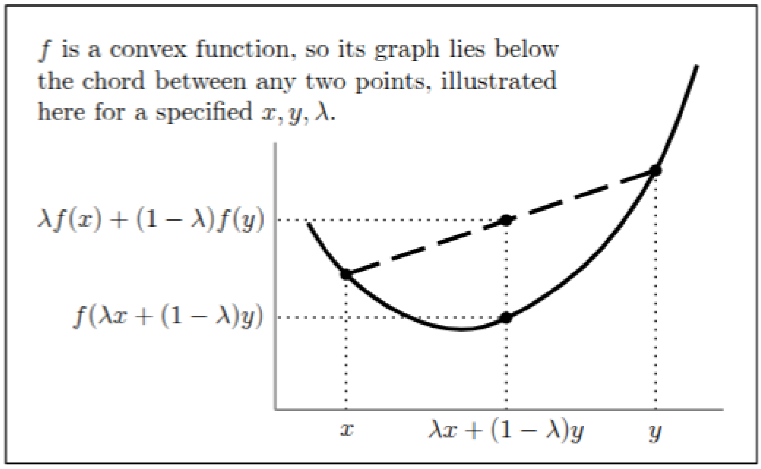
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for every distinct pair $x \neq y$ and every $0 \leq \lambda \leq 1$, which says that the graph of the function lies below the chord connecting any two points on the graph.[1]

If a function is everywhere twice differentiable, then it is convex if and only if (iff) its second derivative is nonnegative: $f''(x) \geq 0$ for all x. A function g is concave if $-g$ is convex. (So concave functions are like upside-down bowls.) A function is *strictly* convex if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

for every distinct pair $x \neq y$ and every $0 < \lambda < 1$, and this can be verified by further showing that $f''(x) > 0$ everywhere, except perhaps at isolated points.


f is a convex function, so its graph lies below the chord between any two points, illustrated here for a specified x, y, λ.
$\lambda f(x) + (1-\lambda)f(y)$
$f(\lambda x + (1-\lambda)y)$
x
$\lambda x + (1-\lambda)y$
y

(a) Prove that $f(x) = x^2$ is strictly convex. (Hint. Show that $f''(x) > 0$ for all x.)

$$\begin{aligned} f'(x) = 2x &\implies f''(x) = 2 > 0 \quad \forall x \\ &\implies f \text{ strictly convex.} \end{aligned}$$

[1] If you fix the points x and y and vary λ from 0 to 1, then $f(\lambda x + (1-\lambda)y)$ traces out the values of the function between x and y. Similarly, $\lambda f(x) + (1-\lambda)f(y)$ traces out the values of the straight line between $f(x)$ and $f(y)$. So $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ is a way to say that the graph of a function lies below the chord connecting any two points on the graph.

(b) Prove that $f(x) = \log x$ is strictly concave for $x > 0$.

$$\begin{aligned} g(x) := -\log x &\implies g' = -\frac{1}{x} \\ &\implies g'' = \frac{1}{x^2} > 0 \quad \forall x > 0 \\ &\implies g \text{ strictly convex} \\ &\implies f \text{ strictly concave.} \end{aligned}$$

(c) Prove that $f(x) = x \log x$ is strictly convex for $x > 0$.

$$f'(x) = 1 + \log x \implies f''(x) = \frac{1}{x} > 0 \quad \forall x > 0$$
$$f \text{ strictly convex.}$$

(d) Prove that $f(x) = ax + b$ is both convex and concave for any choice of a, b.

$$\begin{aligned} f'(x) = a &\implies f''(x) = 0 \geq 0 \quad \forall x \\ &\implies f \text{ convex} \\ g'(x) = -a &\implies f''(x) = 0 \geq 0 \quad \forall x \\ &\implies f \text{ concave} \end{aligned}$$

2. **Convex functions of many variables.** More generally, a convex function $f(x)$, where $x = (x_1, \ldots, x_s)$ is a vector, satisfies the same inequality:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for every distinct pair $x \neq y$ and every $0 \leq \lambda \leq 1$. The intuitions are nearly identical. And g is concave if $-g$ is convex. Verifying that a function of many variables is convex can be a little more tricky, however. One way to do it is to prove that the new function

$$h(\lambda) = f(\lambda x + (1-\lambda)y)$$

is convex for $0 \leq \lambda \leq 1$ and every fixed (and distinct) choice of x and y. This is nice, because h is now a function of one variable, so you might be able to compute the second derivative of h and verify that it is nonnegative: $h''(\lambda) \geq 0$ on $[0,1]$. If you can additionally verify that $h''(\lambda) > 0$ on $(0,1)$, then f is strictly convex (just like in problem 2).

(a) Prove that $f(x) = \sum_{i=1}^{s}(w_i x_i)^2$ is a convex function, where $w = (w_1, \ldots, w_s)$ is a fixed vector of real numbers. (Hint: Show that the function $h(\lambda) = f(\lambda x + (1-\lambda)y) = \sum_{i=1}^{s}(w_i(\lambda x_i + (1-\lambda)y_i))^2$ is convex for every choice of x and y and for $0 \leq \lambda \leq 1$. And show that $h(\lambda)$ is convex by showing that $h''(\lambda) \geq 0$.)

$$\begin{aligned}
h(\lambda) &= \sum_{i=1}^{s} w_i^2(\lambda x_i + (1-\lambda)y_i)^2 \\
h'(\lambda) &= \sum_{i=1}^{s} 2w_i^2(\lambda x_i - \lambda y_i + y_i)(x_i - y_i) \\
h''(\lambda) &= \sum_{i=1}^{s} 2w_i^2 x_i(x_i - y_i) - 2w_i y_i(x_i - y_i) \\
&= \sum_{i=1}^{s} \underbrace{2w_i^2(x_i - y_i)^2}_{\geq 0} \geq 0
\end{aligned}$$

Hence, f is convex.

(b) Prove that $f(x) = -\sum_{i=1}^{s} x_i \log x_i$ is a strictly concave function for $x \in (0,\infty)^s$, meaning that each $x_i > 0$.[2] Note: If x is a pmf, then $f(x) = H(x)$ is the entropy of x, so you have just proved that entropy is a strictly concave function.

$$\begin{aligned}
h(\lambda) &= -\sum_{i=1}^{s}(\lambda x_i + (1-\lambda)y_i)\log(\lambda x_i + (1-\lambda)y_i) \\
&= -\sum_{i=1}^{s} \lambda x_i \log(\lambda x_i + (1-\lambda)y_i) + (1-\lambda)y_i \log(\lambda x_i + (1-\lambda)y_i) \\
h'(\lambda) &= -\sum_{i=1}^{s} x_i \log(\lambda x_i + (1-\lambda)y_i) + \frac{\lambda x_i \cdot (x_i - y_i)}{\lambda x_i + y_i - \lambda y_i} + \frac{y_i \cdot (x_i - y_i)}{\lambda x_i + y_i - \lambda y_i} \\
&\qquad - y_i \log(\lambda x_i + y_i - \lambda y_i) - \frac{\lambda y_i \cdot (x_i - y_i)}{\lambda x_i + y_i - \lambda y_i} \\
&= -\sum_{i=1}^{s}(x_i - y_i)\log(\lambda x_i + y_i - \lambda y_i) + \frac{(x_i - y_i)(\lambda x_i + y_i - \lambda y_i)}{\lambda x_i + y_i - \lambda y_i}
\end{aligned}$$

[2] By defining $0 \log 0 = 0$, we can extend f to a continuous function on $[0,\infty)^s$, so that it is strictly concave on all of $[0,\infty)^s$. You do not have to prove this.

$$\begin{aligned} &= -\sum_{i=1}^{s}(x_i - y_i)\log(\lambda x_i + y_i - \lambda y_i) + (x_i - y_i) \\ h''(\lambda) &= -\sum_{i=1}^{s}\frac{(x_i - y_i)^2}{\lambda x_i + (1-\lambda)y_i} \end{aligned}$$

Now since $x_i, y_i \in (0,\infty)^s$, $x_i \neq y_i$, and $\lambda \in [0,1]$, we have that $\frac{(x_i-y_i)^2}{\lambda x_i + (1-\lambda)y_i} > 0$.

