Partially Observed Markov Decision Processes

From Filtering to Controlled Sensing

Errata & Internet Supplement

Vikram Krishnamurthy
Cornell University, NY, USA.
This version: Monday 13th March, 2017

Errata for book

As of Monday 13th March, 2017, here is a list of errors.

Typo Errors:

1. Page 421. Appendix 17.C. bottom of the page. There are a couple of typo errors in the proof of the UCB algorithm. Also the proof given was too terse (even though the proof is elementary). Below we give a more detailed proof with the two typos corrected. Recall that

$$H_{\theta st} = \hat{c}_{\theta,t} + B\sqrt{\frac{\xi \log t}{s}}, \quad \xi > 1$$

where $\hat{c}_{\theta,t}$ is the arithmetic mean of the rewards from arm θ using s samples until time t. We assume that the rewards $c_n(\theta)$ lie in the interval $[0, B]$ for all arms $\theta \in \Theta$ where B is a known positive real number.

For the reader's convenience, replace the material on page 421 of the book starting with: "It only remains to choose τ..." with the following:

It only remains to choose τ to upper bound the right hand side. Choose

$$\tau = C(\theta^*) = \mathbb{E}\{c_n(\theta^*)\}, \text{ and } u \geq \frac{4\,B^2\,\xi}{\big(C(\theta^*) - C(\theta)\big)^2} \log n \qquad (xx)$$

(Choose u as the smallest integer satisfying the above inequality.) Then

$$\begin{aligned}
\mathbb{P}\big(H_{\theta st} > C(\theta^*)\big) &= \mathbb{P}\big(\hat{c}_{\theta,t} - C(\theta) > -B\sqrt{\frac{\xi \log t}{s}} + C(\theta^*) - C(\theta)\big) \\
&\overset{(a)}{\leq} \mathbb{P}\big(\hat{c}_{\theta,t} - C(\theta) > \frac{1}{2}\big(C(\theta^*) - C(\theta)\big)\big) \\
&\overset{(b)}{\leq} \exp\big(-\frac{s}{2B^2}\big(C(\theta^*) - C(\theta)\big)^2 \overset{(c)}{\leq} \exp\big(-\frac{u}{2B^2}\big(C(\theta^*) - C(\theta)\big)^2 \\
&\overset{(d)}{\leq} n^{-2\xi}
\end{aligned}$$

Also $\mathbb{P}(H_{\theta^* st} \leq C(\theta^*)) \leq t^{-2\xi}$ by Hoeffding's inequality.

Note: Inequality (a) follows since $t \leq n$ and then from (xx),

$$\sqrt{\frac{\xi \log t}{s}} \leq \sqrt{\frac{\xi \log n}{s}} = \sqrt{\frac{u}{s}}\frac{C(\theta^*) - C(\theta)}{2} \leq \frac{C(\theta^*) - C(\theta)}{2}, \text{ since } u \leq s.$$

(b) follows from Hoeffding's inequality[1] using s samples keeping in mind the assumption that $c_n(\theta) \in [0, B]$, (c) follows since $u \leq s$, (d) follows by substituting the expression for u in (xx).

The rest of the proof then follows as in the book.

Minor typos in Book: Here is the list of minor typos so far:

page 59 sec 3.7.2: Here we state two important results (not result). plural.

page 59: two lines before Sec.3.8. $P'\pi_1$ and $P'\pi_2$ – the primes $'$ (denoting transpose) are missing.

page 124, Subsection 6.1.2: indexpolicies! Deterministic Markovian should be omitted. It was meant to be an index command.

page 154: In dynamic programming equation $J_k(\pi)$. The right hand side of the equality: e_1 should be replaced with e_2.

page 159: PROGRAM should be program (lowercase)

page 203: footnote 1: to mean non decreasing

page 206: footnote 4: to mean non decreasing

page 240: replace π_i with $\pi(i)$. and π_j with $\pi(j)$.

page 248: in 11.4.2: L parallel projects (instead of P parallel projects)

page 249: eq (11.17). replace max with argmax.

page 258: in proof of Theorem 12.2.1. Replace $Q_1(\pi, 1)$ with $Q(\pi, 1)$

page 259: 6th and 7th line: re[lace B_{1y} with B_{2y} and B_{2y} with B_{1y}.

page 262: Fig 12.2, $\mathcal{L}(e_3, \bar{\pi}_2)$ and $\mathcal{L}(e_1, \bar{\pi}_1)$ should be swapped.

page 263: replace π^0 with $\tilde{\pi}$.

page 294: Fig 13.1. Private belief π_k should be replaced with η_k.

page 302: (13.39) remove the index $k-1$ in π in two places

page 302: (13.40) remove extra { in second eqn

page 314: replace $\leq$ in (14.4) by $\geq$.

page 350: Sec 15.5 second last line of page $k \in \iota_n$: remove the $\in \iota_n$

page 362: Third last eqn on page, conditioning on $\dot{x}_j$, and $\ddot{x}_j$. there is a missing comma between $\dot{x}_j$, and $\ddot{x}_j$

page 367: Eq (16.9): $Q(i+, u)$ should be replaced by $Q(i+1, u)$

page 412: in footnote replace ρl with $\rho(l)$

page 412: in (17.75) replace θ_k with $\theta_k(l)$

page 444: (B.6) replace comma with full stop.

page 444: B3.1 sentence just above the equation for N_t, missing full stop.

[1] If $\mu_n = \frac{1}{n}\sum_{i=1}^n x_i$ (i.e., μ_n is the arithmetic mean using n samples) where $x_i \in [0, B]$ are i.i.d. with $\mathbb{E}\{x_i\} = \mu$, then

$$P(\mu_n - \mu > \epsilon) \leq \exp\left(-\frac{2n\epsilon^2}{B^2}\right)$$

Preface to Internet Supplement

This document is an *internet supplement* to my book "Partially Observed Markov Decision Processes – From Filtering to Controlled Sensing" published by Cambridge University Press in 2016.[2]

This internet supplement contains exercises, examples and case studies. The material appears in this internet supplement (instead of the book) so that it can be updated. This document will evolve over time and further discussion and examples will be added.

The website `http://www.pomdp.org` contains downloadable software for solving POMDPs and several examples of POMDPs. I have found that by interfacing the POMDP solver with Matlab, one can solve several interesting types of POMDPs such as those with non-linear costs (in terms of the information state) and bandit problems.

I have given considerable thought into designing the exercises and case studies in this internet supplement. They are mainly mini-research type exercises rather than simplistic drill type exercises. Some of the problems are extensions of the material in the book. The exercises are suitable as term projects for a graduate level course on POMDPs; many of these exercises have been used in courses I have taught at UBC.

As can be seen from the content list, this document also contains some short (and in some cases, fairly incomplete) case studies which will be made more detailed over time. These case studies were put in this internet supplement in order to keep the size of the book manageable. As time progresses, I hope to incorporate additional case studies and other pedagogical notes to this document to assist in understanding some of the material in the book. Time permitting, future plans include adding a detailed discussion on structural results for POMDP games; structural results for quasi-variational inequalities, etc.

To avoid confusion in numbering, the equations in this internet supplement are numbered consecutively starting from (1) and not chapter wise. In comparison, the equations in the book are numbered chapterwise.

This internet supplement document is work in progress and will be updated periodically. Having taught this entire book now as ECE 6950 at Cornell University in Fall os 2016, minor typos were picked up as indicated in the errata.

I welcome constructive comments from readers of the book and this internet supplement. Please email me at `vikramk@cornell.edu`

Vikram Krishnamurthy,
2017

[2]Online ISBN:9781316471104 and Hardback ISBN:9781107134607

Contents

Chapter 2

Stochastic State Space Models

1. As a simple drill exercise show that if a Markov chain has 2×2 transition matrix

$$P = \begin{bmatrix} 1-a & b \\ b & 1-b \end{bmatrix}$$

then

$$P^n = \frac{1}{a+b}\begin{bmatrix} b & a \\ b & a \end{bmatrix} + \frac{(1-a-b)^n}{a+b}\begin{bmatrix} a & -a \\ -b & b \end{bmatrix}$$

Therefore the optimal predictor (given by the Chapman Kolmogorov equation) can be evaluated explicitly as $\pi_n = (P^n)'\pi_0$ where $'$ denotes transpose.

2. Theorem 2.4.2 of the book dealt with the stationary distribution and eigenvalues of a stochastic matrix (transition probability matrix of a Markov chain). Parts of Theorem 2.4.2 can be shown via elementary linear algebra.
Statement 2: Define spectral radius $\bar{\lambda}(P) = \max_i |\lambda_i|$
Lemma : $\bar{\lambda}(P) \leq \|P\|_\infty$ where $\|P\|_\infty = \max_i \sum_j P_{ij}$
Proof: For all eigenvalues λ, $|\lambda|\|x\| = \|\lambda x\| = \|Px\| \leq \|P\|\|x\| \implies |\lambda| < \|P\|$.
For a stochastic matrix, $\|P\|_\infty = 1$ and P has an eigenvalue at 1. So $\bar{\lambda} = 1$.
Statement 3: For non-negative matrix P, $P'\pi = \pi$ implies $P'|\pi| = |\pi|$ where $|\pi|$ denotes the vector with element-wise absolute values.
Proof: $|\pi| = |P'\pi| \leq |P'||\pi| = P'|\pi|$ So $P'|\pi| - |\pi| \geq 0$.
But $P'|\pi| - |\pi| > 0$ is impossible, since it implies $1'P'|\pi| > 1'|\pi|$, i.e., $1'|\pi| > 1'|\pi|$.

3. Farkas' lemma is a widely used result in linear algebra. It states: Let M be an $m \times n$ matrix and b an m-dimensional vector. Then only one of the following statements is true:

   (a) There exists a vector $x \in \mathbb{R}^n$ such that $Mx = b$ and $x \geq 0$.

   (b) There exists a vector $y \in \mathbb{R}^m$ such that $M'y \geq 0$ and $b'y < 0$.

   Here $x \geq 0$ means that all components of the vector x are non-negative.
   Use Farkas lemma to prove that every transition matrix P has a stationary distribution. That is, for any $X \times X$ stochastic matrix P, there exists a probability vector π such that $P'\pi = \pi$. (Recall a probability vector π satisfies $\pi(i) \geq 0$, $\sum_i \pi(i) = 1$).

Hint: Write alternative (a) of Farkas lemma as

$$\begin{bmatrix} (P-I)' \\ \mathbf{1}' \end{bmatrix} \pi = \begin{bmatrix} 0_X \\ 1 \end{bmatrix}, \quad \pi > 0$$

Show that this has a solution by demonstrating that alternative (b) does not have a solution.

4. Using the maneuvering target model of Chapter 2.6, simulate the dynamics and measurement process of a target with the following specifications:

Sampling interval	$\Delta = 7$ s
Number of measurements	$N = 50$
Initial target position	$(-500, -500)'$ m
Initial target velocity	$(0.0, 5.0)'$ m/s
Transition probability matrix	$P_{ij} = \begin{cases} 0.9 \text{ if } i = j \\ 0.05 \text{ otherwise} \end{cases}$
Maneuver commands (three)	$fr = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}'$ (straight) $fr = \begin{pmatrix} -1.225 & -0.35 & 1.225 & 0.35 \end{pmatrix}'$ (left turn) $fr = \begin{pmatrix} 1.225 & 0.35 & -1.225 & -0.35 \end{pmatrix}'$ (right turn)
Observation matrix	$C = I_{4\times 4}$
Process noise	$Q = (0.1)^2 I_{4\times 4}$
Measurement noise	$R = \text{diag}(20.0^2, 1.0^2, 20.0^2, 1.0^2)$
Measurement volume V	$[-1000, 1000]$ m in x and y position $[-10.0, 10.0]$ m/s in x and y velocity

5. Simulate the optimal predictor via the composition method. The composition method is discussed in §2.5.2.
6. As should be apparent from an elementary linear systems course, the algebraic Lyapunov equation (2.22) is intimately linked with the stability of a linear discrete time system. Prove that A has all its eigenvalues strictly inside the unit circle iff for every positive definite matrix Q, there exists a positive definite matrix Σ_∞ such that (2.22) holds.
7. Theorem 2.7.2 states that $|\lambda_2| \leq \rho(P)$. That is, the Dobrushin coefficient upper bounds the second largest eigenvalue modulus of a stochastic matrix P. Show that

$$\log |\lambda_2| = \lim_{k\to\infty} \frac{1}{k} \log \rho(P^k)$$

8. Often for sparse transition matrices, $\rho(P)$ is typically equal to 1 and therefore not useful since it provides a trivial upper bound for $|\lambda_2|$. For example, consider a random walk characterized by the tridiagonal transition matrix

$$P = \begin{bmatrix} r_0 & p_0 & 0 & 0 & \cdots & 0 \\ q_1 & r_1 & p_1 & 0 & \cdots & 0 \\ 0 & q_2 & r_2 & p_2 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & q_{X-1} & r_{X-1} & p_{X-1} \\ 0 & \cdots & 0 & 0 & q_X & r_X \end{bmatrix}$$

Then using Property 3 of $\rho(\cdot)$ above, clearly $\sum_l \min\{P_{il}, P_{jl}\} = 0$, implying that $\rho(P) = 1$. So for this example, the Dobrushin coefficient does not say anything about the initial condition being forgotten geometrically fast.

For such cases, it is often useful to consider the Dobrushin coefficient of powers of P. In the above example, clearly every state communicates with every other state in at least X time points. So P^X has strictly positive elements. Therefore $\rho(P^X)$ is strictly smaller than 1 and is a useful bound. Geometric ergodicity follows by consider blocks of length X, i.e.,

$$\|P^{X\prime}\pi - P^{X\prime}\bar{\pi}\|_{\text{TV}} \leq \rho(P^X)\|\pi - \bar{\pi}\|_{\text{TV}}$$

9. Show that the inhomogeneous Markov chain with transition matrix

$$P(2n-1) = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix}, \quad P(2n) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

is weakly ergodic.

10. **Wasserstein distance.** As mentioned in §2.8 of the book, the Dobrushin coefficient is a special case of a more general coefficient of ergodicity. This general definition is in terms of the Wasserstein metric which we now define: Let d be a metric on the state space $\mathcal{X} = \{e_1, e_2, \dots, \}$ where the state space is possibly denumerable. Consider the bivariate random vector $(x, y) \in \mathcal{X} \times \mathcal{X}$ with marginals π_x and π_y, respectively. Define the Wasserstein distance as

$$d(\pi_x, \pi_y) = \inf \mathbb{E}\{d(x, y)\}$$

where the infimum is over the joint distribution of (x, y).

(a) Show that the variational distance is a special case of the Wasserstein distance obtained by choosing $d(x, y)$ as the discrete metric

$$d(x, y) = \begin{cases} 1 & x \neq y \\ 0 & x = y. \end{cases}$$

(b) Define the coefficient of ergodicity associated with the Wasserstein distance as

$$\rho(P) = \sup_{i \neq j} \frac{d(P'e_i, P'e_j)}{d(e_i, e_j)}$$

Show that the Dobrushin coefficient is a special case of the above coefficient of ergodicity corresponding to the discrete metric.

(c) Show that the above coefficient of ergodicity satisfies properties 2, 4 and 5 of Theorem 2.7.2.

11. **Ultrametric transition matrices.** It is trivial to verify that P^n is a stochastic matrix for any integer $n \geq 0$. Under what conditions is $P^{1/n}$ a stochastic matrix? A symmetric ultrametric stochastic matrix P defined in §14.7 of the book satisfies this property.

12. **Composition Method.** In the book, we used the composition method as a simulation based method for implementing the optimal predictor. Recall that to generate samples from $p(x) = \int p(x|y)\, p(y)\, dy$, the composition method algorithm had two steps:

- Generate $Y \sim p(y)$

- Generate $X \sim p(x|Y)$

The proof of the composition method is straightforward as follows: Let W denote the random variable generated by the algorithm. Then

$$\mathbb{P}(W \leq w) = \int_{\mathbb{R}} \mathbb{P}(I(X \leq w)|Y = y)\, p(y)\, dy = \int_{\mathbb{R}} \int_{-\infty}^{w} p(x|y)dx\, p(y)dy = \int_{-\infty}^{w} p(x)dx.$$

Chapter 3

Optimal Filtering

3.1 Problems

1. Standard drill exercises include:
   (a) Compare via simulations the recursive least squares with the Kalman filter
   (b) Compare via simulations the recursive least square and the least mean squares (LMS) algorithm with a HMM filter when tracking a slow Markov chain. Note that Chapter 17.3 of the book gives performance bounds on how well a LMS algorithm can track a slow Markov chain.
   (c) Another standard exercise is to try out variations of the particle filter with different importance distributions and resampling strategies on different models. Compare via simulations the cubature filter, unscented Kalman filter and a particle filter for a bearings only target tracking model.
   (d) A classical result involving the Kalman filter is the so called innovations state space model representation and the associated spectral factorization problem for the Riccati equation, see [2].
   (e) **Posterior Cramer Rao bound.** The posterior Cramer Rao bound [88] for filtering can be used to compute a lower bound to the mean square error. This requires twice differentiability of the logarithm of the joint density. For HMMs, one possibility is to consider the Weiss-Weinstein bounds , see [79]. Chapter 10 of the book gives more useful sample path bounds on the HMM filter using stochastic dominance.
2. **Bayes' rule interpretation of Lasso.**[73] Suppose that the state $x \in \mathbb{R}^X$ is a random variable with prior pdf

$$p(x) = \prod_{j=1}^{X} \frac{\lambda}{2} \exp\left(-\lambda x(j)\right).$$

   Suppose x is observed via the observation equation

$$y = Ax + v, \qquad v \sim \boldsymbol{N}(0, \sigma^2 I)$$

   where A is a known $n \times X$ matrix. The variance σ^2 is not known and has a prior pdf $p(\sigma^2)$. Then show that the posterior of (x, σ^2) given the observation y is of the form

$$p(x, \sigma^2|y) \propto p(\sigma^2)\,(\sigma^2)^{-\frac{n+1}{2}} \exp\big(-\frac{1}{2\sigma^2}\,\mathrm{Lasso}(x, y, \mu)\big)$$

where $\mu = 2\sigma^2\lambda$ and

$$\text{Lasso}(x, y, \mu) = \|y - Ax\|^2 + \mu\|x\|_1.$$

Therefore for fixed σ^2, computing the mode $\hat{x}$ of the posterior is equivalent to computing the minimizer $\hat{x}$ of $\text{Lasso}(x, y, \mu)$.
The resulting Lasso (least absolute shrinkage and selection operator) estimator $\hat{x}$ was proposed in [87] which is one of the most influential papers in statistics since the 1990s. Since $\text{Lasso}(x, y, \mu)$ is convex in x it can be computed efficiently via convex optimization algorithms.

3. Show that if $X \leq Y$ (with probability 1), then $\mathbb{E}\{X|Z\} \leq \mathbb{E}\{Y|Z\}$ for any information Z.
4. Show that for a linear Gaussian system (3.29), (3.30),

$$p(y_k|y_{1:k-1}) = \boldsymbol{N}(y_k - y_{k|k-1}, C_k\Sigma_{k|k-1}C_k' + R_k)$$

where $y_{k|k-1}$ and $\Sigma_{k|k-1}$ are defined in (3.31), (3.32), respectively.
5. **Finite dimensional filters for polynomial systems with Gaussian noise [9, 34].** As discussed in the book, for a linear system with Gaussian noise, the Kalman filter is the optimal filter. Consider the following polynomial system with Gaussian noise:

$$\begin{aligned} x_{k+1} &= A(x_k) + w_k \\ y_k &= x_k + v_k \end{aligned}$$

where w, v are iid Gaussian processes, and $A(x)$ is a polynomial function of the state x. For example, if the state x is a scalar, then for some positive integer p,

$$A(x) = A(0) + A(1)x + \ldots + A(p)x^p.$$

Then it is shown in [9, 34] that the optimal state estimate $\mathbb{E}\{x_k|y_1, \ldots, y_k\}$ can be computed via a finite dimensional filter in terms of quantities derived from the Kalman filter.
The intuition behind the result is as follows: Computing $\mathbb{E}\{x_k|y_{1:k}\}$ requires computing $\mathbb{E}\{x_k^2|y_{1:k}\}$. For the Kalman filter case, this is given by the covariance update. For a higher order polynomial system, one ends with the requirement of computing $\mathbb{E}\{x_k^{p+1}|y_{1:k}\}$ in order to compute $\mathbb{E}\{x_k^p|y_{1:k}\}$ for positive integer p. But because the noise is Gaussian, these higher order moments can be computed explicitly.
6. **Stein's formula**. Recall that the derivation of the Kalman filter used some useful properties of Gaussians. It would be remiss of us not to mention another beautiful formula involving Gaussians, namely Stein's formula. We give here the univariate version: Let $x \sim N(\mu, \sigma^2)$. Then

$$\mathbb{E}\{(x - \mu)\, f(x)\} = \sigma^2\, \mathbb{E}\left\{\frac{df(x)}{dx}\right\}$$

More generally, (x, y) jointly Gaussian implies

$$\text{cov}(f(x), y) = \mathbb{E}\{\frac{df(x)}{dx}\}\ \text{cov}(x, y)$$

7. Simulate in Matlab the HMM filter, and fixed lag smoother. Study empirically how the error probability of the estimates decreases with lag. (The filter is a fixed lag smoother with lag of zero). Please also refer to [30] for a very nice analysis of error probabilities.
8. Consider a HMM where the Markov chain evolves slowly with transition matrix $P = I + \epsilon Q$ where ϵ is a small positive constant and Q is a generator matrix. That is $Q_{ii} < 0$, $Q_{ij} > 0$ and each row of Q sums to zero. Compare the performance of the HMM filter with the recursive least squares algorithm (with an appropriate forgetting factor chosen) for estimating the underlying state.
9. Consider the following Markov modulated auto-regressive time series model:

$$z_{k+1} = A(r_{k+1})\, z_k + \Gamma(r_{k+1})\, w_{k+1} + f(r_{k+1})\, u_{k+1}$$

   where $w_k \sim \boldsymbol{N}(0,1)$, u_k is a known exogenous input. Assume the sequence $\{z_k\}$ is observed. Derive an optimal filter for the underlying Markov chain r_k. (In comparison to a jump Markov linear system, z_k is observed without noise in this problem. The optimal filter is very similar to the HMM filter).
10. Consider a Markov chain x_k corrupted by iid zero mean Gaussian noise and a sinusoid:

$$y_k = x_k + \sin(k/100) + v_k$$

    Obtain a filtering algorithm for extracting x_k given the observations.
11. **Image Based Tracking.** The idea is to estimate the coordinates z_k of the target by measuring its orientation r_k in noise. For example an imager can determine which direction an aircraft's nose is pointing thereby giving useful information about which direction it can move. Assume that the target's orientation evolves according to a finite state Markov chain. (In other words, the imager quantizes the target orientation to one of a finite number of possibilities.) Then the model for the filtering problem is

$$\begin{aligned} z_{k+1} &= A(r_{k+1})\, z_k + \Gamma(r_{k+1})\, w_{k+1} \\ y_k &\sim p(y|r_k) \end{aligned}$$

    Derive the filtering expression for $\mathbb{E}\{z_k|y_1,\ldots,y_k\}$. The papers [85, 45, 22] consider image based filtering.
12. Consider a jump Markov linear system. Via computer simulations, compare the IMM algorithm, Unscented Kalman filter and particle filter.
13. **Radar pulse train de-interleaving.** Consider a radar receiver that receives radar pulses from multiple periodic sources. (This receiver could be viewed as an eavesdropper listening to various radars.) It is of interest to estimate the periods of these sources. For example, suppose:
    - source 1 pulses are received at times

$$2, 7, 12, 17, 22, 27, 32, 37, 42, \ldots, \quad \text{(period = 5, phase =2)}$$

    - source 2 pulses are received at times

$$4, 15, 26, 37, 48, 59, 70, 81, \ldots, \quad \text{(period = 11, phase = 4)}\ .$$

The interleaved signal consists of pulses at times

$$2, 4, 7, 12, 15, 22, 26, 27, 32, 37, 42, \ldots .$$

Notice the above interleaved signal contains time of arrival information only. For example, at time 37, pulses are received from both sources; but it is assumed that there is no amplitude information - so the received signal is simply a time of arrival event at time 37. At the receiver, the interleaved signal (time of arrivals) is corrupted by jitter noise (modeled as iid noise). So the noisy received signal are, for example,

$$2.4,\ 4.1,\ 6.7,\ 11.4,\ 15.5,\ 21.9,\ 26.2,\ 27.5,\ 30.9,\ 38.2,\ 43.6, \ldots .$$

Given this noisy interleaved signal, the de-interleaving problem aims to determine which pulses came from which source. That is, the aim is to estimate estimate the periods (namely, 5 and 11) and phases (namely, 2 and 4) of the 2 sources.
The de-interleaving problem can be formulated as a jump Markov linear system. Define the state $x_k' = (T', \tau_k')$, consists of the periods $T' = (T^{(1)}, \ldots, T^{(N)})$ of the N sources and $\tau_k' = (\tau_k^{(1)}, \ldots, \tau_k^{(N)})$, where $\tau_k^{(i)}$ denotes the last time source i was active up to and including the arrival of the kth pulse. Let $\tau_1 = (\phi^{(1)}, \ldots, \phi^{(N)})$, be the phases of periodic pulse-train sources. Then

$$\tau_{k+1}^i = \begin{cases} \tau_k^i + T^i & \text{if } (k+1)\text{th pulse is due to source } i \\ \tau_k^i & \text{otherwise} \end{cases} \ ; \ \tau_1^i = \phi^{(i)}. \qquad (1)$$

Let $e_i, i = 1, \ldots, N$, be the unit N-dimensional vectors with 1 in the ith position. Let $r_k \in \{1, \ldots, N\}$ denote the active source at time k. Then one can express the time of arrivals as the jump Markov linear system

$$\begin{aligned} x_{k+1} &= A(r_{k+1})x_k + w_k \\ y_k &= C(r_k)x_k + v_k \end{aligned}$$

where

$$A(r_{k+1}) = \begin{bmatrix} I_N & 0_{N\times N} \\ \text{diag}(e_{r_{k+1}}) & I_N \end{bmatrix},\ C(r_k) = \begin{bmatrix} 0_{1\times N} & e_{r_k}' \end{bmatrix}$$

Note that r_k is a periodic process and so has transition probabilities

$$P_{i,i+1} = 1, \ \text{ for } i < M, \text{ and } \ P_{M,1} = 1$$

for some integer M, where M depends on the periods and phases of the sources. v_k denotes the measurement (jitter) noise; while w_k can be used to model time varying periods.
Remark: Obviously, there are identifiability issues; for example, if $\phi^{(1)} = \phi^{(2)}$ and $T^{(1)}$ is a multiple of $T^{(2)}$ then it is impossible to detect source 1.

14. **Narrowband Interference and JMLS.** Narrowband interference corrupting a Markov chain can be modeled as a jump Markov linear system. Narrowband interference can be modeled as an auto-regressive (AR) process with poles close to the unit circle: for example

$$i_k = a\, i_{k-1} + w_k$$

where $a = 1 - \epsilon$ and ϵ is a small positive number. Consider the observation model

$$y_k = x_k + i_k + v_k$$

where x_k is a finite state Markov chain, i_k is narrowband interference and v_k is observation noise. Show that the above model can be represented as a jump Markov linear system.

15. **Bayesian estimation of Stochastic context free grammar.** First some perspective: HMMs with a finite observation space are also called regular grammars. They are a subset of a more general class of models called stochastic context free grammars as depicted by Chomsky's hierarchy in Figure 3.1.

