Instructions:

- Homeworks are to be done individually.
- Typeset your homework in latex using this file as template (e.g. use pdflatex). We do not accept hand-written homeworks. Show your derivations.
- Hand in a *single-sided printed copy* of your homework *before* class. Homework will no longer be accepted once the lecture starts.
- Unless explicitly specified in the questions, you do not need to hand in any code.
- Fill in your name and email below. This will produce a page separated from the rest of your homework. We will do “double blind review.” Do not include identifying information (e.g. your name) in the rest of your homework.

Name:
Email:
1. Consider a sample \( S = ((x_1, y_1), \ldots, (x_n, y_n)) \) where \( x_i \in \mathbb{R}^d \) and \( y_i \in \{-1, 1\} \). Assume the data is linearly separable with large margin, namely there exists \( w \in \mathbb{R}^d \) such that
\[
y_i \langle w, x_i \rangle \geq 1, \; \forall i = 1 \ldots n.
\]
Let \( w^* \) be a vector that has the minimal norm among all vectors that satisfy the above condition. Let \( R = \max_i \|x_i\| \). Define a function
\[
f(w) = \max_{i} 1 - y_i \langle w, x_i \rangle.
\]
(a) Show that \( \min_{w: \|w\| \leq \|w^*\|} f(w) = 0 \).
(b) Show that any \( w \) for which \( f(w) < 1 \) separates the examples in \( S \).
(c) Show how to calculate a subgradient of \( f \).

2. First, describe the subgradient descent algorithm for problem 1, be sure to specify the learning rate \( \eta \). Second, prove that your algorithm can reach a point where all training items are correctly classified. Third, give the number of iterations \( T \) when this is guaranteed to happen.

*Hint: questions 1 and 2 are not about stochastic GD.*

3. Let \( x_1, \ldots, x_n \in \mathbb{N}^d \). That is, \( x_i \) is a \( d \)-dimensional natural number vector (including zero). Assume that these vectors are iid sampled from a multinomial distribution with parameter \( \theta \). Derive the maximum likelihood estimate of \( \theta \). *Hint: introduce a Lagrange multiplier for the constraint that elements of \( \theta \) sums to one.*

4. A book has \( \theta \) pages. I uniformly randomly flip to a page and tell you the page number \( x \). I repeat this \( n \) times to give you independent observations \( x_1, \ldots, x_n \). Derive the maximum likelihood estimate of \( \theta \).