ΕΘΝΙΚΟ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΓΙΑ ΤΟ ΠΡΟΒΛΗΜΑ ONE-ARMED BANDIT

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΛΑΜΠΡΟΥ ΑΠΟΣΤΟΛΟΣ

Μεταπτυχιακό Πρόγραμμα
Στατιστικής και Επιχειρησιακής Έρευνας

ΕΠΙΒΛΕΠΩΝ: ΜΠΟΥΡΝΕΤΑΣ ΑΠΟΣΤΟΛΟΣ

Αθήνα
Νοέμβριος 2014

Θα ήθελα να ευχαριστήσω:

- τον Καθηγητή Μπουρνέτα Απόστολο
  για την βοήθεια που μου προσέφερε καθ΄ όλη την διάρκεια
  της εκπόνησης της διπλωματικής μου εργασίας αλλά και
  για την συμβολή του στην εμπλούτισση των γνώσεων μου
  κατά τη διάρκεια των σπουδών μου στο μεταπτυχιακό πρόγραμμα

- τον Αναπληρωτή Καθηγητή Μηλολιδάκη Κωνσταντίνο
  και την Επίκουρο Καθηγήτρια Λουκία Μελιγκοτίδου
  για την συμμετοχή τους στην τριμελή επιτροπή

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Σκοπός της εργασία αυτής είναι η ανάλυση βέλτιστων πολιτικών και η μελέτη αντίστοιχων υπολογιστικών μεθόδων σε προβλήματα κατηγορίας $one-armed\ bandit$. Μια διαδικασία $one-armed\ bandit$ είναι ειδική περίπτωση μιας γενικότερης κατηγορίας προβλημάτων δυναμικής βελτιστοποίησης, που είναι γνωστή στη βιβλιογραφία ως μοντέλα $multi-armed\ bandit$ και σχετίζονται με την δυναμική κατανομή πόρων σε εναλλακτικές δραστηριότητες υπό συνθήκες ελλιπούς πληροφόρησης εναλλακτικών πειραμάτων. Τα προβλήματα που θα μελετήσουμε αναφέρονται σε πεπερασμένο και άπειρο ορίζοντα.

Στο κεφάλαιο 2 γίνεται μια σύντομη εισαγωγή στο πρόβλημα $multi-armed\ bandit$ ακολουθώντας την προσέγγιση του [Ros83]. Αρχικά ορίζεται η διαδικασία $bandit$, περιγράφεται το πρόβλημα βελτιστοποίησης και η μοντελοποίηση του μέσω δυναμικού προγραμματισμού, και παρουσιάζεται η δομή της βέλτιστς πολιτικής μέσω της βασικής έννοιας του δείκτη $Gittins$.

Στο κεφάλαιο 3 μελετάται η εφαρμογή ενός μοντέλου $one-armed\ bandit$ για τη μελέτη ενός προβλήματος βέλτιστης διακοπής υπό συνθήκες ελλιπούς πληροφόρησης σε πεπερασμένο ορίζοντα. Παρουσιάζεται η μοντελοποίηση του πειράματος μέσω Μπεϋζιανού δυναμικού προγραμματισμού και αναλύεται η δομή της βέλτιστης πολιτικής. Η ανάλυση αυτού του κεφαλαίου βασίζεται στην εργασία [Bur03].

Στο κεφάλαιο 4 μελετάται η ειδική περίπτωση του Κεφαλαίου 3 που αναφέρεται σε δοκιμές $Bernoulli$, για την οποία οι εξισώσεις βελτιστοποίησης και η βέλτιστη πολιτική παίρνουν απλούστερη μορφή που βασίζεται σε κατώφλια. Επίσης υλοποιείται ένας υπολογιστικός αλγόριθμος για τον αριθμητικό υπολογισμό των κατωφλίων, της βέλτιστης πολιτικής και συζητείται η συμπεριφορά του αλγορίθμου ως προς την ακρίβεια.

Στο κεφάλαιο 5 αναλύεται μια εναλλακτική προσέγγιση του Κεφαλαίου 3, υπό άπειρο ορίζοντα και κριτήριο αποπληθωρισμένου αναμενόμενου κέρδους. Το πρόβλημα διατυπώνεται ως ειδική περίπτωση του προβλήματος $multi-armed\ bandit$ για το οποίο η βέλτιστη πολιτική έχει δομή που βασίζεται στον δείκτη $Gittins$. Επίσης παρουσιάζεται ένα μοντέλο βελτισοποίησης (πρόβλημα επανεκκίνησης) για τον υπολογισμό του δείκτη $Gittins$ στην ειδική περίπτωση των δοκιμών $Bernoulli$.Η ανάλυση αυτού του κεφαλαίου βασίζεται στην

εργασία [KD85].

Τέλος στο κεφάλαιο 6 υλοποιούνται δύο ακόμη υπολογιστικοί αλγόριθμοι: πρώτα μια τροποποιημένη μορφή του αλγορίθμου του Κεφαλαίου 4 για την περίπτωση του άπειρου ορίζονταμε αποπληθωρισμένο κέρδος και έπειτα ένας νέος αλγόριθμος υπολογισμού της βέλτιστης πολιτικής μέσω υπολογισμού του δείκτη *Gittins* ως προβλήματος επανεκκίνησης. Γίνεται επίσης σύγκριση των δύο αλγορίθμων ως προς την ακρίβεια των αποτελεσμάτων.

Κεφάλαιο 2

Το πρόβλημα *multi − armed bandit*

2.1 Εισαγωγή

Αρχικά θα ορίσουμε το πλαίσιο των *bandit* διαδικασιών. Έστω ότι έχουμε n ανεξάρτητα διαθέσιμα προγράμματα ($projects$) τα οποία όταν ενεργοποιούνται εξελίσονται σύμφωνα με μια Μαρκοβιανή διαδικασία. Σε κάθε περίοδο αποφασίζουμε να ενεργοποιήσουμε ένα πρόγραμμα βασιζόμενοι σε όλες τις διαθέσιμες πληροφορίες που έχουμε για τις καταστάσεις όλων των προγραμμάτων εκείνη την στιγμή. Αν i είναι η κατάσταση του προγράμματος που επιλέγουμε να ενεργοποιήσουμε, τότε η αναμενόμενη αμοιβή μιας περιόδου είναι $R(i)$ ενώ με πιθανότητα P_{ij} οδηγείται στην κατάσταση j. Τα υπόλοιπα $n-1$ προγράμματα παραμένουν ανενεργά και δεν αλλάζουν κατάσταση. Επιπλέον μας δίνεται η δυνατότητα σε κάποια περίοδο να αποσυρθούμε κερδίζοντας μια αμοιβή M. Η ανάλυση μας θα στηριχθεί στο [Ros83]. Θα αναφέρουμε την περίπτωση που έχουμε ένα διαθέσιμο πρόγραμμα και την επιλογή της απόσυρσης με τελική αμοιβή M. Έπειτα θα περιγράψουμε την βέλτιστη πολιτική όταν έχουμε n ανεξάρτητα προγράμματα σε συνδυασμό με την επιλογή της απόσυρσης και τέλος θα ορίσουμε το $multi-armed\ bandit$ πρόβλημα και θα αναφέρουμε τον δείκτη $Gittins$ ως τρόπο εύρεσης βέλτιστης πολιτικής.

2.2 Διαδικασία *Bandit* με 1 πρόγραμμα

Σε κάθε περίοδο αποφασίζουμε αν θα ενεργοποιήσουμε το μοναδικό διαθέσιμο πρόγραμμα ή αν θα αποσυρθούμε, σταματώντας έτσι το πρόβλημα. Αν επιλέξουμε να συνεχίσουμε και βρισκόμαστε στην κατάσταση i λαμβάνουμε μία ορισμένη αμοιβή $R(i)$ και οδηγούμαστε στην κατάσταση j με πιθανότητα P_{ij}. Σε διαφορετική περίπτωση αν η απόφαση μας είναι να σταματήσουμε λαμβάνουμε τελική αμοιβή M και η διαδικασία τελειώνει. Ορίζουμε ως $V(i:M)$ το μέγιστο αναμενόμενο α-εκπτωτικό κέρδος, όταν η αρχική μας κατάσταση είναι i, όπου α, $0<a<1$, ο εκπτωτικός παράγοντας και M σε άπειρο ορίζοντα η αμοιβή απόσυρσης. Με αυτό τον συμβολισμό δηλώνουμε την εξάρτηση του $V(i:M)$ από την τελική αμοιβή M. Η εξίσωση βελτιστοποίησης διαμορφώνεται ως εξής:

$$V(i,M) = \max[M, R(i) + \alpha \sum_k P_{ik} \cdot V(k,M)] \tag{2.1}$$

Μια βασική ιδιότητα της βέλτιστης πολιτικής είναι ότι αν η επιλογή της απόσυρσης είναι βέλτιστη σε μια δοσμένη κατάσταση με τελική αμοιβή M, τότε είναι επίσης βέλτιστη στην ίδια κατάσταση για οποιαδήποτε τελική αμοιβή μεγαλύτερη του M.

Λήμμα 2.1 *Για δοσμένο i, η $V(i,M) - M$ είναι φθίνουσα ως προς M.*

Απόδειξη. Η απόδειξη δίνεται στο [Ros83], και βασίζεται στο ότι η $V(i,M)$ είναι κοίλη συνάρτηση του M. □

Από την ιδιότητα αυτή προκύπτει το παρακάτω θεώρημα, το οποίο περιγράφει την βέλτιστη πολιτική.

Θεώρημα 2.1 *Στην κατάσταση i ,είναι βέλτιστο να αποσυρθούμε άν $M \geqslant M(i)$ και να συνεχίσουμε αν $M \leqslant M(i)$, όπου $M(i) = \min[M : V(i,M) = M]$.*

Άρα η $M(i)$ είναι εκείνη η τιμή για την οποία υπάρχει αδιαφορία μεταξύ της επιλογής να συνεχίσουμε και της επιλογής να αποσυρθούμε όταν είμαστε στην κατάσταση i. Αυτό το αποτέλεσμα θα μας οδηγήσει και στην εύρεση βέλτιστης πολιτικής για το πρόβλημα όπου έχουμε n ανεξάρτητα διαθέσιμα προγράμματα.

2.3 Διαδικασία *Bandit* με n προγράμματα

Στην συγκεκριμένη περίπτωση αυτό που αλλάζει είναι η διάσταση του προβλήματος καθώς έχουμε n ανεξάρτητα διαθέσιμα προγράμματα. Επομένως σε κάθε περίοδο αφού παρατηρήσουμε την κατάσταση αποφασίζουμε αν θα συνεχίσουμε διαλέγοντας ένα πρόγραμμα το οποίο θα ενεργοποιήσουμε ή αν θα αποσυρθούμε λαμβάνοντας τελική αμοιβή M. Έτσι στην επόμενη περίοδο η κατάσταση από $\underline{i} = (i_1, ..., i_j, ..., i_n)$, καθώς έχουμε n διαδικασίες, θα αλλάξει σε $\underline{i'} = (i_1, ..., k, ..., i_n)$ με πιθανότητα $P_{i_j k}$ αν επιλέξουμε να ενεργοποιήσουμε το πρόγραμμα j. Έτσι έχουμε τις εξισώσεις βελτιστοποίησης

$$V(\underline{i}, M) = \max\left[M, \max_j O^j(V(\underline{i}, M))\right] \tag{2.2}$$

όπου το $O^j(V(\underline{i}, M))$ είναι το αναμενόμενο κέρδος αν για την επόμενη περίοδο ενεργοποιηθεί το πρόγραμμα j και μετά εφαρμοστεί η βέλτιστη πολιτική, δηλαδή

$$O^j(V(\underline{i}, M)) = R(i_j) + \alpha \cdot \sum_k V(i_1, ..., i_{j-1}, k, i_{j+1}, ...i_n : M) P_{i_j, k}.$$

Η μεγαλύτερη δυσκολία αυτού του μοντέλου είναι η επίλυση του. Για παράδειγμα, με την μέθοδο διαδοχικών προσεγγίσεων αν κάθε πρόγραμμα έχει K πιθανές καταστάσεις, τότε στο μοντέλο των n προγραμμάτων θα έχουμε K^n καταστάσεις. Παρ΄ όλα αυτά, το συγκεκριμένο μοντέλο διασπάται σε n διαδικασίες με 1 πρόγραμμα. Συγκεκριμένα η βέλτιστη πολιτική του σύνθετου μοντέλου (2.2) θα έχει την μορφή του απλού μοντέλου (2.1). Αν βρισκόμαστε στην κατάσταση $\underline{i} = (i_1, ..., i_j, ..., i_n)$ είναι βέλτιστο να αποχωρήσουμε αν

$$M(i_j) \leqslant M, \quad \forall j = 1, ..., n$$

και να ενεργοποιήσουμε το πρόγραμμα j αν

$$\max_k M(i_k) = M(i_j) > M.$$

Στην απόδειξη ότι η παραπάνω πολιτική είναι βέλτιστη χρησιμοποιούνται τα παρακάτω λήμματα τα οποία περιλαμβάνουνε ιδιότητες του $V(\underline{i}, M)$ ως συνάρτηση του.

Λήμμα 2.2 *Για σταθερό $\underline{i}$, το $V(\underline{i}, M)$ είναι αύξουσα και κυρτή συνάρτηση του M.*

Απόδειξη. Η απόδειξη δίνεται στο [Ros83]. □

Λήμμα 2.3 *Για αρχική δοσμένη κατάσταση $\underline{i}$, έστω T_M ο βέλτιστος χρόνος αποχώρησης υπό την παραπάνω πολιτική με τελική αμοιβή M. Τότε για όλα τα M για τα οποία υπάρχουν οι $\frac{\partial V(\underline{i},M)}{\partial M}$ έχουμε*

$$\frac{\partial V(\underline{i}, M)}{\partial M} = E(\alpha^{T_M} | X_0 = \underline{i})$$

Απόδειξη. Η απόδειξη παρατίθεται στο [Ros83]. □

Θεώρημα 2.2 *Η συνάρτηση $V(\underline{i}, M)$ δίνεται από τη σχέση*
$V(\underline{i}, M) = C - \int_M^C \prod_{k=1}^n \frac{\partial}{\partial m} V(i_k, m) dm \ , \ M \leqslant C$

Στην κατάσταση $\underline{i}$, η βέλτιστη πολιτική είναι να αποχωρούμε όταν $M(i_j) \leqslant M$ για όλα τα $j = 1, ..., n$ και να συνεχίζουμε το πρόγραμμα j όταν $M(i_j) = \max_k M(i_k) > M$.

Απόδειξη. Η απόδειξη δίνεται στο [Ros83], και βασίζεται στο να δειχθεί ότι η συγκεκριμένη μορφή της $V(\underline{i}, M)$ ικανοποιεί τις εξισώσεις βελτιστοποίησης. □

Έτσι η βέλτιστη πολιτική για το σύνθετο μοντέλο (2.2) προκύπτει από την μελέτη n απλών μοντέλων (2.1) ενεργοποιώντας κάθε φορά το πρόγραμμα με την μέγιστη τιμή αδιαφορίας και αποφασίζοντας έπειτα, συγκρίνοντας την με την τελική αμοιβή M.

Παρακάτω παρουσιάζεται μια εναλλακτική έκφραση της τιμής αδιαφορίας που θα μας δώσει χρήσιμα συμπεράσματα και θα μας βοηθήσει για την συνέχεια στους υπολογισμούς. Έστω ότι έχουμε μόνο ένα πρόγραμμα. Όταν $M = M(i)$ παρουσιάζεται αδιαφορία στο αν θα συνεχίσουμε ή θα αποχωρήσουμε. Επομένως, για οποιαδήποτε θετική στιγμή αποχώρησης T

$$M(i) \geqslant E[\text{εκπτωτικό κέρδος πριν το } T] + M(i)E(\alpha^T)$$

με ισότητα για την βέλτιστη πολιτική συνέχισης.

Έτσι λύνοντας ως προς $(1-\alpha)M(i)$ καταλήγουμε στο παρακάτω αποτέλεσμα

$$(1-\alpha)M(i) = max_{T>0} \frac{E[\text{εκπτωτικό κέρδος πριν το } T]}{E[(1-\alpha^T)/(1-\alpha)]} =$$

$$= \max_{T>0} \frac{E[\text{εκπτωτικό κέρδος πριν το } T]}{E[1 + \alpha + ... + \alpha^{T-1}]}$$

Δηλαδή,

$$(1-\alpha)M(i) = \max_{T>0} \frac{E[\text{εκπτωτικό κέρδος πριν το } T]}{E[\text{εκπτωτικός χρόνος πριν το } T]} \tag{2.3}$$

όπου οι μέσες τιμές είναι με βάση την αρχική κατάσταση i.

Τα παραπάνω αποτελέσματα συνεπάγονται πολύ σημαντική απλούστευση του προβλήματος $multi-armed\ bandit$, καθώς ανάγουν την επίλυση ενός προβλήματος βελιστοποίησης με $n-$διάστατο χώρων καταστάσεων σε n ανεξάρτητα προβλήματα με μονοδιάστατο χώρο καταστάσεων.

Η τιμή αδιαφορία $M(i)$ παίζει πολύ σημαντικό ρόλο στην περιγραφή και βέβαια στον υπολογισμό της βέλτιστης πολιτικής, και αναφέρεται στην βιβλιογραφία ως δείκτης $Gittins$ καθώς η απόδειξη της δομής της βέλτιστης πολιτικής με βάση τους δείκτες $M(i)$ αποδείχθηκε για πρώτη φορά στις εργασίες [GJ74] και [Git79].