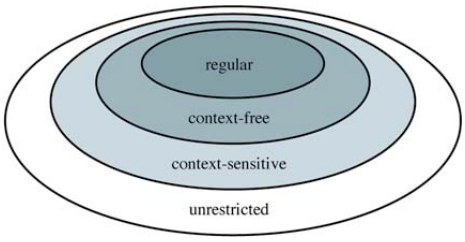

Figure 3.1: The Chomsky hierarchy of languages

Stochastic context free grammars (SCFGs) provide a powerful modeling tool for strings of alphabets and are used widely in natural language processing [58]. For example, consider the randomly generated string $a^n c^m b^n$ where m, n are non-negative integer valued random variables. Here a^n means the alphabet a repeated n times. The string $a^n c^m b^n$ could model the trajectory of a target that moves n steps north and then an arbitrary number of steps east or west and then n steps south, implying that the target performs a U-turn. A basic course in computer science would show (using a pumping lemma) that such strings cannot be generated exclusively using a Markov chain (since the memory n is variable).
If the string $a^n c^m b^n$ was observed in noise, then Bayesian estimation (stochastic parsing) algorithms can be used to estimate the underlying string. Such meta-level tracking algorithms have polynomial computational cost (in the data length) and are useful for estimating *trajectories* of targets (given noisy position and velocity measurements). They allow a human radar operator to interpret tracks and can be viewed as middleware in the human-sensor interface. Such stochastic context free grammars generalize HMMs and facilitate modeling complex spatial trajectories of targets.
Please refer to [58] for Bayesian signal processing algorithms and EM algorithms for stochastic context free grammars. [24, 23] gives examples of meta-level target tracking using stochastic context free grammars.

16. **Kalman vs HMM filter**. A Kalman filter is the optimal state estimator for the linear Gaussian state space model

$$\begin{aligned} x_{k+1} &= Ax_k + w_k, \\ y_k &= C'x_k + v_k. \end{aligned}$$

where w and v are mutually independent iid Gaussian processes.
Recall from (2.28), (2.29) that for a Markov chain with state space $\mathcal{X} = \{e_1, \ldots, e_X\}$ of unit vectors, an HMM can be expressed as

$$\begin{aligned} x_{k+1} &= P' x_k + w_k, \\ y_k &= C' x_k + v_k. \end{aligned}$$

A key difference is that in (2.28), w is no longer i.i.d; instead it is a martingale difference process: $\mathbb{E}\{w_k | x_0, x_1, \ldots, x_k\} = 0$.
From §3.4.4 of the book, it follows that the Kalman filter is the minimum variance linear estimator for the above HMM. Of course the optimal *linear* estimator (Kalman filter) can perform substantially worse than the optimal estimator (HMM filter). Compare the performance of the HMM filter and Kalman filter numerically for the above example.

17. **Interpolation of a HMM.** Consider a Markov chain x_k with transition matrix P where the discrete time clock ticks at intervals of 10 seconds. Assume noisy measurements are obtained of at each time k. Devise a smoothing algorithm to estimate the state of the Markov chain at 5 second intervals. (Note: Obviously on the 5 second time scale, the transition matrix is $P^{1/2}$. For this to be a valid stochastic matrix it is sufficient that P is a symmetric ultrametric matrix or more generally P^{-1} is an M-matrix [36]; see also §14.7 of the book.)
18. **Hierarchical Bayes and Empirical Bayes.** There is much fanfare in machine learning about hierarchical Bayes and empirical Bayes models.
The **hierarchical Bayes** model is of the form

$$\begin{aligned} \Theta &\sim p(\theta) \\ X|\Theta &\sim p(x|\theta) \\ Y|X &\sim p(y|x) \end{aligned} \tag{2}$$

θ is called the hyperparameter and can be thought of as a nuisance parameter. The aim is to compute the posterior $p(x|y)$. Naturally, one can write down Bayes rule for the posterior $p(x|y)$ by integrating away θ. One can then use a MCMC simulation based method to sample from the posterior of $p(x|y)$.
The **empirical Bayes** model is of the form

$$\begin{aligned} X|\Theta &\sim p(x|\theta) \\ Y|X &\sim p(y|x) \end{aligned} \tag{3}$$

So there is no explicit density for the hyperparameter θ. Instead typically the maximum likelihood estimate $\theta^* = \operatorname{argmax}_\theta p(y|\theta)$ is computed. Note

$$p(y|\theta) = \int_{\mathcal{X}} p(y|x)\, p(x|\theta)\, dx$$

The estimate θ^* is then plugged into Bayes rule to evaluate the posterior $p(x|y, \theta^*)$.

3.2 Case Study. Sensitivity of HMM filter to transition matrix

Almost an identical proof to that of geometric ergodicity proof of the HMM filter in §3.7 can be used to obtain expressions for the sensitivity of the HMM filter to the HMM parameters.

Aim: We are interested in a recursion for $\|\pi_k - \underline{\pi}_k\|_1$ when π_k is updated with HMM filter using transition matrix P and $\underline{\pi}_k$ is updated with HMM filter using transition matrix $\underline{P}$. That is, we want an expression for

$$\|T(\pi, y; P) - T(\underline{\pi}, y; \underline{P})\|_1 \text{ in terms of } \|\pi - \underline{\pi}\|_1. \tag{4}$$

Such a bound if useful when the HMM filter is implemented with an incorrect transition matrix $\underline{P}$ instead of actual transition matrix P. The idea is that when P is close to $\underline{P}$ then $T(\pi, y; P)$ is close to $T(\pi, y; P)$.

A special case of (4) is to obtain an expression for

$$\|T(\pi, y; P) - T(\pi, y; \underline{P})\|_1 \tag{5}$$

that is when both HMM filters have the same initial belief π but are updated with different transition matrices, namely P and $\underline{P}$.

The theorem below obtains expressions for both (4) and (5).

Theorem. *Consider a HMM with transition matrix P and state levels g. Let $\epsilon > 0$ denote the user defined parameter. Suppose $\|\underline{P} - P\|_1 \leq \epsilon$, where $\|\cdot\|_1$ denotes the induced 1-norm for matrices.*[1] *Then*

1. *The expected absolute deviation between one step of filtering using P versus $\underline{P}$ is upper bounded as:*

$$\mathbb{E}_y \left| g' \left(T(\pi, y; P) - T(\pi, y; \underline{P}) \right) \right| \;\leq\; \epsilon \sum_y \max_{i,j} g'(I - T(\pi, y; \underline{P})\mathbf{1}')B_y(e_i - e_j) \tag{6}$$

2. *The sample paths of the filtered posteriors and conditional means have the following explicit bounds at each time k:*

$$\|\pi_k - \underline{\pi}_k\|_1 \quad \leq \quad \frac{\epsilon}{\max\{A(\underline{\pi}_{k-1}, y_k) - \epsilon,\, \mu(y_k)\}} + \frac{\rho(\underline{P})\, \|\pi_{k-1} - \underline{\pi}_{k-1}\|_1}{A(\underline{\pi}_{k-1}, y_k)} \tag{7}$$

Here $\rho(\underline{P})$ denotes the Dobrushin coefficient of the transition matrix $\underline{P}$ and $\underline{\pi}_k$ is the posterior computed using the HMM filter with $\underline{P}$, and

$$A(\underline{\pi}, y) = \frac{\mathbf{1}' B_y \underline{P}' \underline{\pi}}{\max_i B_{i,y}}, \quad \mu(y) = \frac{\min_i B_{iy}}{\max_i B_{iy}}. \tag{8}$$

The above theorem gives explicit upper bounds between the filtered distributions using transition matrices $\underline{P}$ and $\bar{P}$. The $\mathbb{E}_y$ in (6) is with respect to the measure $\sigma(\pi, y; P) = \mathbf{1}' B_y P' \pi$ which corresponds to $\mathbb{P}(y_k = y | \pi_{k-1} = \pi)$.

[1] The three statements $\|P'\pi - \underline{P}'\pi\|_1 \leq \epsilon$, $\|\underline{P} - P\|_1 \leq \epsilon$ and $\sum_{i=1}^X \|(P' - \underline{P}')_{:,i}\|_1 \pi(i) \leq \epsilon$ are all equivalent since $\|\pi\|_1 = 1$.

Proof. The triangle inequality for norms yields

$$\begin{aligned}\|\pi_{k+1} - \underline{\pi}_{k+1}\|_{\text{TV}} &= \|T(\pi_k, y_{k+1}; P) - T(\underline{\pi}_k, y_{k+1}; \underline{P})\|_{\text{TV}} \\ &\leq \|T(\pi_k, y_{k+1}; P) - T(\pi_k, y_{k+1}; \underline{P})\|_{\text{TV}} \\ &\quad + \|T(\pi_k, y_{k+1}; \underline{P}) - T(\underline{\pi}_k, y_{k+1}; \underline{P})\|_{\text{TV}}. \end{aligned} \tag{9}$$

Part 1: Consider the first normed term in the right hand side of (9). Applying (3.103) with $\pi = P'\pi_k$ and $\pi^0 = \underline{P}'\pi_k$ yields

$$g'(T(\pi_k, y; P) - T(\pi_k, y; \underline{P})) = \frac{1}{\sigma(\pi, y; P)} g' \left[I - T(\pi, y, \underline{P})\mathbf{1}'\right] B_y(P - \underline{P})'\pi$$

where $\sigma(\pi, y; P) = \mathbf{1}'B_y P'\pi$. Then Lemma 3.7.4(i) yields

$$\begin{aligned} &g'(T(\pi_k, y; P) - T(\pi_k, y; \underline{P})) \\ &\qquad \leq \max_{i,j} \frac{1}{\sigma(\pi, y; P)} g' \left[I - T(\pi, y, \underline{P})\mathbf{1}'\right] B_y(e_i - e_j) \|P'\pi - \underline{P}'\pi\|_{\text{TV}} \end{aligned}$$

Since $\|P'\pi - \underline{P}'\pi\|_{\text{TV}} \leq \epsilon$, taking expectations with respect to the measure $\sigma(\pi, y; P)$, completes the proof of the first assertion.

Part 2: Applying Theorem 3.7.5(i) with the notation $\pi = P'\pi_k$ and $\pi^0 = \underline{P}'\pi_k$ yields

$$\begin{aligned} \|T(\pi_k, y; P) - T(\pi_k, y; \underline{P})\|_{\text{TV}} &\leq \frac{\max_i B_{i,y} \|P'\pi_k - \underline{P}'\pi_k\|_{\text{TV}}}{\mathbf{1}'B_y\underline{P}'\pi_k} \\ \leq \frac{\epsilon}{2} \frac{\max_i B_{i,y}}{\mathbf{1}'B_y\underline{P}'\pi_k} &\leq \frac{\max_i B_{i,y}\, \epsilon/2}{\max\{\mathbf{1}'B_y\underline{P}'\underline{\pi}_k - \epsilon \max_i B_{iy}, \min_i B_{iy}\}}. \end{aligned} \tag{10}$$

The second last inequality follows from the construction of $\underline{P}$ satisfying (10.13b) (recall the variational norm is half the l_1 norm). The last inequality follows from Theorem 3.7.5(ii).

Consider the second normed term in the right hand side of (9). Applying Theorem 3.7.5(i) with notation $\pi = \underline{P}'\pi_k$ and $\pi^0 = \underline{P}'\underline{\pi}_k$ yields

$$\begin{aligned} \|T(\pi_k, y; \underline{P}) - T(\underline{\pi}_k, y; \underline{P})\|_{\text{TV}} &\leq \frac{\max_i B_{i,y} \|\underline{P}'\pi_k - \underline{P}'\underline{\pi}_k\|_{\text{TV}}}{\mathbf{1}'B_y\underline{P}'\underline{\pi}_k} \\ &\leq \frac{\max_i B_{i,y}\, \rho(\underline{P})\, \|\pi_k - \underline{\pi}_k\|_{\text{TV}}}{\mathbf{1}'B_y\underline{P}'\underline{\pi}_k} \end{aligned} \tag{11}$$

where the last inequality follows from the submultiplicative property of the Dobrushin coefficient. Substituting (10) and (11) into the right hand side of the triangle inequality (9) proves the result. □

3.3 Case Study. Reference Probability Method for Filtering

We describe here the so called *reference probability method* for deriving the un-normalized filtering recursion (3.21). The main idea is to start with the joint probability mass function of all observations and states until time k, namely, $p(x_{0:k}, y_{1:k})$. Since this joint density contains all the information we need, it is not surprising that by suitable marginalization

and integration, the filtering recursion and hence the conditional mean estimate can be computed.

Given the relatively straightforward derivations of the filtering recursions given in Chapter 3.3 of the book, the reader might wonder why we present yet another derivation. The reason is that in more complicated filtering problems, the reference probability method gives a systematic way of deriving filtering expressions. It is used extensively in [20] to derive filters in both discrete and continuous time. In continuous time, the reference probability measure is extremely useful – it yields the so called Duncan-Mortenson-Zakai equations for nonlinear filtering.

The Engineering Version

Suppose the state and observation processes $\{x_k\}$ and $\{y_k\}$ are in a probability space with probability measure $\mathbb{P}$. Since the state and observation noise processes are iid, under $\mathbb{P}$, we have the following factorization:

$$\begin{aligned} p(x_{0:k}, y_{1:k}) &= \prod_{n=1}^{k} p(y_n|x_n)\, p(x_n|x_{n-1})\, \pi_0(x_0) \qquad (12) \\ &\propto \prod_{n=1}^{k} p_v\left(D_n^{-1}(x_n)\,[y_n - C_n(x_k)]\right) p_w\left(\Gamma_{n-1}^{-1}(x_{n-1})\;[x_n - A_{n-1}(x_{n-1})]\right)\; \pi_0(x_0) \end{aligned}$$

Starting with $p(x_{0:k}, y_{1:k})$, the conditional expectation of any function $\phi(x_k)$ is

$$\mathbb{E}\{\phi(x_k)|y_{1:k}\} = \frac{\int \phi(x_k) p(x_{0:k}, y_{1:k}) dx_{0:k}}{\int p(x_{0:k}, y_{1:k}) dx_{0:k}} = \frac{\int_{\mathcal{X}} \phi(x_k)\left[\int p(x_{0:k}, y_{1:k}) dx_{0:k-1}\right] dx_k}{\int p(x_{0:k}, y_{1:k}) dx_{0:k}} \qquad (13)$$

The main idea then is to define the term within the square brackets in the numerator as the un-normalized density $q_k(x_k) = \int p(x_{0:k}, y_{1:k}) dx_{0:k-1}$. (Of course then $q_k(x_k) = p(x_k, y_{1:k})$). We now derive the recursion (3.21) for the un-normalized density q_k:

$$\begin{aligned} \int_{\mathcal{X}} \phi(x_k) q_k(x_k) dx_k &= \int_{\mathcal{X}} \phi(x_k) \int p(x_{0:k}, y_{1:k}) dx_{0:k-1} dx_k \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \phi(x_k) p(y_k|x_k) p(x_k|x_{k-1}) \left[\int p(x_{0:k-1}, y_{1:k-1}) dx_{0:k-2}\right] dx_{k-1} dx_k \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \phi(x_k) p(y_k|x_k) p(x_k|x_{k-1}) q_{k-1}(x_{k-1}) dx_{k-1} dx_k \end{aligned}$$

where the second equality follows from (12). Since the above holds for any test function ϕ, it follows that the integrands within the outside integral are equal, thereby yielding the un-normalized filtering recursion (3.21).

Interpretation as Change of Measure

We now interpret the above derivation as the engineering version of the reference probability method.[2] Define a new probability measure $\bar{\mathbb{P}}$ as having associated density

$$q(x_{0:k}, y_{1:k}) = \prod_{n=1}^{k} p_v(y_n)\, p_w(x_n)\, \pi_0(x_0).$$

The above equation is tantamount to saying that under this new measure $\bar{\mathbb{P}}$, the processes $\{x_k\}$ and $\{y_k\}$ are iid sequences with density functions p_w and p_v, respectively. $\bar{\mathbb{P}}$ will be called the *reference probability measure* - under this measure, due to the iid nature of $\{x_k\}$ and $\{y_k\}$, the filtering recursion can be derived conveniently, as we now describe.

Let $\bar{\mathbb{E}}$ denote expectation associated with measure $\bar{P}$, so that for any function $\phi(x_k)$, the conditional expectation is

$$\bar{\mathbb{E}}\{\phi(x_k)|y_{1:k}\} = \int \phi(x_k) q(x_{0:k}, y_{1:k}) dx_{0:k}$$

Obviously, to obtain the expectation $\mathbb{E}\{\phi(x_k)|y_{1:k}\}$ under the probability measure $\mathbb{P}$, it follows from (13) that

$$\begin{aligned}\mathbb{E}\{\phi(x_k)|y_{1:k}\} &= \frac{\int \phi(x_k) \Lambda_k q(x_{0:k}, y_{1:k}) dx_{0:k}}{\int \Lambda_k q(x_{0:k}, y_{1:k}) dx_{0:k}}, \quad \text{where } \Lambda_k = \frac{p(x_{0:k}, y_{1:k})}{q(x_{0:k}, y_{1:k})} \\ &= \frac{\bar{\mathbb{E}}\{\Lambda_k \phi(x_k)|y_{1:k}\}}{\bar{\mathbb{E}}\{\Lambda_k|y_{1:k}\}}\end{aligned} \tag{14}$$

The derivation then proceeds as follows.

$$\begin{aligned}&\int_{\mathcal{X}} q_k(x)\phi(x)dx = \bar{\mathbb{E}}\{\Lambda_k \phi(x_k)|\mathcal{Y}_k\} \qquad \text{(definition of } q_k\text{)} \\ &= \int \frac{p(x_{0:k}, y_{1:k})}{q(x_{0:k}, y_{1:k})} \phi(x_k) q(x_{0:k}, y_{1:k}) dx_{0:k} \\ &= \int \frac{p(x_{0:k-1}, y_{1:k-1})}{q(x_{0:k-1}, y_{1:k-1})} \frac{p(y_k|x_k)p(x_k|x_{k-1})}{\cancel{p_v(y_k)}\cancel{p_w(x_k)}} \phi(x_k) \cancel{p_v(y_k)}\cancel{p_w(x_k)} q(x_{0:k-1}, y_{1:k-1}) dx_{0:k} \\ &= \int \frac{p(x_{0:k-1}, y_{1:k-1})}{q(x_{0:k-1}, y_{1:k-1})} \left[\int_{\mathcal{X}} p(y_k|x_k)p(x_k|x_{k-1})\phi(x_k)dx_k\right] q(x_{0:k-1}, y_{1:k-1}) dx_{0:k-1} \\ &= \bar{\mathbb{E}}\{\Lambda_{k-1} \left[\int_{\mathcal{X}} p(y_k|x_k)p(x_k|x_{k-1})\phi(x_k)dx_k\right] |y_{1:k-1}\} \\ &= \int q_{k-1}(x_{k-1}) \left[\int_{\mathcal{X}} p(y_k|x_k)p(x_k|x_{k-1})\phi(x_k)dx_k\right] dx_{k-1}\end{aligned}$$

where the last equality follows from the definition of q in the first equality.

Since this holds for any test function $\phi(x)$, we have that the material inside the integral in the left and right hand side are equal. So

$$\pi_k(x_k) = p(y_k|x_k) \int_{\mathcal{X}} q_{k-1}(x_{k-1}) p(x_k|x_{k-1}) dx_{k-1}.$$

[2] In continuous time, the change of measure of a random process involves Girsanov's theorem, see [20]. Indeed the Zakai form of the continuous time filters in the appendix of the book can be derived in a fairly straightforward manner using Girsanov's theorem and basic Ito calculus.

Chapter 4

Algorithms for Maximum Likelihood Parameter Estimation

1. A standard drill exercise involves deriving the Cramér-Rao bound in terms of the Fisher information matrix; see wikipedia or any book in statistical signal processing for an elementary description.
2. **Minorization Maximization Algorithm (MM Algorithm).** The EM algorithm is a special case of the MM algorithm[1]; see [37] for a nice tutorial on MM algorithms. MM algorithms constitute a general purpose method for optimization and are not restricted just to maximum likelihood estimation.
   The main idea behind the MM algorithm is as follows: Suppose we wish to compute the maximizer θ^* of a function $\phi(\theta)$. The idea is to construct a minorizing function $g(\theta, \theta^{(m)})$ such that

$$\begin{aligned} g(\theta, \theta^{(m)}) &\leq \phi(\theta) \quad \text{for all } \theta \\ g(\theta^{(m)}, \theta^{(m)}) &= \phi(\theta^{(m)}). \end{aligned} \tag{15}$$

   That is, the minorizing function $g(\theta, \theta^{(m)})$ lies above $\phi(\theta)$ and is a tangent to it at the point $\theta^{(m)}$. Here

$$\theta^{(m)} = \underset{\theta}{\operatorname{argmax}}\, g(\theta^{(m-1)}, \theta)$$

   denotes the estimate of the maximizer at iteration m of MM algorithm.
   The property (15) implies that successive iterations of the MM algorithm yield

$$\phi(\theta^{(m+1)}) \geq \phi(\theta^{(m)}).$$

   In words, successive iterations of the MM algorithm yield increasing values of the objective function which is a very useful property for a general purpose numerical optimization algorithm. This is shown straightforwardly as follows:

$$\begin{aligned} \phi(\theta^{(m+1)}) &= \phi(\theta^{(m+1)}) - g(\theta^{(m+1)}, \theta^{(m)}) + g(\theta^{(m+1)}, \theta^{(m)}) \\ &\overset{a}{\geq} \phi(\theta^{(m+1)}) - g(\theta^{(m+1)}, \theta^{(m)}) + g(\theta^{(m)}, \theta^{(m)}) \\ &\overset{b}{\geq} \phi(\theta^{(m}) - \cancel{g(\theta^{(m)}, \theta^{(m)})} + \cancel{g(\theta^{(m)}, \theta^{(m)})} \end{aligned}$$

[1] MM can also be used equivalently to denote majorization minimization

Inequality (a) follows since $g(\theta^{(m+1)}, \theta^{(m)}) \geq g(\theta^{(m)}, \theta^{(m)})$ by definition since $\theta^{(m+1)} = \text{argmax}_\theta\, g(\theta, \theta^{(m)})$. Inequality (b) follows from (15).
The EM algorithm is a special case of the MM algorithm where

$$g(\theta, \theta^{(m)} = Q(\theta, \theta^{(m)}) - Q(\theta^{(m)}, \theta^{(m)}), \quad \phi(\theta) = \mathcal{L}_N(\theta) - \mathcal{L}_N(\theta^{(m)})$$

Here $\mathcal{L}_N(\theta) = \log p(y_{1:N}|\theta)$ is the log likelihood which we want to maximize to compute the MLE and $Q(\theta, \theta^{(m)})$ is the auxiliary log likelihood defined in (4.18) which is maximized in the M step of the EM algorithm.
Indeed the minorization property (15) was established for the EM algorithm in Lemma 4.3.2 on page 81 of the book by using Jensen's inequality.

3. **EM algorithm in more elegant (abstract) notation.** Let $\{P_\theta\,,\, \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathcal{F})$ all absolutely continuous with respect to a fixed probability measure P_0, and let $\mathcal{Y} \subset \mathcal{F}$. The likelihood function for computing an estimate of the parameter θ based on the information available in $\mathcal{Y}$ is

$$L(\theta) = \mathbb{E}_0[\frac{dP_\theta}{dP_0} \mid \mathcal{Y}]\ ,$$

and the MLE estimate is defined by

$$\widehat{\theta} \in \underset{\theta\in\Theta}{\text{argmax}}\, L(\theta)\ .$$

In general, the MLE is difficult to compute directly, and the EM algorithm provides an iterative approximation method :

Step 1. Set $p = 0$ and choose $\widehat{\theta}_0$.

Step 2. (E–step) Set $\theta' = \widehat{\theta}_p$ and compute $Q(\cdot, \theta')$, where

$$Q(\theta, \theta') = \mathbb{E}_{\theta'}[\log \frac{dP_\theta}{dP_{\theta'}} \mid \mathcal{Y}]\ .$$

Step 3. (M–step) Find

$$\widehat{\theta}_{p+1} \in \underset{\theta\in\Theta}{\text{argmax}}\, Q(\theta, \theta')\ .$$

Step 4. Replace p by $p + 1$ and repeat beginning with Step 2, until a stopping criterion is satisfied.

The sequence generated $\{\widehat{\theta}_p\,,\, p \geq 0\}$ gives non–decreasing values of the likelihood function : indeed, it follows from Jensen's inequality that

$$\log L(\widehat{\theta}_{p+1}) - \log L(\widehat{\theta}_p) \geq Q(\widehat{\theta}_{p+1}, \widehat{\theta}_p) \geq Q(\widehat{\theta}_p, \widehat{\theta}_p) = 0\ ,$$

with equality if and only if $\widehat{\theta}_{p+1} = \widehat{\theta}_p$.

4. **Forward-only EM algorithm for Linear Gaussian Model.** In §4.4 of the book, we described a forward-only EM algorithm for ML parameter estimation of the a HMM. Forward-only EM algorithms can also be constructed for maximum likelihood estimation of the parameters of a linear Gaussian state space model [21]. These involve computing filters for functionals of the state and use Kalman filter estimates.

5. **Sinusoid in HMM.** Consider a sinusoid with amplitude A and phase ϕ. It is observed as
$$y_k = x_k + A\,\sin(k/100 + \phi) + v_k$$
where v_k is an iid Gaussian noise process. Use the EM algorithm to estimate A, ϕ and the parameters of the Markov chain and noise variance.
6. In the forward-only EM algorithm of §4.4, the filters for the number of jumps involves $O(X^4)$ computations at each time while filters for the duration time involve $O(X^3)$ at each time. Is it possible to reduce the computational cost by approximating some of these estimates?
7. Using computer simulations, compare the methods of moments estimator for a HMM in §4.5 with the maximum likelihood estimator in terms of efficiency. That is generate several N point trajectories of an HMM with a fixed set of parameters, then compute the variance of the estimates. (Of course, instances where the MLE the algorithm converges to local maxima should be eliminated from the computation).
8. Non-asymptotic statistical inference using concentration of measure if very popular today. Assuming the likelihood is a Lipschitz function of the observations, and the observations are Markovian, show that the likelihood function concentrates to the Kullback Leibler function.
9. **EM Algorithm for State Estimation.** The EM algorithm was used in Chapter 4 as a numerical algorithm for maximum likelihood *parameter* estimation. It turns out that the EM algorithm can be used for *state* estimation, particularly for a jump Markov linear system (JMLS). Recall from §3.10 that a JMLS has model
$$\begin{aligned} z_{k+1} &= A(r_{k+1})\, z_k + \Gamma(r_{k+1})\, w_{k+1} + f(r_{k+1})\, u_{k+1} \\ y_k &= C(r_k)\, z_k + D(r_k)\, v_k + g(r_k) u_k. \end{aligned}$$
As described in §3.10, the optimal filter for a JMLS is computationally intractable. In comparison for a JMLS, the EM algorithm can be used to estimate the MAP (maximum aposteriori state estimate). system (assuming the parameters of the JMLS are known). Show how one can compute this MAP state estimate $\max_{z_{1:k}, r_{1:k}} P(y_{1:k}|z_{1:k}, r_{1:k})$ using the EM algorithm. In [53] is shown that the resulting EM algorithm involves the cross coupling of a Kalman and HMM smoother. A data augmentation algorithm in similar spirit appears in [19].
10. **Quadratic Convergence of Newton Algorithm.**
We start with some definitions: Given a sequence $\{\theta^{(n)}\}$ generated by an optimization algorithm, the *order* of convergence is p if
$$\beta = \limsup_{n\to\infty} \frac{\|\theta^{(n+1)} - \theta^*\|}{\|\theta^{(n)} - \theta^*\|^p} \text{ exists} \tag{16}$$
Also if $p = 1$ and $\beta < 1$, the sequence is said to converge linearly to θ^* with *convergence ratio (rate)* β. Moreover, the case $p = 1$ and $\beta = 0$ is referred to as superlinear convergence.
(a) Recall that the Newton Raphson algorithm computes the MLE iteratively as
$$\theta^{(n+1)} = \theta^{(n)} + \left(\nabla^2 \mathcal{L}_N(\theta^{(n)})\right)^{-1} \nabla \mathcal{L}_N(\theta^{(n)})$$

The Newton Raphson algorithm has quadratic order of convergence in the following sense. Suppose the log likelihood $\mathcal{L}_N(\theta)$ is twice continuous differentiable and that at a local maximum θ^*, the Hessian $\nabla^2_\theta \mathcal{L}_N$ is positive definite. Then if started sufficient close to θ^*, Newton Raphson converges to θ^* at a quadratic rate. that the model estimates satisfy $\theta^{(n)}$ satisfy

$$\|\theta^{(n+1)} - \theta^*\| \leq \beta \|\theta^{(n)} - \theta^*\|^2$$

for some constant β.
This is shown straightforwardly (see any optimization textbook) as follows:

$$\begin{aligned}\|\theta^{(n+1)} - \theta^*\| &= \|\theta^{(n)} - \theta^* + \left(\nabla^2 \mathcal{L}_N(\theta^{(n)})\right)^{-1} \nabla \mathcal{L}_N(\theta^{(n)})\| \\ &= \|\left(\nabla^2 \mathcal{L}_N(\theta^{(n)})\right)^{-1} \Big(\nabla \mathcal{L}_N(\theta^{(n)}) - \nabla \mathcal{L}_N(\theta^*) - \nabla^2 \mathcal{L}_N(\theta^{(n)})(\theta^{(n)} - \theta^*) \Big)\end{aligned} \tag{17}$$

For $\|\theta^{(n)} - \theta^*\| < \rho$, it is clear from a Taylor series expansion that

$$\|\nabla \mathcal{L}_N(\theta^*) - \nabla \mathcal{L}_N(\theta^{(n)}) - \nabla^2 \mathcal{L}_N(\theta^{(n)})(\theta^* - \theta^{(n)})\| \leq \beta_1 \|\theta^{(n)} - \theta^*\|^2$$

for some positive constant β_1. Also, $\|\left(\nabla^2 \mathcal{L}_N(\theta^{(n)})\right)^{-1}\| \leq \beta_2$.