In particlar, this means that $h''(\lambda) < 0$, so h is stritly concave. Hence, f is strictly concave.

(c) Prove that $f(x) = \sum_{i=1}^{s} x_i \log(x_i/w_i)$ is a strictly convex function for $x \in (0,\infty)^s$, where w is a fixed vector of positive numbers. (The previous footnote applies here, as well.) Note: If x and w are pmfs, then $f(x) = D(x\|w)$ is called the relative entropy of w with respect to x, so you have just proved that relative entropy is a strictly convex function in the first argument.

$$\begin{aligned} h(\lambda) &= \sum_{i=1}^{s}(\lambda x_i + (1-\lambda)y_i)\log\left(\frac{\lambda x_i + (1-\lambda)y_i}{w_i}\right) \\ h'(\lambda) &= \sum_{i=1}^{s}(x_i - y_i)\log\left(\frac{\lambda x_i + (1-\lambda)y_i}{w_i}\right) + \frac{w_i(\lambda x_i + (1-\lambda y_i))(x_i - y_i)}{\lambda x_i + (1-\lambda)y_i} \\ &= \sum_{i=1}^{s}(x_i - y_i)\log\left(\frac{\lambda x_i + (1-\lambda)y_i}{w_i}\right) + w_i(x_i - y_i) \\ h''(\lambda) &= \sum_{i=1}^{s}\frac{w_i(x_i - y_i)^2}{\lambda x_i + (1-\lambda)y_i} \end{aligned}$$

Again, $x_i, y_i \in (0,\infty)^s$ with $x_i \neq y_i$, $w_i \in (0,\infty)^s$, and $\lambda \in [0,1]$. Hence, $\frac{w_i(x_i-y_i)^2}{\lambda x_i + (1-\lambda)y_i} > 0$. So $h''(\lambda) > 0$ and h and f are strictly convex.

3. **Jensen's inequality**. Let f be a convex function. Consider the inequality

$$f\left(\sum_{i=1}^{s} p_i x_i\right) \leq \sum_{i=1}^{s} p_i f(x_i) \qquad \text{for all } x_1, \ldots, x_s \text{ and all pmfs } p = p_{1:s}.$$

(a) Prove the inequality for $s = 1$. (This is trivial.)

We WTS that $f(px) \leq pf(x)$. But $p = p_1$ is a PMF, so $p = 1 \implies f(x) \leq f(x)$. Trivial.

(b) Prove the inequality for $s = 2$. (This is very easy, since f is convex.)

WTS $f(p_1x_1 + p_2x_2) \leq p_1f(x_1) + p_2f(x_2)$. But $p_1 + p_2 = 1$, so equivalently,

$$f(p_1x_1 + (1 - p_1)x_2 \leq p_1f(x_1) + (1 - p_1)f(x_2))$$

which is precisely the condition for convexity of f.

(c) Prove the inequality for $s = 3$. (Hint: Use convexity twice.)

Suppose not. Notice $p_1 + p_2 + p_3 = 1$.

Assume WLOG $p_2, p_3 \neq 0$ (else we reduce to the $s = 1$ case). Let $y = \frac{p_2x_2+p_3x_3}{p_2+p_3}$. Then

$$\begin{aligned}
f(\sum_{i=1}^{3} p_ix_i) &= f(p_1x_1 + (p_2 + p_3)y) \\
&= f(p_1x_1 + (1 - p_1)y) \\
&\leq p_1f(x_1) + (1 - p_1)f(y) && \text{(by convexity of } f) \\
&= p_1f(x_1) + (p_2 + p_3) \cdot f\left(\frac{p_2}{p_2 + p_3}x_2 + \frac{p_3}{p_2 + p_3}x_3\right) \\
&= p_1f(x_1) + (p_2 + p_3) \cdot f\left(\frac{p_2}{p_2 + p_3}x_2 + (1 - \frac{p_2}{p_2 + p_3})x_3\right) \\
&= p_1f(x_1) + (p_2 + p_3)\left[\frac{p_2}{p_2 + p_3}f(x_2) + (1 - \frac{p_2}{p_2 + p_3})f(x_3)\right] && \text{(by convexity of } f) \\
&= p_1f(x_1) + (p_2 + p_3)\left[\frac{p_2}{p_2 + p_3}f(x_2) + \frac{p_3}{p_2 + p_3}f(x_3)\right] \\
&= p_1f(x_1) + p_2f(x_2) + p_3f(x_3) \\
&= \sum_{i=1}^{3} p_if(x_i) \quad \blacksquare
\end{aligned}$$

(d) Assume that the inequality is true for $s = m$. Prove that it must therefore also be true when $s = m+1$.

(Hint: By taking $m = 3$, you can use the previous two parts to prove the $s = m + 1 = 4$ case, and then repeat forever to prove the inequality for any choice of s. This method of proof is called *induction*)

Suppose f convex and

$$f\left(\sum_{i=1}^{m} p_ix_i\right) \leq \sum_{i=1}^{m} p_if(x_i)$$

Notice

$$\begin{aligned}
f\left(\sum_{i=1}^{m+1} p_i x_i\right) &= f\left(\sum_{i=1}^{m} p_i x_i + p_{m+1} x_{m+1}\right) \\
&= f\left(p_{m+1} x_{m+1} + (1-p_{m+1}) \sum_{i=1}^{m} \frac{p_i}{1-p_{m+1}} x_i\right) \\
&= p_{m+1} f(x_{m+1}) + (1-p_{m+1}) f\left(\sum_{i=1}^{m} \frac{p_i}{1-p_{m+1}} x_i\right) && \text{(by convexity)} \\
&\le p_{m+1} f(x_{m+1}) + (1-p_{m+1}) \sum_{i=1}^{m} \frac{p_i}{1-p_{m+1}} f(x_i) && \text{(by assumption)} \\
&= p_{m+1} f(x_{m+1}) + \sum_{i=1}^{m} p_i f(x_i) \\
&= \sum_{i=1}^{m+1} p_i f(x_i)
\end{aligned}$$

By induction, the inequality holds for all s. ■

(e) Let X be a random variable with a pmf $p = p_{1:s}$. Prove Jensen's inequality: $f(\mathbb{E}(X)) \le \mathbb{E}(f(X))$. (Really, there's nothing left to do—just apply the definitions of the expressions on the two sides of the inequality.)

$$\begin{aligned}
f(\mathbb{E}X) &= f\left(\sum_{i=1}^{s} p_i x_i\right) \\
&\le \sum_{i=1}^{s} p_i f(x_i) && \text{(by (d) above)} \\
&= \mathbb{E} f(X) \quad \blacksquare
\end{aligned}$$

4. **Empirical distributions.** The empirical distribution is random, so it has its own distribution, which is confusing. It gets even more confusing when we sample many empirical distributions, so that we can have empirical distributions of empirical distributions. This problem takes you through some of the details and will hopefully make things clearer. Let $X_{1:n} = (X_1, \ldots, X_n)$ be an independent and identically distributed (iid) sequence with common pmf $h = h_{1:s} = (h_1, \ldots, h_s)$. The empirical distribution (or, more accurately, the empirical pmf) of $X_{1:n}$ is the vector $\widehat{p} = \widehat{p}_{1:s}$ defined by

$$\widehat{p}_x = \widehat{p}_x(X_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = x\} = \frac{\#\{i : X_i = x\}}{n}.$$

The second expression emphasizes that $\widehat{p}$ depends on the realization of $X_{1:n}$. The third expression illustrates that $\widehat{p}_x$ is a sum of iid random variables, which is helpful for computing expected values or thinking about the law of large numbers.