Πρέπει εδώ να τονιστεί ότι ο υπολογισμός του δείκτη $Gittins$ με βάση την (2.3) δεν αποτελεί απλούστευση, καθώς είναι ουσιαστικά ισοδύναμος με το πρόβλημα εύρεσης της βέλτιστης πολιτικής απόσυρσης που ορίστηκε στην ενότητα 2.2 (2.1). Πραγματικά η μεγιστοποίηση στην (2.3) γίνεται όχι απλά ως προς της ντετερμινιστικές τιμές του T, αλλά ως προς όλους τους στοχαστικούς χρόνους διακοπής, η κατανομή των οποιών μπορεί να εξαρτάται από την προηγούμενη ιστορία καταστάσεων της διαδικασίας. Επομένως η (2.3) αντιστοιχεί ουσιαστικά στην εύρεση της βέλτιστης πολιτικής διακοπής στο πρόβλημα (2.1), και συγκεκριμένα στην εύρεση μιας συνάρτησης $M(i)$ τέτοιας ώστε για κάθε i, αν λυθεί το πρόβλημα βέλτιστης διακοπής με $M = M(i)$ η βέλτιστη απόφαση στην κατάσταση i είναι αδιάφορη μεταξύ συνέχισης και απόσυρσης.

Στην βιβλιογραφία έχουν προταθεί διάφορες προσεγγίσεις για τον υπολογισμό του δείκτη $Gittins$. Από αυτές θα παρούσιάσουμε στο Κεφάλαιο 5 την προσέγγιση που βασίζεται στο πρόβλημα επανεκκίνησης, προσαρμοσμένη στο συγκεκριμένο πρόβλημα διακοπής δοκιμών $Bernoulli$ υπό ελλιπή πληροφόρηση.

Στα επόμενα Κεφάλαια της εργασίας η μελέτη θα επικεντρωθεί σε διαδικασίες ενός μόνο προγράμματος. Τα προβλήματα αυτού του τύπου αναφέρονται στην βιβλιογραφία ως προβλήματα $one-armed\ bandit$, και είναι ισοδύναμα με τα προβλήματα βέλτιστης διακοπής.

Θα μελετήσουμε ένα πρόβλημα $one-armed\ bandit$ στο πλαίσιο της βελτιστοποίησης υπό ελλιπή πληροφόρηση.

Το πρόβλημα αυτό θα μελετηθεί για τις περιπτώσεις πεπερασμένου και άπειρου ορίζοντα. Πρέπει να τονιστεί ότι η περίπτωση του πεπερασμένου ορίζοντα εισάγει ουσιαστικές δυσκολίες στην ανάλυση. Συγκεκριμένα για την περίπτωση του άπειρου ορίζοντα με κριτήριο αναμενόμενου συνολικού αποπληθωρισμένου κέρδους σχετίζεται με τον υπολογισμό του δείκτη $Gittins$ και είναι χρήσιμη και για το πρόβλημα $multi-armed\ bandit$ όπως συζητήσαμε παραπάνω.

Από την άλλη πλευρά για το κριτήριο του πεπερασμένου ορίζοντα, αφενός η εύρεση της βέλτιστης πολιτικής στο πρόβλημα διακοπής δεν είνι χρονικά ομογενής και δεν ισοδυναμεί με τον υπολογισμό του δείκτη $Gittins$. Αφετέρου, το πρόβλημα $multi-armed\ bandit$ σε πεπερασμένο ορίζοντα, δεν διασπάται γενικά σε επιμέρους μονοδιάστατα προβλήματα και η βέλτιστη πολιτική δεν βασίζεται σε δείκτες και για το λόγο αυτό η επίλυση του προβλήματος έχει στην γενική περίπτωση πολύ μεγάλη πολυπλοκότητα.

Κεφάλαιο 3

Το πρόβλημα $one - armed\ bandit$ για την επιλογή υπό ελλιπή πληροφόρηση

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα αναφερθούμε σε προβλήματα δυναμικής επιλογής μεταξύ δύο πειραμάτων ή γενικότερα στατιστικών πληθυσμών. Το ένα είναι γνωστό σε εμάς, δηλαδή γνωρίζουμε την πιθανότητα επιτυχίας του, ενώ το άλλο που είναι άγνωστο αντιστοιχεί σε μια διεργασία η οποία θα πρέπει να αξιολογηθεί. Τέτοια μοντέλα βρίσκουν εφαρμογή σε κλινικές μελέτες, σε βιομηχανικές διεργασίες, ακόμα και σε τυχερές μηχανές ενός καζίνο (π.χ. κουλοχέρηδες). Τα αποτελέσματα του κάθε πειράματος είναι συσχετισμένα με μια αμοιβή. Σε πρώτη φάση επιθυμούμε να μεγιστοποιήσουμε την αναμενόμενη τιμή της συνολικής αμοιβής που επιτυγχάνεται στον πεπερασμένο ορίζοντα. Θα περιγράψουμε σε αυτό το κεφάλαιο την περίπτωση κατά την οποία τα κέρδη ανήκουν στην μονοπαραμετρική εκθετική οικογένια κατανομών. Ακολουθώντας την προσέγγιση του Μπεϋζιανού δυναμικού πργραμματισμού, θα ορίσουμε μία εκ των προτέρων κατανομή για το άγνωστο *bandit* και θα παρουσιάσουμε την δομή του προβλήματος μεγιστοποίησης. Επισημαίνεται ότι είναι ισοσύναμο με την ελαχιστοποίηση ενός κατάλληλα καθοριμένου προβλήματος ζημιάς. Τέλος θα δώσουμε κάποια βασικά αποτελέματα για την βέλτιστη πολιτική μιας απλής περίπτωσης σε πεπερασμένο ορίζοντα [Bur97]. Η ανάλυση βασίζεται στην εργασία [Bur03].

3.2 Δομή του μοντέλου

Έστω E_1 και E_2 δύο στατιστικά πειράματα. Κάθε πείραμα συνδέεται με μία παράμετρο θ_i η οποία ανήκει σε ένα χώρο Θ, με μία ακολουθία από τυχαίες μεταβλητές $X_i, Y_{i1}, Y_{i2}, ...$ όπου τα Y_{ij} αντιπροσωπεύουν τα αποτελέματα κάθε πειράματος την $j-$οστή φορά που χρησιμοποιήθηκε και τα X_i υποδηλώνουν τα αποτελέσματα του κάθε πειράματος ξεχωριστά. Δοθείσας της τιμής του $\theta_i = \theta$ οι τυχαίες μεταβλητές $X_i, Y_{i1}, Y_{i2}, ...$ είναι ανεξάρτητες και ισόνομες με συνάρτηση πυκνότητας πιθανότητας $f(x|\theta)$ σε σχέση με ένα μη εκφυλισμένο μέτρο v. Έστω τώρα $\mu(\theta)$ και $\sigma^2(\theta)$ η αναμενόμενη τιμή και η διακύμανση, αντίστοιχα, μια τυχαίας μεταβλητής X με συνάρτηση πυκνότητας πιθανότητας $f(x|\theta)$.

Θα κάνουμε τώρα κάποιες υποθέσεις για το μοντέλο μας.

Υπόθεση 3.1 *Η συνάρτηση πυκνότητας πιθανότητας $f(x|\theta)$ ανήκει στην μονοπαραμετρική εκθετική οικογένεια κατανομών με φυσική παράμετρο θ και τύπο*

$$f(x|\theta) = e^{\theta x - \psi(\theta) + s(x)} \tag{3.1}$$

Υπόθεση 3.2 *Ο παραμετρικός χώρος είναι ένα διάστημα της μορφής $\Theta = (\underline{\theta}, \overline{\theta})$, με άκρα που μπορεί να είναι και $\pm\infty$, και ικανοποιούν τις εξής συνθήκες*

$$z_1 = \inf_{\theta\in\Theta} \psi''(\theta) > 0 \tag{3.2}$$

$$z_2 = \sup_{\theta\in\Theta} \psi''(\theta) < \infty \tag{3.3}$$

Υπόθεση 3.3 *Η παράμετρος θ_1 είναι γνωστή εκ των προτέρων και η θ_2 είναι άγνωστη. Ακολουθώντας την Μπεϋζιανή προσέγγιση, υποθέτουμε ότι η θ_2 είναι τυχαία μεταβλητή με εκ των προτέρων κατανομή $H_0(\theta)$, όπου $\theta \in \Theta$.*

Υπόθεση 3.4 *Υποθέτουμε ότι $\underline{\theta} < \theta_1 < \overline{\theta}$, όπου $\underline{\theta}$, $\overline{\theta}$ είναι τέτοια ώστε $(\mu(\underline{\theta}), \mu(\overline{\theta})) = \{\mu(\theta) : \theta \in \Theta\}$.*

Θα χρησιμοποιήσουμε την φυσική αναπαράσταση της παραμέτρου της εκθετικής οικογένειας κατανομών ([Cox74] *p.* 28). Είναι γνωστό από την μονοπαραμετρική οικογένεια κατανομών ότι $\mu(\theta) = \psi'(\theta)$ και ότι $\sigma^2(\theta) = \psi''(\theta)$ και ότι $\mu(\theta)$ είναι αυστηρά αύξουσα συνάρτηση ως προς θ. Επίσης αν $\theta_1 \leq \underline{\theta}$ ($\theta_1 \geq \overline{\theta}$) τότε το πρόβλημα είναι αδιάφορο καθώς θα επιλέγουμε σε κάθε βήμα το E_2 (E_1).

Έστω $t (n = N - t)$ ο αριθμός των δοκιμών οι οποίες έχουν ήδη γίνει (ή μένει να γίνουν). Την στιγμή $t = 0$, έχουμε $X_1 \sim f(x|\theta_1)$ και $X_2 \sim f(x|\theta_2)$ όπου το θ_1 είναι γνωστό και το $\theta_2 \sim H_0(\theta)$.

Ένα δείγμα παρατηρήσεων μεγέθους k_i από το αντίστοιχο πείραμα E_i θα χαρακτηρίζεται από το $d_i(k_i) = (y_{i1}, ..., y_{ik_i})$ για $i = 1, 2$. Ορίζουμε $\underline{k} = (k_1, k_2)$ και $\underline{d}(\underline{k}) = (d_1(k_1), d_2(k_2))$. Εφόσον το θ_1 είναι γνωστό, οι μελλοντικές παρατηρήσεις από το E_1 πείραμα, $Y_{1,k_1+1}, Y_{1,k_1+2}, ...$, δοθέντος του $d_1(k_1)$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με συνάρτηση πυκνότητας πιθανότητας $f(x|\theta_1)$. Αντίστοιχα για το θ_2 οι μελλοντικές παρατηρήσεις $Y_{2,k_1+1}, Y_{2,k_1+2}, ...$ δοθέντος του $d_2(k_2)$ και $\theta_2 = \theta$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με συνάρτηση πυκνότητας πιθανότητας $f(x|\theta)$. Αν μας δίνεται μόνο το $d_2(k_2)$, τότε το θ_2 είναι τυχαία μεταβλητή με εκ των υστέρων κατανομή $H(\theta|d_2(k_2))$ η οποία ορίζεται ως εξής:

$$dH(\theta|d_2(k_2)) = \frac{\widetilde{f}(d_2(k_2)|\theta)dH_0(\theta)}{\widetilde{f}(d_2)(k_2)|H_0} = \frac{f(y_{2,k_2}|\theta)dH(\theta|d_2(k_2-1))}{\int_\Theta f(y_{2,k_2}|\theta)dH(\theta|d_2(k_2-1))}, \tag{3.4}$$

όπου $d_i(k_i) = (d_i(k_i-1), y_{i,k_i})$, $H(\theta|d_2(0)) = H_0(\theta)$ και $\widetilde{f}(d_2(k_2)|\theta)$ (αντίστοιχα $\widetilde{f}(d_2(k_2)|H_0$» δηλώνει την από κοινού συνάρτηση πυκνότητας πιθανότητας του δείγματος $d_2(k_2)$, δεδομένου ότι $\theta_2 = \theta$ (αντίστοιχα αν μας δίνεται η εκ των προτέρων κατανομή H_0).

Επομένως δοθέντος του $d_2(k_2)$ οι μελλοντικές παρατηρήσεις του πειράματος E_2 είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με κατανομή η οποία καθορίζεται από την περιθώρια συνάρτηση πυκνότητας πιθανότητας:

$$f(x|d_2(k_2)) = \int_\Theta f(x|\theta) dH(\theta|d_2(k_2)). \tag{3.5}$$

Ο εκτιμητής *Bayes* για την μ(θ_2) δεδομένου του δείγματος $d_2(k_2)$ είναι ίσος με:

$$\hat{\mu}(d_2(k_2)) = E_{H(\cdot|d_2(k_2))}[\mu(\theta_2)] = E_{f(\cdot|d_2(k_2))}[Y_{2,k_2+1}]. \tag{3.6}$$

Για την περίπτωση της μονοπαραμετρικής εκθετικής οικογένειας κατανομών είναι γνωστό ότι η εκ των υστέρων κατανομή $H(\theta|d_2(k_2))$ και η περιθώρια πυκνότητα $f(x|d_2(k_2))$, (3.4) και (3.5) αντίστοιχα, καθορίζονται μοναδικά από το επαρκές στατιστικά διάνυσμα $(k_2, \overline{y}_{2,k_2})$, όπου $\overline{y}_{2,k} = (1/k)\sum_{j=1}^{k} y_{2,j}$. Επομένως στις (3.4)−(3.6) μπορούμε χωρίς βλάβη της γενικότητας να υπολογίσουμε $d_2(k_2) = (k_2, \overline{y}_{2,k_2})$.

Δεδομένου ότι $d_2(k_2 - 1) = (k_2 - 1, y)$ και $Y_{2,k_2} = y_{2,k}$ καταλήγουμε στο εξής συμπέρασμα για το $d_2(k_2)$:

$$d_2(k_2|d_2(k_2-1), y_{2,k}) = (k, \frac{k-1}{k}y + \frac{1}{k}y_{2,k}) = (k, m(k-1, y, y_{2,k})), \tag{3.7}$$

όπου $m(k-1, y, y_{2,k}) = \frac{ky+x}{k+1}$.

Μια πολιτική με ορίζοντα N ορίζεται ως $\pi = (\pi(0), \pi(1), ..., \pi(N-1))$, όπου

$$\pi(t) = \pi(t|d_1(k_1(t,\pi)), d_2(k_2(t,\pi))) \tag{3.8}$$

είναι ίση με α_1 ή α_2, ανάλογα με το αν την στιγμή t η πολιτική μας υποδεικνύει να πάρουμε δείγμα από το πείραμα E_1 ή E_2 αντίστοιχα έχοντας έτσι το ανάλογο

$$k_i(t,\pi) = \sum_{j=0}^{t-1} \mathbf{1}_{\pi(j)=\alpha_i} \tag{3.9}$$

Η επίδοση μιας πολιτικής π εκφράζεται από το

$$S(t,\pi) = \sum_{j=0}^{t-1} Y_{\pi(j),k_{\pi(j)}(j,\pi)}, \tag{3.10}$$

και τις αναμενόμενες τιμές

$$\mathbf{E}_\theta S(t,\pi) = \mathbf{E}[S(t,\pi)|\theta_2 = \theta] = \mu(\theta_1)\mathbf{E}_\theta k_1(t,\pi) + \mu(\theta)\mathbf{E}_\theta k_2(t,\pi), \tag{3.11}$$

$$M(t, H_0, \pi) = \mathbf{E}_{H_0}[\mathbf{E}_\theta S(t,\pi)] = \mathbf{E}_{f(\cdot|H_0)}[S(t,\pi)]. \tag{3.12}$$

Μία πολιτική π^* είναι βέλτιτη για ορίζοντα N και αρχική εκ των προτέρων κατανομή $H_0(\theta)$, του θ_2, αν και μόνο αν

$$M(N, H_0, \pi^*) = \max_\pi M(N, H_0, \pi), \tag{3.13}$$

όπου το μέγιστο λαμβάνει υπ'οψιν όλες τις διαδοχικές πολιτικές τις οποίες ορίσαμε προηγουμένως.

Μια ισοδύναμη περιγραφή του προβλήματος με βάση την συνάρτηση απώλειας $L(\theta, i)$, η οποία μετρά την αναμενόμενη απώλεια ενός βήματος όταν η άγνωστη παράμετρος είναι ίση με θ και παίρνουμε δείγμα από το πείραμα E_i, είναι η ακόλουθη:

$$L(\theta, i) = \mu^*(\theta) - \mathbf{E}X_i, \tag{3.14}$$

όπου $\mu^*(\theta) = \max(\mu(\theta_1), \mu(\theta))$. Επομένως, ο κίνδυνος του *Bayes* στις πρώτες t παρατηρήσεις είναι ίσο με:

$$R(t, H_0, \pi) = \mathbf{E}_{H_0}[\sum_{j=i}^{t} L(\theta, \pi(j))] = t\mathbf{E}_{H_0}[\mu^*(\theta)] - M(N, H_0, \pi). \tag{3.15}$$

Εφόσον η ποσότητα $t\mathbf{E}_{H_0}[\mu^*(\theta)]$ είναι ανεξάρτητη της πολιτικής η μεγιστοποίση του M είναι ισοδύναμη με την ελαχιστοποίηση του R. Οπότε μπορούμε εναλλακτικά να χρησιμοποιήσουμε και την παρακάτω έκφραση για την βέλτιστη πολιτική π^*:

$$R(N, H_0, \pi^*) = \min_{\pi} R(N, H_0, \pi). \tag{3.16}$$

Η περιγραφή του προβλήματος μέσω του κινδύνου *Bayes* έχει το πλεονέκτημα ότι μπορεί να μοντελοποιήσει γενικότερες μορφές της συνάρτησης απώλειας $L(\theta, i)$.

3.3 Εξιώσεις βελτιστοποίησης και προκαταρτικά αποτελέσματα

Σε αυτή την ενότητα θα αναφέρουμε κάποιες βασικές ιδιότητες και θεωρήματα πάνω στην δομή της βέλτιστης πολιτικής για το πρόβλημα του πεπερασμένου ορίζοντα τα οποία έχουν αναπτυχθεί στην εργασία *Burnetas and Katehakis* [Bur97]. Θα μελετήσουμε το πρόβλημα ως προς n, δηλαδή τα βήματα που μας μένουν ακόμα ή αλλιώς τον αριθμό των δειγμάτων που έχουμε ακόμα να πάρουμε μέχρι το τέλος του πεπερασμένου ορίζοντα N.

