

עיצוב מוקד שירות טלפוני עם לקוחות קצרי רוח

חיבור על מחקר
לשם מילוי חלקי של הדרישות לקבלת תואר
מגיסטר למדעים
בחקר ביצועים וניתוח מערכות

עפר גרנט

הוגש לסנט הטכניון - מכון טכנולוגי לישראל
אדר תשנ"ח חיפה מרץ 1998

המחקר נעשה בהנחייתו של פרופ' אבישי מנדלבאום בפקולטה להנדסת תעשייה וניהול.

ברצוני להודות לפרופ' מנדלבאום על הדרך המוצלחת בה הנחה אותי במהלך המחקר, על סבלנותו הרבה, ונכנותו לעזור בכל עת. תודה גם ל-Dr. Marty Reiman מ-Bell Labs על עזרתו המועילה.

אני מודה לבית הספר ללימודי מוסמכים בטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

תוכן ענינים

4	מבוא
9	1 המודל האנליטי
11	1.1 תיאור כתהליך לידה ומוות
12	1.2 תיאור אלטרנטיבי כרשת דו-צמתית
13	1.3 מדדים בסיסיים - ההסתברות להמתנה וההסתברות לנטישה
15	1.4 ההסתברות לנטישה - קשרים שימושיים
19	2 סקר ספרות
27	3 חישוב מדויק של מדדים במערכת $M/M/N/B+M$ (קצב נטישה θ)
32	4 קירובים
35	5 התנהגות אסימפטוטית של ההסתברות לנטישה
38	6 התנהגות אסימפטוטית בעומס גבוה (Heavy Traffic)
41	6.1 ההסתברות לנטישה וההסתברות להמתנה ($\theta_N \propto N^\gamma$)
41	6.1.1 $\gamma = 0$
44	6.1.2 $\gamma > 0$
47	6.1.3 $\gamma < 0$
50	6.2 התכנסות חלשה של ההתפלגויות הסטציונריות $\{q_N(\infty)\}$
56	6.3 התכנסות חלשה של התהליכים הסטוכסטיים $\{q_N\}$
70	6.4 קירובים להתפלגות הסטציונרית - סיכום

תוכן ענינים המשך:

73	7	עיצוב מוקד שירות - דיון
75	7.1	בדיקת טיב הקירובים
76	7.1.1	תלות במספר המוקדנים
79	7.1.2	יישום
82	7.1.3	קצבי נטישה "קיצוניים"
84	7.2	כללי אצבע
87	7.3	מדד לרמת שירות
89		סיכום / מחקר עתידי
96		נספח: קוד MATLAB לחישוב מדדים מדויקים

רשימת איורים

5	מודל סכמטי של מוקד שירות טלפוני	0.1
11	$\{Q(t), t \geq 0\}$ - דיאגרמת מעברים	1.1
33	$\{q_N(t), t \geq 0\}$ קירובים לתהליך	4.1
81	$M(95)/M(1)/100 + M$ במערכת $P\{Wait > 0\}$ כתלות ב- θ	7.1
82	$M(95)/M(1)/100 + M$ במערכת $P\{Ab\}$ עם קצבי נטישה "איטיים"	7.2
83	$M(95)/M(1)/100 + M$ במערכת $P\{Ab Wait > 0\}$ עם קצבי נטישה "איטיים"	7.3
84	$M(95)/M(1)/100 + M$ במערכת $P\{Ab Wait > 0\}$ עם קצבי נטישה "מהירים"	7.4

רשימת טבלאות

28		$E[f(V, R)]$	מדדים מהצורה	3.1
42		HW	ריכוז תוצאות עבור התנהגות אסימפטוטית בתחום	6.1
76		HW	התנהגות אסימפטוטית של ההסתברות להמתנה בתחום	7.1
77		0.05	מספר מוקדנים מינימלי לקבלת דיוק של	7.2
79		0.01	$\gamma = 0$ - מספר מוקדנים מינימלי לקבלת דיוק של	7.3

תקציר

בעשורים האחרונים ניכר גידול מהיר במספר החברות אשר מספקות שירותים דרך הטלפון, ובהיקף השירותים הניתנים. בעבודה זו אנו מתמקדים במוקדים טלפוניים שמספקים שירותי מידע, שירותים עיסקיים, נותנים שירות ללקוחות או מבצעים מכירות. אלה יכוננו "מוקדי שירות טלפוני". מוקדים מסוג זה גדולים בממוצע (במובן של מספר נותני שירות), ביחס למוקדים טלפוניים ולא-טלפוניים אחרים.