(b) The convergence order and rate of the EM algorithm has been studied in great detail since the early 1980s; there are numerous papers in the area; see [93] and the references therein. The EM algorithm has linear convergence order, i.e., $p = 1$ in (16). Please see [57] and the references therein for examples where EM exhibits superlinear convergence.

Chapter 5

Multi-agent Sensing: Social Learning and Data Incest

5.1 Problems

1. A substantial amount of insight can be gleaned by actually simulating the setup (in Matlab) of the social learning filter for both the random variable and Markov chain case. Also simulate the risk-averse social learning filter discussed in §5.2 of the book.

2. **CVaR Social Learning Filter.** Consider the risk averse social learning discussed in §5.2. Suppose agents choose their actions a_k to minimize the CVaR risk averse measure

$$a_k = \underset{a\in\mathcal{A}}{\operatorname{argmin}}\{\min_{z\in\mathbb{R}} \{z + \frac{1}{\alpha}\mathbb{E}_{y_k}[\max\{(c(x_k, a) - z), 0\}]\}\}$$

   Here $\alpha \in (0, 1]$ reflects the degree of risk-aversion for the agent (the smaller α is, the more risk-averse the agent is). Show that the structural result Theorem 5.5.1 continues to hold for the CVaR social learning filter. Also show that for sufficiently risk-averse agents (namely, α close to zero), social learning ceases and agents always herd.

   Generalize the above result to any coherent risk measure.

3. The necessary and sufficient condition given in Theorem 5.5.3 for exact data incest removal requires that

$$A_n(j, n) = 0 \implies w_n(j) = 0, \text{ where } \quad w_n = T_{n-1}^{-1} t_n,$$

   and $T_n = \operatorname{sgn}((\mathbf{I}_n - A_n)^{-1}) = \begin{bmatrix} T_{n-1} & t_n \\ 0_{1\times n-1} & 1 \end{bmatrix}$ is the transitive closure matrix. Thus the condition depends purely on the adjacency matrix. Discuss what types of matrices satisfy the above condition.

4. Theorem 5.5.3 also applies to data incest where the prior and likelihood are Gaussian. The posterior is then evaluated by a Kalman filter. Compare the performance of exact data incest removal with the covariance intersection algorithm in [17] which assumes no knowledge of the correlation structure (and hence of the network).

5. Consensus algorithms [72] have been extremely popular during the last decade and there are numerous papers in the area. They are non-Bayesian and seek to compute, for example, the average over measurements observed at a number of nodes in a graph. It is worthwhile comparing the performance of the optimal Bayesian incest removal algorithms with consensus algorithms.

6. The data incest removal algorithm in §5.4 of the book arises assumes that agents do not send additional information apart from their incest free estimates. Suppose agents are allowed to send a fixed number of labels of previous agents from whom they have received information. What is the minimum about of additional labels the agents need to send in order to completely remove data incest.

7. Quantify the bias introduced by data incest as a function of the adjacency matrix.

8. Prospect theory (pioneered by the psychologist Kahneman [38] who won the 2003 Nobel prize in economics) is a behavioral economic theory that seek to model how humans make decisions amongst probabilistic alternatives. (It is an alternative to expected utility theory considered in the social learning models of this chapter.) The main features are:
   (a) Preference is an S-shaped curve with reference point $x = 0$
   (b) The investor maximizes the expected value $V(x)$ where V is a preference and x is the change in wealth.
   (c) Decision maker employ decision weight $w(p)$ rather than objective probability p, where the weight function $w(F)$ has a reverse S shape where F is the cumulative probability.

   Construct a social learning filter where the utility function satisfies the above assumptions. Under what conditions do information cascades occur?

9. **Rational Inattention.** Another powerful way for modeling the behavior of (human) decision makers is in terms of rational inattention. See the seminal work of Sims [83] where essentially the ability of the human to absorb information is modeled via the information theoretic capacity of a communication channel.

   The one step version of the rational inattention optimization problem is formulated in [60] and is as follows: Let $x \in \mathcal{X} = \{1, 2, \ldots, X\}$ denote the unknown state of nature with prior $\pi_0(i) = \mathbb{P}(x = i)$. Choosing action $a \in \{1, \ldots, A\}$ results in reward $r(x, a)$ to the decision maker.

   A decision maker aims to optimize both its observation distribution (how it views the world) and its decision to maximize its reward. Denote the observation control as u and decision as a. Then a utility maximizer would seek to compute

$$\begin{aligned} V(\pi_0) = \max_u \mathbb{E}_y\{\max_a\{\mathbb{E}_u\{r_a' x|y\}\} &= \max_u \mathbb{E}_y\{\max_a r_a' T(\pi_0, y, u)\} \\ &= \max_u \sum_y \max_a r_a' \, T(\pi_0, y, u) \, \sigma(\pi_0, y, u) \\ &= \max_u \sum_y \max_a r_a' B_y(u) \pi_0 \end{aligned} \tag{18}$$

where T denotes Bayes' rule:

$$T(\pi, y, u) = \frac{B_y(u)\,\pi}{\sigma(\pi, y, u)}, \quad \sigma(\pi, y, u) = 1' B_y(u)\pi$$

and $B_y(u) = \text{diag}(\mathbb{P}(y|x = 1, u), \ldots, \mathbb{P}(y|x = X, u)$.

In a rational inattention model, an additional term is included for the cost of information. So a rationally inattentive utility maximizer seeks to compute

$$V(\pi) = \max_u \mathbb{E}_y\{\max_a\{\mathbb{E}_u\{r_a' x|y\}\} - \lambda\Big[H(\pi_0) - \sum_y H(T(\pi_0, y, u))\sigma(\pi_0, y, u)\Big] \quad (19)$$

where $H(\pi) = -\sum_{i=1}^X \pi(i)\log\pi(i)$ denotes the entropy and λ is a non-negative scaling constant.

10. There are several real life experiments that seek to understand how humans interact in decision making. See for example [7] and [46]. In [7], four models are considered. How can these models be linked to social learning?

5.2 Social Learning with limited memory

Here we briefly describe a variation of the vanilla social learning protocol (multi-agent system for estimating the state of a random variable). In order to mitigate herding, assume that agents randomly sample only a fixed number of previous actions. The aim below is to describe the resulting setup; see [84] for a detailed discussion.

Let the variable $\theta \in \{1, 2\}$ denote the states. Let $a \in \{1, 2\}$ denote the action alphabet and $y \in \{1, 2\}$ denote the observation alphabet. In this model of social learning with limited memory, it is assumed that each agent (at time $t \geq N + 1$) observes *only* N randomly selected actions from the history $h_t = \{a_1, a_2, \ldots, a_{t-1}\}$. In the periods $t \leq N$, each agent acts according to his private belief. This phase is termed as the seed phase in the model.

Let $z_t^{(1)}$ denote the number of times action 1 is chosen until time t, i.e,

$$z_t^{(1)} = \sum_{j=1}^t I(a_j = 1).$$

Let $\hat{z}_t^{(1)}$ denote the number of times action 1 is chosen in a sample of N randomly observed actions in the past, i.e, $\hat{z}_t^{(1)} = \sum_{j=1}^N I(a_j = 1)$.

The social learning protocol with limited memory is as follows:

1.) *Private belief update*: Agent t makes two observations at each instant $t(> N)$. These observations correspond to a noisy private signal y_t and a sample of N past actions from the history h_t sampled uniformly randomly. Let B_{y_t} and $D_{z_t=k}$ denote the probability of observing y_t and $(\hat{z}_t^{(1)} = k)$ respectively. The private belief is updated as follows.

For each draw from the past, the probability of observing action 1 is $z_t^{(1)}/(t-1)$. So the probability that at time t, action 1 occurs k times in a random sample of N observed actions is

$$\mathbb{P}(\hat{z}_t^{(1)} = k|z_t^{(1)}) = \frac{N!}{(N-k)!k!}\left(\frac{z_t^{(1)}}{t}\right)^k\left(1-\frac{z_t^{(1)}}{t}\right)^{N-k}$$

Therefore, the number of times action '1' is chosen in the sample, $\hat{z}_t^{(1)}$, has a distribution that depends on θ according to:

$$\mathbb{P}(\hat{z}_t^{(1)} = k|\theta) = \sum_{z_t^{(1)}=1}^{t} \mathbb{P}(\hat{z}_t^{(1)} = k|z_t^{(1)})\mathbb{P}(z_t^{(1)}|\theta)$$

After obtaining a private noisy signal y_t, and having observed $(\hat{z}_t^{(1)} = k)$, the belief $\pi_t = [\pi_t(1), \pi_t(2)]'$ where $\pi_t(i) = \mathbb{P}(\theta = i|\hat{z}_t^{(1)}, y_t)$ is updated by agent t as:

$$\pi_t = \frac{B_{y_t}D_{z_t=k}}{\mathbf{1}'B_{y_t}D_{z_t=k}}.$$

Here B and D are the observation likelihoods of y_t and $\hat{z}_t^{(1)}$ given the state:

$$B_{y_t} = \text{diag}(\mathbb{P}(y_t|\theta = i), i \in \{1,2\}), \quad D_{z_t=k} = \begin{bmatrix}\mathbb{P}(\hat{z}_t^{(1)} = k|\theta = 1)\\ \mathbb{P}(\hat{z}_t^{(1)} = k|\theta = 2)\end{bmatrix}$$

2.) *Agent's decision*: With the private belief π_t, the agent t makes a decision as:

$$a_t = \underset{a\in\{1,2\}}{\text{argmin}}\ c_a^T\pi_t$$

where c_a denotes the cost vector.

3.) *Action distribution*: The distribution of actions $\mathbb{P}(z_t^{(1)}|\theta)$ in the two states $\theta = 1, 2$ is assumed to be common knowledge at time t. It is updated after the decision of agent t as follows.

The probability of $(a_t = 1)$ in period t depends on the actual number of '1' actions $z_t^{(1)}$ and on the state according to:

$$\mathbb{P}(a_t = 1|z_t^{(1)} = n, \theta) = \sum_{k=0}^{N}\sum_{i=1}^{2}\mathbb{P}(a_t = 1|y = i, \hat{z}_t^{(1)} = k, z_t^{(1)} = n, \theta)\,\mathbb{P}(y = i|\theta)\,\mathbb{P}(\hat{z}_t^{(1)} = k|z_t^{(1)} = n)$$

where,

$$\mathbb{P}(a_t = 1|y = i, \hat{z}_t^{(1)} = k, z_t^{(1)} = n, \theta) = \begin{cases} 1 & \text{if } c_1^T B_{y=i}D_{z_t=k} < c_2^T B_{y=i}D_{z_t=k};\\ 0 & \text{otherwise.}\end{cases}$$

After agent t takes an action, the distribution is updated as:

$$\mathbb{P}(z_{t+1}^{(1)} = n|\theta) = \mathbb{P}(z_t^{(1)} = n|\theta)(1-\mathbb{P}(a_t = 1|z_t^{(1)} = n, \theta))+\mathbb{P}(z_t^{(1)} = n-1|\theta)\mathbb{P}(a_t = 1|z_t^{(1)} = n, \theta) \tag{20}$$

According to equation (20), the sufficient statistic $\mathbb{P}(z_t^{(1)}|\theta)$ is growing with time t. It is noted that this has $(t-2)$ numbers at time t and hence grows with time. $\mathbb{P}(z_{t+1}^{(1)} = n|\theta)$ in equation (20) is used to compute $D_{z_{t+1}}$.

With the above model, consider the following questions:

1. Show that there is asymptotic herding when $N = 1$.
2. Show that for $N = 2A$, reduction in the historical information will improve social learning. Also, comment on whether there is herding when $N = 2$.
3. Show that as N increases, the convergence to the true state is slower. Hint: Even though more observations are chosen, greater weight on the history precludes the use of private information.

Chapter 6

Fully Observed Markov Decision Processes

6.1 Problems

1. The following nice example from [51] gives a useful motivation for feedback control in stochastic systems. It shows that for stochastic systems, using feedback control can result in behavior that cannot be obtained by an open loop system.
   (a) First, recall from undergraduate control courses that for a deterministic linear time invariant system with forward transfer function $G(z^{-1})$ and negative feedback $H(z^{-1})$, the equivalent transfer function is $\frac{G(z^{-1})}{1+G(z^{-1})H(z^{-1})}$. So an open loop system with this equivalent transfer function is identical to a feedback system.
   (b) More generally, consider the deterministic system

   $$x_{k+1} = \phi(x_k, u_k), y_k = \psi(x_k, u_k)$$

   Suppose the actions are given by a policy of the form

   $$u_k = \mu(x_{0:k}, y_{1:k})$$

   Then clearly, the open loop system,

   $$x_{k+1} = \phi(x_k, \mu(x_{0:k}, y_{1:k})), y_k = \psi(x_k, \mu(x_{0:k}, y_{1:k}))$$

   generates the same state and observation sequences.
   So for a deterministic system (with fully specified model), open and closed loop behavior are identical.
   (c) Now consider a fully observed stochastic system with feedback:

   $$\begin{aligned} x_{k+1} &= x_k + u_k + w_k, \\ u_k &= -x_k \end{aligned} \tag{21}$$

   where w_k is iid with zero mean and variance σ^2 (as usual we assume x_0 is independent of $\{w_k\}$.) Then $x_{k+1} = w_k$ and so $u_k = -w_{k-1}$ for $k = 1, 2, \ldots$. Therefore $\mathbb{E}\{x_k\} = 0$ and $\text{Var}\{x_k^2\} = \sigma^2$.

(d) Finally, consider an open loop stochastic system where u_k is a deterministic sequence:

$$x_{k+1} = x_k + u_k + w_k$$

Then $\mathbb{E}\{x_k\} = \mathbb{E}\{x_0\} + \sum_{n=0}^{k-1} u_k$ and $\text{Var}\{x_k^2\} = \mathbb{E}\{x_0^2\} + k\sigma^2$. Clearly, it is impossible to construct a deterministic input sequence that yields a zero mean state with variance σ^2.

2. **Trading of call options.** An investor buys a call option at a price p. He has N days to exercise this option. If the investor exercises the option when the stock price is x, he gets $x - p$ dollars. The investor can also decide not the exercise the option at all.
Assume the stock price evolves as $x_k = x_0 + \sum_{n=1}^{k} w_n$ where $\{w_n\}$ is in iid process. Let τ denote the day the investor decides to exercise the option. Determine the optimal investment strategy to maximize

$$\mathbb{E}\{(x_\tau - p)I(\tau \leq T)\}.$$

This is an example of a fully observed stopping time problem. Chapter 12 considers more general stopping time POMDPs.
Note: Define $s_k \in \{0, 1\}$ where $s_k = 0$ means that the option has not been exercised until time k. $s_k = 1$ means that the option has been exercised before time k. Define the state $z_k = (x_k, s_k)$.
Denote the action $u_k = 1$ to exercise option and $u_k = 0$ means do not exercise option. Then the dynamics are

$$s_{k+1} = \max\{s_k, u_k\}, \quad x_{k+1} = x_k + w_k$$

The reward at each time k is $r(z_k, u_k, k) = (1 - s_k)u_k(x_k - p)$ and the problem can be formulated as

$$\max_{\mu} \mathbb{E}\{\sum_{k=1}^{N} r(z_k, u_k, k)\}$$

3. Discounted cost problems can also be motivated as stopping time problems (with a random termination time). Suppose at each time k, the MDP can terminate with probability $1 - \rho$ or continue with probability ρ. Let τ denote the random variable for the termination time. Consider the undiscounted cost MDP

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\mu}}\left\{\sum_{k=0}^{\tau} c(x_k, u_k) \mid x_0 = i\right\} &= \mathbb{E}_{\boldsymbol{\mu}}\left\{\sum_{k=0}^{\infty} I(k \leq \tau)\, c(x_k, u_k) \mid x_0 = i\right\} \\ &= \mathbb{E}_{\boldsymbol{\mu}}\left\{\sum_{k=0}^{\infty} \rho^k c(x_k, u_k) \mid x_0 = i\right\}.\end{aligned}$$

The last equality follows since $\mathbb{P}(k \leq \tau) = \rho^k$.

4. We discussed risk averse utilities and dynamic risk measures briefly in §8.6. Also §6.7.3 discussed revealed preferences for constructing a utility function from a dataset. Given a utility function $U(x)$, a widely used measure for the degree of risk aversion is the Arrow-Pratt risk aversion coefficient which is defined as

$$a(x) = -\frac{d^2U/dx^2}{dU/dx}.$$

This is often termed as an absolute risk aversion measure, while $x\,a(x)$ is termed a relative risk aversion measure. Can this risk averse coefficient be used for mean semi-deviation risk, conditional value at risk(CVaR) and exponential risk?

5. A classical result involving utility functions is the following [35, pp.42]: A rational decision maker who compares random variables only according to their means and variances must have preferences consistent with a quadratic utility function. Prove this result.

6.2 Case study. Non-cooperative Discounted Cost Markov games

§6.4 of the book dealt with infinite horizon discounted MDPs. Below we introduce briefly some elementary ideas in non-cooperative infinite horizon discounted Markov games. There are several excellent books in the area [43, 8].

Markov games can be viewed as a multi-agent decentralized extension of MDPs. They arise in a variety of applications including dynamic spectrum allocation, financial models and smart grids. Our aim here is to consider some simple cases where the Nash equilibrium can be obtained by solving a linear programming problem.[1]

Consider the following infinite horizon discounted cost two-payer Markovian game. There are two decision makers (players) indexed by $l = 1, 2$.

- Let $u_k^{(1)} \in \mathcal{U}$ and $u_k^{(2)} \in \mathcal{U}$ denote the action of player 1 and player 2, respectively, at time k. For convenience we assume the same action space for both players.
- The cost incurred by player $l \in \{1, 2\}$ for state x, actions $u^{(1)}, u^{(2)}$ is $c_l(x, u^{(1)}, u^{(2)})$.
- The transition probabilities of the Markov process x depends on the actions of both players:
$$P_{ij}(u^{(1)}, u^{(2)}) = \mathbb{P}(x_{k+1} = j | x_k = i, u_k^{(1)} = u^{(1)}, u_k^{(2)} = u^{(2)})$$
- Define the policies for the stationary (randomized) Markovian policies for two players as $\mu^{(1)}$, $\mu^{(2)}$, respectively. So $u_k^{(1)}$ is chosen from probability distribution $\mu^{(1)}(x_k)$ and $u_k^{(2)}$ is chosen from probability distribution $\mu^{(2)}(x_k)$. For convenience denote the class of stationary Markovian policies as $\boldsymbol{\mu}_S$.
- The cumulative cost incurred by each player $l \in \{1, 2\}$ is

$$J^{(l)}_{\mu^{(1)},\mu^{(2)}}(x) = \mathbb{E}\{ \sum_{k=0}^{\infty} \rho^k c_l(x_k, u_k^{(1)}, u_k^{(2)}) | x_0 = x\} \tag{22}$$

where as usual $\rho \in (0, 1)$ is the discount factor.

The non-cooperative assumption in game theory is that the players are interested in minimizing their individual cumulative costs only; they do not collude.

[1] The reader should be cautious with decentralized stochastic control. The famous Witsenhausen's counterexample formulated in the 1960s shows that even a deceptively simple toy problem in decentralized stochastic control can be very difficult to solve, see `https://en.wikipedia.org/wiki/Witsenhausen\%27s_counterexample`

6.2.1 Nash equilibrium of general sum Markov game

Assume that each player has complete knowledge of the other player's cost function. Then the policies $\mu^{(1)^*}, \mu^{(2)^*}$ of the non-cooperative infinite horizon Markov game constitute a Nash equilibrium if

$$\begin{aligned} J^{(1)}_{\mu^{(1)^*},\mu^{(2)^*}}(x) &\leq J^{(1)}_{\mu^{(1)},\mu^{(2)^*}}(x), \quad \text{for all } \mu^{(1)} \in \boldsymbol{\mu}_S \\ J^{(1)}_{\mu^{(1)^*},\mu^{(2)^*}}(x) &\leq J^{(1)}_{\mu^{(1)^*},\mu^{(2)}}(x), \quad \text{for all } \mu^{(2)} \in \boldsymbol{\mu}_S. \end{aligned} \tag{23}$$

This means that unilateral deviations from $\mu^{(1)^*}, \mu^{(2)^*}$ result in either player being worse off (incurring a larger cost). Since in a non-cooperative game collusion is not allowed, there is no rational reason for players to deviate from the Nash equilibrium (23).

In game theory, two important issues are:

1. *Does a Nash equilibrium exist?* For the above discounted cost game with finite action and state space, the answer is "yes".

   Theorem 1. *A discounted Markov game has at least one Nash equilibrium within the class of Markovian stationary (randomized) policies.*

   The proof is in [26] and involves Kakutani's fixed point theorem.[2]
2. *How can the Nash equilibria be computed?* Define the randomized policy of player 1 (corresponding to $\mu^{(1)}$) and player 2 (corresponding to $\mu^{(2)}$) as

   $$p(i,u^{(1)}) = \mathbb{P}(u^{(1)}_k = u^{(1)} | x_k = i), \quad q(i,u^{(2)}) = \mathbb{P}(u^{(2)}_k = u^{(2)} | x_k = i)$$

   Then for an infinite horizon discounted cost Markov game, the Nash equilibria (p^*, q^*) are global optima of the following non-convex optimization problem:

$$\begin{aligned} &\text{Compute } \max \sum_{l=1}^{2}\sum_{i=1}^{X} \alpha_i \Big(\underline{V}^{(l)}(i) - \sum_{u^{(1)},u^{(2)}} c_l(i,u^{(1)},u^{(2)}) p(i,u^{(1)}) q(i,u^{(2)}) \\ &- \rho \sum_{j\in\mathcal{X}} \sum_{u^{(1)},u^{(2)}} P_{ij}(u^{(1)},u^{(2)}) p(i,u^{(1)}) q(i,u^{(2)}) \underline{V}^{(l)}(j) \Big) \\ &\text{with respect to } (\underline{V}^{(1)}, \underline{V}^{(2)}, p, q) \\ \text{subject to } &\underline{V}^{(1)}(i) \leq \sum_{u^{(2)}} c(i,u^{(1)},u^{(2)}) q(i,u^{(2)}) + \rho \sum_{j\in\mathcal{X}} \sum_{u^{(2)}} P_{ij}(u^{(1)},u^{(2)}) q(i,u^{(2)}) \underline{V}^{(1)}(j), \\ &\underline{V}^{(2)}(i) \leq \sum_{u^{(2)}} c(i,u^{(1)},u^{(2)}) p(i,u^{(1)}) + \rho \sum_{j\in\mathcal{X}} \sum_{u^{(1)}} P_{ij}(u^{(1)},u^{(2)}) p(i,u^{(1)}) \underline{V}^{(2)}(j), \\ &q(i,u^{(2)}) \geq 0, \quad \sum_{u^{(2)}} q(i,u^{(2)}) = 1, \quad i = 1,2,\ldots,X,\ u^{(2)} = 1,\ldots,U \\ &p(i,u^{(1)}) \geq 0, \quad \sum_{u^{(1)}} p(i,u^{(1)}) = 1, \quad i = 1,2,\ldots,X,\ u^{(1)} = 1,\ldots,U. \end{aligned} \tag{24}$$

[2] Existence proofs for equilibria involve using either Kakutani's fixed point theorem (which generalizes Brouwer's fixed point theorem to set valued correspondences) or Tarski's fixed point theorem (which applies to supermodular games). Please see [59] for a nice intuitive visual illustration of these fixed point theorems.

In general, solving the non-convex optimization problem (24) is difficult; there can be multiple global optima (each corresponding to a Nash equilibrium) and multiple local optima. In fact there is a fascinating property that if all the parameters (transition probabilities, costs) are rational numbers, the Nash equilibrium policy can involve irrational numbers. This points to the fact that in general one can only approximately compute the Nash equilibrium.

Proof. First write (24) in more abstract but intuitive notation in terms of the randomized policies p, q as

$$\begin{aligned} & \max \sum_{l=1}^{2} \alpha' \Big(\underline{V}^{(l)} - c_l(p,q) - \rho P(p,q) \underline{V}^{(l)} \Big) \\ \text{subject to } & \underline{V}^{(1)} \le c_1(u^{(1)}, q) + \rho P(u^{(1)}, q) \underline{V}^{(1)}, \quad u^{(1)} = 1, \dots, U \\ & \underline{V}^{(2)} \le c_2(p, u^{(2)}) + \rho P(p, u^{(2)}) \underline{V}^{(2)}, \quad u^{(2)} = 1, \dots, U \\ & p, q \text{ valid pmfs} \end{aligned} \tag{25}$$

It is clear from the constraints that the objective function is always ≤ 0. In fact the maximum is attained when the objective function is zero, in which case the constraints hold with equality. When the constraints hold at equality, they satisfy

$$V_*^{(l)} = \big(I - \rho P(p^*, q^*)\big)^{-1} c_l(p^*, q^*), \quad l = 1, 2.$$

This serves as definition of $V_*^{(l)}$ and is equivalent to saying[3] that $V_*^{(l)}$ is the infinite horizon cost attained by the policies (p^*, q^*). That is,

$$V_*^{(l)} = c_l(p^*, q^*) + \rho P(p^*, q^*) V_*^{(l)} \implies J^{(l)}_{p^*, q^*}(x) = V_*^{(l)}. \tag{26}$$

Also setting $\underline{V}^{(l)} = V_*^{(l)}$, the constraints in (25) satisfy

$$V_*^{(1)} \le c_1(p, q^*) + \rho P(p, q^*) V_*^{(1)}, \quad V_*^{(2)} \le c_2(p^*, q) + \rho P(p^*, q) V_*^{(2)}$$

implying that

$$J^{(1)}_{p, q^*}(x) \ge V_*^{(1)}, \quad J^{(2)}_{p^*, q}(x) \ge V_*^{(2)}. \tag{27}$$

(26) and (27) imply that (p^*, q^*) constitute a Nash equilibrium. □

Remark. The reader should compare the above proof with the linear programming formulation for a discounted cost MDP. In that derivation we started with a similar constraint

$$\underline{V} \le c_1(u^{(1)}) + \rho P(u^{(1)}) \underline{V}. \tag{28}$$

This implies that $\underline{V} < V$ where V denotes the unique value function of Bellman's equation. Therefore the objective was to find $\max \alpha' \underline{V}$ subject to (28). So in MDP case we obtain a linear program. In the dynamic game case, in general, there is no value function to clamp (upper bound) $\underline{V}$.

[3] This holds since from (22), $J^{(l)}_{p^*, q^*}(x) = c_l(p^*, q^*) + \rho P c_l(p^*, q^*) + \rho^2 P^2 c_l(p^*, q^*) + \cdots +$. Indeed a similar expression holds for discounted cost MDPs.

6.2.2 Zero-sum discounted Markov game

With the above brief introduction, the main aim below is to give special cases of *zero-sum* Markov games where the Nash equilibrium can be computed via linear programming. (Recall §6.4.2 of the book shows how a discounted cost MDP can be solved via linear programming.)