(a) Show that $\widehat{p}$ is a pmf for every possible sequence $X_{1:n}$.

It suffices to show that

i. $\sum_x \widehat{p}_x = 1$

ii. $\widehat{p}_x \geq 0$ for all x

Choose any $X_{1:n}$. Then

$$\begin{aligned}\sum_x \widehat{p}_x &= \sum_x \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = x\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{x=1}^{s} \mathbb{1}\{X_i = x\} \\ &= \frac{1}{n}\sum_{i=1}^{n} 1 \\ &= 1\end{aligned}$$

Further, by definition of the counting measure, $\#\{i : X_i = x\} \geq 0$ so for $n > 0$,

$$\widehat{p}_x = \frac{\#\{i : X_i = x\}}{n} \geq 0$$

Hence, $\widehat{p}$ is a pmf for every possible sequence $X_{1:n}$.

(b) Argue that $\widehat{p}$ is random, so it is a random pmf.

$\widehat{p}_x$ is a function of X_i and the constant n. Since X_i are random, $\widehat{p}_x$ is random.

(c) Since $\widehat{p}$ is random, it has an expected value. Compute $\mathbb{E}(\widehat{p})$ in terms of s, h, and n, where $\mathbb{E}$ denotes expected value.[3]

[3] The expected value of a vector is just the vector of the expected values of the components, so, for instance, $\mathbb{E}(\widehat{p}) = (\mathbb{E}(\widehat{p}_1), \ldots, \mathbb{E}(\widehat{p}_s))$.

Since $\mathbb{E}(\widehat{p}) = (\mathbb{E}\widehat{p}_1, \mathbb{E}\widehat{p}_2, \ldots, \mathbb{E}\widehat{p}_s)$, it suffices to compute $\mathbb{E}[\widehat{p}_x]$ for any x:

$$\begin{aligned}
\mathbb{E}[\widehat{p}_x] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = x\}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{X_i = x\}] && \text{(linearity of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{n} [1 \cdot h_x + 0 \cdot (1 - h_x)] \\
&= \frac{1}{n}\sum_{i=1}^{n} h_x \\
&= h_x
\end{aligned}$$

So $\mathbb{E}[\widehat{p}] = h$.

(d) For $n = s^{100}$ what will $\widehat{p}$ be very close to with high probability. Why?

The formula $\widehat{p}_x = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = x\}$ tell us that $\widehat{p}_x$ is the average of n iid Bernoulli random variables with parameter h_x. By the law of large numbers,

$$\widehat{p}_x \xrightarrow{a.s.} \mathbb{E}[\mathbb{1}\{X_i = x\}] = 1 \cdot h_x + 0 \cdot (1 - h_x) = h_x$$

In particular, this means that for $n = s^{100}$, $\widehat{p} \approx h$.

(e) Since $\widehat{p}$ is random, it also has a pmf, say r, given by

$$r_q = \mathbb{P}(\widehat{p} = q)$$

for any pmf q. Give an exact expression for r_q in terms of s, h, q, and n. Make sure to indicate in your expression those q for which $r_q = 0$. For $s = 3$, $h = (1/3, 1/3, 1/3)$, $q = (0.5, 0.3, 0.2)$ and $n = 10$, compute the value of r_q.

$$r_q = \mathbb{P}(\widehat{p} = q) = \binom{n}{nq_1, \ldots, nq_s} \prod_{i=1}^{s} h_i^{nq_i}$$

Hence, for $s = 3$, $h = (1/3, 1/3, 1/3)$, $q = (0.5, 0.3, 0.2)$ and $n = 10$,

$$\begin{aligned}
r_q &= \binom{10}{5, 3, 2} \left(\frac{1}{3}\right)^5 \left(\frac{1}{3}\right)^3 \left(\frac{1}{3}\right)^2 \\
&= \frac{10!}{5!\,3!\,2!} \left(\frac{1}{3}\right)^{10} \\
&= 2520 \cdot \frac{1}{59049} \\
&= \boxed{\frac{280}{6561}}
\end{aligned}$$

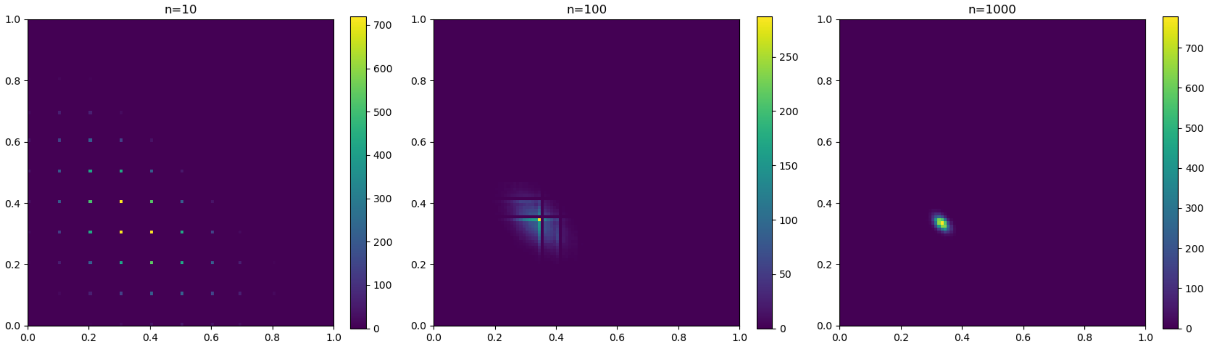
(f) (Do not submit) For the case $s = 3$ and $h = (1/3, 1/3, 1/3)$ and arbitrary n, write a function that generates a random observation from the pmf r. Except for very small n, you can't actually construct the pmf r because there are too many possible candidate pmfs q. But you can still sample from it by simply creating a random $\widehat{p}$.[4] Give an example $\widehat{p}$ that you randomly generated for the cases $n = 10^1, 10^2, 10^4, 10^6$.

- $n = 10 : [0.6, 0.4, 0.0]$
- $n = 100 : [0.31, 0.34, 0.35]$
- $n = 10000 : [0.3312, 0.3369, 0.3319]$
- $n = 1000000 : [0.332956, 0.333168, 0.333876]$

(g) (Do not submit) For $n = 10$, generate $m = 10^5$ iid samples from the pmf r from part (f), i.e., generate 10^5 iid $\widehat{p}$'s.[5] For $q = (0.5, 0.3, 0.2)$ find the fraction of times (out of m) that $\widehat{p} = q$. Your answer should be close to the true value of r_q found in part (e).