האתגר הכללי בתכנון והפעלה של מוקד שירות טלפוני הוא השגת איזון בין יעילות ורמת שירות. בשירות הטלפוני זמן התגובה חייב להיות קצר יחסית (שניות), ובו בזמן היקף הפניות למוקד עשוי להגיע לרבבות ביום. מצב זה מחייב התבססות על ניתוחים של מודלים אנליטיים כדי לנסות ולהשיג את אותו איזון מבוקש. מודל מקובל לניתוח מוקד שירות עם מספר מוקדנים הוא $M/M/N/B$, אך חסר בו מאפיין חשוב: במוקדי שירות, ובפרט בשירות טלפוני, הלקוחות נוטים להיות קצרי רוח, עובדה המתבטאת בכך שלקוחות הממתינים בתור עלולים לעזוב (לנטוש) לפני תחילת שירותם.

מסתבר שנסיון לעצב מוקד על פי תוצאות ניתוח של מודל $M/M/N/B$ (ללא נטישות) יכול לגרור הקצאת "עודף" לא מבוטל של מוקדנים. נטישה של לקוח היא פגיעה ברמת השירות שלו, אך משמעותה גם הורדת עומס מהמערכת ושיפור רמת השירות של הלקוחות האחרים. מכאן שחשוב לספק למתכנן/מנהל מוקד כלים בעזרתם יוכל לבחון את ההשפעה של נטישות על רמת השירות. כמו כן, הוספת תופעת הנטישות למודל $M/M/N/B$ מוסיפה "מימד" למדדים, ומתכנן/מנהל המוקד יכול לכלול בשיקוליו התייחסות ללקוחות "נוטשים", בנוסף ללקוחות "ממתינים".

אנו מטפלים במקרה הפשוט ביותר לניתוח, בו סבלנות הלקוח מתוארת על ידי משתנה מקרי מעריכי שאינו תלוי בתהליכי המופע, השירות ובמצב המערכת. הנחת אי תלות במצב המערכת מתאימה למערכות שירות טלפוניות דווקא, בהן התורים "נסתרים" במובן שללקוח אין אינפורמציה על מצב המערכת, זאת לעומת המצב בבנק למשל, בו הלקוח רואה את התור, והתנהגותו עשויה להיות מושפעת מכך. הסימון בו נשתמש למודל $M/M/N/B$ עם נטישות "מעריכיות" הוא $M/M/N/B + M$.

אנו מציגים שתי סכימות לחישוב מדויק של מדדים במודל $M/M/N/B + M$, תחת ההנחה

שהמערכת נמצאת במצב יציב. בהמשך אנו מתמקדים בניתוח אסימפטוטי של מערכת $M/M/N + M$ בתחום המתאים למערכות שירות גדולות (מספר שרתים גדול, נצילות גבוהה, רמת שירות גבוהה). כאן ההנחה היא שאין מגבלה על קיבולת ההמתנה במערכת. הנחה זו תואמת את המדיניות המקובלת במוקדים רבים לפיה מקצים מספר גדול של קווים לממתנים. הניתוח האסימפטוטי נעשה בשתי רמות:

1. עבור שני מדדים בסיסיים - פרופורציית הלקוחות שנאלצים להמתין בתור ופרופורציית הלקוחות שנוטשים את המוקד לפני תחילת שירותם.

2. ההתפלגות הסטציונרית של התהליך הסטוכסטי של מספר הלקוחות במערכת. כאן נקטנו בשתי גישות - מציאת הגבול החלש של סדרת התפלגויות סטציונריות, ומציאת ההתפלגות הסטציונרית של הגבול החלש של סדרת תהליכים סטוכסטיים (בשתי הדרכים התוצאות כמובן זהות).

ההתנהגות האסימפטוטית שמתקבלת תלויה בערכי הפרמטרים של המערכת והיחסים ביניהם. מטרתנו הראשונה היא להשתמש בתוצאות האסימפטוטיות כקירובים. ניתן לקבוע עבור אילו ערכי פרמטרים מתקבלים הקירובים ה"מוצלחים" ביותר, ובכך מוגדר "תת-מודל" המתאים ביותר לניתוח מוקדי שירות טלפוניים עם לקוחות קצרי רוח. על פי ההתנהגות האסימפטוטית של תת-מודל זה, מתקבלים קירובים למדדים שונים של רמת השירות. על סמך תוצאות אלה אנו מנסחים "כללי אצבע" עבור עיצוב מוקד.