A discounted Markovian game is said to be zero sum[4] if

$$c_1(x, u^{(1)}, u^{(2)}) + c_2(x, u^{(1)}, u^{(2)}) = 0.$$

That is,

$$c(x, u^{(1)}, u^{(2)}) \stackrel{\text{defn}}{=} c_1(x, u^{(1)}, u^{(2)}) = -c_2(x, u^{(1)}, u^{(2)}).$$

For a zero sum game, the Nash equilibrium (23) becomes a saddle point:

$$J_{\mu^{(1)*},\mu^{(2)}}(x) \leq J_{\mu^{(1)*},\mu^{(2)*}}(x) \leq J_{\mu^{(1)},\mu^{(2)*}}(x),$$

that is, it is a minimum in the $\mu^{(1)}$ direction and a maximum in the $\mu^{(2)}$ direction.

A well known result from the 1950s due to Shapley is:

Theorem 2 (Shapley)**.** *A zero sum infinite horizon discounted cost Markov game has a unique value function, even though there could be multiple Nash equilibria (saddle points). Thus all the Nash equilibria are equivalent.*

The value function of the zero-sum game is

$$J_{\mu^{(1)*},\mu^{(2)*}}(i) = V(i)$$

where V satisfies an equation that resembles dynamic programming:

$$V(i) = \text{val}\left[(1-\rho)c(i, u^{(1)}, u^{(2)}) + \rho \sum_j P_{ij}(u^{(1)}, u^{(2)})V(j)\right]_{u^{(1)},u^{(2)}} \tag{29}$$

Here $\text{val}[M]_{u^{(1)},u^{(2)}}$ denotes the value of the matrix[5] game with elements $M(u^{(1)}, u^{(2)})$. Even though for a specific vector V, the val$[\cdot]$ in the right hand side of (29) can be evaluated by solving an LP, it is not useful for the Markov zero sum game, since we have a functional equation in the variable V. So solving a zero sum Markov game is difficult in general.

[4] A constant sum game $c_1(x, u^{(1)}, u^{(2)}) + c_2(x, u^{(1)}, u^{(2)}) = K$ for constant K is equivalent to a zero sum game. Define $\bar{c}_l(x, u^{(1)}, u^{(2)}) = c_l(x, u^{(1)}, u^{(2)}) + K/2$, $l = 1, 2$, resulting in a zero sum game in terms of $\bar{c}_l$.

[5] A zero sum matrix game is of the form: Given a $m \times n$ matrix M, determine the Nash equilibrium

$$(x^*, y^*) = \underset{x}{\text{argmax}}\, \underset{y}{\text{argmin}}\, y'Mx, \qquad \text{where } x, y \text{ are probability vectors}$$

The value of this matrix game is $\text{val}[M] = y^{*\prime}Mx^*$ and is computed as the solution of a linear programming (LP) problem as follows: Clearly $\max_x \min_y y'Mx = \max_x \min_i e_i'Mx$ where e_i, $i = 1, 2, \ldots, m$ denotes the unit m-dimensional vector with 1 in the i-th position. This follows since a linear function is minimized at its extreme points. So the minimization over continuum has been reduced to one over a finite set. Denoting $z = \min_i e_i'Mx$, the value of the game is the solution of the following LP:

$$\text{val}[M] = \begin{cases} \text{Compute } \max z \\ z < e_i'Mx, \quad i = 1, 2, \ldots, m, \\ \mathbf{1}'x = 1, \quad x_j \geq 0, j = 1, 2 \ldots, n \end{cases} \tag{30}$$

Nash Equilibrium as a Non-convex Bilinear Program

To give more insight, as we did in the discounted cost MDP case, let us formulate computing the Nash equilibrium (saddle point) of the zero sum Markov game as an optimization problem. In the MDP case we obtained a LP; for the Markov game (as shown below) we obtain a non-convex bilinear optimization problem.

Define the randomized policy of player 1 (minimizer) and player 2 (maximizer) as

$$p(i,u^{(1)}) = \mathbb{P}(u_k^{(1)} = u^{(1)} | x_k = i), \quad q(i,u^{(2)}) = \mathbb{P}(u_k^{(2)} = u^{(2)} | x_k = i)$$

In complete analogy to the discounted MDP case in (6.23), player 2 optimal strategy q^* is the solution of the bilinear program

$$\begin{aligned} &\max \sum_i \alpha_i \underline{V}(i) \text{ with respect to } (\underline{V}, q) \\ &\text{subject to } \underline{V}(i) \le \sum_{u^{(2)}} c(i,u^{(1)},u^{(2)}) q(i,u^{(2)}) + \rho \sum_{j\in\mathcal{X}} \sum_{u^{(2)}} P_{ij}(u^{(1)},u^{(2)}) q(i,u^{(2)}) \underline{V}(j), \\ &q(i,u^{(2)}) \ge 0, \quad \sum_{u^{(2)}} q(i,u^{(2)}) = 1, \quad i = 1,2,\ldots,X,\ u^{(2)} = 1,2,\ldots,U. \end{aligned} \tag{31}$$

By symmetry, player 1 optimal strategy p^* is the solution of the bilinear program

$$\begin{aligned} &\min \sum_i \alpha_i \underline{V}(i) \text{ with respect to } (\underline{V}, p) \\ &\text{subject to } \underline{V}(i) \ge \sum_{u^{(2)}} c(i,u^{(1)},u^{(2)}) p(i,u^{(1)}) + \rho \sum_{j\in\mathcal{X}} \sum_{u^{(1)}} P_{ij}(u^{(1)},u^{(2)}) p(i,u^{(1)}) \underline{V}(j), \\ &p(i,u^{(1)}) \ge 0, \quad \sum_{u^{(1)}} p(i,u^{(1)}) = 1, \quad i = 1,2,\ldots,X,\ u^{(1)} = 1,2,\ldots,U. \end{aligned} \tag{32}$$

The key difference between the above discounted Markov game problem and the discounted MDP (6.23) is that the above equations are no longer LPs. Indeed the constraints are *bilinear* in $(\underline{V}, q)$ and $(\underline{V}, p)$. So the constraint set for a zero-sum Markov game is non-convex. Despite (31) and (32) being nonconvex, in light of Shapley's theorem all local minima are global minima.

Finally (31) and (32) can be combined into a single optimization problem. To summarize, the (randomized) Nash equilibrium p^*, q^* of a zero-sum Markov game is the solution of the following bilinear (noconvex) optimization problem:

$$\begin{aligned} &\max \sum_i \alpha_i (\underline{V}^{(1)}(i) - \underline{V}^{(2)}(i)) \text{ with respect to } (\underline{V}^{(1)}, \underline{V}^{(2)}, p, q) \\ &\text{subject to } \underline{V}^{(1)}(i) \le \sum_{u^{(2)}} c(i,u^{(1)},u^{(2)}) q(i,u^{(2)}) + \rho \sum_{j\in\mathcal{X}} \sum_{u^{(2)}} P_{ij}(u^{(1)},u^{(2)}) q(i,u^{(2)}) \underline{V}^{(1)}(j), \\ &\underline{V}^{(2)}(i) \ge \sum_{u^{(2)}} c(i,u^{(1)},u^{(2)}) p(i,u^{(1)}) + \rho \sum_{j\in\mathcal{X}} \sum_{u^{(1)}} P_{ij}(u^{(1)},u^{(2)}) p(i,u^{(1)}) \underline{V}^{(2)}(j), \\ &q(i,u^{(2)}) \ge 0, \quad \sum_{u^{(2)}} q(i,u^{(2)}) = 1, \quad i = 1,2,\ldots,X,\ u^{(2)} = 1,2,\ldots,U \\ &p(i,u^{(1)}) \ge 0, \quad \sum_{u^{(1)}} p(i,u^{(1)}) = 1, \quad i = 1,2,\ldots,X,\ u^{(1)} = 1,2,\ldots,U. \end{aligned} \tag{33}$$

Special cases where computing Nash Equilibrium is an LP

We now give two special examples of zero-sum Markov games that can be solved as a linear programming problem (LP); single controller games and switched controller games. In both cases the bilinear terms in (33) vanish and the computing the Nash equilibrium reduces to solving linear programs.

6.2.3 Example 1. Single Controller zero-sum Markov Game

In a single controller Markov game, the transition probabilities are controlled by one player only; we assume that this is player 1. So

$$P_{ij}(u^{(1)}, u^{(2)}) = P_{ij}(u^{(1)}) = \mathbb{P}(x_{k+1} = j | x_k = i, u_k^{(1)} = u^{(1)})$$

Due to this assumption, the bilinear constraint in (31) becomes *linear*, namely

$$\underline{V}(i) \leq \sum_{u^{(2)}} c(i, u^{(1)}, u^{(2)}) q(i, u^{(2)}) + \rho \sum_{j \in \mathcal{X}} P_{ij}(u^{(1)}) \underline{V}(j)$$

since $\sum_{u^{(2)}} q(i, u^{(2)}) = 1$. Therefore (31) is now an LP which can be solved for q^*, namely:

$$\begin{aligned} &\max_{\underline{V}} \sum_i \alpha_i \underline{V}(i) \ \text{ with respect to } (\underline{V}, q) \\ &\text{subject to } \underline{V}(i) \leq \sum_{u^{(2)}} c(i, u^{(1)}, u^{(2)}) q(i, u^{(2)}) + \rho \sum_{j \in \mathcal{X}} P_{ij}(u^{(1)}) \underline{V}(j), \\ &q(i, u^{(2)}) \geq 0, \quad \sum_{u^{(2)}} q(i, u^{(2)}) = 1, \quad i = 1, 2, \ldots, X, \ u^{(2)} = 1, 2, \ldots, U. \end{aligned} \tag{34}$$

Solving the above LP yields the Nash equilibrium policy $\mu^{(2)}$ for player 2.

The dual problem to (34) is the linear program

$$\begin{aligned} \text{Minimize } &\sum_{i \in \mathcal{X}} z(i) \ \text{ with respect to } (z, p) \\ \text{subject to } &p(i, u^{(1)}) \geq 0, \quad i \in \mathcal{X}, u \in \mathcal{U} \\ &\sum_{u^{(1)}} p(j, u^{(1)}) = \rho \sum_i \sum_{u^{(1)}} P_{ij}(u^{(1)}) \, p(i, u^{(1)}) + \alpha_j, \ j \in \mathcal{X}. \\ &z(i) \geq \sum_{u^{(1)}} p(i, u^{(1)}) \, c(i, u^{(1)}, u^{(2)}) \end{aligned}$$

The above dual gives the randomized Nash equilibrium policy p^* for player 1.

6.2.4 Example 2. Switching Controller Markov Game

This is a special case of a zero sum Markov game where the state space $\mathcal{X}$ is partitioned into disjoint sets $S^{(1)}, S^{(2)}$ such that $S^{(1)} \cup S^{(2)} = \mathcal{X}$ and

$$P_{ij}(u^{(1)}, u^{(2)}) = \begin{cases} P_{ij}(u^{(1)}), & i \in S^{(1)} \\ P_{ij}(u^{(2)}), & i \in S^{(2)} \end{cases}$$

So for states in $S^{(1)}$, controller 1 controls that transition matrix, while for states in $S^{(2)}$, controller 2 controls the transition matrix.

Obviously for $i \in S^{(2)}$, (31) becomes an linear program while for $i \in S^{(1)}$, (32) becomes a linear program. As discussed in [26], the Nash equilibrium can be computed by solving a finite sequence of linear programming problems.

Chapter 7

Partially Observed Markov Decision Processes (POMDPs)

Several well studied instances of POMDPs and their parameter files can be found at `http://www.pomdp.org/examples/`

1. Much insight can be gained by simulating the dynamic programming recursion for a 3-state POMDP. The belief state needs to be quantized to a finite grid. We also strongly recommend using the exact POMDP solver in [14] to gain insight into the piecewise linear concave nature of the value function.
2. Implement Lovejoy's suboptimal algorithm and compare its performance with the optimal policy.
3. **Tiger problem**: This is a colorful name given to the following POMDP problem.
   A tiger resides behind one of two doors, a left door (l) and a right door (r). The state $x \in \{l, r\}$ denotes the position of a tiger. The action $u \in \{l, r, h\}$ denotes a human either opening the left door (l), opening the right door (r), or simply hearing (h) the growls of the tiger. If the human opens a door, he gets a perfect measurement of the position of the tiger (if the tiger is not behind the door he opens, then it must be behind the other door). If the human chooses action h then he hears the growls of the tiger which gives noisy information about the tiger's position. Denote the probabilities $B_{ll}(h) = p$, $B_{rr}(h) = q$.
   Every time the human chooses the action to open a door, the problem resets and the tiger is put with equal probability behind one of the doors. (So the transition probabilities for the actions l and r are 0.5).
   The cost of opening the door behind where the tiger is hiding is α, possibly reflecting injury from the tiger. The cost of opening the other door is $-\beta$ indicating a reward. Finally the cost of hearing and not opening a door is γ.
   The aim is to minimize the cost (maximize the reward) over a finite or infinite horizon.

To summarize, the POMDP parameters of the tiger problem are:

$$\begin{aligned}
\mathcal{X} = \{l, r\}, \mathcal{Y} = \{l, r\}, \mathcal{U} = \{l, r, h\}, \\
B(l) = B(r) = I_{2\times 2}, B(h) = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \\
P(l) = P(r) = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, P(h) = I_{2\times 2}, \\
c_l = (\alpha, -\beta)', c_r = (-\beta, \alpha)', c_h = (\gamma, \gamma)'
\end{aligned}$$

4. **Open Loop Feedback Control.** As described in §7.5.5, open loop feedback control is a useful suboptimal scheme for solving POMDPs. Is it possible to exploit knowledge that the value function of a POMDP is piecewise linear and concave in the design of an open loop feedback controller?
5. Finitely transient policies were discussed in §7.6. For a 2-state, 2-action, 2-observation POMDP, give an example of POMDP parameters that yield a finitely transient policy with $n^* = 2$.
6. **Uniform sampling from Belief space.** Recall that the belief space $\Pi(X)$ is the unit $X-1$ dimensional simplex. Show that a convenient way of sampling uniformly from $\Pi(X)$ is to use the Dirichlet distribution

$$\pi_0(i) = \frac{x_i}{\sum_{j=1}^{X} x_j}, \quad \text{where } x_i \sim \text{ unit exponential distribution.}$$

7. **Adaptive Control of a fully observed MDP formulated as a POMDP problem.** Consider a fully observed MDP with transition matrix $P(u)$ and cost $c(i, u)$, where $u \in \{1, 2, \ldots, U\}$ denotes the action. Suppose the true transition matrices $P(u)$ are not known. However, it is known apriori that they belong to a known finite set of matrices $P(u, \theta)$ where $\theta \in \{1, 2, , \ldots, L\}$. As data accumulates, the controller must simultaneously control the Markov chain and also estimate the transition matrices. The above problem can be formulated straightforwardly as a POMDP. Let θ_k denote the parameter process. Since the parameter $\theta_k = \theta$ does not evolve with time, it has identity transition matrix. Note that θ is not known; it is partially observed since we only see the sample path realization of the Markov chain x with transition matrix $P(u, \theta)$.
   Aim: Compute the optimal policy

$$\mu^* = \underset{\mu}{\operatorname{argmin}}\, J_\mu(\pi_0) = \mathbb{E}\{\sum_{k=0}^{N-1} c(x_k, u_k)|\pi_0\}$$

   where π_0 is the prior pmf of θ. The key point here is that as in a POMDP (and unlike an MDP), the action u_k will now depend on the history of past actions and the trajectory of the Markov chain as we will now describe.
   Formulation: Define the augmented state (x_k, θ_k). Since $\theta_k = \theta$ does not evolve, clearly the augmented state has transition probabilities

$$\mathbb{P}(x_{k+1} = j, \theta_{k+1} = m | x_k = i, \theta_k = l, u_k = u) = P_{ij}(u, l)\, \delta(l - m), \quad m = 1, \ldots, L.$$

At time k, denote the history as $\mathcal{H}_k = \{x_0, \ldots, x_k, u_1, \ldots, u_{k-1}\}$. Then define the belief state which is the posterior pmf of the model parameter estimate:

$$\pi_k(l) = \mathbb{P}(\theta_k = l | \mathcal{H}_k), \qquad l = 1, 2. \ldots, L.$$

(a) Show that the posterior is updated via Bayes' formula as

$$\begin{aligned} &\pi_{k+1}(l) = T(\pi_k, x_k, x_{k+1}, u_k)(l) \stackrel{\text{defn}}{=} \frac{P_{x_k,x_{k+1}}(u_k, l)\, \pi_k(l)}{\sigma(\pi_k, x_k, x_{k+1})},\ l = 1, 2. \ldots, L \\ &\text{where } \ \sigma(\pi_k, x_k, x_{k+1}, u_k) = \sum_m P_{x_k,x_{k+1}}(u_k, m)\, \pi_k(m). \end{aligned} \tag{35}$$

Note that π_k lives in the $L-1$ dimensional unit simplex.
Define the belief state as (x_k, π_k). The actions are then chosen as

$$u_k = \mu_k(x_k, \pi_k)$$

Then the optimal policy $\mu_k^*(i, \pi)$ satisfies Bellman's equation

$$\begin{aligned} J_k(i, \pi) &= \min_u Q_k(i, u, \pi), \quad \mu_k^*(i, \pi) = \operatorname*{argmin}_u Q_k(i, u, \pi) \\ Q_k(i, u, \pi) &= c(i, u) + \sum_j J_{k+1}\big(j, T(\pi, i, j, u)\big)\, \sigma(\pi, i, j, u) \end{aligned} \tag{36}$$

initialized with the terminal cost $J_N(i, \pi) = c_N(i)$.

(b) Show that the value function J_k is piecewise linear and concave in π. Also show how the exact POMDP solution algorithms in Chapter 7 can be used to compute the optimal policy.

The above problem is related to the concept of *dual control* which dates back to the 1960s [25]; see also [56] for the use of Lovejoy's suboptimal algorithm to this problem. Dual control relates to the tradeoff between estimation and control: if the controller is uncertain about the model parameter, it needs to control the system more aggressively in order to probe the system to estimate it; if the controller is more certain about the model parameter, it can deploy a less aggressive control. In other words, initially the controller explores and as the controller becomes more certain it exploits. Multi-armed bandit problems optimize the tradeoff between exploration and exploitation.

8. **Optimal Search and Dynamic (Active) hypothesis testing.** In §7.7.4 of the book, we considered the classical optimal search problem where the objective was to search for a non-moving target amongst a finite number of cells. A crucial assumption was that there are no false alarms; if an object is not present in a cell and the cell is searched, the observation recorded is $\bar{F}$ (not found).
A generalization of this problem is studied in [15]. Assume there are $\mathcal{U} = \{1, 2, \ldots, U\}$ cells. When cell u is searched
   - If the target is in cell u then an observation y is generated with pdf or pmf $\phi(y)$.
   - If the target is **not** in cell u, then an observation y is generated with pdf or pmf $\bar{\phi}(y)$. (Recall in classical search $\bar{\phi}(y)$ is dirac measure on the observation symbol $\bar{F}$.)

The aim is to determine the optimal search policy $\boldsymbol{\mu}$ over a time horizon N to maximize

$$J_{\boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{\mu}} \max_{u \in \{1,\ldots,U\}} \pi_N(u)\}$$

at the final time N.

Assume the pdf or pmf $\bar{\phi}(y)$ is symmetric in y, that is $\bar{\phi}(y) = \bar{\phi}(b-y)$ for some real constant b. Then [15, Proposition 3] shows the nice result that the optimal policy is to search either of the two most likely locations given the belief π_k.

The above problem can be viewed as an active hypothesis testing problem, which is an instance of a controlled sensing problem. The decision maker seeks to adaptively select the most informative sensing action for making a decision in a hypothesis testing problem. Active hypothesis testing goes all the way back to the 1959 paper by Chernoff [18]. For a more general and recent take of active hypothesis testing please see [66].

Chapter 8

POMDPs in Controlled Sensing and Sensor Scheduling

1. **Optimal Observer Trajectory for Estimating a Markovian Target.** This problem is identical to the search problem described in §7.7. A target moves in space according to a Markov chain. (For convenience assume X-cells in two dimensional space. A moving observer (sensor) measures the target's state (position) in noise. Assume that the noise depends on the relative distance between the target and the observer. How should the observer move amongst the X-cells in order to locate where the target is? One metric that has been used in the literature [52] is the stochastic observability (which is related to the mutual information) of the target; see also §12.7. The aim of the observer is to move so as to maximize the stochastic observability of the target. As described in §7.7, the problem is equivalent to a POMDP.
   A more fancy version of the setup involves multiple observers (sensors) that move within the state space and collaboratively seek to locate the target. Assume that the observers exchange information about their observations and actions. The problem can again be formulated as a POMDP with a larger action and observation space.
   Suppose the exchange of information between the observers occurs over a noisy communication channel where the error probabilities evolve according to a Markov chain as in §9.6. Formulate the problem as a POMDP.
2. **Risk averse sensor scheduling.** As described in §8.4, in controlled sensing applications, one is interested in incorporating the uncertainty in the state estimate into the instantaneous cost. This cannot be modeled using a linear cost since the uncertainty is minimized at each vertex of the simplex $\Pi(X)$. In §8.4, quadratic functions of the belief were used to model the conditional variance. A more principled alternative is to use dynamic coherent risk measures; recall three examples of such risk measures were discussed in §8.6.
   Discuss how open loop feedback control can be used for a POMDP with dynamic coherent risk measure.
3. **Sensor Usage Constraints.** The aim here is to how the POMDP formulation of a controlled sensing problem can be modified straightforwardly to incorporate sensing constraints on the total usage of particular sensors. Such constraints are often used in sensor resource management.

(a) Consider a N horizon problem where sensor 1 can be used at most L times where $L \leq N$. For notational simplicity, assume that there are two sensors, so $\mathcal{U} = \{1, 2\}$. Assume that there are no constraints on the usage of the other sensors.
For notational convenience we consider rewards denoted as $R(\pi, u) = \sum_{i=1}^{X} R(i, u)\pi(i)$ instead of costs $C(\pi, u)$ expressed in terms of the belief state π. Show that Bellman's equation is given by

$$V_{n+1}(\pi, l) = \max\{R(\pi, 1) + \sum_y V_n(T(\pi, y, 1), l-1)\sigma(\pi, y, 1), \\ R(\pi, 2) + \sum_y V_n(T(\pi, y, 2), l)\sigma(\pi, y, 2)\}$$

with boundary condition $V_n(\pi, 0) = 0$, $n = 0, 1, \ldots, N$.

(b) If the constraint is that sensor 1 needs to be used exactly L times, then show that the following additional boundary condition needs to be included:

$$V_n(\pi, n) = R(\pi, 1) + \sum_y V_{n-1}(T(\pi, y, 1), n-1)\sigma(\pi, y, 1), \quad \text{for } n = 1, \ldots, L.$$

(c) In terms of the POMDP solver software, the constraint for using sensor 1 at most L times is easily incorporated by augmenting the state space. Define the controlled finite state process $r_k \in \{0, 2, \ldots, L\}$ with $(L+1) \times (L+1)$ transition matrices

$$Q(1) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad Q(2) = I.$$

Then define the POMDP with:
- transition matrices $P(1) \otimes Q(1)$ and $P(2) \otimes Q(2)$,
- observation probabilities $p(y|x, r, u) = p(y|x, u)$,
- rewards $R(x, r, u) = R(x, u)$ for $r > 0$ and $R(x, r = 0, u) = 0$.

In the problems for Chapter 12, we consider a simpler version of the above problem for optimal measurement selection of a HMM. In that simpler case, one can develop structural results for the optimal policy.

4. As described in §8.4, in controlled sensing it makes sense to choose a cost that is nonlinear in the belief state π in order to penalize uncertainty in the state estimate. One choice of a nonlinear cost that has zero cost at the vertices of the belief space is

$$C(\pi, u) = \min_{i \in \{1, \ldots, X\}} \pi(i).$$

This cost $C(\pi, u)$ is piecewise linear and concave in $\pi \in \Pi(X)$ where $\Pi(X)$ denotes the belief space.
Since $C(\pi, u)$ is positively homogeneous, show that the value function is piecewise linear and concave for any finite horizon N. Hence the optimal POMDP solvers of Chapter 7 can be used to solve this nonlinear cost POMDP exactly and therefore compute the optimal policy.

Chapter 9

Structural Results for Markov Decision Processes

1. **Supermodularity, Single Crossing Condition & Interval Dominance Order.**
   A key step in establishing structural results for MDPs is to give sufficient conditions for $u^*(x) = \text{argmax}_u \phi(x,u)$ to be increasing in x. In §9.1 of Chapter 9 we gave two conditions, namely supermodularity and the single crossing condition (which is a more general condition than supermodularity). More recently, the interval dominance order has been introduced in [76] as an even more general condition. All three conditions boil down to the following statement:

$$\phi(x+1,u+1) - \phi(x+1,u) \geq \rho(u)\left(\phi(x,u+1) - \phi(x,u)\right) \tag{37}$$

   where $\rho(u)$ is a strictly positive function of u. In particular,
   - Choosing $\rho(u) = 1$ in (37) yields the supermodularity condition.
   - If there exists a fixed positive constant $\rho(u)$ such that (37) holds, then the single crossing condition holds.
   - If there exists a positive function $\rho(u)$ that is increasing[1] in u, then (37) yields the interval dominance order condition (actually this is a sufficient condition for interval dominance, see [76] for details).

   Note that single crossing and interval dominance are ordinal properties in the sense that they are preserved by monotone transformations.
   The sum of supermodular functions is supermodular. Unfortunately, in general, the um of single crossing functions is not single crossing; however, see [77] for some results. Discuss if the interval dominance order holds for sums of functions. Can it be used to develop structural results for an MDP?
2. In general, the sum of single crossing functions is not single crossing. Even a constant plus a single crossing function is not necessarily single crossing. Sketch the curve of a single crossing function which wiggles close to zero. Then adding a positive constant implies that the curve will cross zero more than once. Also the sum of a supermodular plus single crossing is not single crossing. In terms of $\phi(x) = f(x,u+1) - f(x,u)$, supermodular implies $\phi(x)$ is increasing in x. Clearly the sum of an increasing function and a single crossing is not single crossing in general.

