1. We are given a fair coin and want to generate independent samples from a Bernoulli random variable X with mean $p > 0$, i.e. $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$. Find an algorithm that does this, such that the expected number of needed coin flips to generate one sample of X is less or equal than 2.

2. Let $q \in [0,1]$, $n \in \mathbb{N}$ such that nq is an integer in the range $[0,n]$. Show that

$$\frac{2^{nH(q)}}{n+1} \leq \binom{n}{nq} \leq 2^{nH(q)}$$

where $H(q) := -q\log q - (1-q)\log(1-q)$ is the entropy of a Bernoulli random variable with mean q.

3. Let X_1 be a random variable taking values in $\mathcal{X}_1 = \{1,\dots,m\}$ and X_2 be a random variable taking values in $\mathcal{X}_2 = \{m+1,\dots,n\}$, for integers $n > m$. Furthermore, assume X_1 and X_2 to be independent. Define a random variable X as

$$X = X_\theta$$

where θ is a random variable independent of X_1 and independent of X_2 satisfying $\mathbb{P}(\theta = 1) = \alpha$ and $\mathbb{P}(\theta = 2) = 1 - \alpha$ for some $\alpha \in [0,1]$.

   (a) Express $H(X)$ as a function of $H(X_1)$, $H(X_2)$, $H(\theta)$ and α.

   (b) Show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$. Are there any values for α for which equality may hold?

4. The *differential entropy* of a $\mathbb{R}^n$-valued random variable X with density f is defined as

$$h(X) := -\int f(\mathbf{x})\log f(\mathbf{x})\,d\mathbf{x}$$

(with the integration over the support of f). Calculate $h(X)$ when

   (a) X is uniformly distributed on $[0,1]$,

   (b) X is standard normal distributed,

   (c) X is exponential distributed with parameter λ.

5. Let X be a $\mathbb{R}^n$-valued random variable with density f, zero mean and covariance matrix K. Show that

$$h(X) \leq n\log\sqrt{2\pi e} + \log\sqrt{|K|}$$

with equality iff X is multivariate normal.

[Hint: you may use without proof that the information inequality is also valid for densities of continuous multivariate random variables.]

6. **(Strong AEP, replaces Proposition 2.10 of the lecture)** Let $\mathcal{X}$ be a finite set and consider iid copies $\boldsymbol{X} = (X_1,\dots,X_n)$ of a $\mathcal{X}$-valued rv X with pmf p. We fix an order the elements of $\mathcal{X}$ so that the sequence of $(p(x))_{x\in\mathcal{X}}$ is non-decreasing and define the subset $\mathcal{S}_n^\varepsilon \subset \mathcal{X}$ greedily by including the elements one-by-one in the given order until we have $\mathbb{P}\left(\boldsymbol{X} \in \mathcal{S}_n^\varepsilon\right) \geq 1 - \varepsilon$. Show that for any $\varepsilon > 0$ we have

$$(1-2\varepsilon)2^{-n(H(X)+\varepsilon)} \leq \left|\mathcal{S}_n^\varepsilon\right| \leq 2^{-n(H(X)+\varepsilon)},$$

for sufficiently large n.
[Hint: show that $\mathbb{P}(\mathbf{X} \in A \cap B) > 1 - \varepsilon_1 - \varepsilon_2$ for any sets with $\mathbb{P}(\boldsymbol{X} \in A) > 1 - \varepsilon_1$, $\mathbb{P}(\boldsymbol{X} \in B) > 1 - \varepsilon_2$ and use this to estimate $\mathbb{P}\left(\mathbf{X} \in \mathcal{S}_n^\varepsilon \cap \mathcal{T}_n^\varepsilon\right)$.]

7. * **(Optional, revision/outlook on Markov chains)** A Markov chain is a sequence of discrete random variables $(X_n)_{n\geq 1}$ such that for all $x_1,\dots,x_{n+1}\in\mathcal{X}$

$$\mathbb{P}(X_{n+1}=x_{n+1}|X_n=x_n,\dots,X_1=x)=\mathbb{P}(X_{n+1}=x_{n+1}|X_n=x_n).$$

The chain is called homogenous if $p_n(x,y):=\mathbb{P}(X_{n+1}=y|X_n=x)$ does not dependend on n (for every $x,y\in\mathcal{X}$). In this case we call $(p(x,y))_{x,y\in\mathcal{X}}$ the transition matrix of (X_n). A fair die is rolled repeatedly. Which of the following are Markov chains? For those that are, give the transition matrix.

   (a) X_n is the largest roll up to the nth roll,
   (b) X_n is the number of sixes in n rolls,
   (c) X_n is the number of rolls since the most recent six,
   (d) X_n is the time until the next six.

8. * **(Optional, revision/outlook on Markov chains)** Let (X_n) be a Markov chain. Which of the following are Markov chains?

   (a) $(X_{m+n})_{n\geq 1}$ for a fixed $m\geq 0$,
   (b) $(X_{2n})_{n\geq 1}$,
   (c) $(Y_n)_{n\geq 1}$ with $Y_n:=(X_n,X_{n+1})$.

9. * **(Optional, relation to non-linear programming)** Recall Lagrange multipliers: given $f:U\subset\mathbb{R}^n\to\mathbb{R}$ and $g_1,\dots,g_k:U\to\mathbb{R}$ we wish to minimize $f(x)$ subject to constraints $g_i(x)=0$ for $i=1,\dots,k$. Assuming sufficient smoothness of f, $g_1,\dots,g_k$ one can introduce the Lagrangian $\mathcal{L}(x,\lambda_1,\dots,\lambda_k):=f(x)-\sum_{i=1}^k\lambda_i g_i(x)$ and find a minimizer $x\in U$ by solving $\frac{\partial\mathcal{L}}{\partial\lambda_i}=0$, $i=1,\dots,k$.

   (a) Give an informal argument why this works. One way to think about this (wlog $k=1$) is to consider $g^{-1}(0)$, $f^{-1}(v)$ for $v\in\mathbb{R}$ as surfaces. Start with v being much bigger than the minimum constrained by $g(x)=0$ and visualize what happens to the surfaces as v decreases and approaches a constrained minimum. What does this tell us about the relation of the gradient ∇f to ∇g?
   (b) Pick up your favourite analysis textbook and look up the formal proof and the assumptions on $f,g_1,\dots,g_k$.
   (c) Let $X_1,\dots X_n$ be independent, real valued variables with $\mathbb{E}[X_i]=\mu$, $Var(X_i)=\sigma_i^2$. Find $c_1,\dots,c_n$ that minimize $Var\left(\sum_{i=1}^n c_iX_i\right)$ subject to $\mathbb{E}\left[\sum_{i=1}^n c_iX_i\right]=\mu$ for given $\mu\in\mathbb{R}$.