Έστω $P(n, k, y)$ το πρόβλημα μεγιστοποίησης του αναμενόμενου αθροίσματος των παρατηρήσεων σε ένα οριζοντα n ($M(n, H, \pi)$), όταν η αρχική πληροφορία για το θ_2 περιλαμβάνεται στην $H(\theta|(k, y))$ (εκ των υστέρων κατανομή του θ_2 με $d_2(k_2) = (k, y)$). Αντίστοιχα, σαν $Q(n, k, y)$ ορίζουμε το πρόβλημα ελαχιστοποίησης του $R(n, H, \pi)$. Έτσι για τα προβήματα $P(n, k, y)$, $Q(n, k, y)$ έχουμε τις παρακάτω συναρτήσεις βέλτιστης τιμής ως προς το κέρδος και την απώλεια αντίστοιχα:

$$V(n, k, y) = \sup_{\pi} M(n, H(\cdot|(k, y), \pi)), \tag{3.17}$$

$$U(n, k, y) = \inf_{\pi} R(n, H(\cdot|(k, y), \pi)), \tag{3.18}$$

Χρησιμοποιώντας τα βασικά επιχειρήματα μιας Μαρκοβιανής Διαδιδικασίας Αποφάσεων με γενικούς χώρους καταστάσεων και πεπερασμένους χώρους αποφάσεων [DY79] έχουμε τις ακόλουθες προτάσεις:

Πρόταση 3.1 *:*

1. Οι συναρτήσεις $V(n,k,y)$ είναι οι μοναδικές λύσεις της (3.19)

$$V(n,k,y) = \max\left\{r(k,y;\alpha_1) + V(n-1,k,y), r(k,y;\alpha_2) + \mathbf{E}_{f(\cdot|(k,y))}V(n-1,k+1,m(k,y,X_2))\right\} \tag{3.19}$$

όπου $n = 1,2,...,N$, $k = 0,1,...,N-n$, $y \in \Re$, $V(0,k,y) = 0$

2. Οι συναρτήσεις $U(n,k,y)$ είναι οι μοναδικές λύσεις της (3.20)

$$U(n,k,y) = \min\left\{c(k,y;\alpha_1) + U(n-1,k,y), r(k,y;\alpha_2) + \mathbf{E}_{f(\cdot|(k,y))}U(n-1,k+1,m(k,y,X_2))\right\} \tag{3.20}$$

όπου $n = 1,2,...,N$, $k = 0,1,...,N-n$, $y \in \Re$, $U(0,k,y) = 0$

Το αναμενόμενο κέρδος $r(k,y;\alpha)$ και κόστος $c(k,y;\alpha)$, αντίστοιχα, ορίζονται ως εξής:

$$r(k,y;\alpha_1) = \mathbf{E}_{\theta_1}[X_1] = \mu(\theta_1), \tag{3.21}$$

$$\begin{aligned} r(k,y;\alpha_2) &= \mathbf{E}_{f(\cdot|(k,y))}X_2 \\ &= \mathbf{E}_{H(\cdot|(k,y))}[\mathbf{E}_\theta X_2] = \int_\Theta \mu(\theta)dH(\theta|(k,y)) \end{aligned} \tag{3.22}$$

$$\begin{aligned} c(k,y;\alpha_1) &= \mathbf{E}_{H(\cdot|(k,y))}[\mu(\theta^*) - r(k,y;\alpha_1)], \\ &= \int_{\theta\geq\theta_1}(\mu(\theta) - \mu(\theta_1)dH(\theta|(k,y))), \end{aligned} \tag{3.23}$$

$$\begin{aligned} c(k,y;\alpha_2) &= \mathbf{E}_{H(\cdot|(k,y))}[\mu(\theta^*) - r(k,y;\alpha_2)], \\ &= \int_{\theta<\theta_1}(\mu(\theta_1) - \mu(\theta)dH(\theta|(k,y))), \end{aligned} \tag{3.24}$$

Όπως είδαμε στην Πρόταση (3.1) το ελάχιστο άνω φράγμα (*supremum*) και το μέγιστο κάτω φράγμα (*infimum*) επιτυγχάνονται κάτω από μια πολιτική π^*, οπότε μπορούν να αντικατασταθούν από το μέγιστο (*maximum*) και το ελάχιστο (*minimum*), αντίστοιχα.

Ισοδύναμα, οι (3.19) και (3.20) μπορούν να εκφραστούν σαν ένα πρόβλημα βέλτιστης διακοπής, όπου η επιλογή του γνωστού πειράματος σε μια περίοδο και η παραμονή σε αυτή την απόφαση μέχρι το τέλος του ορίζοντα N ορίζεται ως διακοπή. Η απόδειξη είναι μία ειδική περιπτωση όσων δίνονται στο [Bra56] για την περίπτωση του διωνυμικού πληθυσμού.

Πρόταση 3.2 *:*

1. Η (3.19) γράφεται ισοδύναμα:

$$V(n,k,y) = \max\left\{n\mu(\theta_1), r(k,y;\alpha_2) + \mathbf{E}_{f(\cdot|(k,y))}V(n-1,k+1,m(k,y,X_2))\right\} \tag{3.25}$$

2. Η (3.20) γράφεται ισοδύναμα:

$$U(n,k,y) = \min\left\{nc(k,y;\alpha_1), r(k,y;\alpha_2) + \mathbf{E}_{f(\cdot|(k,y))}U(n-1,k+1,m(k,y,X_2))\right\} \tag{3.26}$$

Θα ορίσουμε τώρα κάποιες ποσότητες που θα μας βοηθήσουν στην συνέχεια στην εύρεση της βέλτιστης πολιτικής.

Ορισμός 3.1 *Για* $y = (1/k)\sum_{j=1}^{k} y_{2j}$, *έχουμε*

$$l(\theta, \theta_1|y) = \log \frac{f(y|\theta)}{f(y|\theta_1)}, \tag{3.27}$$

$$\Lambda(k, y) = \int_{\Theta} e^{kl(\theta,\theta_1|y)}) dH_0(\theta), \tag{3.28}$$

$$d(\theta) = \theta - \theta_1, \tag{3.29}$$

$$\delta(\theta) = \mu(\theta) - \mu(\theta_1), \tag{3.30}$$

$$\omega(\theta) = \psi(\theta) - \psi(\theta_1), \tag{3.31}$$

Παρατήρηση 3.1 *Από την ισότητα (3.1) παρατηρούμε ότι*

$$l(\theta, \theta_1|y) = \log \frac{e^{\theta y - \psi(\theta) + s(y)}}{e^{\theta_1 y - \psi(\theta_1) + s(y)}} = d(\theta)y - \omega(\theta) \tag{3.32}$$

$$kl(\theta, \theta_1|y) + l(\theta, \theta_1|y) = (k+1)l(\theta, \theta_1|m(k, y, x)) \tag{3.33}$$

Έτσι, με βάση τους παραπάνω ορισμούς μπορούμε να γράψουμε τις εξισώσεις βελτιστοποίησης ως προς την πυκνότητα $f(x|\theta_1)$ αντί για την περιθώρια πυκνότητα $f(x|(k, y))$.

Πρόταση 3.3 *:*

1. Η (3.25) γράφεται ισοδύναμα:

$$v(n, k, y) = \max\left\{0, q(k, y) + \mathbf{E}_{f(\cdot|\theta_1)} v(n-1, k+1, m(k, y, X_2))\right\} \tag{3.34}$$

όπου $n = 1, 2, ..., N$, $k = 0, 1, ..., N-n$, $y \in \Re$, $v(0, k, y) = 0$,

$$v(n, k, y) = V(n, k, y) - n\mu(\theta_1)\Lambda(k, y) \tag{3.35}$$

και

$$q(k, y) = \int_{\Theta} \delta(\theta) e^{kl(\theta,\theta_1|y)} dH_0(\theta) \tag{3.36}$$

2. Η (3.26) γράφεται ισοδύναμα:

$$u(n, k, y) = \min\left\{n\overline{c}(k, y; \alpha_1), \overline{c}(k, y; \alpha_2) + \mathbf{E}_{f(\cdot|\theta_1)} u(n-1, k+1, m(k, y, X_2))\right\} \tag{3.37}$$

όπου $n = 1, 2, ..., N$, $k = 0, 1, ..., N - n$, $y \in \Re$, $u(0, k, y) = 0$,

$$u(n, k, y) = U(n, k, y)\Lambda(k, y) \tag{3.38}$$

και

$$\overline{c}(k, y; \alpha_1) = \int_{\theta \geq \theta_1} \delta(\theta) e^{kl(\theta, \theta_1 | y)} dH_0(\theta), \tag{3.39}$$

$$\overline{c}(k, y; \alpha_2) = -\int_{\theta \leq \theta_1} \delta(\theta) e^{kl(\theta, \theta_1 | y)} dH_0(\theta). \tag{3.40}$$

Σύμφωνα με τα παραπάνω καταλήγουμε σε δύο θεωρήματα που περιγράφουν την βέλτιστη πολιτική ως προς τα διαστήματα συνέχισης και σταματήματος για $y = (1/k) \sum_{j=1}^{k} y_{2,j}$. Το δεύτερο θεώρημα δίνει μια πιο διαισθητική περιγραφή της βέλτιστης πολιτικής με παράγοντες πληθωρισμού συνυφασμένους με την Μπεϋζιανή εκτίμηση του $\mu(\theta_2)$.

Θεώρημα 3.1 *:*

1. Για κάθε n και k, υπάρχει ένας αριθμός $y_n(k)$ με την ιδιότητα

$$\pi^*(n, k, y) = \begin{cases} \alpha_1 & \text{αν } y < y_n(k) \\ \alpha_2 & \text{αν } y \geq y_n(k) \end{cases} \tag{3.41}$$

όπου $\pi^*(n, k, y)$ είναι η απόφαση που υποδεικνύεται από την βέλτιστη πολιτική στην κατάσταση (n, k, y).

2. Η ακολουθία $y_n(k)$ είναι φθίνουσα ως προς n.

Θεώρημα 3.2 *:*

1. Για κάθε n και k, υπάρχει ένας πραγματικός αριθμός $\epsilon(n, k, y)$ με την ιδιότητα

$$\pi^*(n, k, y) = \begin{cases} \alpha_1 & \text{αν } \mathbf{E}_{H(\cdot|(k,y))}[\mu(\theta_2)] + \epsilon(n, k, y) < \mu(\theta_1) \\ \alpha_2 & \text{αν } \mathbf{E}_{H(\cdot|(k,y))}[\mu(\theta_2)] + \epsilon(n, k, y) \geq \mu(\theta_1) \end{cases} \tag{3.42}$$

όπου $\pi^*(n, k, y)$ είναι η απόφαση που υποδεικνύεται από την βέλτιστη πολιτική στην κατάσταση (n, k, y) και

$$\epsilon(n, k, y) = -\frac{q(k, y_n(k))}{\Lambda(k, y)}. \tag{3.43}$$

2. Οι ποσότητες $\epsilon(n, k, y)$ είναι θετικές και αυξάνουν ως προς n.

Παρατήρηση 3.2 *: Τα κατώφλια $y_n(k)$ αντιπροσωπεύουν το ποσό της άμεσης αμοιβής την οποία είμαστε διατεθειμένοι να θυσιάσουμε ώστε να πάρουμε παραπάνω πληροφορία για το θ_2 η οποία είναι πολύτιμη για τις αποφάσεις που απομένουν.*

Μία ερμηνεία για τις ποσότητες $\epsilon(n, k, y)$ είναι ότι υποδηλώνουν μια θετική προσαύξηση της τρέχουσας εκτίμησης της αμοιβής του πειράματος E_2, $\hat{\mu_2} = \mathbf{E}_{H(\cdot|(k,y))}[\mu(\theta_2)]$, έτσι ώστε να ληφθεί υπ΄ όψιν η αβεβαιότητα που συνδέεται με αυτή.

Κεφάλαιο 4

Υπολογιστική εφαρμογή σε μοντέλα με δοκιμές *Bernoulli*

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα εξετάσουμε μοντέλα που έχουν την ίδια δομή και συμπεριφορά με αυτά του προηγούμενου κεφαλαίου. Θα παρουσιάσουμε την προσαρμογή τους πάνω σε δοκιμές *Bernoulli* και θα δώσουμε τις εξισώσεις βελτιστοποίησης. Θα αναφέρουμε έναν αλγόριθμο για πεπερασμένο ορίζοντα ο οποίος μας δίνει το κέρδος κάτω από την βέτιστη πολιτική και βασικά αποτελέσματα γι΄ αυτήν. Έπειτα, θα κάνουμε κάποιες συγκρίσεις με τα αποτελέσματα του προηγούμενου κεφαλαίου. Τέλος θα υπολογίσουμε προσεγγιστικά τα κατώφλια με βάση τα οποία θα καταγράψουμε την βέλτιστη πολιτική.

4.2 Δομή του μοντέλου

Έχουμε, όπως και στο προηγούμενο κεφάλαιο, δύο διαθέσιμα στατιστικά πειράματα. Το κάθε αποτέλεσμα, $Y_{i1}, Y_{i2}, ...$, και των δύο πειραμάτων ακολουθεί $Bernoulli$ κατανομή με πιθανότητα επιτυχίας p_1, p, όπου p_1 η γνωστή πιθανότητα επιτυχίας και p η άγνωστη. Εμείς θα χρησιμοποιούμε τα X_i που υποδηλώνουν τα αποτελέσματα του κάθε πειράματος ξεχωριστά, με $X_i \sim Bernoulli(p)$. Έτσι έχουμε

$$f(x|\theta) = e^{\theta x - \psi(\theta) + s(x)}, \tag{4.1}$$

με $\theta = \log \frac{p}{1-p}$, $\psi(\theta) = \log(1 + e^{\theta}) = -\log(1-p)$ και $s(x) = 0$.

Το πρώτο πείραμα έχει γνωστή πιθανότητα επιτυχίας $p_1 = 1/2$. Για το δεύτερο πείραμα δεν έχουμε καμία πληροφορία. Θεωρούμε αρχικά εκ των προτέρων κατανομή $Beta(\alpha, \beta)$, όπου $\alpha = \beta = 1$. Παίρνουμε δηλαδή ένα δείγμα δύο παρατηρήσεων με μία επιτυχία και μία αποτυχία έτσι ώστε να μην υπάρξει καμία επιρροή στο άγνωστο πείραμα από εμάς. Ακολουθώντας έτσι την Μπεϋζιανή προσέγγιση η εκ των υστέρων κατανομή που προκύπτει, στο επόμενο βήμα γίνεται εκ των προτέρων ανανεώνοντας έτσι την πληροφορία που μας δόθηκε. Επομένως αν σε κάποιο βήμα η εκ των προτέρων κατανομή μας είναι η $H_n(\theta) = Beta(i, j)$ τότε η εκ των υστέρων κατανομή δεδομένης μιας παρατήρησης $x \in 0, 1$ είναι

$$H_{n+1}(\theta|x) \sim Beta(i + x, j + 1 - x). \tag{4.2}$$

Σε κάθε βήμα αποφασίζουμε ποιο από τα δύο πειράματα θα ενεργοποιήσουμε και αναλόγως λαμβάνουμε και την αντίστοιχη αμοιβή. Η κατάσταση του συστήματος εξαρτάται μόνο από το δεύτερο πείραμα. Ορίζουμε σαν κατάσταση το ζεύγος (α, β), δηλαδή τις επιτυχίες και τις αποτυχίες που έχουμε καταγράψει μέχρι εκείνη την στιγμή για το δεύτερο πείραμα. Το κέρδος ενός βήματος σε περίπτωση που επιλέξουμε το πείραμα με την γνωστή πιθανότητα επιτυχίας είναι σταθερό και ίσο με $\mu(\theta_1) = p_1$. Στην περίπτωση που επιλέξουμε το δεύτερο πείραμα και βρισκόμαστε στην κατάσταση (α, β) τότε λαμβάνουμε ως αμοιβή την $\mu(\theta) = \frac{\alpha}{\alpha+\beta}$.

Παρατήρηση 4.1 *Το κέρδος ενός βήματος στην περίπτωση επιλογής του πρώτου πειράματος είναι το ίδιο για κάθε βήμα. Δεν επηρεάζεται δηλαδή από την κατάσταση του συστήματος.*

Παρατήρηση 4.2 *Αν σε κάποιο βήμα επιλέξουμε το πείραμα με την γνωστή πιθανότητα επιτυχίας τότε η κατάσταση του συστήματος δεν αλλάζει, καθώς δεν θα μας δοθεί κάποια παραπάνω πληροφορία για το δεύτερο πείραμα.*

Σύμφωνα με τα παραπάνω μπορούμε να θεωρήσουμε ότι έχουμε ένα πρόβλημα βέλτιστης διακοπής. Αν μία φορά επιλέξουμε το γνωστό πείραμα τότε και σε κάθε επόμενο βήμα θα επιλέγουμε το ίδιο αφού η κατάσταση του συστήματος θα παραμένει η ίδια και το κέρδος ενός βήματος για το πρώτο πείραμα είναι σταθερό. Έτσι θα οδηγηθούμε σε κατάσταση απορρόφησης και θα παραμείνουμε εκεί μέχρι το τέλος του πεπερασμένου ορίζοντα. Επομένως, όπως και στην (3.25), έχουμε τις παρακάτω εξισώσεις βελτιστοποίησης

$$V_n(\alpha, \beta) = \max\left\{np_1, \frac{\alpha}{\alpha+\beta} + \frac{\alpha}{\alpha+\beta}V_{n-1}(\alpha+1, \beta) + \frac{\beta}{\alpha+\beta}V_{n-1}(\alpha, \beta+1)\right\}. \quad (4.3)$$

4.3 Αλγόριθμος

Ο αλγόριθμός μας, λύνει το πρόβλημα με διαδοχικές προσεγγίσεις αφού του δώσουμε το μέγεθος του πεπερασμένου ορίζοντα που επιθυμούμε, έστω N. Υπολογίζει αρχικά όλα τα $V(\alpha, \beta)$ για κάθε δυνατή κατάσταση ξεκινώντας από το $V_0(\alpha, \beta)$, το οποίο το θέσαμε ίσο μηδέν για κάθε (α,β), και πηγαίνοντας προς το $V_N(1,1)$ αφού $H_0(\theta) = Beta(1,1)$. Σε περίπτωση που οι αμοιβές είναι σε κάποιο βήμα ίσες επιλέγει ως πολιτική την ενεργοποίηση του δεύτερου (άγνωστου) πειράματος. Με αυτό τον τρόπο κερδίζουμε κι άλλη πληροφορία γιατί σε αντίθετη περίπτωση θα οδηγούμασταν στην κατάσταση απορρόφησης κάτι που μπορεί να μην είναι βέλτιστο για την συνέχεια.