רשימת סמלים

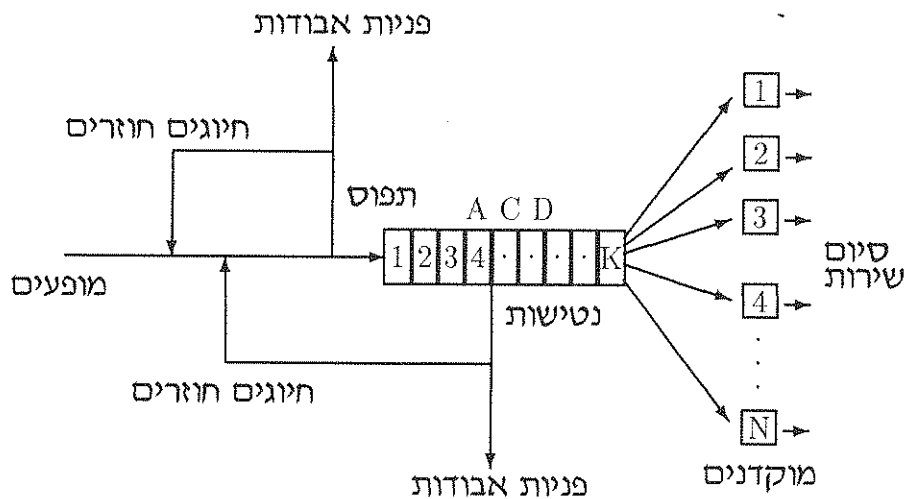
- $\phi(x)$ פונקציית צפיפות של משתנה מקרי נורמלי סטנדרטי
- $\Phi(x)$ פונקציית התפלגות מצטברת של משתנה מקרי נורמלי סטנדרטי
- $\gamma(x, y)$ Incomplete Gamma Function
- $\xrightarrow{d}$ התכנסות חלשה (התכנסות בהתפלגות)
- $\stackrel{d}{=}$ שיוויון בהתפלגות
- $a \wedge b$ המינימום בין a ו- b
- $a \vee b$ המקסימום בין a ו- b

מבוא

אחד המאפיינים של חתך התעסוקה במאה העשרים הוא העליה המתמדת באחוז כוח העבודה המועסק בתחום השירות - כיום 80%-60% במדינות המפותחות [5]. בפרט, בעשורים האחרונים ישנו גידול ניכר במספר החברות אשר מספקות שירותים דרך הטלפון, ובהיקף השירותים הניתנים - לפי אמד אחד היו בארה"ב ב-1994 כ-350,000 חברות שהעסיקו כ-6.5 מיליון עובדים, וקצב הגידול השנתי עמד על כ-20% [9]. אופי השירותים הטלפוניים הניתנים הוא מגוון: מידע, חירום, עסקי, שירות לקוחות ועוד. גם הארכיטקטורה וההיקף של מוקד טלפוני יכולים להיות שונים מאד זה מזה: החל מקו טלפון יחיד שמאויש על ידי מוקדן יחיד ועד לאלפי קוים ש"מזינים" אלפי מוקדנים. יתכן שלקוח יקבל את השירות ממוקדן יחיד, או שיועבר בין מוקדנים שונים לקבלת סדרת שירותים או עד למציאת נותן השירות המתאים.

האתגר הכללי בתכנון והפעלה של מערכת שירות הוא השגת איזון בין יעילות (ניצול משאבים) ורמת שירות. כדי לבצע משימה זו צריך תחילה לקבוע עבור המערכת הספציפית כיצד מוגדרים בה באופן כמותי היעילות ורמת השירות (למשל, יעילות כאחוז הזמן הממוצע שנותני השירות עסוקים, ורמת שירות כאחוז הלקוחות שממתנינים פחות מפרק זמן מסוים עד לתחילת שירותם). לאחר ההגדרה הכמותית יש צורך למדוד את הגדלים הנ"ל כדי להעריך את תפקודה של המערכת ולהחליט האם וכיצד יש צורך לשנותה. גישה זו של מדידה מדויקת אומצה על ידי חברות המספקות שירותים רק בשנות ה-80, למרות שיושמה הרבה קודם בתחום הייצור, למשל. במובן זה, אחד היתרונות של השירות הטלפוני לעומת סוגי שירות רבים אחרים, הוא הקלות היחסית בה ניתן לבצע מדידות של גדלים הקשורים בשירות הניתן - משך השירות, משך ההמתנה, מספר ממתנינים וכדומה. אף על פי כן מעטים עדין המקומות בהם מבצעים מדידות כאלה בצורה מושכלת.

