3. Entropy. The negation of the entropy, \( N(p) = -H(p) \), has Hessian with entries

\[
\frac{\partial^2 N}{\partial p_i \partial p_j} = \begin{cases} 
0 & \text{if } i \neq j, \\
\frac{1}{p_i \ln 2} & \text{if } i = j
\end{cases}
\]

This is a diagonal matrix with positive values on the diagonal. Thus the Hessian is P.S.D., whereupon \( N \) is convex and \( H \) is concave.
4. Regression problem.

(a) Let
\[ X = \begin{pmatrix} \leftarrow x^{(1)} \rightarrow \\
\leftarrow x^{(2)} \rightarrow \\
\left\ldots \right\rightarrow \\
\left\llarrow x^{(n)} \rightarrow \end{pmatrix} \]

Then we can write the Hessian as
\[ H(w) = 2 \sum_{i=1}^{n} x^{(i)} (x^{(i)})^T + 2\lambda I = 2X^T X + 2\lambda I \]

(b) For all \( z \in \mathbb{R}^d \)
\[ z^T H z = z^T (2X^T X + 2\lambda I) z = 2(z^T X^T X z + \lambda z^T I z) = 2||X z||^2 + 2\lambda ||z||^2 \geq 0 \]
Therefore, \( H(w) \) is P.S.D, which means \( L(w) \) is convex.

5. Convex sets.

(a) The circle is not a convex set: for any two points on the circle, the line joining them does not lie on the circle.
(b) The ball is convex.
(c) Hyperplanes are convex.
(d) \( k \)-sparse points are not convex: lines joining two such points can be up to \((2k)\)-sparse.
(e) The set of positive semidefinite matrices is closed under addition and multiplication by positive scalars; therefore it is convex.


(a) We can check that \( \ell_1 \) is a norm by going through the definition, one property at a time:
   i. \( ||x||_1 = \sum_{i=1}^{d} |x_i| \geq 0. \)
   ii. If \( x = 0 \), then \( ||x||_1 = 0. \) If \( \exists i, x_i \neq 0 \), then \( ||x||_1 \geq |x_i| > 0. \) Therefore, \( ||x||_1 = 0 \) if and only if \( x = 0. \)
   iii. For any real-valued \( t \), we have \( ||tx||_1 = \sum_{i=1}^{d} |tx_i| = |t| \sum_{i=1}^{d} |x_i| = |t| ||x||_1 \)
   iv. \( ||x + y||_1 = \sum_{i=1}^{d} |x_i + y_i| \leq \sum_{i=1}^{d} |x_i| + |y_i| = \sum_{i=1}^{d} |x_i| + \sum_{i=1}^{d} |y_i| = ||x||_1 + ||y||_1 \)
(b) Invoking homogeneity and the triangle inequality, we have that for any norm \( f \),
\[ f(\theta x + (1-\theta)y) \leq f(\theta x) + f((1-\theta)y) = |\theta| f(x) + |1-\theta| f(y) = \theta f(x) + (1-\theta)f(y). \]
Thus any norm is a convex function.
(c) Various inequalities relating \( ||x||_1, ||x||, \) and \( ||x||_\infty \):
   i. \( ||x||_1 = \sqrt{\sum_{i=1}^{d} |x_i|^2} = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{d} |x_i||x_j|} \geq \sqrt{\sum_{i=1}^{d} x_i^2} = ||x||. \)
   ii. Let vector \( a = (|x_1|, |x_2|, \ldots, |x_d|), b = (1,1,\ldots,1)_d \)
\[ ||x||_1 = \sum_{i=1}^{d} |x_i| = |a \cdot b| \leq ||a|| ||b|| = \sqrt{\sum_{i=1}^{d} x_i^2} \sqrt{\sum_{i=1}^{d} 1^2} = ||x|| \cdot \sqrt{d}. \]
\[ ||x|| = \sqrt{\sum_{i=1}^{d} x_i^2} \leq \sqrt{d} \cdot \max_i x_i^2 = ||x||_\infty \cdot \sqrt{d}. \]
Therefore, \( ||x||_1 \leq ||x||_1 \leq ||x||_1 \leq ||x|| \cdot \sqrt{d} \leq ||x||_\infty \cdot d. \)
(d) The unit ball \( \{ x : x^T A x \leq 1 \} \) is an ellipsoid.

7. A lower bound for the perceptron. Pick any \( \gamma > 0 \). Consider the following data set in \( \mathbb{R}^d \), where \( d = 1/\gamma^2 \):

- There are \( d \) points, each corresponding to one coordinate direction: \( e_1, e_2, \ldots, e_d \), where \( e_i \) is the vector with all zeros except for a 1 at position \( i \).
- All points have label +1.

These points are correctly classified by the vector \( w^* = (\gamma, \gamma, \ldots, \gamma) \), which has unit length and has margin \( \min_i (w^* \cdot e_i) = \gamma \).

Now suppose the perceptron algorithm is run on this data set, and that it produces a linear separator \( w \). If perceptron does not update on \( e_i \), then \( w_i = 0 \) and \( w \) will not correctly classify \( e_i \). Therefore, there must be at least one update for every data point: a total of \( 1/\gamma^2 \) updates.

8. Small SVM example.

(a) [Diagram of SVM decision boundary]

(b) The margin is \( \sqrt{2} \).

(c) \( w \) lies in the direction \( \begin{pmatrix} 1 \\ 1 \end{pmatrix} \) and has length \( 1/\sqrt{2} \) (since the margin is \( \sqrt{2} \)); therefore, \( w = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} \).

We know that the point \( x_o = (4, 3) \) lies on the decision boundary; solving \( w \cdot x_o + b = 0 \) yields \( b = -7/2 \).

9. Support vectors. The margin decreases if the factor \( C \) is increased.