Η διαδικασία αυτή γίνεται για κάθε $n < N$ δίνοντάς μας έτσι την δυνατότητα να έχουμε όποιο $V_n(1,1)$ επιθυμούμε με μία μόνο εκτέλεση του. Έπειτα εξετάζει ξεχωριστά το κάθε βήμα για να διαπιστώσει από ποια κατάσταση (α, β) και μετά η αμοιβή του γνωστού πειράματος υπερτερεί. Ουσιαστικά σε κάθε βήμα υπολογίζει προσεγγιστικά το $y_n(k)$. Δηλαδή αντί να υπολογίζει το $y_n(k)$ που αναφέραμε στο προηγούμενο κεφάλαιο υπολογίζουμε το $\overline{y} = \frac{\alpha}{\alpha+\beta}$ όπου α, β είναι οι επιτυχίες και οι αποτυχίες αντίστοιχα που έχουμε παρατηρήσει μέχρι εκείνη την στιγμή συμπεριλαμβανομένου και των παρατηρήσεων της αρχικής μας εκ των προτέρων κατανομής. Αφού κάνει όλους αυτούς τους ελέγχους σε κάθε βήμα και για κάθε $n \leq N$ τα αποθηκεύει όλα σε έναν πίνακα για να έχουμε μια ολοκληρωμένη άποψη για την συμπεριφορά και τα χαρακτηριστικά τους. Επομένως, ο αλγόριθμος μας δίνει σε

ένα πίνακα όλα τα $\overline{y}_n(k)$ για κάθε $n \leq N$ και για κάθε k (για τα οποία ισχύει γενικά $\overline{y}_n(k) \leq y_n(k)$), όπου k είναι ο αριθμός των βημάτων που έχουν γίνει.

Σκοπός μας είναι να εξετάσουμε αν υπάρχει ένα κατώφλι το οποίο είναι υπολογιστικά πολύ πιο απλό. Δηλαδή, αν έχουμε μια διαδικασία η οποία αρχικά μας είναι άγνωστη ως προς την πιθανότητα επιτυχίας θα γνωρίζουμε κάθε φορά από ποιο y και πάνω θα πρέπει να την ενεργοποιούμε. Θέλουμε επίσης να δούμε αν με βάση αυτό μπορούμε να διαμορφώσουμε την βέλτιστη πολιτική. Κάτι που φαίνεται να ισχύει, καθώς σε αντιστοιχία με το (3.41) και το (3.42) εδώ έχουμε την παρακάτω βέλτιστη πολιτική

$$\pi^*(n,k,y) = \begin{cases} \alpha_1 & \text{αν } y < \overline{y}_n(k) \\ \alpha_2 & \text{αν } y \geq \overline{y}_n(k) \end{cases} \tag{4.4}$$

και θέτοντας ως $\epsilon(n,k,y) = p_1 - \overline{y}_n(k)$ έχουμε

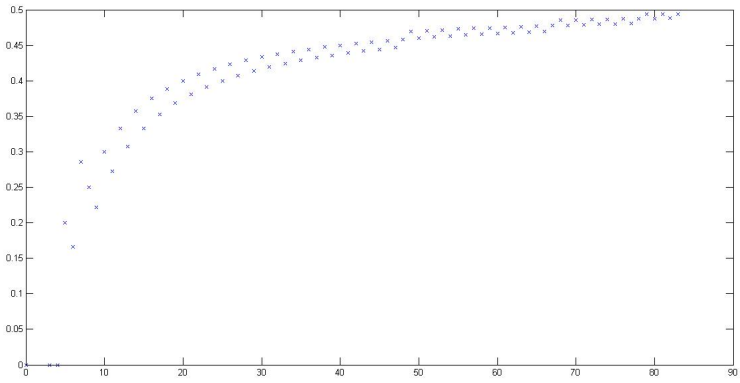
$$\pi^*(n,k,y) = \begin{cases} \alpha_1 & \text{αν } y + \epsilon(n,k,y) < p_1 \\ \alpha_2 & \text{αν } y + \epsilon(n,k,y) \geq p_1 \end{cases} \tag{4.5}$$

όπου $y = \frac{\alpha}{\alpha+\beta}$ και (α, β) η κατάσταση του συστήματος.

Η (4.5) θα μας φανεί πολύ χρήσιμη στην συνέχεια όταν θα κάνουμε την μελέτη για τον άπειρο ορίζοντα καθώς θα εξετάσουμε αν μπορούμε το $y + \epsilon(n,k,y)$ να το θεωρήσουμε ως ένα δείκτη και να δούμε πόσο κοντά βρίσκεται με τον αντίστοιχο δείκτη *Gittins*.

Παρατήρηση 4.3 *Η ακολουθία των $\overline{y}_n(k)$ δεν είναι αύξουσα ούτε ως προς n ούτε ως προς k. Παρ΄ όλα αυτά παρατηρείται μία ανοδική τάση των $\overline{y}_n(k)$. Δηλαδή για έναν τυχαίο ορίζοντα N όσο αυξάνεται το k η τιμή των $\overline{y}_n(k)$ δείχνει γενικά να αυξάνει. Αυτό είναι διαισθητικά αναμενόμενο καθώς όσο περισσότερα βήματα απομένουν τόσο περισσότερο είμαστε διαθετιμένοινα πειραματιστούμε με την άγνωστη διαδικασία.*

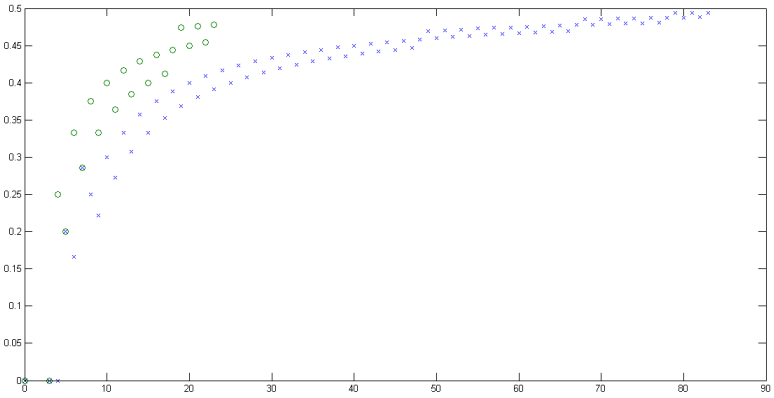
Παράδειγμα 4.1 *Στην παρακάτω εικόνα φαίνεται για $N = 80$ αυτό που αναφέρουμε στην παραπάνω παρατήρηση.*



Στον κάθετο άξονα του γραφήματος απεικονίζονται τα $\overline{y}_n(k)$ και στον οριζόντιο ο αριθμός των βημάτων k που έχουν γίνει, ο οποίος είναι αντιστρόφως ανάλογος του n, δηλαδή του αριθμού των βημάτων που μας απομένουν.

Παρατήρηση 4.4 *Όλα τα $\overline{y}_n(k)$ βρίσκονται κάτω από το $p_1 = 0,5$ και όσο μικραίνει ο ορίζοντας N η γραφική τους παράσταση βρίσκεται πιο πάνω και τείνει να φτάσει πιο γρήγορα στο p_1. Αυτό είναι λογικό καθώς για $y > 0,5$ αποφασίζουμε να ενεργοποίσουμε το άγνωστο πείραμα και όσο πιο λίγα βήματα έχουμε μπροστά μας τόσο πιο γρήγορα καταλήγουμε να χρησιμοποιούμε το γνωστό πείραμα.*

Παράδειγμα 4.2 *Κάτι που φαίνεται και στην παρακάτω εικόνα που έχουμε τις γραφικές παραστάσεις των $\overline{y}_n(k)$ για $N = 80$ (γαλάζια αστέρια) και για $N = 20$ (πράσινοι κύκλοι).*



Στον κάθετο άξονα του γραφήματος απεικονίζονται τα $\overline{y}_n(k)$ και στον οριζόντιο ο αριθμός των βημάτων k που έχουν γίνει, ο οποίος είναι αντιστρόφως ανάλογος του n, δηλαδή του αριθμού των βημάτων που μας απομένουν.

Παράδειγμα 4.3 *Εκτελώντας τον αλγόριθμο για $N = 45$ ενδεικτικά αναφέρουμε ότι*

$$V_{20}(1,1) = 12.5247 \tag{4.6}$$

$$V_{25}(1,1) = 15.5888 \tag{4.7}$$

$$V_{30}(1,1) = 18.6543 \tag{4.8}$$

$$V_{35}(1,1) = 21.7248 \tag{4.9}$$

$$V_{40}(1,1) = 24.8026 \tag{4.10}$$

$$V_{45}(1,1) = 27.8814 \tag{4.11}$$

Παρατήρηση 4.5 *Βλέπουμε με βάση τα αποτελέσματα του παραδείγματος (4.3) ότι οι τελικές αμοιβές ξεπερνούν τις αντίστοιχες που θα λαμβάναμε σε περίπτωση που χρησιμοποιούσαμε μόνο το γνωστό πείραμα, δηλαδή np_1. Βέβαια αν η γνωστή διαδικασία είχε μεγαλύτερη πιθανότητα επιτυχίας θα βλέπαμε να είχε μεγαλύτερη προτίμηση και αντίστοιχα το κέρδος θα ήταν μεγαλύτερο, ενώ τώρα είναι μοιρασμένες οι αποφάσεις (αφού $p_1 = 1/2$). Όταν η κατάσταση του συστήματος έχει πιο πολλές επιτυχίες τότε προτιμάται η άγνωστη ενώ σε αντίθετη περίπτωση προτιμάται η γνωστή. Αυτό είναι φυσιολογικό καθώς η κατάσταση του συστήματος φανερώνει τις επιδόσεις της άγνωστης διαδικασίας.*

Παράδειγμα 4.4 *Έχοντας παρατηρήσει όλα τα παραπάνω θα δώσουμε κάποια αριθμητικά αποτελέσματα για διαφορετικά N, k.*

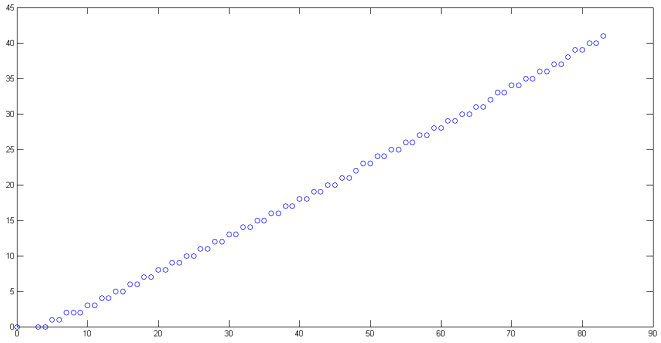
N/k	$k=50$	$k=51$	$k=52$	$k=53$	$k=54$	$k=55$	$k=56$	$k=57$
$N=84$	$0,433$	$0,419$	$0,437$	$0,424$	$0,441$	$0,428$	$0,444$	$0,432$

N/k	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$	$k=12$	$k=13$
$N=32$	$0,454$	$0,434$	$0,458$	$0,440$	$0,461$	$0,444$	$0,464$	$0,448$

N/k	$k=72$	$k=73$	$k=74$	$k=75$	$k=76$	$k=77$	$k=78$	$k=79$
$N=98$	$0,409$	$0,391$	$0,416$	$0,400$	$0,423$	$0,407$	$0,428$	$0,413$

N/k	$k=11$	$k=12$	$k=13$	$k=14$	$k=15$	$k=16$	$k=17$	$k=18$
$N=46$	$0,451$	$0,468$	$0,454$	$0,470$	$0,457$	$0,472$	$0,459$	$0,473$

Παράδειγμα 4.5 *Το παρακάτω διάγραμμα μας δίνει τον αριθμό των επιτυχιών που έχουμε σε κάθε βήμα όταν για πρώτη φορά επιλέγουμε το γνωστό πείραμα. Παρατηρούμε ότι διαγράφει μια αύξουσα πορεία όσο το k αυξάνει.*



Στον κάθετο άξονα του γραφήματος απεικονίζεται ο αριθμός των επιτυχιών την πρώτη φορά που επιλέγουμε το γνωστό πείραμα και στον οριζόντιο ο αριθμός των βημάτων k που

έχουν γίνει, ο οποίος είναι αντιστρόφως ανάλογος του n, δηλαδή του αριθμού των βημάτων που μας απομένουν.

Κώδικας.

$function[Y] = optimalstop(N, p_1)$
$a = 1;$
$b = 1;$
$for\ l = N : -1 : 1$
$v = zeros(l + 2, l + 2);$
$y = zeros(l + 1, l + 1);$
$g = zeros(l + 1, l + 1);$
$for\ j = 2 : l + 2$
$k = -j + 2;$
$for\ i = 1 : l + 3 - j$
$v(i, j) = \max\left\{(j-1) * p_1, \frac{a+l+1-i+k}{a+b+l+k} + ((\frac{a+l+1-i+k}{a+b+l+k} * v(i, j-1) + \frac{b+i-1}{a+b+l+k} * v(i+1, j-1)))\right\};$
$if\ (j-1) * p1 > \frac{a+l+1-i+k}{a+b+l+k} + \frac{a+l+1-i+k}{a+b+l+k} * v(i, j-1) + \frac{b+i-1}{a+b+l+k} * v(i+1, j-1)$
$y(i, j) = \frac{a+l+3-i-j}{a+b+l+2-j};$
$g(i, j) = a + l + 1 - i + k;$
end
end
end
$for\ z = 1 : l$
$Y(N - l + 1, z) = \max(y(:, l + 2 - z));$
$G(N - l + 1, z) = \max(g(:, l + 2 - z));$
end
end
$for\ j = 1 : N$
$for\ i = 1 : N + 1 - j$
$m(j, i) = a + b + i\cdot$
end
end
$plot(m(20, :), Y(20, :), x, m(80, :), Y(80, :), o)$
end □

Κεφάλαιο 5

Πρόβλημα επανεκκίνησης και διαδοχικές κλινικές δοκιμές

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα εξετάσουμε μία άλλη προσέγγιση του προβλήματος $multi-armed\ bandit$. Θα το δούμε ως ένα πρόβλημα επανεκκίνησης όπου γίνεται μία διαφορετική παρατήρηση συγκεκριμένα ότι το m_i είναι η μέγιστη τιμή στην κατάσταση i σε ένα πρόβλημα επανεκκίνησης στο οποίο δύο αποφάσεις είναι διαθέσιμες σε κάθε περίοδο, συνέχιση ή ακαριαία επανεκκίνηση από την κατάσταση i. Αυτή η παρατήρηση μειώνει το πρόβλημα της εύρεσης του m_i στην επίλυση ενός απλού προβλήματος Μαρκοβιανής διαδικασίας αποφάσεων με εκπτωτικό κέρδος. Θα δώσουμε την δομή του και θα αναφέρουμε κάποια βασικά χαρακτηριστικά.

Έπειτα θα παρουσιάσουμε ένα παράδειγμα με διαδοχικές κλινικές δοκιμές μέσα από το οποίο θα δούμε πως μπορούμε να υπολογίσουμε τους δείκτες *Gittins*. Θα έχουμε δηλαδή κάποιες πιθανές θεραπείες για μια συγκεκριμένη ασθένεια. Όταν χρησιμοποιούμε μια τυχούσα θεραπεία, έστω n, αυτή είναι αποτελεσματική με πιθανότητα θ_n και μη αποτελεσματική με πιθανότητα $1-\theta_n$. Σκοπός είναι να βρούμε μια διαδοχική διαδικασία δειγματοληψίας που να μεγιστοποιεί τον αναμενόμενο συνολικό αριθμό των επιτυχημένων θεραπειών. Όταν ο ορίζοντας είναι άπειρος οι εκ των προτέρων κατανομές προσδιορίζονται από τις άγνωστες παραμέτρους εκ των οποίων η μία φανερώνει τον αναμενόμενο εκπτωτικό συνολικό αριθμό των επιτυχημένων θεραπειών ως το μέτρο επίδοσης μιας διαδοχικής διαδικασίας δειγματοληψίας. Επομένως είναι μια εκπτωτική εκδοχή του προβλήματος $multi-armed\ bandit$ που αναλύθηκε από τους *Gittins* και *Jones* [GJ74], (βλέπε επίσης [Whi80], [Whi82] $p.210$). Η αρχική διατύπωση του $multi-armed\ bandit$ προβλήματος και του προβλήματος των διαδοχικών κλινικών δοκιμών οφείλεται στον *Robbins* [Rob52]. Οι *Gittins* και *Jones* έδειξαν ότι υπάρχει ένας δείκτης συσχετισμένος με κάθε κατάσταση του κάθε *bandit* τέτοιος ώστε σε μια βέλτιστη διαδικασία πάντα να χρησιμοποιούμε το *bandit* με τον μεγαλύτερο δείκτη. Σύμφωνα με τους *Katehakis* και *Veinott* [Kat85] όπως θα δούμε υπάρχει ένας νέος χαρακτηρισμός για τον δείκτη που μας βοηθά πολύ στο υπολογιστικό κομμάτι. Η παρακάτω ανάλυση βασίζεται στις εργασίες [Kat85] και [KD85].