איור 0.1: מודל סכמטי של מוקד שירות טלפוני
Schematic model of a telephone call center



בעבודה זו נשתמש במושג "מוקד שירות טלפוני" עבור מוקדים טלפוניים שמספקים שירותי מידע, שירותים עסקיים, נותנים שירות ללקוחות או מבצעים מכירות. מוקדים מסוג זה גדולים בממוצע, ביחס למוקדים טלפוניים ולא-טלפוניים אחרים. סקלת הזמן של השירות הטלפוני חייבת להיות קצרה יחסית - המתנה של שניות, בהשוואה לממשקי שירות אחרים: מכתב - ימים, פקס / דואר אלקטרוני - שעות, שירות פנים-אל-פנים - דקות. כמו כן, היקף הפניות אל ספקי שירותים טלפוניים יכול להיות עצום (למשל, סדר גודל של 40,000 פניות ביום לבזק-144). כדי שניתן יהיה במצב זה לתת שירות ברמה ויעילות גבוהים (למשל, 90% מהלקוחות ממתינים פחות מ-10 שניות והמוקדנים עסוקים למעלה מ-80% מזמנם), יש צורך ביכולת לנתח מוקדי שירות כאלה בצורה מדויקת.

מודל סכמטי מקובל של מוקד שירות טלפוני מופיע באיור 0.1. זהו מודל של מוקד שירות בעל תור יחיד ומספר מוקדנים זהים סטטיסטית העובדים במקביל. המופעים אל המערכת הם לקוחות המנסים ליצור קשר טלפוני. למוקד יש מספר קוים אשר מחוברים ל-ACD (Automatic Call Distributer). ה-ACD מנהל את תור הממתינים ומנתב לקוחות למוקדנים

פנויים. לקוח אשר "מופיע" למערכת כאשר כל הקווים תפוסים שומע צליל תפוס / הודעה. יתכן שהוא ינסה שוב מאוחר יותר (retrial) או שיוותר (lost demand). לקוח שמצליח "להכנס" למערכת, בזמן שכל המוקדנים עסוקים, נאלץ להמתין. אם סבלנותו פוקעת לפני ששירותו מתחיל הוא מנתק ו"נוטש" (abandon). גם אחרי נטישה יתכן נסיון נוסף או ויתור. תחת מודל זה בדרך כלל מניחים שהפרמטרים עליהם שולט מנהל/מתכנן המוקד הם מספר הקווים (K) ומספר המוקדנים (N).

מיד ניתן לראות שהמודל המוצג באיור 0.1 אינו מכסה את כל צורות השירות שניתנות במוקדי שירות טלפוני - השרתים זהים, אין מעבר בין שרתים, ואין אפשרות לתאר שירות שמתקבל בכמה "שלבים" על ידי מעבר בין שרתים שונים בצורה של רשת. כמו כן, במציאות למתכנן/מנהל המוקד יש דרכים נוספות להשפיע על תפקוד המוקד - שימוש במענה קולי, פרסום שמשפיע על זרם הפונים ועוד. זהו מודל פשוט יחסית, אך עדין מתאים למערכות רבות. כפי שמובא בהמשך העבודה, גם עבור מודל זה היה צורך במספר לא מבוטל של הנחות כדי להגיע למודל אנליטי שניתוחו לא מורכב מידי. במקרים בהם הניתוח האנליטי מורכב מאוד, אבל לא רוצים להוסיף הנחות מפשטות, ניתן לעשות שימוש בסימולציה (למשל [9]).

חשוב לציין שהמודל הסכמטי באיור 0.1 והמודלים האנליטיים בהמשך העבודה, מתאימים גם למערכות רבות שאינן מוקדי שירות טלפוניים. מודלים אלה יכולים להתאים לכל מערכת בה יש "לקוחות" שמקבלים "שירות" ולעיתים ממתנינים ב"תור" לשם כך. למשל - מערכות שירות שאינן טלפוניות (בנק, סופרמרקט), מכונות בצומת כבישים, ג'ובים במחשב מרכזי, מוצרים במחסן מלאי ועוד. הסיבה שבעבודה זו ישנה התייחסות מפורשת למוקדי שירות טלפוניים היא שתחומי הפרמטרים בהם נתמקד רלוונטיים במיוחד למערכות כאלה. בפרט, במקרה של קירובים נצפה להתאמה טובה יותר עבור מוקדים עם מספר גדול של מוקדנים שעובדים בעומס גבוה. כמו כן, ישנו הבדל משמעותי בין שירות טלפוני לסוגי שירות אחרים (פרט לסקלת הזמן שצוינה קודם) - בשירות הטלפוני קיימים תורים "נסתרים". את התורים במערכת הלקוח לא יכול לראות, ולכן התנהגותו (החלטה האם ומתי לנטוש למשל) אינה תלויה בגודלם.