- Calculated probability: 0.04301
- $r_q = \frac{280}{6561} \approx 0.0427$

(h) (Do not submit) For $n = 10, 100, 1000$, generate $m = 10^5$ iid $\widehat{p}$'s as in part (g). In each case, make a 2D histogram[6] of the m different $(\widehat{p}_1, \widehat{p}_2)$ pairs over the unit square $[0, 1]^2$. (We can ignore $\widehat{p}_3$ since it is determined from $\widehat{p}_{1:2}$ via $\widehat{p}_3 = 1 - \widehat{p}_1 - \widehat{p}_2$. Also, it is hard to visualize a 3D histogram.) Put the 3 plots together on the same page.[7] Convince yourself of this fact: Each histogram shows a single observation of the empirical distribution of a sample of size m of empirical distributions of samples of size n from the pmf h.



(i) Explain why each histogram in part (h) is an approximate visualization of the pmf r, i.e., of the pmf of the random empirical pmf for the given value of n. Explain how the histograms in part (h) illustrate the law of large numbers. What value is the histogram in part (h) accumulating around for large n?

In part (c), we showed that $\widehat{p} \approx h$ for large n. This gives numerical evidence for the law of large numbers: as n increases, $(\widehat{p}_1, \widehat{p}_2) \to (\frac{1}{3}, \frac{1}{3}) = (h_1, h_2)$. And in fact, this also explains why the historgrams visualize

[4] For this s and h, you can get a random $X_{1:n}$ in Matlab with `x=randi(3,n,1)`. Try it for $n = 10$. Try again and again. Now you can count how many times each value occurs with `histc(x,1:3)` and use `p = histc(x,1:3)/n` to get the fraction of times each value occurs. Try this repeatedly (remember to make a new `x`) until you see how it works. Each new `p` is a new $\widehat{p}$, which is a random observation from the pmf r. You don't really ever need to make `x` explicitly: `p = histc(randi(3,n,1),1:3)/n`. You can use `help` in Matlab to get information about the details of a function, such as `help histc`.

[5] You can write a for loop to make m independent $\widehat{p}$, or you can do it all at once: `p = histc(randi(3,n,m),1:3,1)/n;`. (The semicolon at the end of a command suppresses output to the command window, which is helpful in this case.) The `histc` command counts within each column, so that each column of `p` is an independent $\widehat{p}$.

[6] If you have stored the $\widehat{p}$'s in an $3 \times m$ array `p`, then you can make a 2D histogram in Matlab with `histogram2(p(1,:),p(2,:))`. You can make it better with

```
histogram2(p(1,:),p(2,:),[0:.01:1],[0:.01:1],'normalization','probability'),
```

which will keep the same binning for each n and will also scale the vertical axis so that it is in units of fraction out of m, instead of units of counts.

[7] The Matlab command `subplot` might be useful.

r: r_q is the PMF of the empirical PMF of h. The histograms show the concentrations of randomly sampled empirical distributions of samples from h. Hence, as $\widehat{p} \to h$, with large probability, $r_q \to h$.

5. **The indicator trick.** Given a domain $\mathcal{D}$ and a subset $A \subseteq \mathcal{D}$, the indicator function, $\mathbb{1}_A : \mathcal{D} \to \{0,1\}$, is the function that indicates whether or not $x \in \mathcal{D}$ is in A:

$$\mathbb{1}_A(x) \triangleq \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

$\mathbb{1}_A(x)$ is often (maybe even typically) written as $\mathbb{1}_{x \in A}$.

(a) Let X be a random variable taking values in $\mathcal{D}$ with probability distribution $\mathbb{P}$, so $\mathbb{P}(A) = \mathbb{P}(X \in A)$. Fix $A \in \mathcal{D}$ and let $p = \mathbb{P}(A)$. Notice that $\mathbb{1}_{X \in A}$, being a function of X, is itself a random variable. Compute $\mathbb{E}[\mathbb{1}_{X \in A}]$ in terms of p.

$$\mathbb{E}[\mathbb{1}_{X \in A}] = 1 \cdot \mathbb{P}(X \in A) + 0 \cdot \mathbb{P}(X \notin A) = \mathbb{P}(X \in A) = \boxed{p}$$

(b) Using the same X, A, and p, compute $\mathbb{V}[\mathbb{1}_{X \in A}]$ (the variance of the random variable $\mathbb{1}_{X \in A}$).

$$\mathbb{V}[\mathbb{1}_{X \in A}] = \mathbb{E}[(\mathbb{1}_{X \in A})^2] - \mathbb{E}[\mathbb{1}_{X \in A}]^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = \boxed{p - p^2}$$

(c) Suppose that I have n different letters addressed to n different individuals, each living at a different address. I print n envelopes with the correct addresses, but absent-mindedly place the letters randomly into the envelopes. Each envelope gets exactly one letter. Assuming that every placement of letters into envelopes is equally likely, compute the expected number of letters that go to the right individuals. (Hint: for each $k \in \{1, 2, \ldots, n\}$ define Y_k to indicate the event that the k'th letter goes to the correct address.)

Take letters $X_1, \ldots, X_n$ on envelopes $\{1, \ldots, n\}$ so that $Y_k = \mathbb{1}\{X_k = k\}$ corresponds to the event that the k-th letter ends up in the correct (i.e. k-th) envelope.

Then, X, the number of letters that end up in the correct envelopes, is given by $X = \sum_{k=1}^{n} Y_k$. By linearity of expectation, we have

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{k=1}^{n} Y_k\right] = \sum_{k=1}^{n} \mathbb{E}[Y_k] = \sum_{k=1}^{n} \mathbb{E}[\mathbb{1}\{X_k = k\}] = \sum_{k=1}^{n} \mathbb{P}(X_k = k) = \sum_{k=1}^{n} \frac{1}{n} = \boxed{1}$$

Once you set up (c) correctly, it's quite easy to make the calculation. To appreciate the "indicator trick", think about making the same calculation more directly: first compute the probability, say p_k, that exactly k letters end up in the correct envelopes, for each $k = 1, 2, \ldots, n$. Then get the expected value by computing the sum $\sum_{k=1:n} k p_k$.

6. **On the asymptotics of $C(\widehat{p})$, heuristically.** Most of the discussion of the Gibbs thought experiment was more of an outline than a rigorous derivation. The key insight was the surprising role of entropy in approximating the number of ways to produce an ensemble with a given empirical empirical distribution, $\widehat{p}$. With the help of the Stirling approximation, we concluded that

$$C(\widehat{p}) \approx e^{nH(\widehat{p})} \tag{1}$$

where

$$C(\widehat{p}) \doteq \frac{n!}{(n\widehat{p}_1)! \cdots (n\widehat{p}_s)!} = \binom{n}{n\widehat{p}_1, \ldots, n\widehat{p}_s} \qquad \text{(``the multinomial coefficient'')}$$

and $H(\widehat{p}) = -\sum_{x=1:s} \widehat{p}_x \log \widehat{p}_x$ is Shannon's entropy.