5.2 Πρόβλημα επανεκκίνησης

Είδαμε στο Κεφάλαιο 2 πως ορίζεται το $multi - armed\ bandit\ problem$, πως δουλεύει και πως μπορούμε να βρούμε την βέλτιστη πολιτική με βάση τον δείκτη $Gittins$. Εδώ ο δείκτης ενός προγράμματος σε μια κατάσταση θα χαρακτηριστεί ως η τιμή σε ένα πρόβλημα επανεκκίνησης. Η ακολουθία καταστάσεων είναι μια Μαρκοβιανή αλυσίδα με αριθμήσιμο χώρο καταστάσεων I και με πιθανότητες μεταπήδησης p_{ij}. Έστω $r \equiv (r_i)$ και $P \equiv (\alpha p_{ij})$ όπου α ο συντελεστής αποπληθωρισμού. Για κάθε κατάσταση i, ορίζουμε το p_i να είναι η i-οστή γραμμή του P.

Είναι γνωστό από την θεωρία Μαρκοβιανών διαδικασιών αποφάσεων ότι η μέγιστη τιμή διανύσματος για το πρόβλημα επανεκκίνησης στην i είναι η μοναδική φραγμένη λύση $\upsilon = \upsilon^i = (\upsilon^i_j)$ του

$$\upsilon_j = \max(r_j + P_j\upsilon, r_i + P_i\upsilon),\ j \in I \tag{5.1}$$

Αν θεωρήσουμε μόνο τις στιγμές όπου η διαδικασία είναι στην κατάσταση i, τότε το υ_i μπορεί επίσης να ερμηνευθεί ως η μέγιστη τιμή στην κατάσταση i για την αντίστοιχη ημι-Μαρκοβιανή διαδικασία. Επομένως, το υ_i ικανοποιεί $\upsilon_i = \max_{\tau \geq 1} E[R_j + \alpha^\tau \upsilon_i]$, ή ισοδύναμα $\upsilon_i = m_i$, όπου $\tau + 1$ είναι η πρώτη περίοδος μετά την πρώτη περίοδο που επιλέχθηκε να ξαναεκκινήσει στην κατάσταση i στο πρόβλημα επανεκκίνησης στην i. Επιπλέον ο δείκτης $m = m_i = \upsilon_i$ ικανοποιεί την

$$m = r_i + P_i\upsilon \tag{5.2}$$

Αντικαθιστώντας την (5.2) στην (5.1) παίρνουμε ([Git79], [GJ74], [Whi82]) ότι η $(\upsilon, m) = (\upsilon(m_i), m_i)$ είναι η μοναδική φραγμένη λύση της (5.2) και

$$\upsilon_j = \max(r_j + P_j\upsilon, m),\ j \in I \tag{5.3}$$

όπου $\upsilon(m)$ είναι η μοναδική φραγμένη λύση της (5.3). Έτσι η $\upsilon(m)$ είναι μέγιστη διανυσματική τιμή για το προεξοφλημένο πρόβλημα σταματήματος με αμοιβή σταματήματος m.

Πρόταση 5.1 *Για κάθε κατάσταση i, $m_i = \upsilon^i_i$. Επίσης το m_i είναι η μοναδική λύση της*

$$m_i = r_i + P_i\upsilon(m_i)$$

και η

$$\upsilon^i = \upsilon(m_i)$$

είναι αύξουσα ως προς m_i.

5.3 Σύνολα επανεκκίνησης και συνέχισης

Ακολουθώντας έτσι την βέλτιστη πολιτική σε ένα πρόβλημα επανεκκίνησης, υπάρχουν καταστάσεις στις οποίες όταν βρεθείς είναι βέλτιστο να συνεχίσουμε και αντίστοιχα σε άλλες είναι βέλτιστο να επανεκκινήσουμε στην αρχική κατάσταση.

Το βέλτιστο σύνολο συνέχισης C_i (αντίστοιχα βέλτιστο σύνολο επανεκκίνησης R_i) για το πρόβλημα επανεκκίνησης στην i είναι ένα σύνολο από καταστάσεις j για τις οποίες η συνέχιση είναι βέλτιστη, δηλαδή το $r_j + P_j\upsilon^i - m_i$ είναι μη αρνητικό (αντίστοιχα μη

θετικό). Ο καταλληλότερος χρόνος διακοπής στο πρόβλημα επανεκκίνησης στην i είναι ο αριθμός των περιόδων που απαιτούνται ώστε η κατάσταση του προγράμματος να μην ανήκει στο C_i. Φυσικά είναι διαφορετική η συνέχιση και η επανεκκίνηση για κάθε κατάσταση στο $C_i \cap R_i$.

Πρόταση 5.2 *Για το πρόβλημα επανεκκίνησης στην κατάσταση i, το C_i (αντίστοιχα R_i) είναι το σύνολο από καταστάσεις j για τις οποίες $m_j \geq m_i$ (αντίστοιχα $m_j \leq m_i$).*

Απόδειξη. Από την (5.3) έχουμε

$$||\upsilon(m_j) - \upsilon(m_i)||_\infty \leq |m_j - m_i|$$

έτσι

$$|P_j\upsilon(m_j) - P_j\upsilon(m_i)| \leq \alpha|m_j - m_i|.$$

Άρα από την (5.2) $r_j + P_j\upsilon(m_i) - m_i \geq 0$ (αντίστοιχα ≤ 0) αν και μόνο αν $m_j - m_i \geq 0$ (αντίστοιχα ≤ 0). □

Η παραπάνω πρόταση εξασφαλίζει ότι μια τέτοια πολιτική χρησιμοποιεί τον κανόνα μεγίστου δείκτη σε όλες τις περιόδους. Σε περίπτωση όμως που τα m_i δεν είναι γνωστά θα πρέπει να κάνουμε κάποιες άλλες υποθέσεις.

Έστω λοιπόν ότι το I είναι μερικά διατεταγμένο από την σχέση $\leq$. Ένα υποσύνολο J του I καλείται αύξον (αντίστοιχα φθίνον) αν για $i \in J$ με $i \leq j$ (αντίστοιχα $i \geq j$) στο I συνεπάγεται ότι και το $j \in J$. Καλούμε τον P στοχαστικά μονότονο αν το $P\upsilon$ είναι μη φθίνον στο I όταν το υ είναι μη φθίνον και φραγμένο στο I. Ορίζουμε R_i ως $(R_i\upsilon)_j = \max(r_j + P_j\upsilon, r_i + P_i\upsilon)$ για κάθε $j \in I$ και φραγμένο υ. Παρατηρούμε ότι $R_i\upsilon$ είναι μονότονο ως προς υ. Το επόμενο αποτέλεσμα επεκτείνει σχετική εργασία των *Gittins* [Git79] και *Ross* [Ros83].

Πρόταση 5.3 *Αν I είναι μερικώς διατεταγμένο, η r είναι μη φθίνουσα και P στοχαστικά μονότονος, τότε υ_j^i και m_i είναι αντίστοιχα μη φθίνουσες ως προς i, j. Επιπλέον $C_i \supseteq C_j$ και $R_i \subseteq R_j$ για κάθε $i \leq j$ στο I και C_i (αντίστοιχα R_i) είναι αύξουσα (αντίστοιχα φθίνουσα) για κάθε $i \in I$.*

Απόδειξη. Οι υποθέσεις συνεπάγονται ότι $R_i\upsilon$ είναι μη φθίνουσα στο I αν η υ είναι μη φθίνουσα. Άρα αφού το 0 είναι μη φθίνον στο I, έτσι είναι $R_i^t 0$ και $\lim_{t\to\infty} R_i^t 0 = \upsilon^i$. Τώρα από υπόθεση πάλι, $\upsilon^i = R_i\upsilon^i \leq R_k\upsilon^i$ για $i \leq k$ στο I, έτσι $\upsilon^i \leq \lim_{t\to\infty} R_k^i\upsilon^i = \upsilon^k$ από όπου υ^i είναι δεν φθίνει ως προς i. Οι δύο τελευταίοι ισχυρισμοί ακολουθούν από την προηγούμενη πρόταση. □

5.4 Διαδοχικές κλινικές δοκιμές

Έστω ότι έχουμε N διαθέσιμες θεραπείες για μια συγκεκριμένη ασθένεια. Ορίζουμε σαν $Y_n(k) = 1$ ($Y_n(k) = 0$) το αποτέλεσμα της $n-$οστής θεραπείας την $k-$οστή φορά που χρησιμοποιήθηκε όταν είναι επιτυχημένο (αποτυχημένο). Σε κάθε βήμα πρέπει να αποφασίσουμε ποια θεραπεία θα εφαρμόσουμε στον επόμενο ασθενή βασιζόμενοι στις προηγούμενες παρατηρήσεις. Αρχικά υποθέτουμε ότι η πιθανότητα επιτυχίας του $n-$οστού πειράματος (θ_n) είναι τυχαία μεταβλητή με εκ των προτέρων κατανομή $Beta(a_n, b_n)$. Η θ_n έχει την

ακόλουθη εκ των προτέρων πυκνότητα:

$$g_n(\theta) = \frac{\Gamma(a_n + b_n)}{\Gamma(a_n)\Gamma(b_n)} \theta^{a_n - 1}(1-\theta)^{b_n - 1}, \tag{5.4}$$

για κάθε $\theta \in [0,1]$ και a_n, b_n αυστηρά θετικές σταθερές. Θεωρούμε επίσης ότι $\theta_1, \theta_2, ..., \theta_n$ είναι ανεξάρτητες.

Αν μετά από k δοκιμές χρησιμοποιώντας την $n-$οστή θεραπεία είμαστε στην κατάσταση $x_n(k) = (s_n(k), f_n(k))$, όπου $s_n(k)$ $(f_n(k))$ είναι ο αριθμός των επιτυχιών (αποτυχιών) τότε, η εκ των υστέρων πυκνότητα της θ_n δεδομένου του $x_n(k)$ είναι $Beta(a_n + s_n(k), b_n + f_n(k))$. Επομένως η πληροφορία για τις πρώτες k δοκιμές της $n-$οστής θεραπείας περιλαμβάνονται στην $x_n(k)$.

Επιπλέον, η $\{x_n(k), k \geq 1\}$ είναι Μαρκοβιανή αλυσίδα πάνω στο $S = \{(s,f), \quad s, f = 0, 1, 2, ...\}$ με πιθανότητες μετάβασης

$$P(x_n(k+1) = (s+1, f) | x_n(k) = (s,f))$$

$$= 1 - P(x_n(k+1) = (s, f+1) | x_n(k) = (s,f))$$

$$= P(Y_n(k+1) = 1 | x_n(k) = (s,f)) = \frac{a_n + s}{a_n + b_n + s + f}.$$

Το ζητούμενο είναι να διαμορφώσουμε μια πολιτική π η οποία να μεγιστοποιεί τον αναμενόμενο εκπτωτικό αριθμό των επιτυχιών, δηλαδή να μεγιστοποιεί το

$$w(\pi, \alpha) = \int ... \int E(\sum_{t=1}^{\infty} \alpha^{t-1} Y_{\pi(t)}) g_1(d\theta_1) ... g_N(d\theta_N), \tag{5.5}$$

όπου $Y_{\pi(t)}$ είναι το $Y_n(k)$ αν την στιγμή t η θεραπεία $\pi(t) = n$ χρησιμοποιείται για $k-$οστή φορά και $\alpha \in (0,1)$ είναι ο εκπτωτικός παράγοντας. Μια ερμηνεία του εκπτωτικού παράγοντα είναι ότι το $1 - \alpha$ είναι η πιθανότητα σε μια τυχούσα στιγμή να τερματιστεί όλο το πείραμα.

Εναλλακτικά, έχουμε N Μαρκοβιανές αλυσίδες και το ζητούμενο είναι να ενεργοποιούμε διαδοχικά μία απ΄ όλες, μένοντας οι υπόλοιπες ανενεργές, μεγιστοποιώντας την αναμενόμενη εκπτωτική συνολική αμοιβή. Σε αυτή την περίπτωση, η αναμενόμενη αμοιβή σε μια τυχούσα χρονική στιγμή είναι η αναμενόμενη εκ των υστέρων πιθανότητα επιτυχίας συνδεδεμένη με την κατάσταση την ενεργοποιημένης Μαρκοβιανής αλυσίδας. Δηλαδή, αν είναι ενεργοποιημένη η $n-$οστή αλυσίδα για $k-$οστή φορά με $x_n(k) = (s,f)$ τότε η αντίστοιχη αμοιβή είναι

$$r_n(s,f) = E(Y_n(k+1) | x_n(k) = (s,f)) = \frac{a_n + s}{a_n + b_n + s + f}. \tag{5.6}$$

Με την παραπάνω διατύπωση, οι *Gittins and Jones* [GJ74] έδειξαν ότι το πρόβλημα αυτό μπορεί να διασπαστεί σε N μονοδιάστατα προβλήματα. Κάθε ένα από τα προβλήματα αυτά περιλαμβάνει μία μόνο Μαρκοβιανή αλυσίδα και η λύση του είναι ο υπολογισμός ενός δυναμικού διαδοχικού δείκτη $m_n(s,f)$ συνδεδεμένο με την υπάρχουσα κατάσταση (s,f) της Μαρκοβιανής αλυσίδας. Έτσι σε κάθε περίοδο η βέλτιστη πολιτική για το αρχικό πρόβλημα ενεργοποιεί την αλυσίδα αυτή που έχει τον μέγαλύτερο δείκτη. Βασισμένοι

σε έναν παλαιότερο χαρακτηρισμό του $(1-\alpha)^{-1}m_n(s,f)$, οι *Gittins and Jones* [GJ79] χρησιμοποίησαν έναν αλγόριθμο για να υπολογίσουν τις βέλτιστες πολιτικές.

Όπως είδαμε και στην ενότητα (5.2) οι *Katehakis* και *Veinott* [Kat85] έδωσαν ένα διαφορετικό χαρακτηρισμό για τον δείκτη. Αυτός ο χαρακτηρισμός μετατρέπει τον υπολογισμό του δείκτη σε μια γνωστή μορφή προβλήματος αντικατάστασης [Der70]. Έτσι, αν ονομάσουμε C το σύνολο των πολιτικών που ελέγχουν την $\{x_n(k), k \geq 1\}$ αφήνοντας την να συνεχίσει ή ακαριαία να την επανεκκινήσει από την αρχική της κατάσταση $x_n(1) = (s,f)$ τότε

$$m_n(s,f) = \sup_{\pi} \left\{ E_\pi(\sum_{k=1}^{\infty} r_n(x_n(k)) | x_n(1) = (s,f)) \right\}. \tag{5.7}$$

Θα δείξουμε τώρα ότι η (5.7) μπορεί να χρησιμοποιηθεί για να υπολογίσουμε τους $m_n(s,f)$ με ακρίβεια. Στη συνέχεια θα ασχοληθούμε με μία μόνο θεραπεία. Ο υπολογισμός του δείκτη $m(s,f)$ είναι ουσιαστικά η ίδια διαδικασία με τον υπολογισμό του $m(0,0)$ αλλάζοντας την εκ των προτέρων κατάσταση από (a,b) σε $(a+s, b+f)$, χωρίς βλάβη της γενικότητας. Είναι γνωστό ότι η λύση της (5.7) για μια σταθερή αρχική κατάσταση περιλαμβάνει την λύση των παρακάτω εξισώσεων δυναμικού προγραμματισμού

$$V(s,f) = \max [\frac{a}{a+b} + \alpha \left\{ \frac{a}{a+b} V(1,0) + \frac{b}{a+b} V(0,1) \right\},$$

$$\frac{a+s}{a+b+s+f} + \alpha \left\{ \frac{a+s}{a+b+s+f} V(s+1,f) + \frac{b+f}{a+b+s+f} V(s,f+1) \right\}] \tag{5.8}$$

για κάθε $(s,f) \in S$.

Το γεγονός ότι μέσω της εξίσωσης (5.8) υπολογίζουμε το $m(0,0)$ φανερώνεται από την ύπαρξη των $V(1,0)$ και $V(0,1)$. Δεδομένης της λύσης $V(s,f)$ για κάθε $(s,f) \in S$ της (5.8) έχουμε ότι $m(0,0) = V(0,0)$. Η (5.8) είναι της μορφής $V(s,f) = T_{sf}V$ ή ισοδύναμα

$$V = TV, \tag{5.9}$$

όπου V είναι το διάνυσμα των τιμών $V(s,f)$ και T είναι τελεστής συστολής σε ένα πλήρη μετρικό χώρο. Επομένως έχει μία μοναδική φραγμένη λύση.

Για να υπολογίσουμε την λύση της $V = TV$ πρέπει να θεωρήσουμε το πεπερασμένο υποσύνολο $S_L = \{(s,f) \in S : s+f \leq L\}$ και τα παρακάτω συστήματα εξισώσεων

$u_L(s,f) = T_{sf}u_L,$ αν $s+f \leq L,$

$u_L(s,f) = \frac{a+s}{a+b+s+f}\frac{1}{1-\alpha},$ αν $s+f = L$

$U_L(s,f) = T_{sf}U_L,$ αν $s+f \leq L,$

$U_L(s,f) = \frac{1}{1-\alpha},$ αν $s+f = L.$

Για λόγους συμβολισμού θα χρησιμοποιήσουμε τις παρακάτω εξισώσεις,

$$u_L = T_1 u_L, \tag{5.10}$$

$$U_L = T_2 u_L. \tag{5.11}$$

Οι μετασχηματισμοί T, T_1, T_2 είναι μονότονες συστολές, [Ber76], και γι αυτό οι διαδοχικές προσεγγίσεις συγκλίνουν στα μοναδικά σταθερά σημεία για κάθε αρχικό σημείο $V^{(0)}$, $u_L^{(0)}$, $U_L^{(0)}$. Επομένως έχουμε

$$\lim_{n\to\infty} V^{(n)} = \lim_{n\to\infty} TV^{(n-1)} = V, \tag{5.12}$$

$$\lim_{n\to\infty} u_L^{(n)} = \lim_{n\to\infty} T_1 u_L^{(n-1)} = u_L, \tag{5.13}$$

$$\lim_{n\to\infty} U_L^{(n)} = \lim_{n\to\infty} T_2 U_L^{(n-1)} = U_L \tag{5.14}$$

Επιπλέον, αν τα αρχικά σημεία $V^{(0)}$, $u_L^{(0)}$, $U_L^{(0)}$ επιλεχθούν καταλλήλως η σύγκλιση στην (5.12) επέρχεται από πάνω ή από κάτω και στις (5.13) ((5.14)) από κάτω (πάνω) αντίστοιχα.