מודל מקובל למוקד שירות טלפוני הוא ה-Erlang-C (נהוג לסמנו $M/M/N$), בו מתעלמים מהנטישות ומחיוגים חוזרים. בעבודה זו נתרכז בתופעת הנטישות, ונתעלם מחיוגים חוזרים. ההנחה היא שברוב המקרים פרק הזמן שעובר עד לביצוע חיוג חוזר, גדול יחסית לזמן הממוצע בין מופעים עוקבים למוקד. לכן החיוגים החוזרים, שמהווים חלק מהמופעים למוקד, אינם "מקלקלים" את התכונות של תהליך המופע אותן נניח בהמשך. התייחסות לנטישות במודל של תור היתה כבר ב-1937 על ידי Palm [29], בהקשר של מתקשרים אל מרכזיית טלפונים ידנית. מאז שנות ה-50 הופיעו עבודות רבות שמטפלות באספקטים שונים של הנושא. החשיבות בהוספת טיפול בנטישות, מעבר לכך שהמודל מדויק יותר, היא העובדה שכיום יש מודעות גבוהה לתופעת הנטישות בקרב מתכנני/מנהלי מוקדי שירות טלפוניים. יותר מכך, רבים מתייחסים אל אחוזי הלקוחות הנוטשים כאחד המרכיבים המרכזיים של רמת השירות. משמעות אופרטיבית של השימוש במודל עם נטישות מופיעה בתחילת פרק 7.

ישנם מדדים שונים הקשורים ברמת השירות, על פיהם ניתן להגדיר אם השירות טוב או אינו טוב. הגדרות אלה תלויות בין השאר באופי השירות: במוקד חירום חשובה מאד הזמינות - להמנע ממצב שמטלפן מקבל "תפוס". במוקד עסקי יתכן שירצו להקטין מאד את מספר הלקוחות שהמתינו זמן רב לקבלת שירות. להלן מספר מדדים בסיסיים אשר רלוונטים למוקד שירות טלפוני במצב יציב:

1. אחוז הנוטשים = ההסתברות לנטישה ($P\{Ab\}$).
2. אחוז הממתנים יותר מ- t יחידות זמן בתור = ההסתברות שלקוח ממתין יותר מ- t יחידות זמן בתור ($P\{Wait > t\}$).
3. אחוז הנוטשים מאלה שממתנים יותר מ- t יחידות זמן בתור = ההסתברות שלקוח נוטש בהנתן שממתין יותר מ- t יחידות זמן בתור ($P\{Ab|Wait > t\}$).
4. ממוצע זמן המתנה בתור עד לנטישה = תוחלת זמן ההמתנה בהנתן נטישה ($E[Wait|Ab]$).
5. ממוצע זמן ההמתנה בתור של לקוחות שהמתינו יותר מ- t יחידות זמן ואז נטשו = תוחלת זמן ההמתנה בהנתן המתנה מעל t יחידות זמן ונטישה ($E[Wait|Wait > t; Ab]$).
6. התוחלת של פונקציה כלשהי g של זמן ההמתנה ($E[g(Wait)]$), למשל לחישוב מומנטים מעבר לתוחלת.

שאר העבודה מאורגן באופן הבא: בפרק 1 נתאר את המודל האנליטי הבסיסי בו נטפל, ונפרט מספר תוצאות אלמנטריות. פרק 2 הוא סקר של הספרות הרלוונטית לעבודה. בפרק 3 נפרט שתי סכימות כלליות לחישוב מדדים מדויקים במודל האנליטי. בפרק 4 מתחיל החלק המרכזי של העבודה שמטרתו למצא קירובים למדדים שונים במודל. פרק 4 מהווה מבוא לנושא זה. בפרקים 5-6 ישנם ניתוחים אסימפטוטיים של ההסתברות לנטישה וההסתברות להמתנה, והוכחת התכנסות חלשה של התהליך הסטוכסטי של מספר הלקוחות במערכת וההתפלגות הסטציונרית שלו. פרק 7 בוחן את הקירובים שהתקבלו, ומופיעים בו כללי אצבע לעיצוב מוקד שירות טלפוני.

פרק 1

המודל האנליטי

בפרק הקודם הוצג מודל סכמטי של מוקד שירות טלפוני. כדי לנתח את המודל באופן אנליטי, יש צורך לעבור מהתיאור הסכמטי לתיאור אנליטי וזאת על ידי הנחות כלשהן על התהליכים במודל. כמקובל במידול מערכות, אנו מפשטים או מקרבים את המציאות כדי שנקבל מודל אנליטי בר ניתוח. אנו נטפל במודל בסיסי של מוקד שירות עם לקוחות קצרי רוח. במודל זה מופע הלקוחות הוא אל תור יחיד אשר "מזין" מספר מוקדנים. לקוח אשר ממתין בתור נוטש (ולא חוזר) אם זמן ההמתנה עובר את סף הסבלנות שלו. לקוח אשר מסיים את שירותו עוזב. בנוסף נניח את ההנחות הבאות:

