**Massachusetts Institute of Technology**
**6.435 Theory of Learning and System Identification**

Prof. Dahleh, Prof. Mitter — Homework 3
Out 3/19 — Due T, 4/3

## 1 [Exchangeability and Pólya's Urn]

(a) Random variables $X_1, \cdots, X_m$ are said to be exchangeable if, for all permutations $\boldsymbol{\pi}$ of $\{1, \cdots, m\}$, we have:

$$\mathbf{P}(X_1, \cdots, X_m) = \mathbf{P}(X_{\pi_1}, \cdots, X_{\pi_m}).$$

Consider an urn containing $m$ balls, $r$ of which are marked '1' and the rest '0'. We extract balls from this urn, one at a time without replacement, and let $X_i$ be the marking of the $i^{\text{th}}$ extracted ball. Since we run out in $m$ steps, we obtain a finite sequence $X_1, \cdots, X_m$. Are these random variables independent? Show that they are exchangeable.

(b) Random variables $\{X_n\}$, in an infinite sequence, are said to be exchangeable if, for all finite $m$, $X_1, \cdots, X_m$ are exchangeable.

Let us start with the same urn as in part (a), and proceed with a sequence of extractions, such that after each extraction we replace the extracted ball together with $c > 0$ balls of the same marking. Since we never run out, we obtain an infinite sequence $\{X_n\}$. Show that these random variables are exchangeable.

## 2 [De Finetti's Theorem]

This problem guides through a proof [1] of the zero-one case of de Finetti's theorem, which can be stated as follows:

> To every infinite sequence of exchangeable random variables $\{X_n\}$ having values in $\{0, 1\}$, there corresponds a probability distribution $F$ over $[0, 1]$, such that:
>
> $$p_{k,n} = \mathbf{P}(X_1 = 1, \cdots, X_k = 1, X_{k+1} = 0, X_n = 0) = \int_0^1 \theta^k (1-\theta)^{n-k} F(\mathrm{d}\theta). \qquad (1)$$

(a) De Finetti's theorem can be interpreted as the existence of a prior for a random parameter $\Theta$. Indeed, if $\Theta$ has prior $F$ and if, given $\Theta = \theta$, $X_1, X_2, \cdots$ are independent, Bernoulli, with parameter $\theta$, show that (1) holds.

Note that the theorem is not true for a finite sequence of exchangeable random variables $X_1, \cdots, X_m$. Nevertheless, as a first step, we express $p_{k,n}$ for every finite $m$, and $k \leq n \leq m$, in parts (b) and (c). We then extend to the infinite case, in parts (d) and (e).

[1] Due to Heath and Sudderth, 1976.

(b) Show that for $m$ finite, and $k \leq n \leq m$, we can express $p_{k,n}$ as follows:

$$p_{k,n} = \sum_{r=0}^{m} \mathbf{P}(X_1 = 1, \cdots, X_k = 1, X_{k+1} = 0, X_n = 0 | E_r) \cdot q_r, \qquad (2)$$

where $E_r$ is the event $\{\sum_{j=1}^{m} X_j = r\}$ and $q_r = \mathbf{P}(E_r)$, for $r = 0, \cdots, m$.

(c) Show that:

$$\mathbf{P}(X_1 = 1, \cdots, X_k = 1, X_{k+1} = 0, X_n = 0 | E_r) = \frac{\langle r \rangle_k \langle m - r \rangle_{n-k}}{\langle m \rangle_n},$$

where $\langle a \rangle_i = \prod_{j=0}^{i-1} (a - j)$. [Hint: Consider different cases. Some of these have zero probability. For the others, use exchangeability to reduce finding the probability to a counting problem.]

(d) Show that the above converges to $\theta^k (1-\theta)^{n-k}$, uniformly for all $\theta = r/m$. That is, show that for all $\epsilon > 0$, there exists $M$, such that for all $m > M$ we have:

$$\left| \frac{\langle \theta m \rangle_k \langle (1-\theta) m \rangle_{n-k}}{\langle m \rangle_n} - \theta^k (1-\theta)^{n-k} \right| < \epsilon, \qquad \forall \theta = \frac{r}{m}.$$

(e) Rewrite (2) as follows:

$$p_{k,n} = \int_0^1 \frac{\langle \theta m \rangle_k \langle (1-\theta) m \rangle_{n-k}}{\langle m \rangle_n} F_m(\mathrm{d}\theta),$$

where $F_m$ is a staircase probability distribition on $[0, 1]$ with jumps of $q_r$ at $r/m$. Complete the proof of de Finetti's theorem, using the result of part (d) and the following restricted statement of Helly's selection theorem:

"Every sequence $\{F_m\}$ of probability distributions on $[0, 1]$ contains a subsequence $F_{m_1}, F_{m_2}, \cdots$ that converges uniformly to a limit $F$. That is, for all $\epsilon > 0$, there exists $M$ such that for all $m_i > M$, $|F(x) - F_{m_i}(x)| < \epsilon$, for all $x$."