The Stirling approximation uses $k^k e^{-k}\sqrt{2\pi k}$ to approximate $k!$, with bounds

$$e^{\frac{1}{12k+1}} \leq \frac{k!}{k^k e^{-k}\sqrt{2\pi k}} \leq e^{\frac{1}{12k}} \qquad \text{for all } k \geq 1 \tag{2}$$

If we write $\mathcal{S}(k)$ for $k^k e^{-k}\sqrt{2\pi k}$ and define $\mathcal{S}(0) = 1$, then $\mathcal{S}(k)$ is exact when $k = 0$ (since $0! = 1$), and $\mathcal{S}(k)$ is an excellent bound in the sense of (2) for all other k. For convenience, we further extend the definition of $\mathcal{S}$ by declaring that $\mathcal{S}$ evaluated on *any* expression means replacing every factorial $k!$ in the expression by $\mathcal{S}(k)$, and in particular

$$\mathcal{S}(C(\widehat{p})) = \frac{n^n e^{-n}\sqrt{2\pi n}}{\prod_{x=1}^{s} n_x^{n_x} e^{-n_x}\sqrt{2\pi n_x}}$$

where $n_x \triangleq n\widehat{p}_x = \#\{k : X_k = x\}$.

The purpose of this problem is to give some justification for the following step that was taken in class (without explanation) while deriving (1): for fixed s and empirical probability $\widehat{p}$,

$$\log \frac{C(\widehat{p})}{\mathcal{S}(C(\widehat{p}))} = O(\frac{1}{n})$$

To this end, let $\mathcal{I} = \{x \in \{1, 2, \ldots, s\} \big| \widehat{p}_x \neq 0\}$ and show that

$$\left| \log \frac{C(\widehat{p})}{\mathcal{S}(C(\widehat{p}))} \right| \leq \frac{1}{12n}\left(1 + \sum_{x\in\mathcal{I}} \frac{1}{\widehat{p}_x}\right)$$

$$\frac{C(\widehat{p})}{S(C(\widehat{p}))} = \frac{n!}{\prod_{x\in\mathcal{I}}(n\widehat{p}_x)!} \cdot \frac{\prod_{x\in\mathcal{I}}(n\widehat{p}_x)^{n\widehat{p}_x} e^{n\widehat{p}_x}\sqrt{2\pi n\widehat{p}_x}}{n^n e^{-n}\sqrt{2\pi n}} = \frac{n!}{S(n)} \frac{\prod_{x\in\mathcal{I}} S(n\widehat{p}_x)}{\prod_{x\in\mathcal{I}}(n\widehat{p}_x)!} = \frac{n!}{S(n)} \prod_{x\in\mathcal{I}} \frac{S(n\widehat{p}_x)}{(n\widehat{p}_x)!}$$

$$\begin{aligned}
\left|\log \frac{C(\widehat{p})}{S(C(\widehat{p}))}\right| &= \left|\log \frac{n!}{S(n)} + \sum_{x\in\mathcal{I}} \log \frac{S(n\widehat{p}_x)}{(n\widehat{p}_x)!}\right| \\
&\leq \left|\log \frac{n!}{S(n)}\right| + \sum_{x\in\mathcal{I}} \left|\log \frac{S(n\widehat{p}_x)}{(n\widehat{p}_x)!}\right| && \text{(triangle inequality)} \\
&\leq \frac{1}{12n} + \sum_{x\in\mathcal{I}} \frac{1}{12n\widehat{p}_x + 1} && \text{(by Stirling approximation)} \\
&\leq \frac{1}{12n} + \sum_{x\in\mathcal{I}} \frac{1}{12n\widehat{p}_x} \\
&= \frac{1}{12n}\left(1 + \sum_{x\in\mathcal{I}} \frac{1}{\widehat{p}_x}\right) \quad \blacksquare
\end{aligned}$$

7. **Refresher: working with Gaussian random variables (GRVs).** Given a continuous RV X with probability density function f, we will write $X \sim \mathcal{N}(\mu, \sigma^2)$ (alternatively $f \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that X has a Gaussian (aka normal) distribution with mean μ and variance σ^2):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

(a) **Sums of independent GRVs.** Linear combinations of *independent* Gaussian random variables are Gaussian. Formally, if $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, for $k = 1, \ldots, n$, are n independent Gaussian RVs, and if $a_l \in \mathbb{R}$, for $l = 0, 1, \ldots, n$, are $n+1$ constants, then $X \triangleq a_0 + \sum_{k=1}^{n} a_k X_k$ is also Gaussian. Give explicit formulas for the mean (μ) and variance (σ^2) of X in terms of the a_k's, μ_k's, and σ_k's.

Since X is a Gaussian, $\mathbb{E}[X] = \mu$. Hence,

$$\begin{aligned}
\mu = \mathbb{E}[x] = \mathbb{E}\left[a_0 + \sum_{k=1}^{n} a_k X_k\right] \\
= a_0 + \sum_{k=1}^{n} a_k \mathbb{E}[X_k] \qquad \text{(linearity)} \\
= a_0 + \sum_{k=1}^{n} a_k \mu_k
\end{aligned}$$

Further, $\mathbb{V}[X] = \sigma^2$. Hence,

$$\begin{aligned}
\sigma^2 = \mathbb{V}[X] = \mathbb{V}\left[a_0 + \sum_{k=1}^{n} a_k X_k\right] \\
= \mathbb{V}\left[\sum_{k=1}^{n} a_k X_k\right] \\
= \sum_{i,j}^{n} a_i a_j \mathrm{cov}(X_i, X_j) \\
= \sum_{k=1}^{n} a_k^2\, \mathbb{V}[X_k] \qquad \text{(since } X_i \text{ are independent)} \\
= \sum_{k=1}^{n} a_k^2 \sigma_k^2
\end{aligned}$$

(b) **Complete the square.** Assume that X is a continuous RV with pdf

$$f(x) = ce^{-\frac{1}{4}x^2 + \frac{1}{2}x} \qquad x \in \mathbb{R}$$

Show that $X \sim \mathcal{N}(\mu, \sigma^2)$, and give explicit values for μ, σ^2, and the normalizing constant c. (Hint: complete the square.)

With PDF,

$$\begin{aligned}
f(x) = ce^{-\frac{1}{4}x^2 + \frac{1}{2}x} \\
= ce^{-\frac{1}{4}(x^2 - 2x + 1) + \frac{1}{4}} \\
= ce^{1/4} \cdot e^{-\frac{(x-1)^2}{4}}
\end{aligned}$$

we have that $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\boxed{c = \frac{e^{-1/4}}{\sqrt{4\pi}}, \mu = 1, \sigma^2 = 2}$.

(c) **A neat trick.** Follow these steps to show that

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{x=-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = 1$$

for all μ and $\sigma^2 > 0$:

i. Use a change of variables to conclude that it would be sufficient to show

$$\int_{x=-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi} \tag{3}$$

Let $u = \frac{x-\mu}{\sigma}$. Then it suffices to show

$$\begin{aligned} \frac{1}{\sigma}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx &= \sqrt{2\pi} \\ \frac{1}{\sigma}\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\,\sigma du &= \sqrt{2\pi} \\ \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\,du &= \sqrt{2\pi} \end{aligned}$$

ii. Argue that, instead of showing (3), it would be sufficient to show

$$\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dxdy = 2\pi \tag{4}$$

Let $I = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx$. Since it suffices to show $I = \sqrt{2\pi}$, we may write

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy = 2\pi \end{aligned}$$

iii. Change the double integral in (4) to polar coordinates and then integrate the left-hand side to get 2π.