- N מוקדנים בלתי תלויים (ב"ת), זהים סטטיסטית, אשר מספקים שירות שנמשך זמן המפולג $exp(\mu)$ (כלומר ממוצע זמן שירות הוא $1/\mu$).
- מופע הלקוחות הוא על פי תהליך פואסון הומוגני עם פרמטר λ (כלומר קצב מופע λ). זמני השירות ב"ת בתהליך המופע.
- מדיניות השירות היא FCFS (First-Come-First-Served).
- המוקדנים אינם בטלים כל עוד יש לקוחות בתור - לקוח אשר מופיע כאשר יש מוקדנים פנויים משורת מיד.
- סבלנותו של כל לקוח (משך הזמן שהוא מוכן להמתין בתור לפני נטישה) היא משתנה מקרי (מ"מ) מעריכי עם ממוצע $1/\theta$, ב"ת בלקוחות האחרים.
- במערכת יש מקום ל- B לקוחות (כלומר יתכנו לכל היותר $K = B - N$ ממתנים). לקוח אשר מופיע כאשר כל מקומות ההמתנה מאוישים, עוזב.

למעשה יש פה שילוב של נטישות במערכת $M/M/N/B$. מכיוון שסבלנותו של לקוח היא משתנה מקרי מעריכי, מתקבלת מערכת קלה יחסית לניתוח אנליטי מדויק. נסמן מערכת זו ב- $M/M/N/B+M$ (על פי הסימונים שב-[2] Baccelli and Hebuterne מערכת $G/G/N/B$ עם נטישות, כאשר התפלגות הזמן עד לנטישה היא כללית מסומנת $G/G/N/B+G$).

חשוב לציין שבמודל $M/M/N/B+M$ מניחים שסבלנותו של לקוח לא תלויה במקומו בתור. סביר שבתורים בהם לקוחות "עומדים", לקוח בקדמת התור יהיה בדרך כלל בעל סבלנות רבה יותר מאשר לקוח בסוף התור, אשר מעריך שיהיה עליו להמתין זמן רב ולכן עלול לנטוש במהרה. המודל הנ"ל מתאים יותר למצב הקיים היום במערכות טלפוניות בהן קיימים תורים "נסתרים", וללקוח אין אינפורמציה על מצב המערכת, בפרט הוא אינו יודע את מיקומו בתור.

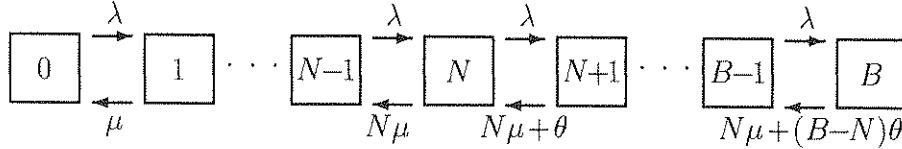
נשים לב שעבור $B = \infty$ זו מערכת $M/M/N+M$. כידוע, במערכת $M/M/N$ כאשר $\frac{\lambda}{N\mu} \geq 1$ אין מצב יציב. תכונה חשובה של מערכת $M/M/N+M$ היא שתמיד קיים מצב יציב - חסום סטוכסטית מלעיל ומלרע על ידי מערכות $M/M/\infty$ (בהן ההתפלגות הסטוכסטית של מספר הלקוחות במערכת היא פואסונית) עם קצבי שירות $\mu \wedge \theta$ ו- $\mu \vee \theta$ בהתאמה (כאן " $a \wedge b$ " הוא המינימום בין המספרים a ו- b , ו- " $a \vee b$ " המקסימום ביניהם). לכן במקרה זה הדיון בהתפלגות סטוכסטית אינו מוגבל על ידי ערכי הפרמטרים. עבור מודל עם סבלנות מהתפלגות כללית F התנאי לקיום מצב יציב הוא (ראה [2]):

$$(1.1) \quad \lambda(1 - F(\infty)) < N\mu .$$

תנאי זה מבטא את הדרישה שהמוקדנים יוכלו לפחות לעמוד בעומס של אותם לקוחות שסבלנותם אינה מוגבלת.

הערה: כאשר $B < \infty$ תמיד קיים מצב יציב (עם ובלי נטישות).