[1] Recall that in the book we use increasing in the weak sense to mean non-decreasing

3. **Invariance of optimal policy to costs.** Recall that Theorem 9.3.1 require that the MDP costs satisfy assumptions (A1) and (A3) for the optimal policy to be monotone. Show that for a discounted cost infinite horizon MDP, assumption (A1) and (A3) can be relaxed as follows:
   There exists a single vector $\phi \in \mathbb{R}^X$ such that for every action $u \in \mathcal{U}$,
   (A1') $(I - \rho P(u))\phi$ is a vector with increasing elements. (Recall ρ is the discount factor.)
   (A3') $(P(u+1) - P(u))\phi$ is a vector with decreasing elements.
   In other words the structure of the transition matrix is enough to ensure a monotone policy and no assumptions are required on the cost (of course the costs are assumed to be bounded)
   Hint: Define the new value function $\bar{V}(i) = V(i) - \phi(i)$. Clearly the optimal policy remains unchanged and $\bar{V}$ satisfies Bellman's equation

$$\bar{V}(i) = \min_u \{c(i,u) - \phi(i) + \rho \sum_j \phi(j) P_{ij}(u) + \rho \sum_j \bar{V}(j) P_{ij}(u)\}$$

   where $\rho \in (0,1)$ denotes the discount factor.
4. **Myopic lower bound to optimal policy.** Recall that supermodularity of the transition matrix (A4) was a key requirement for the optimal policy to be monotone. In particular, Theorem 9.3.1 shows that $Q(i,u)$ is submodular, i.e., $Q(i,u+1) - Q(i,u)$ is decreasing in i. Sometimes supermodularity of the transition matrix is too much to ask for. Consider instead of (A4) the relaxed condition
   (A4') $P_i(u+1) \geq_s P_i(u)$ for each row i.
   Show that (A4') together with (A1), (A2) implies that

$$\sum_j P_{ij}(u+1) V(j) \leq \sum_j P_{ij}(u) V(j)$$

   Define the myopic policy $\underline{\mu}(i) = \operatorname{argmin}_u c(i,u)$. Show that under (A1), (A2), (A4'), $\mu^*(i) \geq \underline{\mu}(i)$. In other words, the myopic policy $\underline{\mu}$ forms a lower bound to the optimal policy μ^*.
5. **Monotone policy iteration algorithm.** Suppose an MDP has a monotone policy. If the MDP parameters are known, then the policy iteration algorithm of §6.4.2 can be used. If the policy μ_{n-1} at iteration $n-1$ is monotone then show that under the assumptions of (A1), (A2) of Theorem 9.3.1, the policy evaluation step yields $J_{\mu_{n-1}}$ as a decreasing vector. Also show that under (A1)-(A4), (a similar proof to Theorem 9.3.1) implies that the policy improvement step yields μ_n that is monotone. So the policy iteration algorithm will automatically be confined to monotone policies if initialized by a monotone policy.
6. **Monotone Policies for MDPs with vector-valued states**. Consider a MDP with vector-valued states $x_k = [x_k(1), \ldots, x_k(L)]'$. Here each component $x_k(l)$ takes on X values in a finite set. Obviously, the vector state x_k has state space with cardinality X^L. So one can consider an equivalent X^L state scalar valued MDP.
   However, in many cases, the setup of the MDP is in terms of the evolution of individual elements of the state; for example in the extreme case where each component of the state evolves independently according to a finite state Markov chain. By using multi-variate first order dominance, give sufficient conditions on the transition matrix

and costs so that the optimal policy $\mu^*(\mathbf{i})$ is increasing in $\mathbf{i}$ where $\mathbf{i} = (i_1, i_2, \ldots, i_L)$: that is

$$\mathbf{i} \geq \mathbf{j} \implies \mu^*(\mathbf{i}) \geq \mu^*(\mathbf{j}).$$

Here $\mathbf{i} \geq \mathbf{j}$ denotes the partial order $i_1 \geq j_1, i_2 \geq j_2, \ldots, i_L \geq j_L$. To give some perspective, for multi-variate POMDPs, the book develops the multivariate MLR order (called the TP2 stochastic order) - such orders are closed under Bayesian updates and hence ideally suited for developing structural results for POMDPs.

7. **Stochastic knapsack problem.** Consider the following version of the stochastic knapsack problem;[2] see [81] and also [16]. A machine must operate for T time points. Suppose that one specific component of the machine fails intermittently. This component is replaced when it fails. There are U-possible brands one can choose to replace this component when it fails. Brand $u \in \{1, 2 \ldots, U\}$ costs c_u and has an operating lifetime that is exponentially distributed with rate λ_u. The aim is to minimize the expected total cost incurred by replacing the failed component so that the machine operates for T time points.
Suppose a component has just failed. Let t denote the remaining time left to operate the machine. The optimal policy for deciding which of the U possible brands to choose the replacement satisfies Bellman's equation

$$\begin{aligned} Q(t,u) &= c(u) + \int_0^t V(t-\tau)\, \lambda_u e^{-\lambda_u \tau} d\tau, \quad Q(0,u) = 0, \\ V(t) &= \min_{u \in \{1,2,\ldots,U\}} Q(t,u), \quad \mu^*(t) = \operatorname*{argmin}_{u \in \{1,2,\ldots,U\}} Q(t,u) \end{aligned}$$

Show that if $\lambda_u c(u)$ is decreasing with u, then $Q(t,u)$ is submodular. In particular, show that

$$\frac{d}{dt} Q(t,u) = \lambda_u c(u)$$

Therefore, the optimal policy $\mu^*(t)$ has the following structure: Use brand 1 when the time remaining is small, then switch to brand 2 when the time increases, then brand 3, etc.
Generalize the above result to the case when time k is discrete and the brand u has life time pmf $p(k,u)$, $k = 0, 1 \ldots$. Then Bellman's equation reads

$$\begin{aligned} Q(n,u) &= c(u) + \sum_{k=0}^{n} V(n-k)\, p(k,u) \\ V(n) &= \min_{u \in \{1,2,\ldots,U\}} Q(n,u), \quad \mu^*(n) = \operatorname*{argmin}_{u \in \{1,2,\ldots,U\}} Q(n,u) \end{aligned}$$

What are sufficient conditions in terms of submodularity of the lifetime pmf $p(k,u)$ for the optimal policy to be monotone?

8. **Monotonicity of optimal policy with respect to horizon.** Show that the following result holds for a finite horizon MDP. If $Q_n(i,u)$ is supermodular in (i,u,n)

[2] The classical NP hard knapsack problem deals with U items with costs $c(1), c(2), \ldots, c(U)$ and lifetimes $t_1, t_2, \ldots t_U$. The aim is to compute the minimum cost subset of these items whose total lifetime is at most T.

then $V_n(i) = \max_u Q_n(i,u)$ is supermodular in i,u. Note that checking supermodularity with respect to (i,u,n) is pairwise: so it suffices to check supermodularity with respect to (i,u), (i,n) and (u,n).
With the above result, consider a finite horizon MDP satisfies the assumptions (A1)-(A4) of §9.3. Under what further conditions is $\mu_n^*(i)$ is increasing in n for fixed i? What does this mean intuitively?

9. **Monotone Discounted Cost Markov Games.** In § 6.2 on page 31 of this internet supplement we briefly described the formulation of infinite horizon discounted cost Markov games. Below we comment briefly on structural results for the Nash equilibrium of such games.
Consider the infinite horizon discounted cumulative cost of (22). The structural results developed in this chapter for MDPs extend straightforwardly to infinite horizon discounted cost Markov games. The assumptions (A1) to (A4) of §9.3 of the book need to be extended as follows:

   (A1) Costs $c(x,u,u^-)$ are decreasing in x and u^-. Here u^- denotes the actions of others players.

   (A2) $P_i(u,u^-) \leq_s P_{i+1}(u,u^-)$ for each i and fixed u,u^-. Here $P_i(u,u^-)$ denotes the i-th row of the transition matrix for action u,u^-.

   (A3) $c(x,u,u^-)$ is submodular in (x,u) and (u,u^-)

   (A4) $P_{ij}(u,u^-)$ is tail-sum supermodular in (i,u,u^-). That is,

$$\sum_{j\geq l} \big(P_{ij}(u+1,u^-) - P_{ij}(u,u^-)\big) \text{ is increasing in } i.$$

   Theorem 1. *Under conditions (A1)-(A4), there exists a pure Nash equilibrium* $(\mu^{(1)^*}, \mu^{(2)^*})$ *such that the pure policies* $\mu^{(1)^*}$ *and* $\mu^{(2)^*}$ *are increasing in state* i.

   Contrast this with the case of a general Markov game (§6.2 of this internet supplement) where one can only guarantee the existence of a randomized Nash equilibrium in general.
   The proof of the above theorem is as follows. First for any increasing fixed policy $\mu^{(2)}$ for player 2, one can show via an identical proof to Theorem 9.3.1, the optimal policy $\mu^{(1)^*}(x,\mu^{(2)}(x))$ is increasing in x. Similarly, for any increasing fixed policy $\mu^{(1)}$ for player 1, $\mu^{(2)^*}(x,\mu^{(1)}(x))$ is increasing in x. These are obtained as the solution of Bellman's equation. In game theory, these are called best response strategies. Therefore the vector function $[\mu^{(1)^*}(x), \mu^{(2)^*}(x)]$ is increasing in x. It then follows from Tarski's fixed point theorem[3] that such a function has a fixed point. Clearly this fixed point is a Nash equilibrium since any unilateral deviation makes one of the players worse off.
   Actually for submodular games a lot more holds. The smallest and largest Nash equilibria are pure (non-randomized) and satisfy the monotone property of the above theorem. These can be obtained via a best response algorithm the simply iterates the best responses $\mu^{(1)^*}(x,\mu^{(2)}(x))$ and $\mu^{(2)^*}(x,\mu^{(1)}(x))$ until convergence. There are numerous papers and books in the area.

[3] Let X denote a compact lattice and $f : X \to X$ denote an increasing function. Then there exists a fixed point $x^* \in X$ such that $f(x^*) = x^*$

Chapter 10

Structural Results for Optimal Filters

1. In the structural results presented in the book, we have only considered first order stochastic dominance and monotone likelihood ratio dominance (MLR) since they are sufficient for our purposes. Naturally there are many other concepts of stochastic dominance [65]. Show that

$$\text{MLR} \implies \text{Hazard rate order} \implies \text{first order} \implies \text{second order}$$

   Even though second order stochastic dominance is useful for concave decreasing functions (such as the value function of a POMDP), just like first order dominance, it cannot cope with conditioning (Bayes' rule).
2. Consider a reversible Markov chain with transition matrix P, initial distribution π_0 and stationary distribution π_∞. Suppose $\pi_0 \leq_r \pi_\infty$. Show that if P has rows that are first order increasing then $\pi_n \leq_r \pi_\infty$.
3. **TPn matrix.** A key assumption (F2) in the structural results is that the transition matrix P is TP2. More generally, suppose $n = 2, 3, \ldots$. Then a $X \times X$ matrix P is said to be totally positive of order n (denoted as TPn) if for each $k \leq n$, all the $k \times k$ minors of P are non-negative.
4. **TP2 matrix properties.**[1] §10.5 gave some useful properties of TP2 matrices. Suppose the $X \times X$ stochastic matrix P is TP2.
   (a) Show that this implies that the elements satisfy

$$P_{11} \geq P_{21} \geq \cdots \geq P_{X1}$$
$$P_{1X} \geq P_{2X} \geq \cdots \geq P_{XX}$$

   (b) Suppose P has no null columns. Show that if $P_{ij} = 0$, then either $P_{kl} = 0$ for $k \leq i$ and $l \geq j$, or $P_{kl} = 0$ for $k \geq i$ and $l \leq j$.
   (c) Show that

$$e_1'(P^n)'e_1 \downarrow n, \qquad e_X'(P^n)'e_1 \uparrow n.$$

[1] Note that a TP2 matrix does not need to be a square matrix; we consider P to be square here since it is a transition probability matrix.

Also show that for each n,

$$e_1'(P^n)'e_i \downarrow i, \qquad e_X'(P^n)'e_i \uparrow i$$

Please see [40] for several other interesting properties of TP2 matrices.

5. MLR dominance is intimately linked with the TP2 property. Show that

$$\pi_1 \leq_r \pi_2 \iff \begin{bmatrix} \pi_1' \\ \pi_2' \end{bmatrix} \text{ is TP2 .}$$

6. **Properties of MLR dominance.** Suppose X and Y are random variables and recall that $\geq_r$ denotes MLR dominance.[2]
   (a) Show that $X \geq_r Y$ is equivalent to

$$\{X|X \in A\} \geq_s \{Y|Y \in A\}$$

   for all events A with $P(X \in A) > 0$ and $P(Y \in A) > 0$ where $\geq_s$ denotes first order dominance. This property is due to [92].
   (b) Show that $X \geq_r Y$ implies that $g(X) \geq_r g(Y)$ for any increasing function g.
   (c) Show that $X \geq_r Y$ implies that $\max\{X, c\} \geq_r \max\{Y, c\}$ for any positive constant c.
   (d) Under what conditions does $X \geq_r Y$ imply that $-X \leq_r -Y$?

   Do the above two properties hold for first order dominance?

7. **MLR monotone optimal predictor.** Consider the HMM predictor given by the Chapman Kolmogorov equation $\pi_k = P'\pi_{k-1}$. Show that if P is a TP2 matrix and $\pi_0 \leq_r \pi_1$, then $\pi_0 \leq_r \pi_1 \leq_r \pi_2 \leq_r \ldots$.

8. **MLR constrained importance sampling.** One of the main results of this chapter was to construct reduced complexity HMM filters that provably form lower and upper bounds to the optimal HMM filter in the MLR sense. In this regard, consider the following problem. Suppose it is known that $\underline{P}'\pi \leq_r P'\pi$. Then given the reduced complexity computation of $\underline{P}'\pi$, how can this be exploited to compute $P'\pi$?
   It is helpful to think of the following toy example: Suppose it is known that $x'p \leq 1$ for a positive vector x and probability vector p. How can this constraint be exploited to actually compute the inner product $x'p$? Obviously from a deterministic point of view there is little one can do to exploit this constraint. But one can use constrained important sampling: one simple estimator is as follows:

$$\frac{1}{N}\sum_{i=1}^{N} x_i I(x_i \leq 1)$$

   where index i is simulated iid from probability vector p. In [49] a more sophisticated constrained importance sampling approach is used to estimate $P'\pi$ by exploiting the constraint $\underline{P}'\pi \leq_r P'\pi$.

[2] Stochastic dominance is a property of the distribution of a random variable and has nothing to do with the random variable itself. Therefore in the book, we defined stochastic dominance in terms of the pdf or pmf. Here to simplify notation we use the random variable instead of its distribution.

9. **Posterior Cramer Rao bound.** The posterior Cramer Rao bound [88] for filtering can be used to compute a lower bound to the mean square error. This requires twice differentiability of the logarithm of the joint density. For HMMs, one possibility is to consider the Weiss-Weinstein bounds , see [79]. Alternatively, the analysis of [30] can be used. Compare these with the sample path bounds for the HMM filter obtained in this chapter.
10. The shifted likelihood ratio order is a stronger order than the MLR order. Indeed, $p > q$ in the shifted likelihood ratio order sense if p_i/q_{i+j} is increasing in i for any j. (If $j = 0$ it coincides with the standard MLR order.) What additional assumptions are required to preserve the shifted likelihood ratio order under Bayes' rule? Show that the shifted likelihood ratio order is closed under convolution. How can this property be exploited to bound an optimal filter?
11. In deriving sample path bounds for the optimal filter, we did not exploit the fact that $T(\pi, y)$ increases with y. How can this fact be used in bounding the sample path of an optimal filter?
12. **Neyman-Pearson Detector** Here we briefly review elementary Neyman-Pearson detection theory and show the classical result that MLR dominance results in a threshold optimal detector.
    Given the observation x of a random variable, we wish to decide if x is from pdf f or g. To do this, we construct a decision policy $\phi(x)$. The detector decides

$$\begin{array}{ll} f & \text{if } \phi(x) = 0 \\ g & \text{if } \phi(x) = 1 \end{array} \tag{38}$$

    The performance of the decision policy ϕ in (38) is determined in terms of two metrics:
    (a) $\mathcal{P} = \mathbb{P}(\text{ reject } f | f \text{ is true })$
    (b) $\mathcal{Q} = \mathbb{P}(\text{ reject } f | f \text{ is false })$
    Clearly for the decision policy $\phi(\cdot)$ in (38),

$$\mathcal{P} = \int_{\mathbb{R}} f(x)\phi(x)dx, \quad \mathcal{Q} = \int_{\mathbb{R}} g(x)\phi(x)dx.$$

    The well known Neyman-Pearson detector seeks to determine the optimal decision policy ϕ^* that maximizes $\mathcal{Q}$ subject to the constraint $\mathcal{P} \leq \alpha$ for some user specified $\alpha \in (0, 1]$. The main result is

    Theorem (Neyman-Pearson lemma)**.** *Amongst all decision rules ϕ such that $\mathcal{P} \leq \alpha$, the decision rule ϕ^* which maximizes $\mathcal{Q}$ is given by*

$$\phi^*(x) = \begin{cases} 0 & \frac{f(x)}{g(x)} \geq c \\ 1 & \frac{f(x)}{g(x)} < c \end{cases}$$

    where c is chosen so that $\mathcal{P} = \alpha$.

    Proof. Clearly for any $x \in \mathbb{R}$,

$$\big(\phi^*(x) - \phi(x)\big)\big(cg(x) - f(x)\big) \geq 0.$$

Please verify the above inequality by showing that if $\phi^*(x) = 1$ then both the terms in the above product are nonnegative; while if $\phi^*(x) = 0$, then both the terms are nonpositive. Therefore,

$$c\Big(\int \phi^*(x)g(x)dx - \int \phi(x)g(x)dx\Big) \geq \int \phi^*(x)f(x)dx - \int \phi(x)f(x)dx$$

The right hand side is non-negative since by construction $\int \phi^*(x)f(x)dx = \alpha$, while $\int \phi(x)f(x)dx \leq \alpha$. □

Threshold structure of optimal detector. Let us now give conditions so that the optimal Neyman-Pearson decision policy is a threshold policy: Suppose now that f MLR dominates g, that is $f(x)/g(x) \uparrow x$. Then clearly

$$\phi^*(x) = \begin{cases} 0 & x \geq x^* \\ 1 & x < x^* \end{cases} \tag{39}$$

where threshold x^* satisfies

$$\int_{-\infty}^{x^*} f(x)dx = \alpha$$

Thus if $f \geq_r g$, then the optimal detector (in the Neyman-Pearson sense) is the threshold detector (39).

Chapter 11

Monotonicity of Value Function for POMDPs

1. Theorem 11.2.1 is the main result of the chapter and it gives conditions under which the value function of a POMDP is MLR decreasing. Condition (C) was the main assumption on the possibly non-linear cost. Give sufficient conditions for a quadratic cost $1 - \pi'\pi + c_u'\pi$ to satisfy (C). Under what conditions does the entropy $-\sum_i \pi(i) \log \pi(i) + c_u'\pi$ satisfy (C).
2. The shifted likelihood ratio order is a stronger order than the MLR order. Indeed, $p > q$ in the shifted likelihood ratio order sense if p_i/q_{i+j} is increasing in i for any j. If $j = 0$ it coincides with the standard MLR order. (Recall also the problem in the previous chapter which says that the shifted likelihood ratio order is closed under convolution.) By using the shifted likelihood ratio order, what further results on the value function $V(\pi)$ can one get by using Theorem 11.2.1.
3. Theorem 11.3.1 gives sufficient conditions for a 2-state POMDP to have a threshold policy. We have assumed that the observation probabilities are not action dependent. How should the assumptions and proof be modified to allow for action dependent observation probabilities?
4. How can Theorem 11.3.1 be modified if dynamic risk measures of §8.6 are considered? (see also §13.3).
5. **Finite dimensional characterization of Gittins index for POMDP bandit** [50]: §11.4 dealt with POMDP multi-armed bandit problem. Consider a POMDP bandit where the Gittins index (11.18) is characterized as the solution of Bellman's equation (11.19). Since the value function of a POMDP is piecewise linear and concave (and therefore a finite dimensional characterization), it follows that a value iteration algorithm for (11.19) that characterizes the Gittins index also has a finite dimensional characterization. Obtain an expression for this finite dimensional characterization for the Gittins index (11.18) for a horizon N value iteration algorithm.
6. §11.4 of the book deals with structural results for POMDP bandits. Consider the problem where several searchers are looking for a stationary target. Only one searcher can operate at a given time and the searchers cannot receive state estimate information from other searchers or a base-station. The base station simply sends a 0 or 1 signal to each searcher telling them when to operate and when to shut down. When

it operates, the searcher obtains moves according to a Markov chain and obtains noisy information about the target. Show how the problem can be formulated as a POMDP multi-armed bandit.
Show how a radar seeking to hide its emissions (low probability of intercept radar) can be formulated approximately as a POMDP bandit.

7. How does the structural result for the Gittins index for a POMDP bandit specialize to that of a full observed Markov decision process bandit problem?
8. Consider Problem 7 on page 39 of Chapter 7 where optimal adaptive control of a fully observed MDP was formulated as a POMDP. Give conditions that ensure that the value function $J_k(i,\pi)$ is MLR decreasing in π and also monotone in i. What are the implications of this monotonicity in terms of dual control (i.e., exploration vs exploitation)?
9. **Optimality of Threshold Policy for 2-state POMDP** Recall that Theorem 11.3.1 in the book gave sufficient conditions for the optimal policy of a 2-state POMDP to be a threshold. Consider the proof of Theorem 11.3.1 in Appendix 11.A of the book. The last step involved going from (11.28) to a simpler expression via tedious but elementary steps. Here we specify what these steps are.
Start with (11.28) in the book:

$$\begin{aligned} I_3 &= \left[\sigma(\bar{\pi},y,2)+\sigma(\bar{\pi},y,1)\frac{T(\bar{\pi},y,1)-T(\pi,y,2)}{T(\pi,y,2)-T(\bar{\pi},y,2)}+\sigma(\pi,y,1)\frac{T(\pi,y,2)-T(\pi,y,1)}{T(\pi,y,2)-T(\bar{\pi},y,2)}\right] \\ &= \frac{I_{31}+I_{32}+I_{33}}{\sigma(\pi,y,2)\left(T(\pi,y,2)-T(\bar{\pi},y,2)\right)} \\ I_{31} &= \sigma(\pi,y,2)\sigma(\bar{\pi},y,1)\left(T(\bar{\pi},y,1)-T(\pi,y,2)\right) \\ I_{32} &= \sigma(\pi,y,2)\sigma(\bar{\pi},y,2)\left(T(\pi,y,2)-T(\bar{\pi},y,2)\right) \\ I_{33} &= \sigma(\pi,y,2)\sigma(\pi,y,1)\left(T(\pi,y,2)-T(\pi,y,1)\right) \end{aligned} \tag{40}$$

The second element of HMM predictors $P(a)'\pi$ and $(P(a)'\bar{\pi})$ are denoted by b_{a2}, $b_{a1}, a=1,2$ respectively. Here b_{a2} is defined as follows

$$b_{a2} = (1-\pi(2))P_{12}(a)+\pi(2)P_{22}(a). \tag{41}$$

Consider the following simplification of the term I_{31} by using b_{a2} and b_{a1}.

$$\begin{aligned} I_{31} =& (B_{1y}(1-b_{22})+B_{2y}b_{22})B_{2y}b_{11}-(B_{1y}(1-b_{11})+B_{2y}b_{11})B_{2y}b_{22} \\ =& B_{1y}B_{2y}(b_{11}-b_{22}) \end{aligned} \tag{42}$$

Similarly, I_{32} and I_{33} are simplified as follows

$$I_{32} = B_{1y}B_{2y}(b_{22}-b_{21}), I_{33} = B_{1y}B_{2y}(b_{22}-b_{12}) \tag{43}$$

Substituting (42), (43) in (40) yields the following

$$I_3 = B_{1y}B_{2y}\frac{b_{11}+b_{22}-b_{21}-b_{12}}{\sigma(\pi,y,2)\left(T(\pi,y,2)-T(\bar{\pi},y,2)\right)} \tag{44}$$

Substituting (41) for b_{ij} and some trivial algebraic manipulations yield the following

$$I_3 = B_{1y}B_{2y}(\pi(2) - \bar{\pi}(2))\frac{P_{22}(2) - P_{12}(2) - (P_{22}(1) - P_{12}(1))}{\sigma(\pi, y, 2)\,(T(\pi, y, 2) - T(\bar{\pi}, y, 2))}. \tag{45}$$

10. Consider the following special case of a POMDP. Suppose the prior belief $\pi_0 \in \Pi(X)$ is known. From time 1 onwards, the state is fully observed. How can the structural results in this chapter be used to characterize the optimal policy?

Chapter 12

Structural Results for Stopping Time POMDPs

12.1 Problems

Most results in stopping time POMDPs in the literature use the fact that the stopping set is convex (namely, Theorem 12.2.1). Recall that the only requirements of Theorem 12.2.1 are that the value function is convex and the stopping cost is linear. Another important result for finite horizon POMDP stopping time problems is the nested stopping set property $\mathcal{S}_0 \subseteq \mathcal{S}_1 \subseteq \mathcal{S}_2 \ldots$. The following exercises discuss both these aspects.

1. **Nested stopping set structure.** Consider the stopping time POMDP dynamic programming equation

$$V(\pi) = \min\{c_1'\pi, c_2'\pi + \sum_y V(T(\pi, y, u))\sigma(\pi, y, u)\}.$$

Define the stopping set as

$$\mathcal{S} = \{\pi : c_1'\pi \leq c_2'\pi + \sum_y V(T(\pi, y, u))\sigma(\pi, y, u)\} = \{\pi : \mu^*(\pi) = 1 \text{ (stop) }\}$$

Recall the value iteration algorithm is

$$V_{n+1}(\pi) = \min\{c_1'\pi, c_2'\pi + \sum_y V_n(T(\pi, y, u))\sigma(\pi, y, u)\}, \quad V_0(\pi) = 0.$$

Define the stopping sets $\mathcal{S}_n = \{\pi : c_1'\pi \leq c_2'\pi + \sum_y V_n(T(\pi, y, u))\sigma(\pi, y, u)\}$. Show that the stopping sets satisfy $\mathcal{S}_0 \subseteq \mathcal{S}_1 \subseteq \mathcal{S}_2 \ldots$ implying that

$$\mathcal{S} = \cup_n \mathcal{S}_n$$

2. **Explicit characterization of stopping set.** Theorem 12.2.1 showed that for a stopping time POMDP, the stopping set $\mathcal{S}$ is convex. By imposing further conditions, the set $\mathcal{S}$ can be determined explicitly. Consider the following set of belief states

$$\mathcal{S}^o = \{\pi : c_1'\pi \leq c_2'\pi + c_1'P'\pi\} \tag{46}$$

Suppose the transition matrix P and observation probabilities B of the stopping time POMDP satisfy the following property:

$$\pi \in \mathcal{S}^o \implies T(\pi, y) \in \mathcal{S}^o, \quad \forall y \in \mathcal{Y}. \tag{47}$$

(a) Prove that $\mathcal{S}^o = \mathcal{S}$. Therefore, the hyperplane $c_1'\pi = c_2'\pi + c_1'P'\pi$ determines the stopping set $\mathcal{S}$.
The proof proceeds in two steps: First prove by induction on the value iteration algorithm that for $\pi \in \mathcal{S}^o$, $V_n(\pi) = c_1'\pi$, for $n = 1, 2 \ldots$.
Second, consider a belief π such that the optimal policy goes one step and then stops. This implies that the value function is $V(\pi) = c_2'\pi + c_1'P'\pi$. Therefore clearly $c_2'\pi + c_1'P'\pi < c_1'\pi$. This implies that $\pi \notin \mathcal{S}^o$. So for any belief π such that $\mu^*(\pi)$ goes one step and stops, then $\pi \notin \mathcal{S}^o$. Therefore, for any belief π such that $\mu^*(\pi)$ goes more than one step and stops, then $\pi \notin \mathcal{S}^o$.
The two steps imply that $\mathcal{S}^o = \mathcal{S}$.Therefore that the stopping set is explicitly given by the polytope in (46).

(b) Give sufficient conditions on P and B so that condition (47) holds for a stopping time POMDP.

3. Show that an identical, proof to Theorem 12.2.1 implies that the stopping sets $\mathcal{S}_n$, $n = 1, 2, \ldots$ are convex for a finite horizon problem.
4. **Choosing a single sample from a HMM.** Suppose a Markov chain x_k is observed in noise sequentially over time as $y_k \sim B_{x_k,y}$, $k = 1, 2 \ldots, N$. Over a horizon of length N, I need to choose a single observation y_k to maximize $\mathbb{E}\{y_k\}$, $k \in 1, \ldots, N$. If at time k I decide to choose observation y_k, then I get reward $\mathbb{E}\{y_k\}$ and the problem stops. If I decide not to choose observation y_k, then I can use it to update my estimate of the state and proceed to the next time instant. However, I am not allowed to choose y_k at a later time.
   (a) Which single observation should I choose?
   Show that Bellman's equation becomes

$$V_{n+1}(\pi) = \max_{u \in \{1,2\}} \{r'\pi, \sum_y V_n(T(\pi, y))\sigma(\pi, y)\}$$

   where the elements of r are $r(i) = \sum_y yB_{iy}$, $i = 1, \ldots, X$. Here $u = 1$ denotes choose an observation, while $u = 2$ denotes do not choose an observation.
   (b) Show using an identical proof to Theorem 12.2.1 that the region of the belief space $\mathcal{S}_n = \{\pi : \mu^*(\mu) = 1\}$ is convex. Moreover if ((F1),(F2)) hold, show that e_1 belongs to $\mathcal{S}_n$. Also show that $\mathcal{S}_0 \subseteq \mathcal{S}_1 \subseteq \mathcal{S}_2 \ldots$.
   (c) **Optimal Channel sensing.** Another interpretation of the above problem is as follows: The quality x_k of a communication channel is observed in noise. I need to transmit a packet using this channel. If the channel is in state x, I incur a cost $c(x)$ for transmission. Given N slots, when should I transmit?
5. **Optimal measurement selection for a Hidden Markov Model (Multiple stopping problem)**. The following problem generalizes the previous problem as follows. I need to choose the best L observations of a Hidden Markov model in a horizon of length N where $L \leq N$? If I select observation k then I get a reward $\mathbb{E}\{y_k\}$, if I reject the observation then I get no reward. In either case, I use the

observation y_k to update my belief state. (This problem is also called the multiple stopping problem in [67].) Show that Bellman's dynamic programming recursion reads:

$$V_{n+1}(\pi, l) = \max\{r'\pi + \sum_y V_n(T(\pi, y), l-1)\sigma(\pi, y), \\ \sum_y V_n(T(\pi, y), l)\sigma(\pi, y)\}, \quad n = 1, \dots, N$$

with initial condition $V_n(\pi, 0) = 0$, $n = 0, 1, \dots$ and boundary conditions

$$V_n(\pi, n) = r'\pi + \sum_y V_{n-1}(T(\pi, y), n-1)\sigma(\pi, Y), \quad n = 1, \dots, L.$$

The boundary condition says that if I have only n time points left to make n observations, then I need to make an observation at each of these n time points. Obtain a structural result for the optimal measurement selection policy. (Notice that the actions do no affect the evolution of the belief state π, they only affect l, so the problem is simpler than a full blown POMDP.)