Ένας αλγόριθμος που υπολογίζει το $V(0,0)$ βασισμένος στην (5.12) περιλαμβάνει άπειρο αριθμό μεταβλητών. Παρ' όλα αυτά οι παρακάτω προτάσεις μας επιτρέπουν να χρησιμοποιήσουμε τις (5.14) (5.13) οι οποίες περιλαμβάνουν πεπερασμένο αριθμό μεταβλητών.

Πρόταση 5.4 *Για τις ισότητες (5.9), (5.10), (5.11) έχουμε*

$$\frac{a+s}{a+b+s+f}(1-\alpha)^{-1} \le V(s,f) \le (1-\alpha)^{-1} \tag{5.15}$$

για κάθε $(s,f) \in S$

και

$$u_L(s,f) \le V(s,f) \le U_L(s,f) \tag{5.16}$$

για κάθε (s,f) τέτοια ώστε $s+f \le S$.

Η πρώτη ανισότητα στην (5.15) προκύπτει από το γεγονός ότι το αριστερό μέρος είναι η αναμενόμενη εκπτωτική αμοιβή που επιτυγχάνεται από την υπο-βέλτιστη πολιτική με βάση την οποία δεν επανεκκινούμε στην κατάσταση (0,0) όταν η αρχική μας κατάσταση είναι (s,f). Αντίστοιχα, η δεύτερη ανισότητα προκύπτει από το γεγονός ότι το δεξιό μέρος είναι η αναμενόμενη εκπτωτική αμοιβή που επιτυγχάνεται όταν όλες οι αμοιβές ενός βήματος αντικατασταθούν από το άνω φράγμα τους που είναι το 1. Οι ανισώσεις στην (5.16) προκύπτουν από την (5.15), τις ισότητες (5.12), (5.13), (5.14) και την μονοτονία των μετασχηματισμών T, T_1, T_2.

Πρόταση 5.5 *Για κάθε $\epsilon > 0$ υπάρχει ένα $L_0 = L(\epsilon)$ τέτοιο ώστε*

$$U_L(0,0) - u_L(0,0) \le \epsilon, \tag{5.17}$$

για κάθε $L \ge L_0$.

Απόδειξη. Λόγω της (5.16) αρκεί να δείξουμε ότι για οποιεσδήποτε θετικές σταθερές ϵ_1, ϵ_2 υπάρχουν $L_1 = L(\epsilon_1), L_2 = L(\epsilon_2)$ τέτοια ώστε

$$U_L(0,0) - V(0,0) \le \epsilon_1, \tag{5.18}$$

για κάθε $L \geq L_1$,

$$V(0,0) - u_L(0,0) \leq \epsilon_2, \tag{5.19}$$

για κάθε $L \geq L_2$. Θα το δείξουμε μόνο για το ϵ_1 αφού και για το ϵ_2 είναι ακριβώς το ίδιο.

Αν πάρουμε $U_L^{(0)} = V^{(0)} = \frac{1}{(1-\alpha)}$ στις (5.12), (5.14) τότε για κάθε L και για όλα τα $n \leq L$ έχουμε ότι

$$U_L^{(n)}(0,0) = V^{(n)}(0,0), \tag{5.20}$$

και η σύγκλιση στις (5.12) και (5.14) επέρχεται από πάνω. Έτσι χρησιμοποιώντας την (5.12) και το γεγονός ότι $V(s,f) \geq 0$ έχουμε

$$V^{(n)}(0,0) - V(0,0) \leq \alpha \sup_{(s,f)} \left\{ V^{(n-1)}(s,f) - V(s,f) \right\} \leq ...$$

$$\leq \alpha^n \sup_{(s,f)} \left\{ V^{(0)}(s,f) - V(s,f) \right\} \leq \alpha^n (1-\alpha)^{-1}. \tag{5.21}$$

Από τις (5.20), (5.21) προκύπτει ότι για κάθε L και για όλα τα $n \leq L$ έχουμε

$$U_L^{(n)}(0,0) - V(0,0) \leq \alpha^n (1-\alpha)^{-1}. \tag{5.22}$$

Με παρόμοια επιχειρήματα και χρησιμοποιώντας την (5.14) συνεπάγεται ότι για κάθε $n \geq 1$

$$U_L^{(n)}(0,0) - U_L(0,0) \leq \alpha^n (1-\alpha)^{-1}. \tag{5.23}$$

Συνδυάζοντας έτσι τις (5.22), (5.23) ολοκληρώνεται η απόδειξη της (5.18). □

Πρόταση 5.6 *Είχαμε υποθέσει από την αρχή ότι μια κλινική δοκιμή μπορεί να έχει δύο μόνο αποτελέσματα, την επιτυχία και την αποτυχία. Η μεθοδολογία που περιγράψαμε παραπάνω επεκτείνεται ευθέως στην περίπτωση που το αποτέλεσμα μιας δοκιμής μπορεί να κατηγοριοποιηθεί σε c κλάσεις, με $c \geq 2$. Τότε η παράμετρος θ_n, θα είναι ένα διάνυσμα της μορφής $(\theta_n^1, ..., \theta_n^c)$ όπου η θ_n^i είναι η πιθανότητα το αποτέλεσμα της δοκιμής να ανήκει στην $i-$ οστή κλάση. Η Βήτα εκ των προτέρων κατανομή θα αντικατασταθεί από την Dirichlet εκ των προτέρων κατανομή και ο χώρος των καταστάσεων θα είναι ο $S = \{(s_1, s_2, ..., s_c), s_i = 0, 1, ...\}$, όπου το s_i δηλώνει τον αριθμό των δοκιμών που είχαν αποτέλεσμα το οποίο ανήκει στην $i-$ οστή κλάση, με $(1 \leq i \leq c)$. Η αμοιβή είναι μια δοσμένη συνάρτηση της κλάσης, βλέπε επίσης [Gla78].*

Κεφάλαιο 6

Υπολογισμοί και συγκρίσεις δεικτών σε άπειρο ορίζοντα

6.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιάσουμε δύο αλγορίθμους. Με τον πρώτο υπολογίζουμε τον δείκτη *Gittins* θεωρώντας το $multi-armed\ bandit\ problem$ σαν ένα πρόβλημα επανεκκίνησης. Θα χρησιμοποιήσουμε τις διαδοχικές προσεγγίσεις για ορίζοντα 200 βημάτων ο οποίος μπορεί να θεωρηθεί προσεγγιστικά άπειρος. Έτσι σε κάθε βήμα θα ελέγχει αν πρέπει να συνεχίσει ή να επανεκκινήσει στην κατάσταση που του έχει δοθεί από εμάς και για την οποία θέλουμε να υπολογίσουμε τον δείκτη.

Ο δεύτερος αλγόριθμος έχει την ίδια ακριβώς δομή με αυτόν του Κεφαλαίου 4 με την διαφορά ότι εκτελείται για μεγάλο ορίζοντα και θα υπάρχει και ο εκπτωτικός παράγοντας α. Θα τον χρησιμοποιήσουμε για να υπολογίζουμε όλα τα $\overline{y}_n(k)$. Έτσι έχοντας τον δείκτη *Gittins* από τον προηγούμενο αλγόριθμο και το αντίστοιχο $\overline{y}_n(k)$ θα μπορέσουμε να κάνουμε την σύγκριση και να δούμε αν οι δύο δείκτες είναι κοντά. Και στους δύο αλγορίθμους θα χρησιμοποιήσουμε δοκιμές *Bernoulli*. Ουσιαστικά έχουμε το μοντέλο με ένα γνωστό και ένα άγνωστο *bandit* και θα το μελετήσουμε από δύο διαφορετικές οπτικές. Και στις δύο περιπτώσεις η σύγκριση γίνεται με το γνωστό *bandit*.

Σκοπός αυτού του κεφαλαίου είναι να διαπιστώσουμε κατά πόσο τα αποτελέσματα υπολογισμού των πολιτικών είναι κοντά.

6.2 Υπολογισμός δείκτη *Gittins*

Όπως είδαμε και στο προηγούμενο κεφάλαιο το $multi-armed\ bandit\ problem$ μπορεί να χαρακτηριστεί και σαν ένα πρόβλημα επανεκκίνησης. Πάνω σε αυτή την παρατήρηση θα βασιστούμε για να μπορέσουμε να υπολογίσουμε τον δείκτη *Gittins*.

Έχουμε λοιπόν το παρακάτω μοντέλο. Έστω J οι διαθέσιμες θεραπείες και θ_j η πιθανότητα επιτυχίας της $j-$οστής θεραπείας. Εμείς θα απομωνόσουμε μια τυχούσα θεραπεία και με βάση αυτή θα προχωρήσουμε στις περαιτέρω συγκρίσεις. Χωρίς περιορισμό της γενικότητας θα θεωρήσουμε ότι η εκ των προτέρων κατανομή της τυχαίας μεταβλητής θ είναι η $Beta(0,0)$, όπου θ η άγνωστη πιθανότητα επιτυχίας της τυχούσας θεραπείας με $0 \leq \theta \leq 1$. Δηλαδή, θεωρούμε ότι δεν έχουμε καμία πληροφορία για το τυχαίο *bandit*.

Έτσι το μόνο στοιχείο που θα δέχεται ο αλγόριθμός μας είναι η κατάσταση για την οποία θέλουμε να υπολογίσουμε τον δείκτη *Gittins*.

Σε κάθε βήμα θα ελέγχουμε αν το βέλτιστο είναι να συνεχίσουμε την ακολουθιακή πορεία ή αν θα πρέπει να επανεκκινήσουμε στην κατάσταση που του έχουμε δώσει αρχικά. Επομένως ξεκινάμε έχοντας σαν εκ των προτέρων κατανομή την $Beta(a,b)$, όπου τα a,b έχουν δοθεί από εμάς. Συνολικά ο αλγόριθμός μας θα κάνει 200 βήματα, γι αυτό και μπορούμε να θεωρήσουμε ότι ο ορίζοντας μας είναι άπειρος.

Έστω ότι βρισκόμαστε σε τυχαίο βήμα n και έχουμε παρατηρήσει x επιτυχίες και y αποτυχίες. Δηλαδή η κατάσταση του συστήματος μας είναι $(a+x, b+y)$. Η εκ των προτέρων κατανομή γι αυτό το βήμα θα είναι πλέον η $Beta(a+x, b+y)$. Έτσι οι εξισώσεις βελτιστοποίησης θα έχουν διαμορφωθεί ως εξής:

$$V(a,b) = \max\Big[\frac{a}{a+b} + \alpha\left\{\frac{a}{a+b}V(a+1,b) + \frac{b}{a+b}V(a,b+1)\right\},$$

$$\frac{a+x}{a+b+x+y} + \alpha\left\{\frac{a+x}{a+b+x+y}V(a+x+1,b+y) + \frac{b+y}{a+b+x+y}V(a+x,b+y+1)\right\}\Big] \tag{6.1}$$

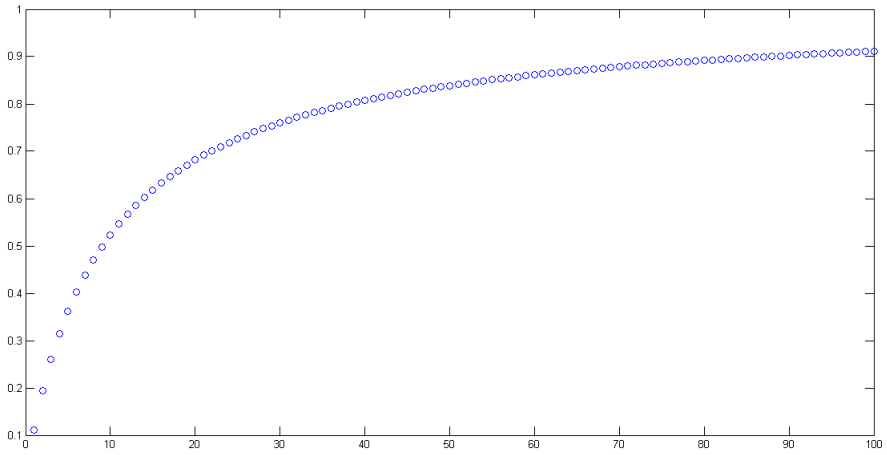
Ο πρώτος όρος της μεγιστοποίησης παραμένει σταθερό σε κάθε βήμα και αντιπροσωπεύει την επανεκκίνηση στην δοσμένη κατάσταση. Ο δεύτερος όρος μας δίνει το κέρδος σε περίπτωση που επιλέξουμε να συνεχίσουμε. Το α είναι ο εκπτωτικός παράγοντας και στις συγκρίσεις μας θα είναι $\alpha = \frac{8}{10}$.

Η επίλυση των παραπάνω εξισώσεων θα πραγματοποιηθεί με την μέθοδο των διαδοχικών προσεγγίσεων. Κοιτάμε το πρόβλημα από το τέλος προς την αρχή λόγω του δυναμικού προγραμματισμού. Σε κάθε βήμα ουσιαστικά μεταβάλλουμε τον ορίζοντα και υπολογίζουμε όλα τα δυνατά V. Αυτό συμβαίνει καθώς υπάρχει πάντα η πιθανότητα να επανεκκινήσουμε στην αρχική κατάσταση. Επομένως σε κάθε επόμενο βήμα μελετάμε το ίδιο πρόβλημα μειώνοντας ωστόσο την περίοδο κατά ένα βήμα. Γι αυτό το λόγο οι δυνατές καταστάσεις αυξάνονται πάρα πολύ γρήγορα. Σκοπός μας είναι να μπορούμε να υπολογίσουμε τον δείκτη για κάθε κατάσταση έτσι ώστε να τον συγκρίνουμε με τα αντίστοιχα αποτελέσματα του επόμενου αλγορίθμου. Μέσα από την παραπάνω διαδικασία θα μας δωθεί το $V(a,b)$ σε ορίζοντα 200 βημάτων. Όμως ισχύει ότι $m(a,b) = V(a,b)$, κάτι που φανερώνεται από την ύπαρξη των $V(a+1,b)$, $V(a,b+1)$ σε κάθε βήμα. Έτσι καταλήγουμε στο ζητούμενο. Το τελικό αποτέλεσμα του πρώτου αλγόριθμου θα είναι πολλαπλασιασμένο με (1-α), δηλαδή την πιθανότητα τερματισμού του τυχαίου *bandit*. Άρα σαν δείκτη θα χρησιμοποιούμε από εδώ και πέρα το $m(a,b)$ που είναι ο δείκτης *Gittins* πολλαπλασιασμένος με (1-α).

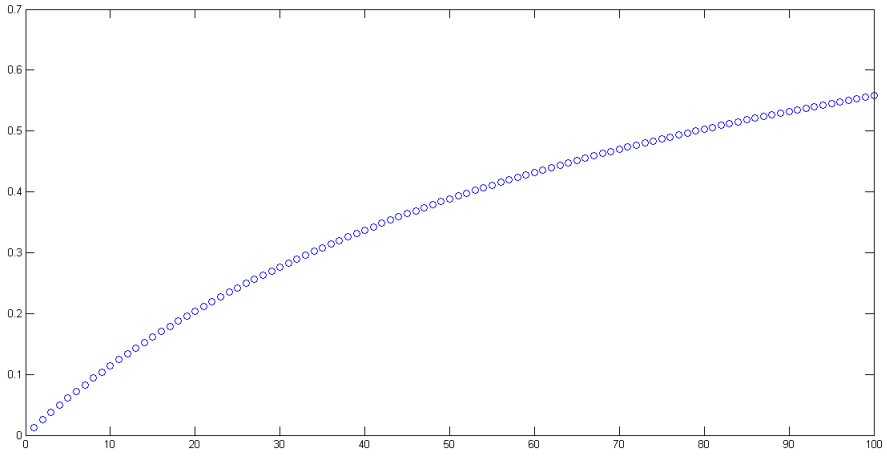
Θα δούμε τώρα κάποιες βασικές παρατηρήσεις για τον δείκτη *Gittins*.

Παρατήρηση 6.1 *Ο δείκτης είναι γνησίως αύξουσα συνάρτηση ως προς τον αριθμό των επιτυχιών όταν οι αποτυχίες μένουν σταθερές και γνησίως φθίνουσα συνάρτηση ως προς τον αριθμό των αποτυχιών όταν οι επιτυχίες μένουν σταθερές.*

Παράδειγμα 6.1 *Για* $b = 10$ *και* $a = 1, ..., 100$.

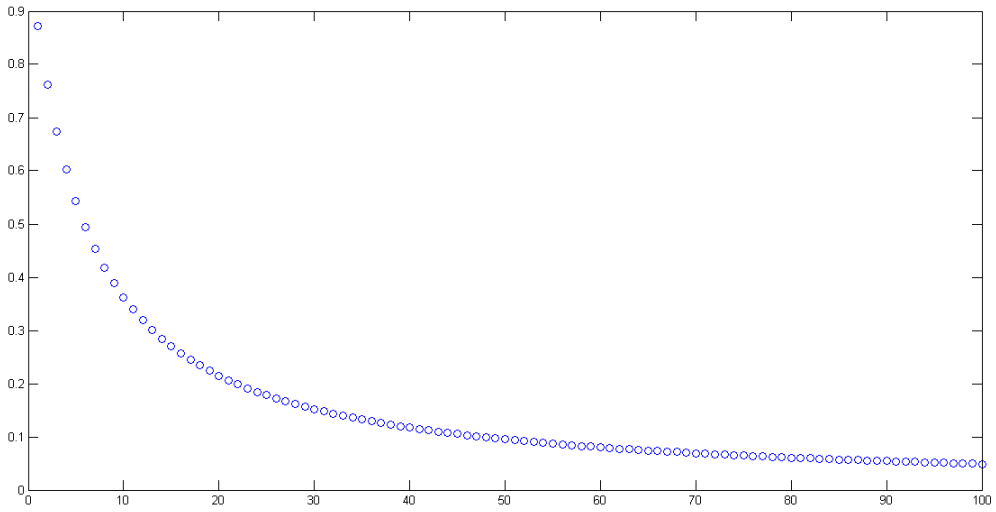


Παράδειγμα 6.2 *Για* $b = 80$ *και* $a = 1, ..., 100$.