איור 1.1: $\{Q(t), t \geq 0\}$ - דיאגרמת מעברים
 $\{Q(t), t \geq 0\}$ - Transition diagram



1.1 תיאור תהליך לידה ומוות

את מצב המערכת בזמן t נתאר על ידי מספר הלקוחות במערכת (מקבלים שירות או ממתנינים) ונסמנו $Q(t)$. מכאן שמצב המערכת לאורך זמן יתואר על ידי התהליך הסטוכסטי $Q = \{Q(t), t \geq 0\}$. מכיוון שזמני השירות, סבלנות הלקוחות והזמנים הבין-מופעיים הם כולם ב"ת ומפולגים מעריכית, הרי ש- Q הוא תהליך לידה ומוות עם קצבי לידה (λ_k) ומוות (μ_k) הנתונים על ידי:

$$(1.2) \quad \lambda_k = \begin{cases} \lambda, & 0 \leq k < B \\ 0, & \text{otherwise} \end{cases}; \quad \mu_k = \begin{cases} (N \wedge k)\mu + [k - N]^+\theta, & k \leq B \\ 0, & \text{otherwise} \end{cases}$$

תיאור התהליך Q באמצעות דיאגרמת מעברים מופיע באיור 1.1.

פתרון משוואות המצב היציב נותן:

$$(1.3) \quad \pi_k = \begin{cases} \frac{(\lambda/\mu)^k}{k!} \pi_0, & 0 \leq k \leq N \\ \prod_{j=N+1}^k \left(\frac{\lambda}{N\mu + (j-N)\theta} \right) \frac{(\lambda/\mu)^N}{N!} \pi_0, & N < k \leq B \end{cases}$$

כאשר

$$(1.4) \quad \pi_0 = \left[\sum_{k=0}^N \frac{(\lambda/\mu)^k}{k!} + \sum_{k=N+1}^B \prod_{j=N+1}^k \left(\frac{\lambda}{N\mu + (j-N)\theta} \right) \frac{(\lambda/\mu)^N}{N!} \right]^{-1}.$$

1.2 תיאור אלטרנטיבי כרשת דו-צמתית

נציג דרך אחרת לתיאור מערכת $M/M/N + M$, בצורה של רשת. כאן מצב המערכת יתואר על ידי וקטור דו מימדי - מספר הלקוחות במערכת שמקבלים שירות או יקבלו שירות בעתיד, ומספר הלקוחות במערכת שנוטשים לפני תחילת שירותם. עבור כל לקוח שמגיע "מוגרל" מראש, בהתאם למצב המערכת, האם הוא מקבל שירות או שהוא נוטש. תוצאת הגרלה זו קובעת לאיזה מבין שני "מוקדי שירות" הוא מגיע. מוקד אחד הוא המוקד ה"מקורי" בו יש N מוקדנים שנותנים שירות בקצב μ , ולא קיימת בו תופעת נטישות. המוקד השני מכיל נוטשים בלבד, ולכן הוא מערכת מטיפוס $M/M/\infty$ עם קצב "שירות" θ . היתרון בהצגה דו-מימדית זו הוא שמצב המערכת מכיל יותר אינפורמציה - לא רק את מספר הלקוחות במערכת, אלא גם כמה מהם יקבלו/לא יקבלו שירות.

נתמקד בחלק של המערכת בו נמצאים מקבלי השירות. המצב בו מבצעים "הגרלה" לפני שהלקוח נכנס למערכת מתאים למודל בו הלקוח מקבל אינפורמציה על מצב המערכת בעת הגעתו, והוא צריך להחליט האם הוא מצטרף בכל זאת לתור. תחת מודל זה אנו מניחים שלקוחות אשר החליטו להצטרף כבר לא נוטשים. המערכת שמתקבלת היא $M_{Q^1}/M/N$, כאשר Q^1 מסמן שתהליך המופע (כאן הכוונה לכניסה ממש לחלק זה של המערכת) תלוי במצב המערכת (state dependent). שים לב שקיימים גדלים בסיסיים אשר לא משתנים במערכת זו לעומת ההצגה המקורית, למשל: תוחלת זמן ההמתנה של לקוחות שמקבלים שירות, ופרופורציית הנוטשים.

כפי שמופיע ב-[8] Boxma and de Waal, התהליך הסטוכסטי של מספר הלקוחות מקבלי השירות, $\{Q^1(t), t \geq 0\}$, הוא תהליך לידה ומוות עם קצבים:

$$(1.5) \quad \lambda_k = \begin{cases} \lambda, & k < N \\ \lambda(1 - P\{Ab|Q^1 = k \text{ on arrival}\}), & k \geq N \end{cases} ; \quad \mu_k = (N \wedge k)\mu,$$

כאשר את הגודל $1 - P\{Ab|Q^1 = k \text{ on arrival}\}$ ניתן לחשב על פי השיקול הבא: ראשית, לקוח אשר נמצא בראש התור לא נוטש בהסתברות $\frac{N\mu}{\theta + N\mu}$ (על פי תחרות בין $N + 1$ מ"מ מעריכיים ב"ת: N זמני שירות + סבלנות הממתין). כלומר הוא צריך להמתין עד שאחד המוקדנים מסיים טיפול בלקוח. כעת אם הוא נמצא במקום ה- n בתור הוא צריך להמתין שיסתיים הטיפול ב- n לקוחות. בגלל תכונת חוסר הזכרון של מ"מ מעריכי, ההסתברות שהמתין לא נוטש במשך

