Assigned reading: Chapter 12, first 3.5 pages (through the statements of Theorems 12.1 and 12.2) of Prof. Raginsky’s notes, and the supplementary notes part 7.

(Problems and solutions)

1. **[Concave-convex properties of mutual information.]**
   For simplicity, consider finite sets \(X\) and \(Y\). A probability mass function (pmf) on \(X\): \(p_X\), and a conditional pmf, \(p_{Y|X}\), determine a joint pmf on \(X \times Y\) by \(p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)\). In turn, the joint pmf \(p_{X,Y}\) determines the mutual information, \(I(X;Y)\).

   (a) Show \(I(X;Y)\) is a concave function of \(p_X\) (for \(p_{Y|X}\) fixed). Hint: One approach is to use \(I(X;Y) = \min_Q D(P_{Y|X}||Q|p_X)\), where \(D(P_{Y|X}||Q|p_X) = \sum_i D(P_{Y|X} = i||Q)p_X(i)\), for any pmf \(Q\) on \(Y\).

   **Solution:** See the hint. The expression given shows \(D(P_{Y|X}||Q|p_X)\) is a linear function of \(p_X\) for each \(Q\) fixed. The hint together with the fact the minimum of linear functions is a concave function establishes the result.

   (b) Show \(I(X;Y)\) is a convex function of \(p_{Y|X}\) (for \(p_X\) fixed). Hint: One approach is to use the fact \(I(X;Y) = D(p_{X,Y}||p_X \otimes p_Y)\) and use the joint convexity of \(D(p||q)\) you proved in problem set 1.

   **Solution:** Since \(p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)\) and \(p_Y(y) = \sum_x p_X(x)p_{Y|X}(y|x)\), we see that \((p_{X,Y}, p_X \otimes p_Y)\) is linear in \(p_{Y|X}\) for \(p_X\) fixed. Since the composition of a linear function and a convex function is convex, by the hint it follows that \(I(X;Y)\) is a convex function of \(p_{Y|X}\).

2. **[Min max excess loss for interval classifiers under Massart condition]**
   Let \(X = [0,1]^2\) and let \(\mathcal{F}\) denote the space of axis parallel rectangle classifiers. Given a probability measure \(P\) on \(X \times \{0,1\}\), let \(\eta(x) = P(Y = 1|X = x)\). Consider the set of probability measures \(\mathcal{P}(h)\) on \(X \times \{0,1\}\) such that \(|\eta(x) - 0.5| \geq h/2\) for all \(x \in X\). Let \(R_n(h, \mathcal{F})\) denote the min max excess risk. For a given \(h \in (0,1)\) fixed, is \(nR_n(h, \mathcal{F})\) bounded over all \(n\)? Justify your answer.

   **Solution:** No. The class \(\mathcal{F}\) is \((N,D)\) rich for \(D = 1\) and all \(N \geq 1\). To see this, fix \(N \geq 1\). Let \(\bar{x}_i = (\frac{i}{N}, \frac{N-i}{N})\) for \(1 \leq i \leq N\). Then for each \(i\), the classifier \(f_i\) corresponding to the the rectangle with corners at \((0,0)\) and \(\bar{x}_i\) assigns label 1 to \(\bar{x}_i\) and 0 to all \(\bar{x}_j\) with \(j \neq i\).

   Therefore, by Massart’s lower bound on \(R_n(h, \mathcal{F})\) we have \(R_n(h, \mathcal{F}) \geq c(1-h)\frac{1}{nh} [1 + \log nh^2]\) for all \(n\) sufficiently large. So for \(h\) fixed, \(R_n(h, \mathcal{F}) \geq c_h n \log n\) for some \(c_h > 0\).

3. **[Implicit regularization of gradient descent –background for next problem]**
   Suppose \(X = \mathbb{R}^{m \times m}\) and \(\mathcal{F}\) is a set mappings from \(X \to \mathbb{R}\) denoted by \(A \mapsto \langle F,A \rangle\) where \(\langle F,A \rangle = \sum_{i,j} F_{ij} A_{ij}\) such that \(F\) is a positive semidefinite matrix. For \(1 \leq r \leq m\), let \(\mathcal{F}_r\) denote the subset of \(\mathcal{F}\) such that \(\text{rank}(F) \leq r\). Any \(F \in \mathcal{F}_r\) can be represented by \(F = UU^\top\) where \(U\) is a matrix of dimensions \(m \times r\).

   (a) Find the gradient of the mapping \(U \mapsto f(U) = \langle UU^\top, A \rangle\) as a function of \(U\). Hint: Express the gradient \(\nabla f(U)\) as a matrix with the same dimensions as \(U\) such that for any \(W\), also of the same dimensions, \(f(U + \epsilon W) = f(U) + \epsilon \langle \nabla f(U), W \rangle + o(\epsilon)\).

   **Solution:** Observe that

   \[
   \begin{align*}
   f(U + \epsilon W) &= \langle (U + \epsilon W)(U + \epsilon W)^\top, A \rangle \\
   &= f(U) + \epsilon \langle UW^\top, A \rangle + \epsilon \langle WU^\top, A \rangle + o(\epsilon) \\
   &= f(U) + \epsilon \langle WU^\top, A + A^\top \rangle \\
   &= f(U) + \epsilon \langle (A + A^\top)U, W \rangle \\
   &= f(U) + \epsilon \langle A + A^\top \rangle U, W \\
   
   \end{align*}
   \]

   Therefore, \(\nabla f(U) = (A + A^\top)U\).
(b) Find the gradient of the mapping $U \mapsto (\langle UU^T, A \rangle - y)^2$

**Solution:** Let $g(U) = (\langle UU^T, A \rangle - y)^2 = (F(U) - y)^2$. By the chain rule and part (a), $\nabla g(U) = 2(f(U) - y)(A + A^T)U$.

4. [Exploring stochastic gradient descent]

This is a python programming problem. Preliminaries are introduced in the previous problem. We suppose $X_1, \ldots, X_n$ are iid $\mathbb{R}^{m \times m}$ matrices with independent $N(0, 1)$ entries. Observations are generated by the linear operation: $y_i = \langle F^*, A_i \rangle$ where $F^* = U^*U^{*T}$. i.e. the realizable case. We run (batch) gradient descent for attempting to minimize the empirical loss as a function of $U$. The problem is nonconvex and will often have multiple global minima, all with zero loss because we consider only the realizable case. See the python notebook for details. You can see a static version at http://nbviewer.jupyter.org/urls/courses.engr.illinois.edu/ece543/ece543_PythonProblem7.ipynb?flush_cache=true and download the ipynb file from the static version or directly from https://courses.engr.illinois.edu/ece543/ece543_PythonProblem7.ipynb.