Let $x^2 + y^2 = r^2$, hence

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy \\ &= \int_{0}^{2\pi}\int_{0}^{\infty} re^{-\frac{r^2}{2}}\,dr\,d\theta \\ &= \int_{0}^{2\pi}\left[-e^{-\frac{r^2}{2}}\right]_0^{\infty}\,d\theta \\ &= \int_{0}^{2\pi} 1\,d\theta = 2\pi \quad \blacksquare \end{aligned}$$

8. **Large deviation, continuous random variables.** Consider $X_{1:n} \triangleq (X_1, \ldots, X_n)$ drawn *iid* from the standard normal pdf (prbabilitiy density function) $\mathcal{N}(0,1)$, and let σ_n be the standard deviation of the sample mean, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The goal is to use experiments to make an informed conjecture about the limit as $n \to \infty$ of the empirical distribution $\widehat{p}(X_{1:n})$, given that $\overline{X}$ turns out to be, quite unexpectedly, more than 3.8 standard deviations bigger than its mean, i.e. $\overline{X} > 3.8\sigma_n$. Specifically, with n=10,000:

   (a) Calculate $\sigma_{10{,}000}$[8] and the exact probability of the unexpected event: $\mathbb{P}(\overline{X} > 3.8\sigma_n)$. Exact means without using the central limit theorem or any Monte Carlo experimentation (though you will need a computer, or at least an old-fashioned lookup table).

$$\mathbb{E}[\overline{X}_{10000}] = \mathbb{E}\left[\frac{1}{10000}\sum_{i=1}^{10000} X_i\right] = \frac{1}{10000}\sum_{i=1}^{10000}\mathbb{E}[X_i] = 0$$

$$\begin{aligned}\mathbb{V}[\overline{X}_{10000}] &= \mathbb{V}\left[\frac{1}{10000}\sum_{i=1}^{10000} X_i\right] \\ &= \frac{1}{10000^2}\sum_{i=1}^{10000}\mathbb{V}[X_i] = \frac{1}{10000}\end{aligned}$$

   Hence,

$$\sigma_{10000} = \sqrt{\frac{1}{10000}} = \frac{1}{100}$$

   and

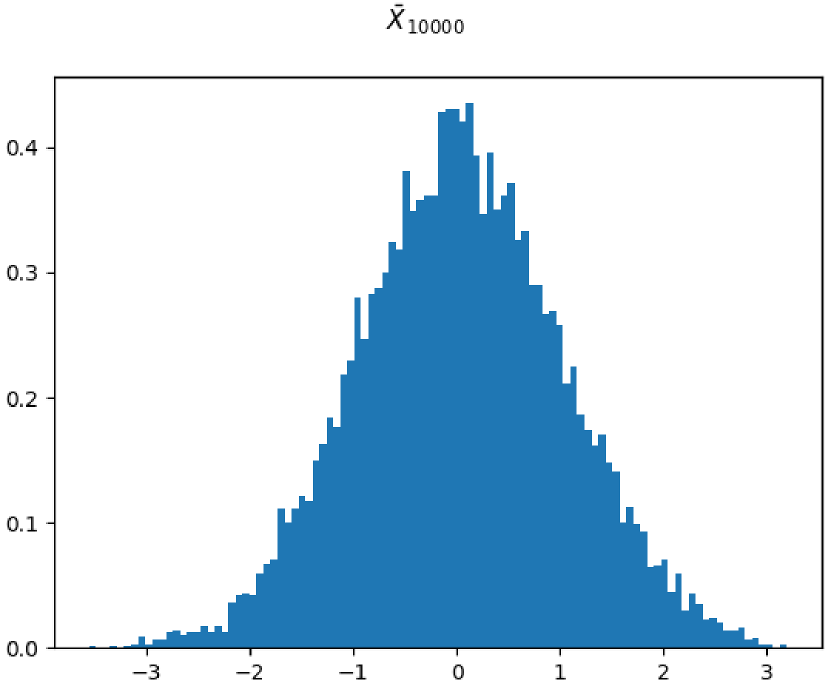
$$\begin{aligned}\mathbb{P}(\overline{X} > 3.8\sigma_{10000}) &= \mathbb{P}(\overline{X} > 0.038) \\ &= 1 - \mathbb{P}(\overline{X} \leq 0.038) \\ &= 1 - \mathbb{P}\left(\frac{\overline{X} - 0}{\sigma_{10000}} \leq \frac{0.038 - 0}{\sigma_{10000}}\right) \\ &= 1 - \mathbb{P}(Z \leq 0.38) \\ &= 1 - 0.99993 \\ &= \boxed{0.00007}\end{aligned}$$

   (b) One way to sample from the conditional distribution on $X_{1:n}$ (given that $\overline{X} > 3.8\sigma_n$) is to make repeated, independent, draws of all 10,000 *iid* $\mathcal{N}(0,1)$ random variables. About how many draws would you expect to need before getting a sample with $\overline{X} > 3.8\sigma_{10{,}000}$?

   Since $\mathbb{P}(\overline{X} > 3.8\sigma_{10000}) = 0.00007$, we would expect to need $\frac{1}{0.00007} \approx 14286$ draws before getting a sample with $\overline{X} > 3.8\sigma_{10000}$.

   (c) (Do not submit) Perform the experiment described in (b). Once you have a sample, display the normalized histogram, and compute the sample mean, the sample standard deviation, and the sample kurtosis.

[8] Handy things to remember when computing variances of sums of independent random variables:
(i) If X and Y are independent and α and β are scalars, and if $W = \alpha X + \beta Y$, then $\sigma_W^2 = \alpha^2\sigma_X^2 + \beta^2\sigma_Y^2$
(ii) For any random variable X with mean $\mu_X = \mathbb{E}[X]$, $\sigma_X^2 = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2$



After 8941 draws, we obtained a sample with:

- Mean: 0.04337
- Standard deviation: 0.9919
- Kurtosis: -0.0791

(d) Let $h(x)$ be the standard normal density function, and define the set B by

$$B = \big\{g : g \text{ is a pdf on } \mathbb{R} \text{ , } E_g[X] > 3.8\sigma_{10,000}\big\}1$$

and define the density function $f(x)$ by

$$f = \arg\min_{g \in B} D(g||h) \tag{5}$$

where h is the standard normal density function and

$$D(g||h) = \int_{-\infty}^{\infty} g(x) \log\big(\frac{g(x)}{h(x)}\big) dx$$

Use your observations from (c) to make a conjecture about the solution to (5). (You can check your answer if you can solve problem 11..)

Since the histogram in (c) looks like a normal curve, we would expect the solution to (5) to look like $q \sim \mathcal{N}(0.38, 1)$

For 2610 or for extra credit:

These problems create careful bounds for some of the approximations used in class.

9. **Rigorous bounds on** $C(\widehat{p})$. Fix a positive integer n. Let $p = p_{1:s}$ be a pmf with the property that np is a vector of integers, and let $H(p)$ denote the entropy of p computed using $\log_e$ (i.e. $H(p) = -\sum_{i=1}^s p_i \log_e p_i$).

   (a) Prove that there are at most $(n+1)^s$ possible choices of p (i.e., pmfs with the property that np is a vector of integers.)

   Let $n_i = np_i \in \mathbb{Z}$.

   Notice

$$\sum_{i=1}^s p_i = 1 \implies \sum_{i=1}^s n_i = n$$

   Further, since p is a PMF, $0 \le p_i \le 1 \implies 0 \le n_i \le n \quad \forall i$.

   This tells us that

$$\left\{ n_i \in \mathbb{Z} : 0 \le n_i \le n, \sum_{i=1}^s n_i = n \right\} \subseteq \{n_i \in \mathbb{Z} : 0 \le n_i \le n\} = \{0, 1, 2, \dots, n\}$$

   Ignoring the constraint $\sum_{i=1}^s n_i = n$ for now, we see that there are $(n+1)^s$ possible choices of p (by assigning a random integer in $[0, n]$ to each of the s elements of p).