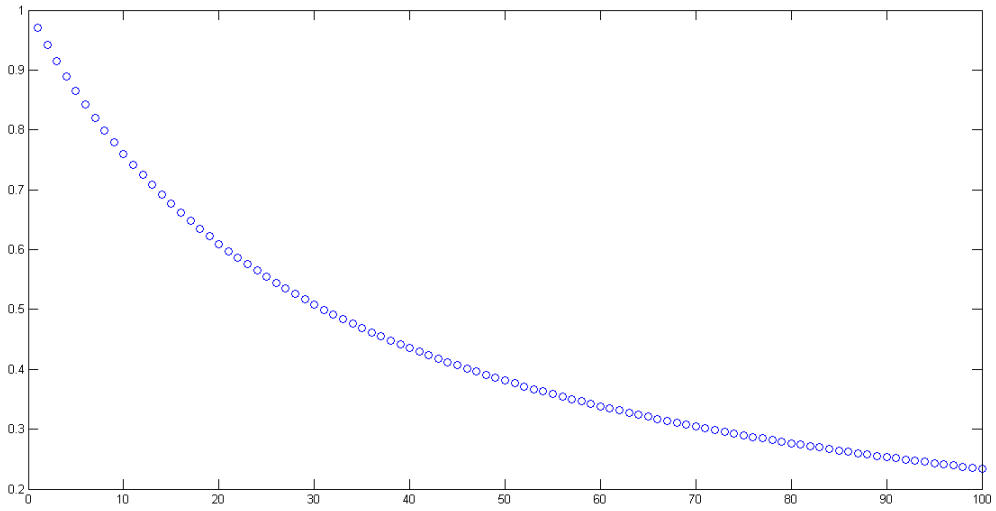


Στα παραπάνω γραφήματα στον κάθετο άξονα απεικονίζεται η τιμή του δείκτη και στον οριζόντιο η τιμή του a, δηλαδή των επιτυχιών.

Παράδειγμα 6.3 *Για* $a = 5$ *και* $b = 1, ..., 100$.



Παράδειγμα 6.4 *Για* $a = 30$ *και* $b = 1, ..., 100$.



Στα παραπάνω γραφήματα στον κάθετο άξονα απεικονίζεται η τιμή του δείκτη και στον οριζόντιο η τιμή του b, δηλαδή των αποτυχιών.

Η παραπάνω παρατήρηση οφείλεται στο ότι ο δείκτης ουσιαστικά δηλώνει το πόσο αξιόπιστο είναι ένα *bandit*. Γι αυτό άλλωστε επιλέγουμε πάντα εκείνο το *bandit* με τον μεγαλύτερο δείκτη. Άρα όσο περισσότερες επιτυχίες έχουμε παρατηρήσει, τόσο πιο πιθανό είναι να επιλέξουμε το συγκεκριμένο *bandit* στο επόμενο βήμα.

Παρατήρηση 6.2 *Ο ρυθμός αύξησης του δείκτη είναι μεγαλύτερος για μικρές τιμές του* k.

Δηλαδή έστω ότι βρισκόμαστε στην κατάσταση (1,1), όπου $m(1,1) = 0.6413$. Τότε αν πάμε στις καταστάσεις (2,1), (3,1), (4,1) θα παρατηρήσουμε ότι

$$m(2,1) = 0.7596,$$

$$m(3,1) = 0.8157,$$

και

$$m(4,1) = 0.8492.$$

Ενώ όταν είμαστε στην κατάσταση (20,20), όπου $m(20,20) = 0.5120$ παρατηρούμε ότι

$$m(21,20) = 0.5240,$$

$$m(22,20) = 0.5353,$$

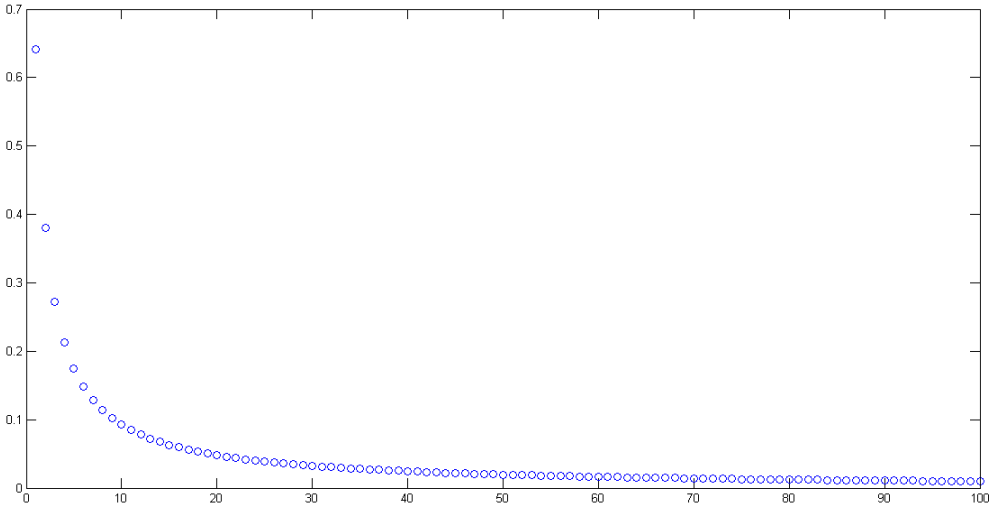
και

$$m(23,20) = 0.5462.$$

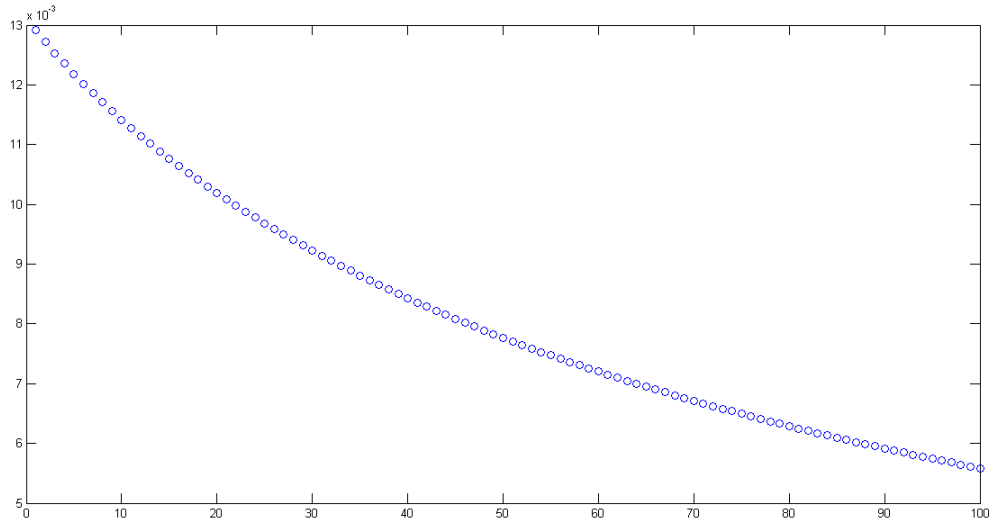
Επομένως παρατηρούμε ότι η πληροφορία παίζει πολύ σημαντικό ρόλο στον δείκτη. Όταν έχουμε παραπάνω πληροφορία δεν προκύπτει μεγάλη αλλαγή στην τιμή του δείκτη έπειτα από μία δοκιμή, ενώ όταν έχουμε λίγες δοκιμές μια επιτυχία ή αποτυχία επιφέρει σημαντικές αλλαγές. Άλλωστε, αν το μελετήσουμε καθαρά από στατιστικής πλευράς ένα δείγμα πολλών παρατηρήσεων δεν επηρεάζεται τόσο πολύ από μία ακομά παρατήρηση. Χρειάζεται πλήθος παρατηρήσεων οι οποίες έχουν κοινά χαρακτηριστικά ούτως ώστε να υπάρξει αλλαγή. Δηλαδή στην περίπτωση που εξετάζουμε, όταν το k (όπου k οι δοκιμές που έχουν παρατηρηθεί) είναι μεγάλο θα πρέπει να υπάρξει πλήθος επιτυχιών (ή αποτυχιών) για να παρατηρηθεί σημαντική αύξηση (ή μείωση αντίστοιχα) στην τιμή του δείκτη. Η συγκεκριμένη παρατήρηση γίνεται πιο εύκολα κατανοητή αν δούμε τα παραδείγματα (6.1)-(6.4) όπου για μικρό a (b) η τιμή του δείκτη αυξάνεται (μειώνεται) πιο απότομα απ΄ ότι για μεγάλο.

Θέλοντας να δούμε και κάποια άλλα χαρακτηριστικά του δείκτη θα μελετήσουμε τον λόγο $\frac{m(a,b)}{a}$, κρατώντας το b σταθερό. Γι αυτό το λόγο δημιουργήσαμε έαν αλγόριθμο ο οποίος δέχεται από εμάς την τιμή του b ,δηλαδή των αποτυχιών, και μας δίνει την γραφική παράσταση του $\frac{m(a,b)}{a}$ σε σχέση με το a. Ουσιαστικά θέλουμε να παρατηρήσουμε τι ακριβώς κάνει το πηλίκο $\frac{m(a,b)}{a}$ αυξάνοντας την τιμή του a από 1 μέχρι 100 κρατώντας σταθερή την τιμή των αποτυχιών. Έτσι θα δώσουμε κάποια παραδείγματα παρακάτω για διάφορα b.

Παράδειγμα 6.5 *Για* $b = 5$ *και* $a = 1, ..., 100$.



Παράδειγμα 6.6 *Για* $b = 80$ *και* $a = 1, ..., 100$.

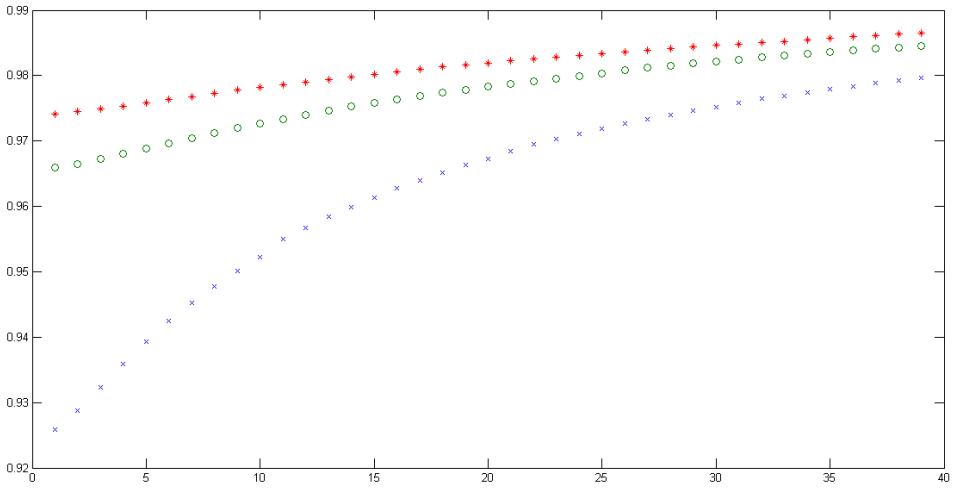


Στα παραπάνω διαγράμματα στον κάθετο άξονα έχουμε το πηλίκο $\frac{m(a,b)}{a}$ και στον οριζοντιο το a.

Παρατήρηση 6.3 *Παρατηρούμε ότι και στις δύο περιπτώσεις ο λόγος φθίνει, δηλαδή υπερισχύει το a αντί του δείκτη. Αυτό είναι φυσιλογικό καθώς* $m(a,b) \leq 1$ *για κάθε* (a,b). *Για μικρό b η μείωση είναι πιο απότομη.*

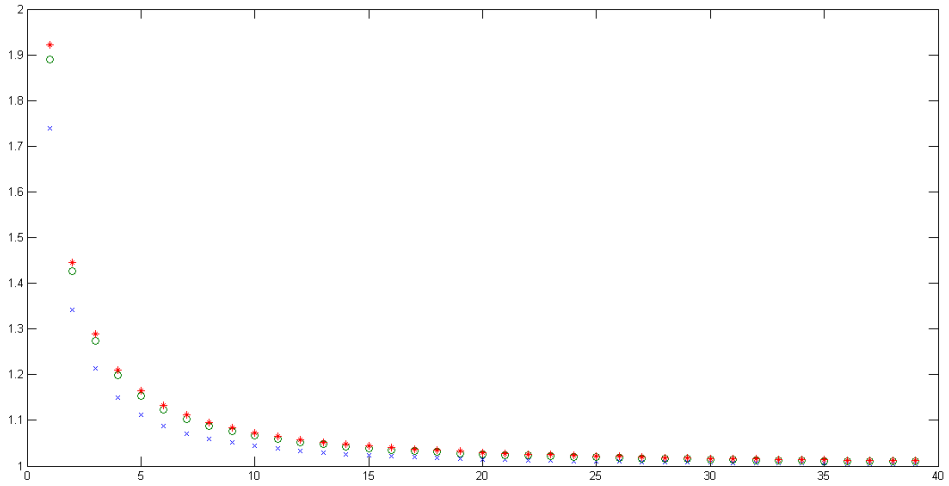
Εκτός του λόγου $\frac{m(a,b)}{a}$ είναι ενδιαφέρον να εξετάσουμε την πορεία που διαγράφει ο λόγος διαδοχικών δεικτών *Gittins*. Γι αυτό παρακάτω δίνονται κάποια γραφήματα για διάφορες τιμές των (a,b).

Παράδειγμα 6.7 *Στο παρακάτω γράφημα απεικονίζονται οι καμπύλες των λόγων* $\frac{m(a,b+1)}{m(a,b)}$ *όταν το b μεταβάλλεται από την τιμή 1 μέχρι 40 και το a μένει σταθερό στις τιμές* $a = 10$, $a = 25$, $a = 34$.



Στο παραπάνω διάγραμμα με (x) απεικονίζεται η καμπύλη όταν το $a = 10$, με (o) απεικονίζεται η καμπύλη όταν το $a = 25$ και με $(*)$ απεικονίζεται η καμπύλη όταν το $a = 34$. Στον κάθετο άξονα βρίσκονται οι λόγοι $\frac{m(a,b+1)}{m(a,b)}$ και στον οριζόντιο το b.

Παράδειγμα 6.8 *Μπορούμε επίσης να μελετήσουμε τον λόγο* $\frac{m(a,b+1)}{m(a,b)}$ *όταν η τιμή του* b *παραμένει σταθερή και μεταβάλλουμε το* a*. Στο παρακάτω γράφημα με* (x) *απεικονίζεται η καμπύλη όταν το* $b = 10$*, με* (o) *απεικονίζεται η καμπύλη όταν το* $b = 25$ *και με* $(*)$ *απεικονίζεται η καμπύλη όταν το* $b = 34$



Στον κάθετο άξονα βρίσκονται οι λόγοι $\frac{m(a+1,b)}{m(a,b)}$ και στον οριζόντιο το a. Η τιμή του a μεταβάλλεται από 1 μέχρι 40.

Παρατήρηση 6.4 *Με βάση τα δύο προηγούμενα γραφήματα συμπεραίνουμε ότι ο λόγος διαδοχικών δεικτών (*$\frac{m(a+1,b)}{m(a,b)}$ *ή* $\frac{m(a+1,b)}{m(a,b)}$*) τείνει προς την μονάδα. Είτε φθίνει και καταλήγει προσεγγιστικά στην μονάδα (*$\frac{m(a+1,b)}{m(a,b)}$*) είτε αυξάνει και σταθεροποιείται προσεγγιστικά στην μονάδα (*$\frac{m(a+1,b)}{m(a,b)}$*). Άρα παρατηρούμε ότι όσο αυξάνονται οι επαναλήψεις, είτε επιτυχίες είτε αποτυχίες, η τιμή του δείκτη σταθεροποιείται.*

Παρατήρηση 6.5 *Τέλος, όπως έχουμε δει στην Πρόταση (5.1) και στην (5.3) το πρόβλημα επανεκκίνησης μπορεί να χαρακτηριστεί ως ένα εκπτωτικό πρόβλημα σταματήματος με αμοιβή σταματήματος την τιμή του δείκτη. Άρα ο κάθε δείκτης, στο πρόβλημα επανεκκίνησης (5.3), μας δίνει ταυτόχρονα και την αμοιβή που θα είχαμε αν επιλέγαμε να σταματήσουμε σε μια τυχούσα κατάσταση* (a, b)*.*

Κώδικας.
$function[m] = restart(a, b)$
$c = \frac{8}{10}$;
$v = zeros(20502, 202)$;
$k = 0$;
$for\ j = 2 : 202$
$i1 = 0$;
$i2 = 0$;
$for\ l = 202 - j : -1 : 1$
$for\ i = 1 : l + 1$
$v(i+i1, j) = max[\frac{a+l+1-i}{a+b+l} + c*\left\{\frac{a+l+1-i}{a+b+l} * v(i + i1 + i2, j - 1) + \frac{b+i-1}{a+b+l} * v(i + i1 + i2 + 1, j - 1)\right\},$
$\frac{a}{a+b} + c * \left\{\frac{a}{a+b} * v(20502 - k - 1, j - 1) + \frac{b}{a+b} * v(20502 - k, j - 1)\right\}]$;
end
$i1 = i1 + l + 1$;
$i2 = i2 + 1$;
end
$k = k + 204 - j$;
end
$v(1, 202) = \frac{a}{a+b} + c * \left\{\frac{a}{a+b} * v(1, 201) + \frac{b}{a+b} * v(2, 201)\right\}$;
$m = v(1, 202) * (1 - c)$;
end □

6.3 Δοκιμές *Bernoulli* σε προσεγγιστικά άπειρο ορίζοντα

Στο Κεφάλαιο 4 παρουσιάσαμε έναν αλγόριθμο σε πεπερασμένο ορίζοντα ο οποίος υπολογίζει όλα τα $\overline{y}_n(k)$ και με την βοήθεια αυτών δώσαμε μια άλλη έκφραση της βέλτιστης πολιτικής. Θέλοντας όμως ένα μέτρο σύγκρισης με τον δείκτη *Gittins* θα πρέπει να μελετήσουμε την περίπτωση του άπειρου ορίζοντα. Γι αυτό τον λόγο σε αυτή την ενότητα θα καταγράψουμε έναν αλγόριθμο στον οποίο χρησιμοποιούμε μεγάλο αριθμό βημάτων έτσι ώστε να θεωρήσουμε ότι είναι προσεγγιστικά άπειρος ο ορίζοντας μας.