6. **Separable POMDPs.** Recall that the action space is denoted as $\mathcal{U} = \{1, 2, \dots, U\}$. In analogy to [35, Chapter 7.4], define a POMDP to be separable if: the exists a subset $\bar{\mathcal{U}} = \{1, 2 \dots, \bar{U}\}$ of the action space $\mathcal{U}$ such that for $u \in \bar{\mathcal{U}}$
   (a) The cost is additively separable: $c(x, u) = \phi(u) + g(x)$ for some scalars $\phi(u)$ and $g(x)$.
   (b) The transition matrix $P_{ij}(u)$ depends only on j. That is the process evolves independently of the previous state.

   Assuming that the actions $u \in \bar{\mathcal{U}}$ are ordered so that $\phi(1) < \phi(2) < \dots < \phi(\bar{U})$, clearly it is never optimal to pick actions $2, \dots, \bar{U}$. So solving the POMDP involves choosing between actions $\{1, \bar{U}+1, \dots, U\}$. So from Theorem 12.2.1, the set of beliefs where the optimal policy $\mu^*(\pi) = 1$ is convex.
   Solving for the optimal policy for which the actions $\{\bar{U}+1, \dots, U\}$ arise is still as complex as a solving a standard POMDP. However, the bounds proposed in Chapter 14 can be used.
   Consider the special case of the above model where $\bar{\mathcal{U}} = \mathcal{U}$ and instead of (a), $c(x, u)$ are arbitrary costs. Then show that the optimal policy is a linear threshold policy.

12.2 Case Study: Bayesian Nash equilibrium of one-shot global game for coordinated sensing

This section gives a short description of Bayesian global games. The ideas involve MLR dominance of posterior distributions and supermodularity and serves as a useful illustration of the structural results developed in the chapter.

We start with some perspective: Recall that in the classical Bayesian social learning, agents act sequentially in time. The global games model that has been studied in economics during the last two decades, considers multiple agents that act simultaneously by predicting the behavior of other agents. The theory of global games was first introduced in [13] as a tool for refining equilibria in economic game theory; see [62] for an excellent exposition.

Global games represent a useful method for decentralized coordination amongst agents; they have been used to model speculative currency attacks and regime change in social systems, see [62, 39, 4]. Applications in sensor networks and cognitive radio appear in [41, 42].

12.2.1 Global Game Model

Consider a continuum of agents in which each agent i obtains noisy measurements $Y^{(i)}$ of an underlying state of nature X. Here

$$Y^{(i)} = X + W^{(i)}, \quad X \sim \pi,\ W^{(i)} \sim p_W(\cdot)$$

Assume all agents have the same noise distribution p_W. Based on its observation $y^{(i)}$, each agent takes an action $u^i \in \{1, 2\}$ to optimize its expected reward

$$R(X, \alpha, u = 2) = X + f(\alpha), \quad R(X, u = 1) = 0 \tag{48}$$

Here $\alpha \in [0, 1]$ denotes the fraction of agents that choose action 2 and $f(\alpha)$ is a user specified function. We will call f the congestion function for reasons explained below.

As an illustrative example, suppose x (state of nature) denotes the quality of a social group and $y^{(i)}$ denotes the measurement of this quality by agent i. The action $u^i = 1$ means that agent i decides not to join the social group, while $u^i = 2$ means that agent i joins the group. The utility function $R(u^i = 2, \alpha)$ for joining the social group depends on α, where α is the fraction of people who decide to join the group. If $\alpha \approx 1$, i.e., too many people join the group, then the utility to each agent is small since the group is too congested and agents do not receive sufficient individual service. On the other hand, if $\alpha \approx 0$, i.e., too few people join the group, then the utility is also small since there is not enough social interaction. In this case the congestion function $f(\alpha)$ would be chosen as a quasi-concave function of α (that increases with α up to a certain value of α and then decreases with α).

Since each agent is rational, it uses its observation $y^{(i)}$ to predict α, i.e., the fraction of other agents that choose action 2. The main question is: *What is the optimal strategy for each agent i to maximize its expected reward?*

12.2.2 Bayesian Nash Equilibrium

Let us now formulate this problem: Each agent chooses its action $u \in \{1, 2\}$ based on a (possibly randomized) strategy $\mu^{(i)}$ that maps the current observation $Y^{(i)}$ to the action u. In a global game we are interested in *symmetric strategies*, i.e., where all choose the same strategy denoted as μ. That is, each agent i deploys the strategy

$$\mu : Y^{(i)} \to \{1, 2\}.$$

(Of course, the action $\mu(Y^{(i)})$ picked by individual agents i depend on their random observation $Y^{(i)}$. So the actions picked are not necessarily identical even though the strategies are identical).

Let $\alpha(x)$ denote the fraction of agents that select action $u = 2$ (go) given the quality of music $X = x$. Since we are considering an infinite number of agents that behave

independently, $\alpha(x)$ is also (with probability 1) the conditional probability that an agent receives signal $Y^{(i)}$ and decides to pick $u = 2$, given X. So

$$\alpha(x) = P(\mu(Y) = 2|X = x). \tag{49}$$

We can now define the Bayesian Nash equilibrium (BNE) of the global game. For each agent i given its observation $Y^{(i)}$, the goal is to choose a strategy to optimize its local reward. That is, agent i seeks to compute strategy $\mu^{(i),*}$ such that

$$\mu^{(i),*}(Y^{(i)}) \in \{1 \text{ (stay) }, 2 \text{ (go) }\} \text{ maximizes } \mathbb{E}[R(X, \alpha(X), \mu^{(i)}(Y^{(i)}))|Y^{(i)}]. \tag{50}$$

Here $R(X, \alpha(X), u)$ is defined as in (48) with $\alpha(X)$ defined in (49).

If such a strategy $\mu^{(i),*}$ in (50) exists and is the same for all agents i, then they constitute a *symmetric* BNE for the global game. We will use the notation $\mu^*(Y)$ to denote this symmetric BNE.

Remark: Since we are dealing with an incomplete information game, players use randomized strategies. If a BNE exists, then a pure (non-randomized) version exists straightforwardly (see Proposition 8E.1, pp.225 in [59]). Indeed, with $y^{(i)}$ denoting realization of random variable $Y^{(i)}$,

$$\mathbb{E}[R(X, \alpha(X), \mu(Y^{(i)}))|Y^{(i)} = y^{(i)}] = \sum_{u=1}^{2} \mathbb{E}[R(X, \alpha(X), u)|Y^{(i)} = y^{(i)}]P(u|Y^{(i)} = y^{(i)}).$$

Since a linear combination is maximized at its extreme values, the optimal (BNE) strategy is to choose $P(u^*|Y^{(i)} = y^{(i)}) = 1$ where

$$u^* = \mu^*(y^{(i)}) = \underset{u\in\{1,2\}}{\operatorname{argmax}}\, \mathbb{E}[R(X, \alpha(X), u)|Y^{(i)} = y^{(i)}]. \tag{51}$$

For notational convenience denote

$$R(y, u) = \mathbb{E}[R(X, \alpha(X), u)|Y^{(i)} = y^{(i)}]$$

12.2.3 Main Result. Monotone BNE

With the above description, we will now give sufficient conditions for the BNE $\mu^*(y)$ to be monotone increasing in y (denoted $\mu^*(y) \uparrow y$). This implies that the BNE is a threshold policy of the form:

$$\mu^*(y) = \begin{cases} 1 & y \le y^* \\ 2 & y > y^* \end{cases}$$

Before proving this monotone structure, first note that $\mu^*(y) \uparrow y$ implies that $\alpha(x)$ in (49) becomes

$$\alpha(x) = P(y > y^*|X = x) = P(x + w > y^*) = P(w > y^* - x) = 1 - F_W(y^* - x)$$

Clearly from (51), a sufficient condition for $\mu^*(y) \uparrow y$ is that

$$R(y, u) = \int R(x, \alpha(x), u)\, p(x|y)dx$$

is supermodular in (y, u) that is

$$R(y, u+1) - R(y, u) \uparrow y.$$

Since $R(X, u = 0)$ it follows that $R(y, 1) = 0$. So it suffices that $R(y, 2) \uparrow y$.

1. What are sufficient conditions on the noise pdf $p_W(\cdot)$, and congestion function $f(\cdot)$ in (48) so that $R(y, 2) \uparrow y$ and so BNE $\mu^*(y) \uparrow y$?
   Clearly sufficient conditions for $R(y, 2) \uparrow y$ are:
   (a) $p(x|y)$ is MLR increasing in y,
   (b) $\mathbb{R}(x, \alpha(x), 2)$ is increasing in x.
   But we know that $p(x|y)$ is MLR increasing in y if the noise distribution is such that $p_W(y - x)$ is TP2 in x, y
   Also $R(x, \alpha(x), 2)$ is increasing in x if its derivative wrt x is positive. That is,

$$\frac{d}{dx} R(x, \alpha(x), 2) = 1 + \frac{df}{d\alpha}\frac{d\alpha}{dx} = 1 + \frac{df}{d\alpha} p_W(y^* - x) > 0$$

   To summarize: The BNE $\mu^*(y) \uparrow y$ if the following two conditions hold:
   (a) $p(y|x) = p_W(y - x)$ is TP2 in (x, y)
   (b)

$$\frac{df}{d\alpha} > -\frac{1}{p_W(y^* - x)}$$

   Note that a sufficient condition for the second condition is that

$$\frac{df}{d\alpha} > -\frac{1}{\max_w p_W(w)}$$

2. Suppose W is uniformly distributed in $[-1.1]$. Then using the above conditions show that a sufficient condition on the congestion function $f(\alpha)$ for the BNE to be monotone is that $df/d\alpha > -2$.
3. Suppose W is zero mean Gaussian noise with variance σ^2. Then using the above conditions show that a sufficient condition on the congestion function $f(\alpha)$ for the BNE to be monotone is that $df/d\alpha > -\sqrt{2\pi}\sigma$.

12.2.4 One-shot HMM Global Game

Suppose that $X_0 \sim \pi_0$, and given X_0, X_1 is obtained by simulating from transition matrix P. The observation for agent i is obtained as the HMM observation

$$Y^{(i)} = X_1 + W^{(i)}, \quad W^{(i)} \sim p_W(\cdot).$$

In analogy to the above derivation, characterize the BNE of the resulting one-shot HMM global game. (This will require assuming that P is TP2.)

Chapter 13

Stopping Time POMDPs for Quickest Change Detection

1. For classical detection theory, a "classic" book is the multi-volume [89].
2. As mentioned in the book, there are two approaches to quickest change detection: Bayesian and minimax. Chapter 13 of the book deals with Bayesian quickest detection which assumes that the change point distribution is known (e.g. phase distribution). The focus of Chapter 13 was to determine the structure of the optimal policy of the Bayesian detector by showing that the problem is a special case of a stopping time POMDP. [91] uses nonlinear renewal theory to analyze the performance of the optimal Bayesian detector.
   The minimax formulation for quickest detection assumes that the change point is either deterministic or has an unknown distribution. For an excellent starting point on performance analysis of change detectors with minimax formulations please see [86] and [75]. The papers [54, 64] gives a lucid description of the analysis of change detection in this framework.
3. **Shiryaev Detection Statistic.** In the classical Bayesian formulation of quickest detection described in §12.2, a two state Markov chain is considered to model geometric distributed change times. Recall (12.6), namely,

$$P = \begin{bmatrix} 1 & 0 \\ 1 - P_{22} & P_{22} \end{bmatrix}, \ \pi_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \tau^0 = \inf\{k : x_k = 1\}. \tag{52}$$

   where $1 - P_{22}$ is the parameter of the geometric prior.
   In classical detection theory, the belief state π_k is written in terms of the *Shiryaev detection statistic* r_k which is defined as follows:

$$r_k \stackrel{\text{defn}}{=} \frac{1}{1 - P_{22}} \times \frac{\pi_k(2)}{1 - \pi_k(2)} \tag{53}$$

   Clearly r_k is an increasing function of $\pi_k(2)$ and so all the monotonicity results in the chapter continue to hold. In particular Corollary 12.2.2 in the book holds for r_k implying a threshold policy in terms of r_k.
   In terms of the Shiryaev statistic r_k, it is straightforward to write the belief state update (HMM filter for 2 state Markov chain) as a function of the likelihood ratio as

follows:

$$\boxed{r_k = \frac{1}{1-p}\,(r_{k-1}+1)\,L(y_k)} \tag{54}$$

where

$$p = 1 - P_{22}, \quad L(y_k) = \frac{B_{2y_k}}{B_{1y_k}} \text{ (likelihood ratio)}$$

In (54) by choosing $p \to 0$, the Shiryaev detection statistic converges to the so called *Shiryaev-Roberts detection statistic*. Note that as $p \to 0$ (equivalently $P_{22} \to 1$), the Markov chain becomes a slow Markov chain. We have analyzed in detail how to track the state of such a slow Markov chain via a stochastic approximation algorithm in Chapter 17 of the book.

The Shiryaev-Roberts detector for change detection reads:

(a) Update the Shiryaev-Roberts statistic

$$r_k = (r_{k-1}+1)\,L(y_k)$$

(b) If $r_k \geq r^*$ then stop and declare a change. Here r^* is a suitably chosen detection threshold.

Please see [74] for a nice survey description of minimax change detection and also the sense in which the above Shiryaev-Roberts detector is optimal.

4. **Classical Bayesian sequential detection.** This problem shows that classical Bayesian sequential detection is a trivial case of the results developed in Chapter 12. Consider a random variable $x \in \{1, 2\}$. So we have a degenerate Markov chain with transition matrix $P = I$. Given noisy observations $y_k \sim B_{xy}$, accumulated over time, the aim is to decide if the underlying state is either 1 or 2.

Taking stop action 1 declares that the state is 1 and stops. Taking stop action 2 declares that the state is 2 and stops. Taking action 3 at time k simply takes another measurement y_{k+1}. The misclassification costs are:

$$c(x = 2, u = 1) = c(x = 1, u = 2) = L.$$

The cost of taking an additional measurement is $c(x, u = 3) = C$. What is the optimal policy $\mu^*(\pi)$?

Since $P = I$, show that the dynamic programming equation reads

$$V(\pi) = \min\{\pi_2 L,\ \pi_1 L,\ C + \sum_y V(T(\pi, y))\sigma(\pi, y)\}$$

$$T(\pi, y) = \frac{B_y \pi}{\mathbf{1}' B_y \pi}, \quad \sigma(\pi, y) = \mathbf{1}' B_y \pi,$$

where $\pi = [\pi(1), \pi(2)]'$ is the belief state. Note that $y \in \mathcal{Y}$ where $\mathcal{Y}$ can be finite or continuum (in which case $\sum$ denotes integration over $\mathcal{Y}$).

From Theorem 12.2.1 we immediately know that the stopping sets

$$\mathcal{R}_1 = \{\pi : \mu^*(\pi) = 1\}, \text{ and } \mathcal{R}_2 = \{\pi : \mu^*(\pi) = 2\}$$

are convex sets. Since the belief state is two dimensional, it belongs to a one dimensional simplex. In terms of the second component $\pi(2)$, $\mathcal{R}_1$ and $\mathcal{R}_2$ are intervals

in the unit interval $[0,1]$. Clearly $\pi(2) = 0 \in \mathcal{R}_1$ and $\pi(2) = 1$ in $\mathcal{R}_2$. Therefore $\mathcal{R}_1 = [0, \pi_1^*]$ and $\mathcal{R}_2 = [\pi_2^*, 1]$ for some $\pi_1^* \leq \pi_2^*$. So the continue region is $[\pi_1^*, \pi_2^*]$. Of course, Theorem 12.2.1 in the book is much more general since it does not require $X = 2$ states and x_k can evolve according to a Markov chain with transition matrix P (whereas in the simplistic setting above, x is a random variable).

5. **Stochastic Ordering of Passage Times for Phase-Distribution.** In quickest detection, we formulated the change point τ^0 to have a *phase type (PH) distribution*. A systematic investigation of the statistical properties of PH-distributions can be found in [71]. The family of all PH-distributions forms a dense subset for the set of all distributions [71] i.e., for any given distribution function F such that $F(0) = 0$, one can find a sequence of PH-distributions $\{F_n, n \geq 1\}$ to approximate F uniformly over $[0, \infty)$. Thus PH-distributions can be used to approximate change points with an arbitrary distribution. This is done by constructing a multi-state Markov chain as follows: Assume state '1' (corresponding to belief e_1) is an absorbing state and denotes the state after the jump change. The states $2, \ldots, X$ (corresponding to beliefs $e_2, \ldots, e_X$) can be viewed as a single composite state that x resides in before the jump. To avoid trivialities, assume that the change occurs after at least one measurement. So the initial distribution π_0 satisfies $\pi_0(1) = 0$. The transition probability matrix is of the form

$$P = \begin{bmatrix} 1 & 0 \\ \underline{P}_{(X-1)\times 1} & \bar{P}_{(X-1)\times(X-1)} \end{bmatrix}. \tag{55}$$

The *first passage time* τ^0 to state 1 denotes the time at which x_k enters the absorbing state 1:

$$\tau^0 = \min\{k : x_k = 1\}. \tag{56}$$

As described in §13.1 of the book, the distribution of τ^0 is determined by choosing the transition probabilities $\underline{P}, \bar{P}$ in (55). The distribution of the absorption time to state 1 is denoted by

$$\nu_k = \mathbb{P}(\tau^0 = k)$$

and given by

$$\nu_0 = \pi_0(1), \quad \nu_k = \bar{\pi}_0' \bar{P}^{k-1} \underline{P}, \quad k \geq 1, \tag{57}$$

where $\bar{\pi}_0 = [\pi_0(2), \ldots, \pi_0(X)]'$.

Definition. Increasing Hazard Rate: A pmf p is said to be increasing hazard rate (IHR) if

$$\frac{\bar{F}_{i+1}}{\bar{F}_i} \downarrow i, \quad \text{where } \bar{F}_i = \sum_{j=i}^{\infty} p_j$$

Aim. Show that if the transition matrix P in (55) is TP2 and initial condition $\pi_0 = e_X$, then the passage time distribution ν_k in (57) satisfies the increasing hazard rate (IHR) property; see [82] for a detailed proof.

6. **Order book high frequency trading and social learning.** Agent based models for high frequency trading with an order book have been studied a lot recently [6]. Agents trade (buy or sell) stocks by exploiting information about the decisions of previous agents (social learning) via an order book in addition to a private (noisy) signal they receive on the value of the stock. We are interested in the following:

(1) Modeling the dynamics of these risk averse agents, (2) Sequential detection of a market shock based on the behavior of these agents.
The agents perform social learning according to the protocol in §13.4.1 of the book. A market maker needs to decide based on the actions of the agents if there is a sudden change (shock) in the underlying value of an asset. Assume that the shock occurs with a phase distributed change time. The individual agents perform social learning with a CVaR social learning filter as in §5.2 of the book. The market maker aims to determine the shock as soon as possible.
Formulate this decision problem as a quickest detection problem. Simulate the value function and optimal policy. Compare it with the market maker's optimal policy obtained when the agents perform risk neutral social learning. See [44] for details.

Chapter 14

Myopic Policy Bounds for POMDPs and Sensitivity

1. To obtain upper and lower bounds to the optimal policy, the key idea was to change the cost vector but still preserve the optimal policy. [48] gives a complete description of this idea. What if a nonlinear cost was subtracted from the costs thereby still keeping the optimal policy the same. Does that allow for larger regions of the belief space where the upper and lower bounds coincide? Is it possible to construct different transition matrices that yield the same optimal policy?
2. **First order dominance of Markov chain sample paths.** In §10.2 of the book we defined the importance concept of copositive dominance to say that if two transition matrices P_1 and P_2 satisfy $P_1 \preceq P_2$ (see Definition 10.2.3), then the one step ahead predicted belief satisfies the MLR dominance property

$$P_1'\pi \leq_r P_2'\pi.$$

   If we only want first order stochastic dominance, then the following condition suffices: Let U denote the $X \times X$ dimensional triangular matrix with elements $U_{ij} = 0, i > j$ and $U_{ij} = 1, i \leq j$.
   (a) Show the following result:

$$P_1 U \geq P_2 U \implies P_1'\pi_1 \geq_s P_2'\pi_2 \text{ if } \pi_1 \geq_s \pi_2.$$

   (b) Consider the following special case of a POMDP. Suppose the prior belief $\pi_0 \in \Pi(X)$ is known. From time 1 onwards, the state is fully observed. How can the structural results in this chapter be used to characterize the optimal policy?
3. In [55] it is assumed that one can construct a POMDP with observation matrices $B(1), B(2)$ such that (i) $T(\pi, y, 2) \geq_r T(\pi, y, 1)$ for each y and (ii) $\sigma(\pi, 2) \geq_s \sigma(\pi, 1)$. Prove that it is impossible to construct an example that satisfies (i) and (ii) apart from the trivial case where $B(1) = B(2)$. Therefore Theorem 14.3.1 does not apply when the transition probabilities are the same and only the observation probabilities are action dependent. For such cases, Blackwell dominance is used.
4. **Extensions of Blackwell dominance idea to POMDPs.**
   Blackwell dominance was used in §14.7 of the book to construct myopic policies that bound the optimal policy of a POMDP. Below we show that Blackwell dominance is

quite finicky when it comes to POMDP structural results. In particular, even with minor changes in the definition, the proof of Theorem 14.7.1 can break down.
Recall that observation matrix $B(2)$ Blackwell dominates $B(1)$ (meaning that $B(2)$ is more accurate than $B(1)$)

$$B(2) \succeq B(1) \text{ if } B(1) = B(2)R$$

for some stochastic matrix R. In Theorem 14.7.1 we assumed that the POMDP has dependency structure $x \to y^{(2)} \to y^{(1)}$. That is, the observation distributions are $B_{x,y^{(2)}}(2) = p(y^{(2)}|x)$, $B_{x,y^{(1)}}(1) = p(y^{(1)}|x)$, and $R_{y^{(2)},y^{(1)}} = p(y^{(1)}|y^{(2)})$.

(a) **Standard Case**: Recall the proof of Theorem 14.7.1 which is written element wise below for improved clarity: The j-th element of the updated belief using the HMM filter is

$$\begin{aligned} T_j(\pi, y^{(1)}, 1) &= \frac{\sum_{y^{(2)}} \sum_i \pi(i) P_{ij}\, p(y^{(2)}|j)\, p(y^{(1)}|y^{(2)})}{\sum_m \sum_{y^{(2)}} \sum_i \pi(i) P_{im}\, p(y^{(2)}|m)\, p(y^{(1)}|y^{(2)})} \\ &= \frac{\sum_{y^{(2)}} \sum_i \pi(i) P_{ij} p(y^{(2)}|j) \frac{\sum_l \sum_m \pi(l) P_{lm} p(y^{(2)}|m)}{\sum_l \sum_m \pi(l) P_{lm}\, p(y^{(2)}|m)}\, p(y^{(1)}|y^{(2)})}{\sum_m \sum_{y^{(2)}} \sum_i \pi(i) P_{im} p(y^{(2)}|m)\, p(y^{(1)}|y^{(2)})} \\ &= \frac{\sum_{y^{(2)}} T_j(\pi, y^{(2)}, 2)\, \sigma(\pi, y^{(2)}, 2) p(y^{(1)}|y^{(2)})}{\sum_{y^{(2)}} \sigma(\pi, y^{(2)}, 2)\, p(y^{(1)}|y^{(2)})} \end{aligned}$$

Then clearly $\frac{\sigma(\pi,y^{(2)},2)\, p(y^{(1)}|y^{(2)})}{\sum_{y^{(2)}} \sigma(\pi,y^{(2)},2)\, p(y^{(1)}|y^{(2)})}$ is a probability measure w.r.t $y^{(2)}$.

(b) **Pre-multiplication definition**: Consider now the following minor modification of the definition of Blackwell dominance. Suppose $B(1) = R\, B(2)$ (that is R premultiplies $B(2)$) instead of the standard definition $B(1) = B(2)\, R$. One would still expect that $B(1)$ is more noisy than $B(2)$. However, the proof of Theorem 14.7.1 no longer holds and there seems no obvious way to salvage it. Of course, if $B(1) = R\, B(2)$ and we make the additional assumption that $B(2) \succeq R$, then clearly $B(2) \succeq B(1)$ and the proof of Theorem 14.7.1 holds.

(c) **State dependent Blackwell dominance**: Consider next the more general POMDP where $p(y^{(1)}|y^{(2)}, x)$ depends on the state x. (In Theorem 14.7.1 and example (a) above this was functionally independent of x.) Then

$$T_j(\pi, y^{(1)}, 1) = \frac{\sum_{y^{(2)}} T_j(\pi, y^{(2)}, 2)\, \sigma(\pi, y^{(2)}, 2)\, p(y^{(1)}|y^{(2)}, j)}{\sum_{y^{(2)}} \sum_m \sigma(\pi, y^{(2)}, 2)\, p(y^{(1)}|y^{(2)}, m)}$$

Now

$$\frac{\sigma(\pi, y^{(2)}, 2)\, p(y^{(1)}|y^{(2)}, j)}{\sum_{y^{(2)}} \sum_m \sigma(\pi, y^{(2)}, 2)\, p(y^{(1)}|y^{(2)}, m)}$$

is no longer a probability measure w.r.t. $y^{(2)}$ since j in the numerator is a fixed index. The proof of Theorem 14.7.1 no longer holds.

(d) Next consider the case where the observation distribution is $p(y_k^{(2)}|x_k, x_{k-1})$ and $p(y^{(1)}|y^{(2)})$. Then the proof of Theorem 14.7.1 continues to hold.

5. **Blackwell dominance implies higher channel capacity.** Show that if $B(1)$ Blackwell dominates $B(2)$, i.e., $B(2) = B(1)Q$ for some stochastic matrix Q, then the capacity of a channel with likelihood probabilities given by $B(1)$ is higher than that with likelihood probabilities $B(2)$. Please see [78] and references therein for a discussion of the relation of Blackwell dominance and channel capacity.
6. **Positively homogeneous concave value function**. Recall a positively homogeneous function $\phi(\cdot)$ satisfies $\phi(\alpha x) = \alpha\phi(x)$ for any $\alpha \geq 0$.
   It is easily to prove that a positively homogeneous function $\phi(\cdot)$ is concave iff

$$\phi(\sum_i x_i) \geq \sum_i \phi(x_i)$$

   Since the value function $V(\pi)$ of a POMDP is positively homogeneous and concave, it follows that

$$\sum_y V(T(\pi, y, u))\sigma(\pi, y, u) = \sum_y V(B_y(u)P'(u)\pi) \geq V(\sum_y B_y(u)P'(u)\pi) = V(P'(u)\pi)$$

   The above result is also obtained as a special case of Blackwell dominance of $B(u) \succeq \frac{1}{Y}\mathbf{1}$ where $\frac{1}{Y}\mathbf{1}$ is the non-informative observation probability matrix.
7. **Combining Copositive and Blackwell Dominance.** Recall that the structural result involving Blackwell dominance deals with action dependent observation probabilities but assumes identical transition matrices for the various actions. Show that copositive dominance and Blackwell dominance can be combined to deal with a POMDP with action dependent transition and observation probabilities of the form:
   Action $u = 1$: P^2, B.
   Action $u = 2$: P, B^2.
   Give numerical examples of POMDPs with the above structure.