[Hints: [1]Recall that (2) applies for all $m$. [2]If a sequence converges to a limit, so do all of its subsequences. [3]If $|F(x) - F_m(x)| < \epsilon$ for all $x$ then $\left| \int h(\theta) F(\mathrm{d}\theta) - \int h(\theta) F_m(\mathrm{d}\theta) \right| < \epsilon \int h(\theta) \mathrm{d}\theta$, for $h(\theta) \geq 0$.]

(f) As an illustration, consider the following experimental setting. An ordinary thumbtack is thrown into the air in the center of a large smooth wooden floor. It can come to rest in one of two ways, called "heads" and "tails". This experiment is repeated a great many times, ensuring that subsequent trials are identical.

We ask: "What is the probability that there are seven heads in the first ten flips?" Discuss how you would approach this problem, emphasizing the role of de Finetti's theorem in the answer to this question.

## 3 [Non-Separable SVM]

We would like to extend the notion of the optimal margin separating hyperplane to nonseparable data, i.e. when data is such that any hyperplane results in misclassified samples. The idea is to obtain a method that will find a hyperplane which minimizes the number of misclassifications yet gives a wide margin over the correctly classified samples.

Say we have some hyperplane specified by the normal $\psi$ and offset $b$. Recall that when a data point $(x_i, y_i)$ is properly classified, we can write $y_i(\psi' x_i - b) \geq 1$. When $(x_i, y_i)$ is misclassified, we observe that we can still write $y_i(\psi' x_i - b) \geq 1 - \xi_i$, for some $\xi_i > 0$, which we call a *slack* variable. Since positive slack indicates a misclassified sample, a natural attempt to reduce misclassification is to incorporate the total slack into the optimization problem of the separable setting:

$$\min_{\psi, b, \xi^\ell} \left[ \frac{1}{2}\psi'\psi + C \sum_{i=1}^{\ell} \xi_i \right] \qquad \text{s.t.} \qquad y_i(\psi' x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

We assume that the constant $C$ is predefined. There are justifications of why some choices of $C$ may work better than others, but we do not elaborate that here[†]. Notice that this formulation effectively relaxes the notion of margin, whence the terminology "soft margin SVM".

(a) If the data is in fact separable, show that the solution is the same as that of the original optimization, if $C$ is large enough.

(b) Write down the Lagrangian $\mathcal{L}(\psi, b, \xi, \alpha^\ell \beta^\ell)$ of the new problem, with multipliers $\alpha^\ell$ and $\beta^\ell$ for the first and second sets of constraints respectively.

(c) Show that the optimal $\psi^\circ$ has the same expression as before, and that therefore the notion of support vectors still applies and the problem still depends only on the inner products $x_i' x_j$, and not individual $x_i$'s.

(d) Show that $\alpha^\ell$ is constrained within $[0, C]^\ell$.

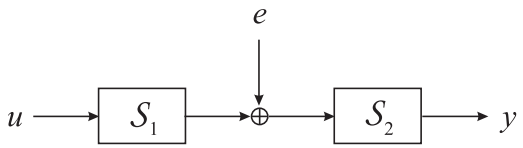
(e) Show that both properly classified and misclassified samples can be support vectors, and that $\alpha_i^\circ = C$ if the $i^{\text{th}}$ sample is a misclassified support vector.

(f) Show that $\psi^{\circ\prime}\psi^\circ + C\sum_{i=1}^{\ell} \xi_i^\circ = \sum_{i=1}^{\ell} \alpha_i^\circ$. [Hint: Use alignment conditions for both $\alpha_i$ and $\beta_i$, $i = 1, \ldots, \ell$.]

(g) [†][optional] Suggest how one could justify letting $C$ be inversely proportional to $\ell$.