   Including the the constraint $\sum_{i=1}^s n_i = n$, we see that there will in fact be many fewer than $(n+1)^s$ possible choices of p.

   (b) Prove that

$$\binom{n}{np_1, \dots, np_s} \le e^{nH(p)}.$$

   Hint: Let $X_{1:n}$ be iid p and express $\mathbb{P}(\widehat{p}(X_{1:n}) = p)$ in terms of $H(p)$ and the multinomial coefficient (and then note that probabilities are always ≤ 1).

   Recall

$$\begin{aligned}
\mathbb{P}(\widehat{p} = p) &= \binom{n}{np_1 \cdots np_s} \prod_{x=1}^s p_x^{np_x} \\
\log \mathbb{P}(\widehat{p} = p) &= \log \binom{n}{np_1 \cdots np_s} + \sum_{x=1}^s np_x \log p_x \\
&= \log \binom{n}{np_1 \cdots np_s} - nH(p)
\end{aligned}$$

   But then

$$\mathbb{P}(\widehat{p} = p) = e^{-nH(p)} \binom{n}{np_1 \cdots np_s} \implies e^{nH(p)} \mathbb{P}(\widehat{p} = p) = \binom{n}{np_1 \cdots np_s}$$

   However, $\mathbb{P}(\widehat{p} = p) \le 1$, so

$$e^{nH(p)} \ge \binom{n}{np_1 \cdots np_s}$$

   (c) Prove that

$$\frac{1}{(n+1)^s} e^{nH(p)} \le \binom{n}{np_1, \dots, np_s}.$$

Hint: Let $X_{1:n}$ be iid p, and prove that $\mathbb{P}(\widehat{p}(X_{1:n}) = p) \geq \mathbb{P}(\widehat{p}(X_{1:n}) = q)$ for all valid pmfs q. Use this to immediately obtain $1 = \sum_q \mathbb{P}(\widehat{p}(X_{1:n}) = q) \leq \sum_q \mathbb{P}(\widehat{p}(X_{1:n}) = p)$, and then use part (a). You may find this bound useful: For any two positive integers, a and b, $a!/b! \geq b^{a-b}$.

As before,

$$\begin{aligned}
\mathbb{P}(\widehat{p} = p) &= \binom{n}{np_1 \cdots np_s} \prod_{x=1}^{s} p_x^{np_x} = n! \prod_{x=1}^{s} \frac{p_x^{np_x}}{(np_x)!} \\
\mathbb{P}(\widehat{p} = q) &= \binom{n}{nq_1 \cdots nq_s} \prod_{x=1}^{s} p_x^{nq_x} = n! \prod_{x=1}^{s} \frac{p_x^{nq_x}}{(nq_x)!}
\end{aligned}$$

Then, consider

$$\frac{\mathbb{P}(\widehat{p} = p)}{\mathbb{P}(\widehat{p} = q)} = \prod_{x=1}^{s} \frac{p_x^{np_x}}{(np_x)!} \cdot \frac{(nq_x)!}{p_x^{nq_x}} \geq \prod_{x=1}^{s} p_x^{np_x - nq_x} (np_x)^{nq_x - np_x} = \prod_{x=1}^{s} n^{nq_x - np_x}$$

Taking the log,

$$\begin{aligned}
\log \frac{\mathbb{P}(\widehat{p} = p)}{\mathbb{P}(\widehat{p} = q)} &\geq \sum_{x=1}^{s} (nq_x - np_x) \log n \\
&= n \log n \sum_{x=1}^{s} (q_x - p_x) = 0 \\
&\implies \frac{\mathbb{P}(\widehat{p} = p)}{\mathbb{P}(\widehat{p} = q)} \geq 1 \\
&\implies \mathbb{P}(\widehat{p} = p) \geq \mathbb{P}(\widehat{p} = q)
\end{aligned}$$

In particular, this means that

$$1 = \sum_{q} \mathbb{P}(\widehat{p} = q) \leq \sum_{q} \mathbb{P}(\widehat{p} = p)$$

But from (a),

$$1 \leq \sum_{q} \mathbb{P}(\widehat{p} = p) = \sum_{q} \binom{n}{np_1 \cdots np_s} \prod_{x=1}^{s} p_x^{np_x} = (n+1)^s \binom{n}{np_1 \cdots np_s} \prod_{x=1}^{s} p_x^{np_x}$$

$$\begin{aligned}
\frac{e^{nH(p)}}{(n+1)^s} &\leq \binom{n}{np_1 \cdots np_s} e^{nH(p)} \prod_{x=1}^{s} p_x^{np_x} \\
&= \binom{n}{np_1 \cdots np_s} \exp\left(-\sum_{x=1}^{s} np_x \log p_x \right) \prod_{x=1}^{s} p_x^{np_x}
\end{aligned}$$

$$\begin{aligned}
\log \frac{e^{nH(p)}}{(n+1)^s} &\leq \log \binom{n}{np_1 \cdots np_s} - \sum_{x=1}^{s} np_x \log p_x + \sum_{x=1}^{s} np_x \log p_x \\
\log \frac{e^{nH(p)}}{(n+1)^s} &\leq \log \binom{n}{np_1 \cdots np_s} \\
\frac{e^{nH(p)}}{(n+1)^s} &\leq \binom{n}{np_1 \cdots np_s} \quad \blacksquare
\end{aligned}$$

10. **Relative entropy and likelihood bounds on the empirical distribution.** Let $X_{1:n}$ be iid $h = h_{1:s}$, let $q = q_{1:s}$ be a pmf with the property that nq is a vector of integers, and let

$$D(q\|h) \doteq \sum_{i=1}^{s} q_i \log_e \frac{q_i}{h_i}$$

denote relative entropy computed using $\log_e$. Use the previous problem to prove that

$$\frac{1}{(n+1)^s} e^{-nD(q\|h)} \leq \mathbb{P}(\widehat{p}(X_{1:n}) = q) \leq e^{-nD(q\|h)}.$$

Since

$$\mathbb{P}(\widehat{p} = q) = \binom{n}{nq_1 \cdots nq_s} \prod_{x=1}^{s} h_x^{nq_x}$$

by 9.b,

$$\frac{1}{(n+1)^s} e^{nH(q)} \prod_{x=1}^{s} h_x^{nq_x} \leq \mathbb{P}(\widehat{p} = q) \leq e^{nH(q)} \prod_{x=1}^{s} h_x^{nq_x}$$

so it suffices to show that

$$e^{-nD(q\|h)} = e^{nH(q)} \prod_{x=1}^{s} h_x^{nq_x}$$

By definition,

$$\begin{aligned} e^{-nD(q\|h)} &= \exp(-n \sum_{i=1}^{s} q_i \log \frac{q_i}{h_i}) \\ &= \exp(\sum_{i=1}^{s} -nq_i \log q_i + nq_i \log h_i) \\ &= e^{nH(q)} \exp\left(\sum_{i=1}^{s} nq_i \log h_i\right) \\ &= e^{nH(q)} \prod_{x=1}^{s} h_x^{nq_x} \end{aligned}$$

so indeed,

$$\frac{1}{(n+1)^s} e^{-nD(q\|h)} \leq \mathbb{P}(\widehat{p}(X_{1:n}) = q) \leq e^{-nD(q\|h)} \quad \blacksquare$$

11. **Calculus of variations, with constraints.** Consider again the set up in problem (8.). Find a closed-form solution to the optimization problem defined in equation (5). Here's an outline, in case you're not familiar with the calculus of variations: (i) introduce Lagrange multipliers, which will be multiplying integrals instead of sums; (ii) replace every instance of $g(x)$ by $g(x)+\epsilon\eta(x)$, where $\eta(x)$ is an arbitrary function; (iii) evaluate the derivative, with respect to ϵ, of the resulting expression at $\epsilon = 0$; (iv) use the fact that η is arbitrary.