Chapter 15

Part IV. Stochastic Approximation and Reinforcement Learning

Here we present three case studies of stochastic approximation algorithms. The first case study deals with online HMM parameter estimation and extends the method described in Chapter 17. The second case study deals with reinforcement learning of equilibria in repeated games. The third case study deals with discrete stochastic optimization (recall §17.4 gave two algorithms) and provides a simple example of such an algorithm.

15.1 Case Study. Online HMM parameter estimation

Recall from Chapter 17 that estimating the parameters of a HMM in real time is motivated by adaptive control of a POMDP. The parameter estimation algorithm can be used to estimate the parameters of the POMDP for a fixed policy; then the policy can be updated using dynamic programming (or approximation) based on the parameters and so on.

This case study outlines several algorithms for recursive estimation of HMM parameters. The reader should implement these algorithms in Matlab to get a good feel for how they work.

Consider the loss function for N data points of a HMM or Gaussian state space model:

$$J_N(\theta) = \mathbb{E}\{\sum_{k=1}^{N} c_\theta(x_k, y_k, \pi_k^\theta)\} \tag{58}$$

where x_k denotes the state, y_k denotes the observation, π_k^θ denotes the belief state, and θ denotes the model variable.

The aim is to determine the model θ that maximizes this loss function.

An offline gradient algorithm operates iteratively to minimize this loss as follows:

$$\theta^{(l+1)} = \theta^{(l)} - \epsilon \nabla_\theta J_N(\theta)\big|_{\theta=\theta^{(l)}} \tag{59}$$

The notation $|_{\theta=\theta^{(l)}}$ above means that the derivatives are evaluated at $\theta = \theta^{(l)}$.

An offline Newton type algorithm operates iteratively as follows:

$$\theta^{(l+1)} = \theta^{(l)} - \left[\nabla_\theta^2 J_N(\theta)\right]^{-1} \nabla_\theta J_N(\theta)\big|_{\theta=\theta^{(l)}} \tag{60}$$

15.1.1 Recursive Gradient and Gauss-Newton Algorithms

A recursive online version of the above gradient algorithm (59) is

$$\boxed{\begin{aligned}\theta_k &= \theta_{k-1} - \epsilon\, \nabla_\theta c_\theta(x_k, y_k, \pi_k^\theta)\big|_{\theta=\theta_{k-1}} \\ \pi_k^{\theta_{k-1}} &= T(\pi_{k-1}^{\theta_{k-1}}, y_k; \theta_{k-1})\end{aligned}} \tag{61}$$

where $T(\pi_{k-1}^{\theta_{k-1}}, y_k; \theta_{k-1})$ is the optimal filtering recursion at time k using prior $\pi_{k-1}^{\theta_{k-1}}$, model θ_{k-1} and observation y_k. The notation $|_{\theta=\theta_{k-1}}$ above means that the derivatives are evaluated at $\theta = \theta_{k-1}$ Finally, ϵ is a small positive step size.

The recursive Gauss Newton algorithm is an online implementation of (60) and reads

$$\boxed{\begin{aligned}\theta_k &= \theta_{k-1} - \mathcal{I}_k^{-1}\, \nabla_\theta c_\theta(x_k, y_k, \pi_k^\theta)\big|_{\theta=\theta_{k-1}} \\ \mathcal{I}_k &= \mathcal{I}_{k-1} + \epsilon \nabla^2 c_\theta(x_k, y_k, \pi_k^\theta)\big|_{\theta=\theta_{k-1}} \\ \pi_k^{\theta_{k-1}} &= T(\pi_{k-1}^{\theta_{k-1}}, y_k; \theta_{k-1})\end{aligned}} \tag{62}$$

Note that the above recursive Gauss Newton is a stochastic approximation algorithm with a matrix step size $\mathcal{I}_k$.

15.1.2 Justification of (61)

Before proceeding with examples, we give a heuristic derivation of (61). Write (59) as

$$\theta_k^{(k)} = \theta_k^{(k-1)} - \epsilon \nabla_\theta J_N(\theta)\big|_{\theta=\theta_k^{(k-1)}}$$

Here the subscript k denotes the estimate based on observations $y_{1:k}$. The superscript (k) denotes the iteration of the offline optimization algorithm.

Suppose that at each iteration k we collect one more observation. Then the above algorithm becomes

$$\theta_k^{(k)} = \theta_{k-1}^{(k-1)} - \epsilon \nabla_\theta J_k(\theta)\big|_{\theta=\theta_{k-1}^{(k-1)}} \tag{63}$$

Introduce the convenient notation

$$\theta_k = \theta_k^{(k)}.$$

Next we use the following two crucial approximations:

- First, make the inductive assumption that θ_{k-1} minimized $J_{k-1}(\theta)$ so that

  $$\nabla_\theta J_{k-1}(\theta)\big|_{\theta=\theta_{k-1}} = 0$$

  Then from (58) it follows that

  $$\nabla_\theta J_k(\theta)\big|_{\theta=\theta_{k-1}} = \nabla_\theta \mathbb{E}\{c_\theta(x_k, y_k, \pi_k^\theta)\}\big|_{\theta=\theta_{k-1}} \tag{64}$$

- Note that evaluating the right hand side of (64) requires running a filter and its derivates wrt θ from time 0 to k for fixed model θ_{k-1}. We want a recursive approximation for this. It is here that the second approximation is used. We revlaute the filtering recursion using the a sequence of available model estimates θ_t, $t = 1, \ldots, k$ at each time t. In other words, we make the approximation

  $$\pi_k^{\theta_{k-1}} = T(\pi_{k-1}^{\theta_{k-1}}, y_k; \theta_k), k = 1, 2, , \ldots, \tag{65}$$

To summarize, introducing approximations (64) and (65) in (63) yields the online gradient algorithm (61). The derivation of the Gauss-Newton algorithm is similar.

15.1.3 Examples of online HMM estimation algorithm

With the algorithms (61) and (62) we can obtain several types of online HMM parameter estimators by choosing different loss functions J in (58). Below we outline two popular choices.

1. Recursive EM algorithm[1]

Recall from the EM algorithm, that the auxiliary likelihood for fixed parameter $\bar{\theta}$ is

$$Q_n(\theta, \bar{\theta}) = \mathbb{E}\{\log(p(x_{0:n}, y_{1:n}|\theta)|y_{1:n}, \bar{\theta})\} = \mathbb{E}\{\sum_{k=1}^{n} \log p_\theta(x_k, y_k|x_{k-1})|y_{1:n}, \bar{\theta}\}$$

With θ^o denoting the true model, θ denoting the model variable, and $\bar{\theta}$ denoting a fixed model value, define

$$J_n(\theta, \bar{\theta}) = \mathbb{E}_{y_{1:n}}\{Q_n(\theta, \bar{\theta})|\theta^o\}.$$

To be more specific, for a HMM, from (4.26), in the notation of (58),

$$c_\theta(y_k, \bar{\theta}) = \sum_{i=1}^{X} \pi^{\bar{\theta}}_{k|n}(i) \ \log B^\theta_{iy_k} + \sum_{i=1}^{X}\sum_{j=1}^{X} \pi^{\bar{\theta}}_{k|n}(i,j) \log P^\theta_{ij}. \tag{66}$$

where P^θ denotes the transition matrix and B^θ is the observation matrix and $\bar{\theta}$ is a fixed model for which the smoothed posterior $\pi^{\bar{\theta}}_{k|n}$ is computed.

Note that c_θ is a reward and not a loss; our aim is to maximize J_n. The idea then is to implement a Gauss-Newton stochastic gradient algorithm for maximizing $J_n(\theta, \bar{\theta})$ for fixed model $\bar{\theta}$, then update $\bar{\theta}$ and so on. This yields the following *recursive EM algorithm*:

1. For $k = n\Delta + 1, \dots (n+1)\Delta$ run

$$\begin{aligned} \theta_k &= \theta_{k-1} + \mathcal{I}_k^{-1} \sum_i \nabla_\theta c_\theta(y_k, \bar{\theta}_{n-1}) \pi_k^{\bar{\theta}_n}(i) \\ \mathcal{I}_k &= \mathcal{I}_{k-1} + \epsilon \sum_i \nabla^2 c_\theta(i, y_k, \pi_k^\theta)\big|_{\theta=\theta_{k-1}} \pi_k^{\bar{\theta}_n}(i) \\ \pi_k^{\bar{\theta}_{n-1}} &= T(\pi_{k-1}^{\bar{\theta}_{n-1}}, y_k; \theta_{k-1}) \quad \text{(HMM filter update)} \end{aligned} \tag{67}$$

   Here $\pi^{\bar{\theta}}_{k|n}$ and $\pi^{\bar{\theta}}_{k|n}(i,j)$ in (66) are replaced by filtered estimates $\pi^{\bar{\theta}}_k$ and $\pi^{\bar{\theta}}_{k-1}(i) P^{\bar{\theta}}_{ij} B^{\bar{\theta}}_j y_k$.

2. Then update $\bar{\theta}_{n+1} = \theta_{(n+1)\Delta}$, set n to $n+1$ and go to step 1.

To ensure that the transition matrix estimates are a valid stochastic matrix, one can parametrize it in terms of spherical coordinates, see (16.18).

As an illustrative example, suppose we wish to estimate the X-dimension vector of state levels $g = (g(1), g(2), \dots, g(X))'$ of a HMM in zero mean Gaussian noise with known variance σ^2. Assume the transition matrix P is known. Then $\theta = g$ and

$$c_\theta(y_k, \bar{\theta}) = -\frac{1}{2\sigma^2} \sum_i \pi_k^{\bar{\theta}}(i) \big(y_k - g(i)\big)^2 + \text{ constant}$$

[1] This name is a misnomer. More accurately the algorithm below is a stochastic approximation algorithm that seeks to approximate the EM algorithm

2. Recursive Prediction Error (RPE)

Suppose g is the vector of state levels of the underlying Markov chain and P the transition matrix. Then the model to estimate is $\theta = (g, P)$. Offline prediction error methods seek to find the model θ that minimizes the loss function

$$J_N(\theta) = \mathbb{E}\{\sum_{k=1}^{N}(y_k - g'\pi_{k|k-1})^2\}$$

So squared prediction error at each time k is

$$c_\theta(x_k, \theta_k, \pi_k^\theta) = \left(y_k - g'P'\pi_{k-1}^\theta\right)^2 \tag{68}$$

Note that unlike (58) there is no conditional expectation in the loss function. Note the key difference compared to the recursive EM. In the recursive EM $c_\theta(x_k, y_k)$ is functionally independent of π^θ and hence the recursive EM does not involve derivatives (sensitivity) of the HMM filter. In comparison, the RPE cost (68) involves derivatives of π_{k-1}^θ with respect to θ. Then the derivatives with respect to θ can be evaluated as in §17.2.

3. Recursive Maximum likelihood

This was discussed in §17.2. The cost function is

$$c_\theta(x_k, \theta_k, \pi_k^\theta) = \log\left[\mathbf{1}'B_{y_k}(\theta)\pi_{k|k-1}^\theta\right]$$

Recursive versions of the method of moment estimation algorithm for the HMM parameters is presented in [61].

15.2 Case Study. Reinforcement Learning of Correlated Equilibria

This case study illustrates the use of stochastic approximation algorithms for learning the correlated equilibrium in a repeated game. Recall in Chapter 17 we used the ordinary differential equation analysis of a stochastic approximation algorithm to characterize where it converges to. For a game, we will show that the stochastic approximation algorithm converges to a differential inclusion (rather than a differential equation). Differential inclusions are generalization of ordinary differential equations (ODEs) and arise naturally in game-theoretic learning, since the strategies according to which others play are unknown. Then by a straightforward Lyapunov function type proof, we show that the differential inclusion converges to the set of correlated equilibria of the game, implying that the stochastic approximation algorithm also converges to the set of correlated equilibria.

15.2.1 Finite Game Model

Consider a finite action static game[2] comprising two players $l = 1, 2$ with costs $c_l(u^{(1)}, u^{(2)})$ where $u^{(1)}, u^{(2)} \in \{1, \ldots, U\}$. Let p and q denote the randomized policies (strategies) of

[2]For notational convenience we assume two players with identical action spaces. All the results below straightforwardly generalize to multiple players and non-identical action spaces.

the two players: $p(i) = \mathbb{P}(u^{(1)} = i)$ and $q(i) = \mathbb{P}(u^{(2)} = i)$. So p, q are U dimensional probability vectors that live in the $U-1$ dimensional unit simplex Π. Then the policies (p^*, q^*) constitute a Nash equilibrium if the following inequalities hold:

$$\boxed{\begin{aligned}\sum_{u^{(1)},u^{(2)}} c_1(u^{(1)},u^{(2)})\, p^*(u^{(1)})\, q^*(u^{(2)}) &\le \sum_{u^{(2)}} c_1(u,u^{(2)})\, q^*(u^{(2)}), \quad u = 1,\dots,U \\ \sum_{u^{(1)},u^{(2)}} c_2(u^{(1)},u^{(2)})\, p^*(u^{(1)})\, q^*(u^{(2)}) &\le \sum_{u^{(1)}} c_2(u^{(1)},u)\, p^*(u^{(1)}), \quad u = 1,\dots,U.\end{aligned}} \tag{69}$$

Equivalently, (p^*, q^*) constitute a Nash equilibrium if for all policies $p, q \in \Pi$,

$$\begin{aligned}\sum_{u^{(1)},u^{(2)}} c_1(u^{(1)},u^{(2)})\, p^*(u^{(1)})\, q^*(u^{(2)}) &\le \sum_{u^{(1)},u^{(2)}} c_1(u^{(1)},u^{(2)})\, p(u^{(1)})\, q^*(u^{(2)}) \\ \sum_{u^{(1)},u^{(2)}} c_2(u^{(1)},u^{(2)})\, p^*(u^{(1)})\, q^*(u^{(2)}) &\le \sum_{u^{(1)},u^{(2)}} c_2(u^{(1)},u^{(2)})\, p^*(u^{(1)})\, q(u^{(2)})\end{aligned} \tag{70}$$

The first inequality in (70) says that if player 1 cheats and deploys policy p instead of p^*, then it is worse off and incurs an higher cost. The second inequality says that same thing for player 2. So in a non-cooperative game, since collusion is not allowed, there is no rational reason for any of the players to unilaterally deviate from the Nash equilibrium p^*, q^*.

By a standard application of Kakutani's fixed point theorem, it can be shown that for a finite action game, at least one Nash equilibrium always exists. However, computing it can be difficult since the above constraints are bilinear and therefore nonconvex.

15.2.2 Correlated Equilibrium

The Nash equilibrium assumes that the player's act independently. The correlated equilibrium is a generalization of the Nash equilibrium. The two players now choose their action from the joint probability distribution $\pi(u^{(1)}, u^{(2)})$ where

$$\pi(i,j) = \mathbb{P}(u^{(1)} = i, u^{(2)} = j).$$

Hence the actions of the players are correlated. Then the policy π^* is said to be a correlated equilibrium if

$$\boxed{\begin{aligned}\sum_{u^{(2)}} c_1(u^{(1)},u^{(2)})\, \pi^*(u^{(1)},u^{(2)}) &\le \sum_{u^{(2)}} c_1(u,u^{(2)})\, \pi^*(u^{(1)},u^{(2)}) \\ \sum_{u^{(1)}} c_2(u^{(1)},u^{(2)}) \pi^*(u^{(1)},u^{(2)}) &\le \sum_{u^{(1)}} c_2(u^{(1)},u)\, \pi^*(u^{(1)},u^{(2)})\end{aligned}} \tag{71}$$

Define the set of correlated equilibria as

$$\mathcal{C} = \left\{ \pi : (71) \text{ holds and } \pi(u^{(1)},u^{(2)}) \ge 0, \sum_{u^{(1)},u^{(2)}} \pi(u^{(1)},u^{(2)}) = 1 \right\} \tag{72}$$

Remark: In the special case where the players act independently, the correlated equilibrium specializes to a Nash equilibrium. Independence implies the joint distribution $\pi^*(u^{(1)}, u^{(2)})$ becomes the product of marginals: so $\pi^*(u^{(1)}, u^{(2)}) = p^*(u^{(1)})q^*(u^{(2)})$. Then clearly (71) reduces to the definition (69) of a Nash equilibrium. Note that the set of correlated equilibria specified by (72) is a convex polytope in π.

Why Correlated Equilibria?

John F. Nash proved in his famous paper [69] that every game with a finite set of players and actions has at least one mixed strategy Nash equilibrium. However, as asserted by Robert J. Aumann [3] in the following extract from [5], "Nash equilibrium does make sense if one starts by assuming that, for some specified reason, each player knows which strategies the other players are using." Evidently, this assumption is rather restrictive and, more importantly, is rarely true in any strategic interactive situation. He adds:

> *"Far from being inconsistent with the Bayesian view of the world, the notion of equilibrium is an unavoidable consequence of that view. It turns out, though, that the appropriate equilibrium notion is not the ordinary mixed strategy equilibrium of Nash (1951), but the more general notion of correlated equilibrium."*
> – Robert J. Aumann

This, indeed, is the very reason why correlated equilibrium [5] best suits and is central to the analysis of strategic decision-making.

There is much to be said about correlated equilibrium; see Aumann [5] for rationality arguments. Some advantages that make it ever more appealing include:

1. *Realistic:* Correlated equilibrium is realistic in multi-agent learning. Indeed, Hart and Mas-Colell observe in [32] that for most simple adaptive procedures, "... there is a natural coordination device: the common history, observed by all players. It is thus reasonable to expect that, at the end, independence among players will not obtain;"

2. *Structural Simplicity:* The correlated equilibria set constitutes a compact convex polyhedron, whereas the Nash equilibria are isolated points at the extrema of this set [70]. Indeed from (72), the set of correlated equilibria is a convex polytope in π.

3. *Computational Simplicity:* Computing correlated equilibrium only requires solving a linear feasibility problem (linear program with null objective function) that can be done in polynomial time, whereas computing Nash equilibrium requires finding fixed points;

4. *Payoff Gains:* The coordination among agents in the correlated equilibrium can lead to potentially higher payoffs than if agents take their actions independently (as required by Nash equilibrium) [5];

5. *Learning:* There is no natural process that is known to converge to a Nash equilibrium in a general non-cooperative game that is not essentially equivalent to exhaustive search. There are, however, natural processes that do converge to correlated equilibria (the so-called law of conservation of coordination [33]), e.g., regret-matching [31].

Existence of a centralized coordinating device neglects the distributed essence of social networks. Limited information at each agent about the strategies of others further

[3] Robert J. Aumann was awarded the Nobel Memorial Prize in Economics in 2005 for his work on conflict and cooperation through game-theoretic analysis. He is the first to conduct a full-fledged formal analysis of the so-called infinitely repeated games.

complicates the process of computing correlated equilibria. In fact, even if agents could compute correlated equilibria, they would need a mechanism that facilitates coordinating on the same equilibrium state in the presence of multiple equilibria—each describing, for instance, a stable coordinated behavior of manufacturers on targeting influential nodes in the competitive diffusion process [90]. This highlights the significance of adaptive learning algorithms that, through repeated interactive play and simple strategy adjustments by agents, ensure reaching correlated equilibrium. The most well-known of such algorithms, fictitious play, was first introduced in 1951 [80], and is extensively treated in [27]. It, however, requires monitoring the behavior of all other agents that contradicts the information exchange structure in social networks. The focus below is on the more recent regret-matching learning algorithms [11, 12, 31, 32].

Figure 15.1 illustrates how the various notions of equilibrium are related in terms of the relative size and inclusion in other equilibria sets. As discussed earlier in this subsection, dominant strategies and pure strategy Nash equilibria do not always exist—the game of "Matching Pennies" being a simple example. Every finite game, however, has at least one mixed strategy Nash equilibrium. Therefore, the "nonexistence critique" does not apply to any notion that generalizes the mixed strategy Nash equilibrium in Figure 15.1. A Hannan consistent strategy (also known as "universally consistent" strategies [29]) is one that ensures, no matter what other players do, the player's average payoff is asymptotically no worse than if she were to play any *constant* strategy for in all previous periods. Hannan consistent strategies guarantee no asymptotic external regrets and lead to the so-called "coarse correlated equilibrium" [63] notion that generalizes the Aumann's correlated equilibrium.

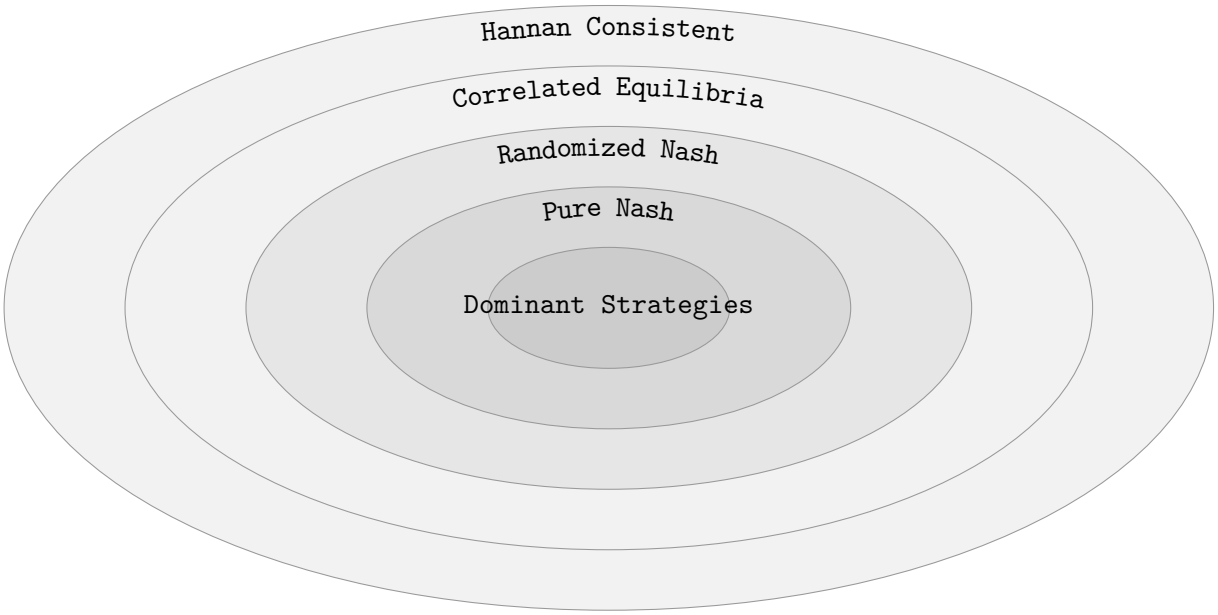

Figure 15.1: Equilibrium notions in non-cooperative games. Enlarging the equilibria set weakens the behavioral sophistication on the player's part to distributively reach equilibrium through repeated plays of the game.

15.2.3 Reinforcement Learning Algorithm

To describe the learning algorithm and the concept of regret, it is convenient to deal with rewards rather than costs. Each agent l has utility reward $r_l(u^{(l)}, u^{-l})$ where $u^{(l)}$ denotes the action of agent l and u^{-l} denotes the action of the other agents. The action space for each agent l is $\{1, 2, \ldots, U\}$. Define the inertia parameter

$$\mu \geq U\big(\max r_l(u, u^{-l}) - \min r_l(u, u^{-l})\big) \tag{73}$$

Each agent then runs the regret matching Algorithm I. Algorithm I assumes that once a decision is made by an agent, it is observable by all other agents. However, agent l does not know the utility function of other agents. Therefore, a learning algorithms such as Algorithm I is required to learn the correlated equilibria.

The assumption that the actions of each agent are known to all other agents can be relaxed; see [32] for "blind" algorithms that do not require this.

Algorithm I Regret Matching Algorithm for Learning Correlated Equilibrium of Game

Each agent l with utility reward $r_l(u^{(l)}, u^{-l})$ independently executes the following:

1. **Initialization:** Choose action $u_0^{(l)} \in \{1, \ldots, U\}$ arbitrarily. Set $R_1^l = 0$.
2. Repeat for $n = 1, 2, \ldots$, the following steps:
   Choose Action: $u_n^{(l)} \in \{1, \ldots, U\}$ with probability

$$\mathbb{P}(u_n^{(l)} = j | u_{n-1}^{(l)} = i, R_n^l) = \begin{cases} \frac{|R_n^l(i,j)|^+}{\mu} & j \neq i, \\ 1 - \sum_{m \neq i} \frac{|R_n^l(i,m)|^+}{\mu} & j = i \end{cases} \tag{74}$$

   where inertia parameter μ is defined in (73) and $|x|^+ \stackrel{\text{defn}}{=} \max\{x, 0\}$.
   Regret Update: Update the $U \times U$ regret matrix R_{n+1}^l as

$$R_{n+1}^l(i,j) = R_n^l(i,j) + \epsilon \left(I\{u_n^{(l)} = i\}(r_l(j, u_n^{-l}) - r_l(i, u_n^{-l})) - R_n^l(u,j) \right). \tag{75}$$

   Here $\varepsilon \ll 1$ denotes a constant positive step size.

Discussion and Intuition of Algorithm I

1. Adaptive Behavior: In (75), ϵ serves as a forgetting factor to foster adaptivity to the evolution of the non-cooperative game parameters. That is, as agents repeatedly take actions, the effect of the old underlying parameters on their current decisions vanishes.
2. Inertia: The choice of μ guarantees that there is always a positive probability of playing the same action as the last period. Therefore, μ can be viewed as an "inertia" parameter: A higher μ yields switching with lower probabilities. It plays a significant role in breaking away from bad cycles. It is worth emphasizing that the speed of convergence to the correlated equilibria set is closely related to this inertia parameter.
3. Better-reply vs. Best-reply: In light of the above discussion, the most distinctive feature of the regret-matching procedure, that differentiates it from other works such as [28], is that it implements a better-reply rather than a best-reply strategy[4]. This inertia assigns

[4]This has the additional effect of making the behavior continuous, without need for approximations [31].

positive probabilities to any actions that are just better. Indeed, the behavior of a regret-matching decision maker is very far from that of a rational decision maker that makes optimal decisions given his (more or less well-formed) beliefs about the environment. Instead, it resembles the model of a reflex-oriented individual that reinforces decisions with "pleasurable" consequences [32].

We also point out the generality of Algorithm I, by noting that it can be easily transformed into the well-known *fictitious play* algorithm by choosing $u_{n+1}^{(l)} = \arg\max_k R_{n+1}^l(i,j)$ deterministically, where $u_n^{(l)} = i$, and the extremely simple *best response* algorithm by further specifying $\epsilon = 1$.

4. Computational Cost: The computational burden (in terms of calculations per iteration) of the regret-matching algorithm does not grow with the number of agents and is hence scalable. At each iteration, each agent needs to execute two multiplications, two additions, one comparison and two table lookups (assuming random numbers are stored in a table) to calculate the next decision. Therefore, it is suitable for implementation in sensors with limited local computational capability.

5. Global performance metric Finally, we introduce a metric for the global behavior of the system. The global behavior z_n at time k is defined as the empirical frequency of joint play of all agents up to period k. Formally,

$$z_n = \sum_{\tau \leq k} (1-\epsilon)^{k-\tau} e_{\mathbf{u}_\tau} \tag{76}$$

where $e_{\mathbf{u}_\tau}$ denotes the unit vector with the element corresponding to the joint play $\mathbf{u}_\tau$ being equal to one. Given z_n, the average payoff accrued by each agent can be straightforwardly evaluated, hence the name global behavior. It is more convenient to define z_n via the stochastic approximation recursion

$$z_n = z_{k-1} + \epsilon \left[e_{\mathbf{u}_n} - z_{k-1} \right]. \tag{77}$$

The global behavior z_n is a system "diagnostic" and is only used for the analysis of the emergent collective behavior of agents. That is, it does not need to be computed by individual agents. In real-life application such as smart sensor networks, however, a network controller can monitor z_n and use it to adjust agents' payoff functions to achieve the desired global behavior.