## 4 [System Identification]

The goal of this problem is to introduce the basic elements of system indentification. Specifically, we look at the parametric estimation of LTI systems, emphasizing the role and effects of model and input selection. We are given data $(u_i, y_i)$ for $i = 1, \cdots, \ell$. Departing from the i.i.d. case, we assume $u$ and $y$ to be related by the setup:




We typically have control over how the input, $u$, is selected, and will consider two different choices in this problem. LTI systems $\mathcal{S}_1$ and $\mathcal{S}_2$ are characterized by their transfer functions $H_1$ and $H_2$. $e$ is an exogenous noise process, which we assume to be white zero-mean Gaussian with variance $\lambda^2$.

We will tackle this problem parametrically, using the first-order version of an archetypal class of parametrized models, known as ARX (autoregressive with external input):

$$y_i + ay_{i-1} = bu_{i-1} + e_i, \qquad u_0 = y_0 = 0,\ e_i \sim \mathcal{N}(0, \lambda^2),\ \text{white}.$$

Note that this model set corresponds to $H_1(z) = bz^{-1}$ and $H_2(z) = 1/(1 + az^{-1})$. Our task, then, is to estimate the parameter vector $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$.

(a) It can be shown that, asymptotically, the ML estimate $\hat{\theta}$ reduces to:

$$\hat{\theta} = \operatorname{argmin} \sum_{i=1}^{\ell} e_i^2,$$

which, if we write $e_i = y_i + ay_{i-1} - bu_{i-1}$, has the natural interpretation of reducing the discrepancy between the model's prediction and the data, within the limit of noise. Show that $\hat{\theta}$ is the solution of the following matrix equation, if it exists (all sums over $i = 1, \ldots, \ell$):

$$\begin{pmatrix} \frac{1}{N}\sum y_{i-1}^2 & -\frac{1}{N}\sum u_{i-1}y_{i-1} \\ -\frac{1}{N}\sum u_{i-1}y_{i-1} & \frac{1}{N}\sum u_{i-1}^2 \end{pmatrix} \hat{\theta} = \begin{pmatrix} -\frac{1}{N}\sum y_{i-1}y_i \\ \frac{1}{N}\sum u_{i-1}y_i \end{pmatrix}.$$

[Hint: Write the dynamics in matrix form as $e^\ell = y^\ell - \Phi \cdot \theta$, where $\Phi$ is a $\ell \times 2$ matrix of appropriately arranged inputs and outputs.]

In order to study the effects of mismodeling without excessively altering the analysis, we will derive our results for models of the form $H_1(z) = bz^{-1}/(1 + cz^{-1})$ and $H_2(z) = (1 + cz^{-1})/(1 + az^{-1})$, which we can write as:

$$y_i + ay_{i-1} = bu_{i-1} + e_i + ce_{i-1}, \qquad u_0 = y_0 = 0.$$

We are interested in how the estimate behaves when $c = 0$ (correct model) and $c \neq 0$ (incorrect model), under different input choices. [Note that, when $c \neq 0$, the method from part (a) is in fact minimizing $\sum \tilde{e}_i^2$, where $\tilde{e}_i = e_i + ce_{i-1}$.]

(b) Assume $u$ is chosen to be white zero-mean Gaussian with variance $\sigma^2$, uncorrelated with $e$. Show that a true model with $c \neq 0$ induces a bias on the estimate. More specifically, show that the asymptotic ($\ell \to \infty$) value of $\hat{\theta}$ is given by:

$$\hat{a} = a - \frac{c(1-a^2)\lambda^2}{b^2\sigma^2 + (1+c^2-2ac)\lambda^2}, \qquad \hat{b} = b.$$

Compute the numerical asymptotic values for $a = -0.8, b = 1$ for the two cases $c = 0$ and $c = -0.8$. Discuss how the choice of model impacts the estimation process. [Hint: All processes are stationary and ergodic: statistics are time invariant and time averages converge to ensemble averages, i.e. expectations.]

(c) Now, assume $u$ is a step with amplitude $\sigma$, i.e. $u_0 = 0$, $u_i = \sigma$ for $i = 1, \ldots, \ell$. Show that the new asymptotic $\hat{\theta}$ is given by:

$$\hat{a} = a - \frac{c(1-a^2)}{1+c^2-2ac}, \qquad \hat{b} = b - \frac{bc(1-a)}{1+c^2-2ac}.$$

Compute the numerical asymptotic values for the cases of part (b). What role does $\sigma$ play? In some sense white noise is a "richer" input than a step. Discuss that, and how the choice of input affects the estimation problem, under both correct and incorrect modeling. [Hint: Ergodicity and stationarity still hold, with a deterministic signal being a degenerate case. Note that although $y$ doesn't converge to a steady state, $\mathbf{E}[Y]$ does.]