Η δομή του μοντέλου που θα εξετάσουμε είναι ίδια με αυτή που έχουμε αναφέρει στο Κεφάλαιο 4 με την μόνη διαφορά ότι εδώ έχουμε και τον εκπτωτικό παράγοντα α, με $\alpha = \frac{8}{10}$. Έστω ότι έχουμε δύο διαθέσιμα πειράματα. Το πρώτο έχει γνωστή πιθανότητα επιτυχίας, $p_1 = 1/2$. Για το δεύτερο δεν έχουμε κάποια πληροφορία και γι αυτό θα θεωρήσουμε μια εκ των προτέρων κατανομή. Θεωρούμε λοιπόν την $Beta(a, b)$ με $a = b = 1$. Έτσι σε κάθε βήμα βλέποντας την κατάσταση του συστήματος αποφασίζουμε ποιο από τα δύο πειράματα θα ενεργοποιήσουμε. Η κατάσταση του συστήματος όπως και στον πεπερασμένο ορίζοντα εξαρτάται μόνο από το άγνωστο πείραμα. Άρα και στην περίπτωση του άπειρου ορίζοντα έχουμε ένα πρόβλημα βέλτιστης διακοπής καθώς αν σε κάποιο βήμα επιλέξουμε το γνωστό πείραμα δεν θα υπάρξει κάποια μεταβολή στην κατάσταση του συστήματος και η επιλογή του γνωστού πειράματος θα παραμείνει βέλτιστη μέχρι το τέλος.

Έστω ότι βρισκόμαστε στο τυχαίο βήμα n (δηλαδή μένουν ακόμα n βήματα) τότε η

τελική αμοιβή σε περίπτωση που επιλέξουμε το γνωστό πείραμα θα είναι $\sum_{j=0}^{n-1} \alpha^j p_1$. Όσον αφορά το άγνωστο πείραμα δεν αλλάζει κάτι σε σχέση με αυτά που είχαμε περιγράψει στο Κεφάλαιο 4. Επομένως σε κάθε βήμα η εκ των υστέρων κατανομή του προηγούμενου βήματος γίνεται εκ των προτέρων συμπεριλαμβάνοντας έτσι την πληροφορία που μας δίνεται σε κάθε βήμα. Εδώ το κέρδος ενός βήματος για το άγνωστο πείραμα είναι $\mu(\theta) = \frac{a}{a+b}$ εφόσον βρισκόμαστε στην κατάσταση (a, b). Άρα οι εξισώσεις βελτιστοποίησης διαμορφώνονται ως εξής:

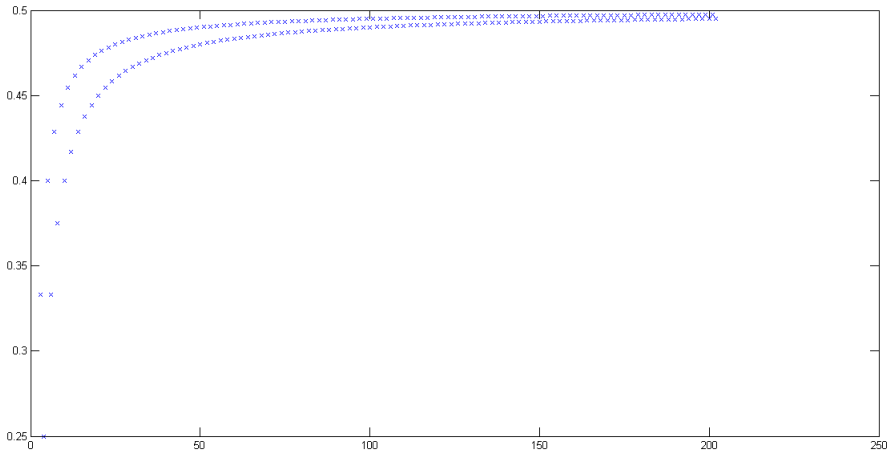
$$V_n(a,b) = \max \left\{ \sum_{j=0}^{n-1} \alpha^j p_1, \frac{a}{a+b} + \alpha[\frac{a}{a+b} V_{n-1}(a+1,b) + \frac{b}{a+b} V_{n-1}(a,b+1)] \right\}. \tag{6.2}$$

Ο αλγόριθμος έχει την ίδια δομή όπως και στην περίπτωση του πεπερασμένου ορίζοντα. Δηλαδή με την βοήθεια των διαδοχικών προσεγγίσεων λύνει τις παραπάνω εξισώσεις βελτιστοποίησης ξεκινώντας πάντα από το τέλος και πηγαίνοντας προς την αρχή του προβλήματος. Σε κάθε βήμα αποφασίζει ποιο από τα δύο πειράματα θα ενεργοποιήσει. Σε περίπτωση που προκύπτουν ίσες αμοιβές από τα δύο πειράματα τότε επιλέγει το άγνωστο πείραμα ως βέλτιστη απόφαση.

Επιπλέον σε κάθε βήμα ψάχνει και βρίσκει την κατάσταση εκείνη στην οποία για πρώτη φορα επιλέξαμε το γνωστό πείραμα. Αποθηκεύει έπειτα το κατώφλι $\overline{y}_n(k)$ για το κάθε βήμα ξεχωριστά. Έτσι στο τέλος έχουμε σε ένα πίνακα όλα τα κατώφλια τα οποία θα μας χρειαστούν για τις περαιτέρω συγκρίσεις με τον δείκτη *Gittins*. Η βέλτιστη πολιτική διαμορφώνεται και στο συγκεκριμένο μοντέλο όπως και στην (4.4).

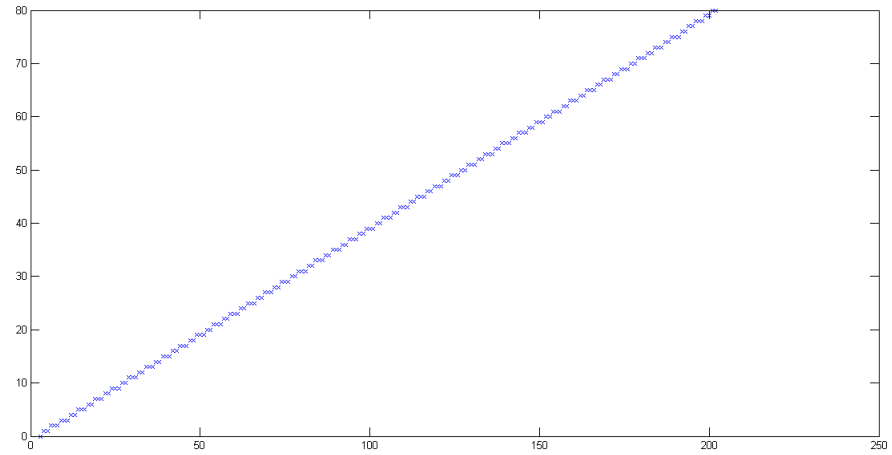
Παρατήρηση 6.6 *Η ακολουθία των $\overline{y}_n(k)$ δεν είναι αύξουσα ούτε ως προς n ούτε ως προς k. Παρ΄ όλα αυτά παρατηρείται μία ανοδική τάση των $\overline{y}_n(k)$. Όπως και στον πεπερασμένο ορίζοντα η ακολουθία των $\overline{y}_n(k)$ βρίσκεται όλη κάτω από το $p_1 = 0,5$.*

Παράδειγμα 6.9 *Στην παρακάτω εικόνα φαίνεται για $N = 200$ αυτό που αναφέρουμε στην παραπάνω παρατήρηση.*



Στον κάθετο άξονα του γραφήματος απεικονίζονται τα $\overline{y}_n(k)$ και στον οριζόντιο ο αριθμός των βημάτων k που έχουν γίνει, ο οποίος είναι αντιστρόφως ανάλογος του n, δηλαδή του αριθμού των βημάτων που μας απομένουν.

Παράδειγμα 6.10 *Το παρακάτω διάγραμμα μας δίνει τον αριθμό των επιτυχιών που έχουμε σε κάθε βήμα όταν για πρώτη φορά επιλέγουμε το γνωστό πείραμα. Παρατηρούμε ότι διαγράφει μια φθίνουσα πορεία όσο το k αυξάνει (καθώς μειώνονται δηλαδή τα βήματα που μας απομένουν).*



Στον κάθετο άξονα του γραφήματος απεικονίζεται ο αριθμός των επιτυχιών την πρώτη φορά που επιλέγουμε το γνωστό πείραμα και στον οριζόντιο ο αριθμός των βημάτων k που

έχουν γίνει, ο οποίος είναι αντιστρόφως ανάλογος του n, δηλαδή του αριθμού των βημάτων που μας απομένουν.

Παρατήρηση 6.7 *Με βάση όλα τα παραπάνω παρατηρούμε ότι και στον προσεγγιστικά άπειρο ορίζοντα οι ιδιότητες των κατωφλίων παραμένουν οι ίδιες με μικρές διαφορές λόγω του εκπτωτικού παράγοντα.*

Κώδικας.

$function[Y] = optimalstop1(N, p1)$
$c = \frac{8}{10}$;
$a = 1$;
$b = 1$;
$for\ l = N : -1 : 1$
$v = zeros(l + 2, l + 2)$;
$y = zeros(l + 1, l + 1)$;
$g = zeros(l + 1, l + 1)$;
$for\ j = 2 : l + 2$
$k = -j + 2; 4$
$for\ i = 1 : l + 3 - j$
$v(i,j) = max\left\{\frac{(1-c^(j-1))*p1*c}{1-c}, \frac{a+l+1-i+k}{a+b+l+k} + c*[\frac{a+l+1-i+k}{a+b+l+k} * v(i,j-1) + \frac{b+i-1}{a+b+l+k} * v(i+1,j-1)]\right\}$;
$if\ \frac{(1-c^(j-1))*p1*c}{1-c} > \frac{a+l+1-i+k}{a+b+l+k} + c*[\frac{a+l+1-i+k}{a+b+l+k} * v(i,j-1) + \frac{b+i-1}{a+b+l+k} * v(i+1,j-1)]$
$y(i,j) = \frac{a+l+3-i-j}{a+b+l+2-j}$;
$g(i,j) = a + l + 1 - i + k$;
end
end
end
$for\ z = 1 : l$
$Y(N - l + 1, z) = max(y(:, l + 2 - z))$;
$G(N - l + 1, z) = max(g(:, l + 2 - z))$;
end
end
$for\ j = 1 : N$
$for\ i = 1 : N + 1 - j$
$m(j,i) = a + b + i$;
end
end
end □

6.4 Σύγκριση

Στις δύο προηγούμενες ενότητες υπολογίσαμε τον δείκτη *Gittins* και τα κατώφλια $\overline{y}_n(k)$. Ο δείκτης *Gittins* υπολογίστηκε για ένα τυχαίο άγνωστο *bandit* βλέποντας το πρόβλημα $multi - armed\ bandit$ ως ένα πρόβλημα επανεκκίνησης. Για τα κατώφλια έχουμε ότι

$\overline{y}_n(k) = \frac{x}{x+y}$ όπου (x, y) είναι η κατάσταση στην οποία για πρώτη φορά είναι βέλτιστο να ενεργοποιήσουμε το γνωστό πείραμα.

Σκοπός είναι να εξετάσουμε πόσο κοντά βρίσκονται αυτές οι δύο προσεγγίσεις του προβλήματος $one - armed\ bandit$. Για την σύγκριση θα χρησιμοποιήσουμε την ποσότητα $\overline{y} + p_1 - \overline{y}_n(k)$ όπου $\overline{y} = \frac{x}{x+y}$ για κάθε κατάσταση (x, y) που μας ενδιαφέρει.

Ενδεικτικά θα αναφέρουμε τα παρακάτω αποτελέσματα:

Παράδειγμα 6.11 *Για την κατάσταση (10,10) έχουμε:*

$$m(10, 10) = 0,5231 > p_1$$

$$\overline{y} + p_1 - \overline{y}_{200}(20) = 0,5 + 0,5 - 0,4500 = 0,5500 > p_1$$

Παράδειγμα 6.12 *Για την κατάσταση (8,9) έχουμε:*

$$m(8, 9) = 0,4972 < p_1$$

$$\overline{y} + p_1 - \overline{y}_{200}(17) = 0,4705 + 0,5 - 0,4706 = 0,4999 < p_1$$

Παράδειγμα 6.13 *Για την κατάσταση (10,30) έχουμε:*

$$m(10, 30) = 0,2604 < p_1$$

$$\overline{y} + p_1 - \overline{y}_{200}(40) = 0,2500 + 0,5 - 0,4750 = 0,2750 < p_1$$

Παράδειγμα 6.14 *Για την κατάσταση (25,15) έχουμε:*

$$m(25, 15) = 0,6367 > p_1$$

$$\overline{y} + p_1 - \overline{y}_{200}(40) = 0,6250 + 0,5 - 0,4750 = 0,6500 > p_1$$

Παράδειγμα 6.15 *Για την κατάσταση (16,15) έχουμε:*

$$m(16,15) = 0,5315 > p_1$$

$$\overline{y} + p_1 - \overline{y}_{200}(40) = 0,5333 + 0,5 - 0,4839 = 0,5494 > p_1$$

Παρατήρηση 6.8 *Εξετάζοντας τα παραπάνω αποτελέσματα παρατηρούμε ότι αν είχαμε να αποφασίσουμε αν θα ενεργοποιήσουμε το άγνωστο bandit ή ένα γνωστό με πιθανότητα επιτυχίας $p_1 = \frac{1}{2}$ θα παίρναμε και με τις δύο προσεγγίσεις πάντα την ίδια απόφαση. Άρα και οι δύο δείκτες μας οδηγούν στην ίδια πολιτική.*

Παρατήρηση 6.9 *Με βάση όλα τα παραπάνω μπορούμε να ισχυριστούμε ότι το πρόβλημα $one-armed\ bandit$ με ενα γνωστό και με ένα άγνωστο πείραμα είναι ένα πρόβλημα $two-armed\ bandit$ όπου το γνωστό bandit έχει σταθερό δείκτη.*

Βιβλιογραφία

[Ber76] Bertsekas, D. P.: *Dynamic Programming and Stochastic Control.* Academic Press, 1976.

[Bra56] Bradt, R.N., Johnson S.M. & Karlin S.: *On sequential designs for maximizing the sum of n observations.* In *Annals of Mathematical Statistics 27: 1060-1074.*, 1956.

[Bur97] Burnetas, A.N. & Katehakis, M.N.: *On the finite horizon one-armed bandit problem.* Stochastic Analysis and Applications 16(1) 845-859, 1997.

[Bur03] Burnetas, A.N. & Katehakis, M.N.: *Analysis for the finite horizon one-armed bandit problem.* Probability in the Engineering and Informational Sciences 17, n.1, pp. 53-82, 2003.

[Cox74] Cox, D.R. & Hinkley, D.V: *Theoretical statistics.* New York: Chapman & Hall, 1974.

[Der70] Derman, C.: *Finite State Markovian Decision Processes.* Academic Press, 1970.

[DY79] Dynkin, E. B. and A. A. Yushkevich: *Controlled Markov Processes.* Springer–Verlag, 1979.

[Git79] Gittins, J. C.: *Bandit processes and dynamic allocation indices.* In *Journal of the Royal Statistics Society B 41: 148-164*, volume 41, pages 335–340, 1979.

[GJ74] Gittins, J. C. and D. M. Jones: *A dynamic allocation index for the sequential design of experiments.* In J. Gani, K. Sarkadi and I. Vince (editors): *Progress in Statistics*, volume 1, pages 241–266. North Holland, 1974.

[GJ79] Gittins, J. C. and D. M. Jones: *A dynamic allocation index for the discounted multiarmed bandit problem.* Biometrika, 66:561–565, 1979.

[Gla78] Glazebrook, K. D.: *On the optimal allocation of two or more treatments in a controlled clinical trial.* Biometrika, 65:335–340, 1978.

[Kat85] Katehakis, M.N. $ Veinott Jr., A.F.: *The multi-armed bandi problem: Decomposition and computation.* Mathematics of Operations Research, 1985.

[KD85] Katehakis, M. N. and C. Derman: *Computing optimal sequential allocation rules in clinical trials.* In Ryzin, J. van (editor): *Adaptive Statistical Procedures*

and Related Topics, volume 8, pages 29–39. I.M.S. Lecture Notes – Monograph Series, 1985.

[Rob52] Robbins, H.: *Some aspects of the sequential design of experiments.* Bull. Amer. Math. Monthly, 58:527–536, 1952.

[Ros83] Ross, S. M.: *Introduction to Stochastic Dynamic Programming.* Academic Press, 1983.

[Whi80] Whittle, P.: *Multi-armed bandits and the gittins index.* 42:143–149, 1980.

[Whi82] Whittle, P.: *Optimization Over Time*, volume 1. J. Wiley, New York, 210-220, 1982.