15.2.4 Ordinary Differential Inclusion Analysis of Algorithm I

Recall from Chapter 17 that the dynamics of a stochastic approximation algorithm can be characterized by an ordinary differential equation obtained by averaging the equations in the algorithm. In particular, using Theorem 17.1.1 of Chapter 17, the estimates generated by the stochastic approximation algorithm converge weakly to the averaged system corresponding to (75) and (77), namely,

$$\begin{aligned} \frac{dR(i,j)}{dt} &= \mathbb{E}_\pi \left\{ I(u_t = i)(r_l(j, u^{-l}) - r_l(i, u^{-l})) - R(i,j) \right\} \\ &= \sum_{u^{-l}} \left[\pi(i|u^{-l}) \left(r_l(j, u^{-l}) - r_l(i, u^{-l}) \right) \right] \pi(u^{-l}) - R(i,j) \\ \frac{dz}{dt} &= \pi(i|u^{-l})\, \pi(u^{-l}) - z \end{aligned} \tag{78}$$

where $\pi(u^{(l)}, u^{-l}) = \pi(u^{-l}|u^{(l)})\pi(u^{(l)})$ is the stationary distribution of the Markov process $(u^{(l)}, u^{-l})$.

Next note that the transition probabilities in (74) of $u_n^{(l)}$ given R_n are conditionally independent of u_n^{-l}. So given R_n, $\pi(i|u^{-l}) = \pi(i)$. So given the transition probabilities in (74), clearly the stationary distribution $\pi(u^{(l)})$ satisfies the linear algebraic equation

$$\pi(i) = \pi(i)\Big[1 - \sum_{j\neq i} \frac{|R(j,i)|^+}{\mu}\Big] + \sum_{j\neq i} \pi(j) \frac{|R(i,j)|^+}{\mu}.$$

which after cancelling out $\pi(i)$ on both sides yields

$$\sum_{i\neq j} \pi(i)|R(i,j)|^+ = \sum_{i\neq j} \pi(j)|R(j,i)|^+ \tag{79}$$

Therefore the stationary distribution π is functionally independent of the inertia parameter μ.

Finally note that as far as player l is concerned, the strategy $\pi(u^{-l})$ is not known. All is known is that $\pi(u^{-l})$ is a valid pmf. So we can write the averaged dynamics of the regret matching Algorithm I as

$$\left.\begin{aligned} &\frac{dR(i,j)}{dt} \in \sum_{u^{-l}} \Big[\pi(i)\left(r_l(j,u^{-l}) - r_l(i,u^{-l})\right)\Big]\pi(u^{-l}) - R(i,j) \\ &\frac{dz}{dt} \in \pi(i)\,\pi(u^{-l}) - z \\ &\sum_{i\neq j} \pi(i)|R(i,j)|^+ = \sum_{i\neq j} \pi(j)|R(j,i)|^+ \end{aligned}\right\} \pi(u^{-l}) \in \text{ valid pmf} \tag{80}$$

The above averaged dynamics constitute an algebraically constrained ordinary differential inclusion.[5] We refer the reader to [10, 11] for an excellent exposition of the use of differential inclusions for analyzing game theoretical type learning algorithms.

Remark: The asymptotics of a stochastic approximation algorithm is typically captured by an ordinary differential equation (ODE). Here, although agents observe u^{-l}, they are oblivious to the strategies $\pi(u^{-l})$ from which u^{-l} has been drawn. Different strategies $\pi(u^{-l})$ result in different trajectories of R_n. Therefore, R_t and z_t are specified by a differential inclusions rather than ODEs .

15.2.5 Convergence of Algorithm I to the set of correlated equilibria

The previous subsection says that the regret matching Algorithm I behaves asymptotically as an algebraically constrained differential inclusion (80). So we only need to analyze the behavior of this differential inclusion to characterize the behavior of the regret matching algorithm.

[5] Differential inclusions are a generalization of the concept of ordinary differential equations. A generic differential inclusion is of the form $dx/dt \in \mathcal{F}(x,t)$, where $\mathcal{F}(x,t)$ specifies a family of trajectories rather than a single trajectory as in the ordinary differential equations $dx/dt = F(x,t)$.

Theorem 1. *Suppose every agent follows the "regret-matching"Algorithm I. Then as $t \to \infty$: (i) $R(t)$ converges to the negative orthant in the sense that*

$$\text{dist}\left[R(t), \mathbb{R}^-\right] = \inf_{\boldsymbol{r} \in \mathbb{R}^-} \left\|R(t) - \boldsymbol{r}\right\| \Rightarrow 0; \tag{81}$$

(ii) $z(t)$ converges to the correlated equilibria set $\mathcal{C}$ in the sense that

$$\text{dist}[z(t), \mathcal{C}] = \inf_{\boldsymbol{z} \in \mathcal{C}} \|z(t) - \boldsymbol{z}\| \Rightarrow 0. \tag{82}$$

The proof below shows the simplicity and elegance of the ordinary differential equation (inclusion) approach for analyzing stochastic approximation algorithm. Just a few elementary lines based on the Lyapunov function yields the proof.

Proof. Define the Lyapunov function

$$V(R) = \frac{1}{2}\big(\text{dist}[R, \mathbb{R}^-]\big)^2 = \frac{1}{2}\sum_{i,j}\big(|R(i,j)|^+\big)^2. \tag{83}$$

Evaluating the time-derivative and substituting for $dR(i,j)/dt$ from (80) we obtain

$$\begin{aligned}
\frac{d}{dt}V(R) &= \sum_{i,j}|R(i,j)|^+ \cdot \frac{d}{dt}R(i,j) \\
&= \sum_{i,j}|R(i,j)|^+\Big[(r_l(j,u^{-l}) - r_l(i,u^{-l}))\pi(i) - R(i,j)\Big] \\
&= \underbrace{\sum_{i,j}|R(i,j)|^+(r_l(j,u^{-l}) - r_l(i,u^{-l}))\pi(i)}_{=0 \text{ from (79)}} - \sum_{i,j}|R(i,j)|^+R(i,j) \\
&= -2V(R).
\end{aligned} \tag{84}$$

In the last equality we used

$$\sum_{i,j}|R(i,j)|^+R(i,j) = \sum_{i,j}\big(|R(i,j)|^+\big)^2 = 2V(R). \tag{85}$$

This completes the proof of the first assertion, namely that Algorithm I eventually generates regrets that are non-positive.

To prove the second assertion, from Algorithm I, the elements of the regret matrix are

$$\begin{aligned}
R_k(i,j) &= \epsilon\sum_{\tau \le k}(1-\epsilon)^{k-\tau}\Big[r_l(j,u_\tau^{-l}) - r_l(u_\tau^{(l)}, u_\tau^{-l})\Big]I(u_\tau^{(l)} = i) \\
&= \sum_{u^{-l}} z(i,u^{-l})[r_l(j,u^{-l}) - r_l(i,u^{-l})]
\end{aligned} \tag{86}$$

where $z(i,u^{-l})$ denotes the empirical distribution of agent l choosing action i and the rest playing u^{-l}. On any convergent subsequence $\{z_{\underline{k}}\}_{\underline{k} \ge 0} \to \pi$, then

$$\lim_{k\to\infty} R_k(i,j) = \sum_{u^{-l}} \pi(i,u^{-l})[r_l(j,u^{-l}) - r_l(i,u^{-l})] \tag{87}$$

where $\pi(i, u^{-l})$ denotes the probability of agent l choosing action i and the rest playing u^{-l}. The first assertion of the theorem proved that the regrets converge to non-positive values (negative orthant). Therefore (87) yields that

$$\sum_{u^{-l}} \pi(i, u^{-l})[r_l(j, u^{-l}) - r_l(i, u^{-l})] \leq 0$$

implying that π is a correlated equilibrium. □

15.2.6 Extension to switched Markov games

Consider the case now where rewards $r_l(u^{(l)}, u^{-l})$ evolve according to an unknown Markov chain θ_n. Such a time varying game can result from utilities in a social network evolving with time or the number of players changing with time. The reward for agent l is now $r_l(u^{(l)}, u^{-l}, \theta_n)$. The aim is to track the set of correlated equilibria $\mathcal{C}(\theta_n)$; that is use the regret matching algorithm I so that agents eventually deploy strategies from $\mathcal{C}(\theta_n)$. If θ_n evolves with transition matrix $I + \epsilon^2 Q$ (where Q is a generator), then it is on a slower time scale than the dynamics of the regret matching Algorithm I. Then a more general proof in the spirit of Theorem 17.3.3 yields that the regret matching algorithm can track the time varying correlated equilibrium set $\mathcal{C}(\theta_n)$. Moreover, in analogy to §17.3.4, if the transition matrix for θ_n is $I + \epsilon Q$, then the asymptotic dynamics are given by a switched Markov differential inclusion, see [47, 68].

15.3 Stochastic Search-Ruler Algorithm

We discuss two simple variants of Algorithm 25 that require less restrictive conditions for convergence than condition (O). Assume $c_n(\theta)$ are uniformly bounded for $\theta \in \Theta$. Neither of the algorithms given below are particularly novel; but they are useful from a pedagogical point of view.

It is convenient to normalize the objective (17.43) as follows: Let $\alpha \leq c_n(\theta) \leq \beta$ where α denotes a finite lower bound and $\beta > 0$ denotes a finite upper bound. Define the normalized costs $m_n(\theta)$ as

$$m_n(\theta) = \frac{c_n(\theta) - \alpha}{\beta - \alpha}, \quad \text{where } 0 \leq m_n(\theta) \leq 1. \tag{88}$$

Then the stochastic optimization problem (17.43) is equivalent to

$$\theta^* = \arg\min_{\theta \in \Theta} m(\theta) \text{ where } m(\theta) = \mathbb{E}\{m_n(\theta)\} \tag{89}$$

since scaling the cost function does not affect the minimizing solution. Recall $\Theta = \{1, 2, \ldots, S\}$.

Define the loss function

$$Y_n(\theta, u_n) = I\left(m_n(\theta) - u_n\right) \text{ where } I(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \tag{90}$$

Here u_n is a independent uniform random number in $[0, 1]$. The uniform random number u_n is a stochastic ruler against which the candidate $m_n(\theta)$ is measured. The result was

originally used in devising stochastic ruler optimization algorithms [1] – although here we propose a more efficient algorithm than the stochastic ruler. Applying Algorithm 25 to the cost function $\mathbb{E}\{Y_n(\theta, u_n)\}$ defined in (90) yields the following stochastic search-ruler algorithm:

Algorithm II Stochastic Search-Ruler

Identical to Algorithm 25 with $c_n(\theta_n)$ and $c_n(\tilde{\theta}_n)$ replaced by $Y_n(\theta_n, u_n)$ and $Y_n(\tilde{\theta}_n, \tilde{u}_n)$. Here u_n and $\tilde{u}_n$ are independent uniform random numbers in $[0, 1]$.

Analogous to Theorem 17.4.1 we have the following result:

Theorem 1. *Consider the discrete stochastic optimization problem (17.43). Then the Markov chain $\{\theta_n\}$ generated by Algorithm II has the following property for its stationary distribution π_∞:*

$$\frac{\pi_\infty(\theta^*)}{\pi_\infty(\theta)} = \frac{m(\theta)}{m(\theta^*)} \frac{(1 - m(\theta^*))}{(1 - m(\theta))} > 1. \tag{91}$$

The theorem says that Algorithm II is attracted to set the global minimizers $\mathcal{G}$. It spends more time in $\mathcal{G}$ than any other candidates. The restrictive condition (O) is not required for Algorithm II to be attracted to $\mathcal{G}$. Theorem 1 gives an explicit representation of the discriminative power of the algorithm between the optimizer θ^* and any other candidate θ in terms of the normalized expected costs $m(\theta)$ and $m(\theta^*)$. Algorithm II is more efficient than the stochastic ruler algorithm of [3] when the candidate samples are chosen with equal probability. The stochastic ruler algorithm of [3] has asymptotic efficiency $\pi(\theta^*)/\pi(\theta) = (1 - m(\theta^*))/(1 - m(\theta))$. So Algorithm II has the additional improvement in efficiency due to the additional multiplicative term $m(\theta)/m(\theta^*)$ in (91).

Variance reduction using common random numbers: A more efficient implementation of Algorithm II can be obtained by using variance reduction based on common random numbers (discussed in Appendix A.2.4 of the book) as follows: Since u_n is uniformly distributed in $[0, 1]$, so is $1-u_n$. Similar to Theorem 1 it can be shown that the optimizer θ^* is the minimizing solution of the following stochastic optimization problem $\theta^* = \arg\min_\theta \mathbb{E}\{Z_n(\theta, u_n)\}$ where

$$Z_n(\theta, u_n) = \frac{1}{2} \left[Y_n(\theta, u_n) + Y_n(\theta, 1 - u_n) \right] \tag{92}$$

where the normalized sample cost $m_n(\theta)$ is defined in (89). Applying Algorithm II with $Z_n(\theta_n, u_n)$ and $Z_n(\tilde{\theta}_n, u_n)$ replacing $Y_n(\theta_n, u_n)$ and $Y_n(\tilde{\theta}_n, u_n)$, respectively, yields the variance reduced search-ruler algorithm.

In particular, since the indicator function $I(\cdot)$ in (90) is a monotone function of its argument, it follows that $\mathrm{Var}\{Z_n(\theta, u_n)\} \leq \mathrm{Var}\{Y_n(\theta, u_n)\}$. As a result one would expect that the stochastic optimization algorithm using Z_n would converge faster.

Proof. We first show that θ^* defined in (89) is the minimizing solution of the stochastic optimization problem $\theta^* = \arg\min_\theta \mathbb{E}\{Y_n(\theta, u_n)\}$. Using the smoothing property of conditional expectations (3.11) yields

$$\begin{aligned} \mathbb{E}\{I\left(m_n(\theta) - u_n\right)\} &= \mathbb{E}\{\mathbb{E}\{I\left(m_n(\theta) - u_n\right) | m_n(\theta)\}\} \\ &= \mathbb{E}\{\mathbb{P}(u_n < m_n(\theta))\} = \mathbb{E}\{m_n(\theta)\} = m(\theta) \end{aligned}$$

The second equality follows since expectation of an indicator function is probability, the third equality holds because u_n is a uniform random number in [0,1] so that $\mathbb{P}(u_n < a) = a$ for any a in $[0, 1]$.

Next we show that the state process $\{\theta_n\}$ generated by Algorithm II is a homogeneous, aperiodic, irreducible, Markov chain on the state space Θ with transition probabilities

$$P_{ij} = P(\theta_n = j | \theta_{n-1} = i) = \frac{1}{S-1} m(i)(1 - m(j)).$$

That the process $\{\theta_n\}$ is a homogeneous aperiodic irreducible Markov chain follows from its construction in Algorithm II – indeed θ_n only depends probabilistically on θ_{n-1}. From Algorithm II, given candidate i and its associated cost $Y_n(i, u_n)$, candidate j is accepted if its associated cost $\tilde{Y}_n(j, \tilde{u}_n)$ is smaller. So

$$\begin{aligned} P_{ij} &= \frac{1}{S-1} P(\tilde{Y}_n(j, \tilde{u}_n) < Y_n(i, u_n)) \\ &= \frac{1}{S-1} P(m_n(j) < \tilde{u}_n) P(m_n(i) > u_n) \end{aligned}$$

Finally, for this transition matrix, it is easily verified that

$$\pi_\infty(\theta) = \kappa(1 - m(\theta)) \prod_{j \neq \theta} m(j) \tag{93}$$

is the invariant distribution where κ denotes a normalization constant. Hence

$$\frac{\pi_\infty(\theta^*)}{\pi_\infty(\theta)} = \frac{m(\theta)}{m(\theta^*)} \frac{(1 - m(\theta^*))}{(1 - m(\theta))} = \frac{1/m(\theta^*) - 1}{1/m(\theta) - 1} > 1$$

since $m(\theta^*)$ is the global minimum and therefore $m(\theta^*) < m(\theta)$ for $\theta \in \Theta - \mathcal{G}$. □

Bibliography

[1] M. Alrefaei and S. Andradottir. A modification of the stochastic ruler method for discrete stochastic optimization. *European Journal of Operational Research*, 133:160–182, 2001.

[2] B. D. O. Anderson and J. B. Moore. *Optimal filtering.* Prentice Hall, Englewood Cliffs, New Jersey, 1979.

[3] S. Andradottir. Accelerating the convergence of random search methods for discrete stochastic optimization. *ACM Transactions on Modelling and Computer Simulation*, 9(4):349–380, Oct. 1999.

[4] G.M. Angeletos, C. Hellwig, and A. Pavan. Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks. *Econometrica*, 75(3):711–756, 2007.

[5] R. J. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55(1):1–18, 1987.

[6] M. Avellaneda and S. Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, Apr 2008.

[7] B. Bahrami, K. Olsen, P. Latham, A. Roepstorff, G. Rees, and C. Frith. Optimally interacting minds. *Science*, 329(5995):1081–1085, 2010.

[8] T. Basar and G. J. Olsder. *Dynamic Noncooperative Game Theory.* SIAM Series in Classics in Applied Mathematics, 1991.

[9] M. Basin. On optimal filtering for polynomial system states. *Journal of dynamic systems, measurement, and control*, 125(1):123–125, 2003.

[10] M. Benaim, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

[11] M. Benaim, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions, Part II: Applications. *Mathematics of Operations Research*, 31(3):673–695, 2006.

[12] A. Cahn. General procedures leading to correlated equilibria. *International Journal of Game Theory*, 33(1):21–40, Dec. 2004.

[13] H. Carlsson and E. van Damme. Global games and equilibrium selection. *Econometrica*, 61(5):989–1018, Sept. 1993.

[14] A. R. Cassandra. Tony's POMDP page. http://www.cs.brown.edu/research/ai/pomdp/index.html.

[15] D. Castañon. Optimal search strategies in dynamic hypothesis testing. *Systems, Man and Cybernetics, IEEE Transactions on*, 25(7):1130–1138, 1995.

[16] K. Chen and S. Ross. An adaptive stochastic knapsack problem. *European Journal of Operational Research*, 239(3):625–635, 2014.

[17] L. Chen, P. O. Arambel, and R. K. Mehra. Estimation Under Unknown Correlation: Covariance Intersection Revisited. *IEEE Transactions on Automatic Control*, 47(11):1879–1882, 11 2002.

[18] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[19] A. Doucet, A. Logothetis, and V. Krishnamurthy. Stochastic sampling algorithms for state estimation of jump Markov linear systems. *IEEE Transactions on Automatic Control*, 45(2):188–202, Feb. 2000.

[20] R. J. Elliott, L. Aggoun, and J. B. Moore. *Hidden Markov Models – Estimation and Control*. Springer-Verlag, New York, 1995.

[21] R. J. Elliott and V. Krishnamurthy. New finite dimensional filters for estimation of discrete-time linear Gaussian models. *IEEE Transactions on Automatic Control*, 44(5):938–951, May 1999.

[22] J. Evans and R.J. Evans. Image-enhanced multiple model tracking. *Automatica*, 35(11):1769–1786, 1999.

[23] M. Fanaswala and V. Krishnamurthy. Syntactic models for trajectory constrained track-before-detect. *IEEE Transactions on Signal Processing*, 62(23):6130–6142, 2014.

[24] M. Fanaswalla and V. Krishnamurthy. Detection of anomalous trajectory patterns in target tracking via stochastic context-free grammars and reciprocal process models. *IEEE Journal on Selected Topics Signal Processing*, 7(1):76–90, Feb. 2013.

[25] A. A. Fel'dbaum. *Optimal control systems*. Academic Press, 1965.

[26] J. Filar and K. Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[27] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games*. MIT Press, 1998.

[28] D. Fudenberg and D. K. Levine. Conditional universal consistency. *Games and Economic Behavior*, 29(1):104–130, Oct. 1999.

[29] D. Fudenberg and D.K. Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, 1995.

[30] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

[31] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[32] S. Hart and A. Mas-Colell. A reinforcement procedure leading to correlated equilibrium. In G. Debreu, W. Neuefeind, and W. Trockel, editors, *Economic Essays: A Festschrift for Werner Hildenbrand*, pages 181–200. Springer, 2001.

[33] S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to nash equilibrium. *American Economic Review*, 93(5):1830–1836, December 2003.

[34] M. Hernandez-Gonzalez and M. Basin. Discrete-time filtering for nonlinear polynomial systems over linear observations. *Intenrational journal of systems science*, 45(7):1461–1472, 2014.

[35] D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research*, volume 2. McGraw-Hill, 1984.

[36] N. Higham and L. Lin. On pth roots of stochastic matrices. *Linear Algebra and its Applications*, 435(3):448–463, 2011.

[37] D. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[38] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, pages 263–291, 1979.

[39] L. Karp, I.H. Lee, and R. Mason. A global game with strategic substitutes and complements. *Games and Economic Behavior*, 60:155–175, 2007.

[40] J. Keilson and A. Kester. Monotone matrices and monotone Markov processes. *Stochastic Processes and their Applications*, 5(3):231–241, 1977.

[41] V. Krishnamurthy. Decentralalized activation in dense sensor networks via global games. *IEEE Transactions on Signal Processing*, 56(10):4936–4950, 2008.

[42] V. Krishnamurthy. Decentralized spectrum access amongst cognitive radios-an interacting multivariate global game-theoretic approach. *IEEE Transactions on Signal Processing*, 57(10):3999–4013, Oct. 2009.

[43] V. Krishnamurthy and F. Vazquez Abad. Gradient based policy optimization of constrained unichain Markov decision processes. In S. Cohen, D. Madan, and T. Siu, editors, *Stochastic Processes, Finance and Control: A Festschrift in Honor of Robert J. Elliott*. World Scientific, 2012. http://arxiv.org/abs/1110.4946.

[44] V. Krishnamurthy and S. Bhatt. Sequential detection of market shocks with risk-averse cvar social sensors. *IEEE Journal Selected Topics in Signal Processing*, 2016.

[45] V. Krishnamurthy and R.J. Elliott. Filters for estimating Markov modulated poisson processes and image based tracking. *Automatica*, 33(5):821–833, May 1997.

[46] V. Krishnamurthy and W. Hoiles. Online reputation and polling systems: Data incest, social learning and revealed preferences. *IEEE Transactions Computational Social Systems*, 1(3):164–179, Jan. 2015.

[47] V. Krishnamurthy, M. Maskery, and G. Yin. Decentralized activation in a ZigBee-enabled unattended ground sensor network: A correlated equilibrium game theoretic analysis. *IEEE Transactions on Signal Processing*, 56(12):6086–6101, December 2008.

[48] V. Krishnamurthy and U. Pareek. Myopic bounds for optimal policy of POMDPs: An extension of Lovejoy's structural results. *Operations Research*, 62(2):428–434, 2015.

[49] V. Krishnamurthy and C. Rojas. Reduced complexity HMM filtering with stochastic dominance bounds: A convex optimization approach. *IEEE Transactions on Signal Processing*, 62(23):6309–6322, 2014.

[50] V. Krishnamurthy and B. Wahlberg. POMDP multiarmed bandits – structural results. *Mathematics of Operations Research*, 34(2):287–302, May 2009.

[51] P. R. Kumar and P. Varaiya. *Stochastic systems – Estimation, Identification and Adaptive Control*. Prentice-Hall, New Jersey, 1986.

[52] A. Logothetis and A. Isaksson. On sensor scheduling via information theoretic criteria. In *Proc. American Control Conf.*, pages 2402–2406, San Diego, 1999.

[53] A. Logothetis and V. Krishnamurthy. Expectation maximization algorithms for MAP estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing*, 47(8):2139–2156, August 1999.

[54] G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, pages 1897–1908, 1971.

[55] W. S. Lovejoy. Some monotonicity results for partially observed Markov decision processes. *Operations Research*, 35(5):736–743, Sept.-Oct. 1987.

[56] W. S. Lovejoy. Suboptimal policies with bounds for parameter adaptive decision processes. *Operations Research*, 41(3):583–599, 1993.

[57] J. Ma, L. Xu, and M. I. Jordan. Asymptotic convergence rate of the em algorithm for gaussian mixtures. *Neural Computation*, 12(12):2881–2907, 2000.

[58] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[59] A. Mas-Colell, M. Whinston, and J. Green. *Microeconomic Theory*. Oxford, 1995.

[60] F. Matejka and A. McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *The American Economic Review*, 105(1):272–298, 2014.

[61] R. Mattila, V. Krishnamurthy, and B. Wahlberg. Recursive identification of chain dynamics in hidden markov models using non-negative matrix factorization. In *Proceedings of IEEE CDC 2015*, 2015.

[62] S. Morris and H. S. Shin. Global games: Theory and applications. In *Advances in Economic Theory and Econometrics: Proceedings of Eight World Congress of the Econometric Society*, pages 56–114. Cambridge University Press, 2000.

[63] H. Moulin and J.-P. Vial. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4):201–221, 1978.

[64] G. B. Moustakides. Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14:1379–1387, 1986.

[65] A. Muller and D. Stoyan. *Comparison Methods for Stochastic Models and Risk*. Wiley, 2002.

[66] M. Naghshvar and T. Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013.

[67] T. Nakai. The problem of optimal stopping in a partially observable markov chain. *Journal of Optimization Theory and Applications*, 45(3):425–442, 1985.

[68] O. Namvar, V. Krishnamurthy, and G. Yin. Distributed tracking of correlated equilibria in regime switching noncooperative games. *IEEE Transactions on Automatic Control*, 58(10):2435–2450, 2013.

[69] J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, Sep. 1951.

[70] R. Nau, S. Canovas, and P. Hansen. On the geometry of Nash equilibria and correlated equilibria. *International Journal of Game Theory*, 32(4):443–453, 2004.

[71] M. F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications*. Marcel Dekker, N.Y., 1989.

[72] R. Olfati-Saber and R. M. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):1520–1533, Sept. 2004.

[73] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[74] A. Polunchenko and A. Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and computing in applied probability*, 14(3):649–684, 2012.

[75] H. V. Poor and O. Hadjiliadis. *Quickest Detection*. Cambridge University Press, 2008.

[76] J. Quah and B. Strulovici. Comparative statics, informativeness, and the interval dominance order. *Econometrica*, 77(6):1949–1992, 2009.

[77] J. Quah and B. Strulovici. Aggregating the single crossing property. *Econometrica*, 80(5):2333–2348, 2012.

[78] M. Raginsky. Shannon meets blackwell and le cam: Channels, codes, and statistical experiments. In *Proceedings of 2011 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 1220–1224. IEEE, 2011.

[79] I. Rapoport and Y. Oshman. A Cramér-Rao-type estimation lower bound for systems with measurement faults. *IEEE Transactions on Automatic Control*, 50(9):1234–1245, 2005.

[80] J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, Sep. 1951.

[81] S. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, San Diego, California., 1983.

[82] J. G. Shanthikumar. DFR property of first-passage times and its preservation under geometric compounding. *The Annals of Probability*, pages 397–406, 1988.

[83] C. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.

[84] L. Smith and P. Sorensen. Informational herding and optimal experimentation. Economics Papers 139, Economics Group, Nuffield College, University of Oxford, 1997.

[85] D. D. Sworder, P. F. Singer, D. Doria, and R. G. Hutchins. Image-enhanced estimation methods. *Proceedings of the IEEE*, 81(6):797–812, June 1993.

[86] A. Tartakovsky and G. Moustakides. State-of-the-art in bayesian changepoint detection. *Sequential Analysis*, 29(2):125–145, 2010.

[87] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[88] P. Tichavsky, C. H. Muravchik, and A. Nehorai. Posterior Cramér-Rao bounds for discrete-time nonlinear filtering. *IEEE Transactions on Signal Processing*, 46(5):1386–1396, May 1998.

[89] H.L. Van Trees. *Detection, Estimation and Modulation Theory*. John Wiley & Sons, 1968.

[90] V. Tzoumas, C. Amanatidis, and E. Markakis. A game-theoretic analysis of a competitive diffusion process over social networks. In *Internet and Network Economics*, volume 7695, pages 1–14. Springer, 2012.

[91] V. Veeravalli and T. Banerjee. Quickest change detection. *Academic press library in signal processing: Array and statistical signal processing*, 3:209–256, 2013.

[92] W. Whitt. Multivariate monotone likelihood ratio and uniform conditional stochastic order. *Journal Applied Probability*, 19:695–701, 1982.

[93] L. Xu and M. I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

Index