We want to find a closed form solution to

$$f = \arg\min_{g \in B} D(g \| h)$$

where

- $B = \{g : g \text{ pdf on } \mathbb{R},\ \mathbb{E}_g[X] > 3.8\sigma_{10{,}000}\}$
- h is the standard normal density function, i.e. $h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- $D(g\|h) = \int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{h(x)}\, dx$

By definition of B, we have constraints

- $\int_{-\infty}^{\infty} g(x)\, dx = 1$
- $\int_{-\infty}^{\infty} xg(x)\, dx > 3.8\sigma_{10{,}000}$

Hence, we seek g which maximizes

$$D(g\|h) + \gamma \int_{-\infty}^{\infty} g(x)\, dx + \lambda \int_{-\infty}^{\infty} xg(x)\, dx$$

Taking derivatives WRT g and setting to zero, we have

$$\begin{aligned}
&\frac{\partial}{\partial g(x)} \left[D(g\|h) + \gamma \int_{-\infty}^{\infty} g(x)\, dx + \lambda \int_{-\infty}^{\infty} xg(x)\, dx \right] \\
&\quad = \frac{\partial}{\partial g(x)} \left[\int_{-\infty}^{\infty} g(x) \log \frac{g(x)}{h(x)}\, dx + \gamma \int_{-\infty}^{\infty} g(x)\, dx + \lambda \int_{-\infty}^{\infty} xg(x)\, dx \right] \\
&\quad = \frac{\partial}{\partial g(x)} \left[\int_{-\infty}^{\infty} g(x) \left(\log \frac{g(x)}{h(x)} + \gamma + \lambda x \right) \right]
\end{aligned}$$

12. **Proof of the LDP.**[9] Let $\mathcal{S} \triangleq \{q \in [0,1]^s : \sum_x q_x = 1\}$ be the s-dimensional probability simplex, i.e., the set of all probability mass functions (pmfs) over the sample space $\{1, \dots, s\}$. For any two pmfs $\alpha, \beta \in \mathcal{S}$, define $\|\alpha - \beta\| = \max_{x=1:s} |\alpha_x - \beta_x|$. Fix $h \in \mathcal{S}$, let $X_1, X_2, \dots$ be iid random variables with common pmf h, and define $\widehat{p} = \widehat{p}(X_{1:n})$ to be the empirical distribution. Suppose $B \subseteq \mathcal{S}$ and define

$$D(B\|h) \triangleq \inf_{q \in B} D(q\|h).$$

(Previous problems are useful throughout this problem.)

(a) Prove that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p}(X_{1:n}) \in B) \leq -D(B\|h).$$

Since B is finite, our results above hold and we have

$$\begin{aligned} \mathbb{P}(\widehat{p} \in B) &= \sum_{q \in B} \mathbb{P}(\widehat{p} = q) \\ &\leq e^{-D(q\|h)} && \text{by problem 10} \end{aligned}$$

so

$$\frac{1}{n} \log \mathbb{P}(\widehat{p} \in B) = -D(q\|h)$$

Then since $\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p} \in B) = 0$,

$$\limsup f = \liminf f = \lim f$$

we have

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p} \in B) &\leq \liminf_{n\to\infty} -D(q\|h) \\ \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p} \in B) &\leq \inf_{q \in B} -D(q\|h) \end{aligned}$$

(b) Suppose B is an open set[10] and $h_i > 0$ for all $i = 1, \dots, s$. Prove that there exists a sequence $q^1, q^2, \dots$ of pmfs in B (using superscripts for indices, not powers) with the properties that $D(q^n\|h) \to D(B\|h)$ as $n \to \infty$ and that $\mathbb{P}(\widehat{p}(X_{1:n}) = q_n) > 0$ for all n sufficiently large.

Let $q^* = \arg\inf_{q \in B} D(B\|h)$.

Suppose $q^* \in B$. But then since B is open, $\exists q^\dagger \in B_\varepsilon(q^*)$ such that $D(q^\dagger\|h) \leq D(q^*\|h)$, which contradicts that q^* is a lower bound. Hence, $q^* \notin B$.

Similarly, suppose $q \notin B$ and $q \notin \Omega$, the boundary of B. Then, again, there exists $q^\dagger \in B_\varepsilon(q)$ such that

$$D(q^*\|h) \leq D(q^\dagger) \leq D(q\|h) \quad \forall q \in B$$

which contradicts that q^* is the greatest lower bound.

Hence, $q^* \in \Omega \implies q^* \in \overline{B}$ since $\overline{B} = B \cup \Omega$ is the smallest closed set containing B. By an equivalent definition of the closure, $\overline{B} = B \cup X$ where X is the set of limit points of B.

[9] This is a really hard problem, even though you will be guided through it step by step. If you can work through every step, then great. But even if you cannot, it will be well worth your time to give it a try. And of course there will be plenty of opportunities for partial credit.

[10] This means that B is open relative to the topology on $\mathcal{S}$. B cannot be open in $\mathbb{R}^s$ since it is a subset of $\mathcal{S}$. To avoid confusion, we might say that B is relatively open in $\mathcal{S}$, meaning that $B = \mathcal{S} \cap O$ for some open set $O \subseteq \mathbb{R}^s$.

Since $q^* \notin B$, $q^* \in X$ so q^* is a limit of point of B. Hence, there exists a sequence $q^n \in B$ such that $q^n \to q^*$ as $n \to \infty$.

Then, by definition of q^*, $q^n \to q^* \implies D(q^n \| h) \to D(q^* \| h) = D(B \| h)$ as $n \to \infty$.

Finally, to show $\mathbb{P}(\widehat{p} = q_n) > 0$, it suffices to show $q_n \in \mathbb{Q}^s$ for all n.

(c) Prove that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p}(X_{1:n}) \in B) \geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p}(X_{1:n}) = q_n) = -D(B\|h)$$

(Combining (a) and (c) gives

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\widehat{p}(X_{1:n}) \in B) = -D(B\|h),$$

which is the first part of the LDP.)

(d) Prove that

$$\lim_{n\to\infty} \mathbb{P}\left(D(\widehat{p}(X_{1:n})\|h) < D(B\|h) + \delta \;\middle|\; \widehat{p}(X_{1:n}) \in B\right) = 1$$

for all $\delta > 0$.

(e) Assume that $p^* \triangleq \arg\min_{q \in \overline{B}} D(q\|h)$ is unique. Show that $D(p^*\|h) = D(B\|h)$ and that

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widehat{p}(X_{1:n}) - p^*\| < \epsilon \;\middle|\; \widehat{p}(X_{1:n}) \in B\right) = 1$$

for every $\epsilon > 0$. This is the second part of the LDP. (Hint: Begin by showing that for each $\epsilon > 0$, there is a $\delta > 0$ such that the event in part (d) is a subset of the event in part (e).)