השירות של הלקוח ה- i שלפניו, $i = 1, 2, \dots, n$, היא גם כן $\frac{N\mu}{\theta + N\mu}$. מכאן, ועל פי האי תלות שקיימת מקבלים

$$(1.6) \quad 1 - P\{Ab|Q^1 = k \text{ on arrival}\} = \left(\frac{N\mu}{\theta + N\mu}\right)^{k-N+1}.$$

ההתפלגות הסטציונרית π^1 נתונה על ידי:

$$(1.7) \quad \pi_k^1 = \begin{cases} \frac{(\lambda/\mu)^k}{k!} \pi_0^1, & 0 \leq k \leq N \\ \left(\frac{\lambda}{N\mu}\right)^{k-N} \left(1 + \frac{\theta}{N\mu}\right)^{-(k-N)(k-N+1)/2} \cdot \frac{(\lambda/\mu)^N}{N!} \pi_0^1, & N < k \end{cases}$$

כאשר

$$(1.8) \quad \pi_0^1 = \left[\sum_{k=0}^N \frac{(\lambda/\mu)^k}{k!} + \sum_{k=N+1}^{\infty} \left(\frac{\lambda}{N\mu}\right)^{k-N} \left(1 + \frac{\theta}{N\mu}\right)^{-(k-N)(k-N+1)/2} \cdot \frac{(\lambda/\mu)^N}{N!} \right]^{-1}.$$

הצגת המודל בצורה זו היא בעלת יתרון חישובי שמתואר בחלק הבא (§1.3).

1.3 מדדים בסיסיים - ההסתברות להמתנה וההסתברות לנטישה

שני מדדים בסיסיים במודל $M/M/N/B + M$ הם ההסתברות להמתנה וההסתברות לנטישה. להלן דרכים אלמנטריות לחישוב גדלים אלה:

1. ההסתברות שלקוח יצטרך להמתין בתור לפני תחילת שירותו ($P\{Wait > 0\}$).

המופעים למערכת הם פואסוניים, לכן על פי תכונת PASTA גודל זה שווה לפרופורציית הזמן שכל המוקדנים עסוקים, פרט למקרה בו המערכת מלאה ולקוח שמופיע עוזב מבלי שנכנס אליה. כלומר

$$(1.9) \quad P\{Wait > 0\} = \sum_{n=N}^{B-1} \pi_n.$$

2. ההסתברות שלקוח נכנס למערכת ונוטש לפני תחילת שירותו ($P\{Ab\}$).

החישוב מתבסס על ההצגה

$$(1.10) \quad P\{Ab\} = \sum_{n=0}^{B-1} P\{Ab|n \text{ in system on arrival}\} \pi_n.$$

נסמן $P\{Ab|n\} = P\{Ab|n \text{ in system on arrival}\}$, ונחשב את $P\{Ab|n\}$.
 על פי הנחות המודל, אם לקוח מופיע כאשר יש מוקדנים פנויים הוא משורת מיד ולכן
 ההסתברות שהוא נוטש היא 0. כמו כן, אם לקוח מופיע והמערכת מלאה - הוא עוזב, אך
 זו אינה נחשבת נטישה. עבור המקרים בהם $B > n > N$ קיים הקשר הרקורסיבי

$$(1.11) \quad 1 - P\{Ab|n\} = \frac{(n - N)\theta + N\mu}{(n - N + 1)\theta + N\mu} (1 - P\{Ab|n - 1\}) ,$$

אשר מתקבל מהשיקול שהלקוח לא נוטש רק אם הוא לא נוטש לפני שהוא מתקדם בתור,
 ומשיקולים של חוסר זכרון ותחרות בין מ"מ מעריכיים ב"ת, בדומה לאלה שמובילים ל-
 (1.6). מכאן על ידי הצבה מקבלים

$$(1.12) \quad 1 - P\{Ab|n\} = (1 - P\{Ab|N\}) \prod_{k=N+1}^n \frac{(k - N)\theta + N\mu}{(k - N + 1)\theta + N\mu} , \quad B > n > N .$$

שים לב שהמכפלה כאן היא טלסקופית. לאחר הצבת הקשר

$$(1.13) \quad 1 - P\{Ab|N\} = \frac{N\mu}{\theta + N\mu} ,$$

שמתקבל מהתחרות בין N זמני השירות וסבלנותו של הלקוח, נקבל

$$(1.14) \quad P\{Ab|n\} = \begin{cases} \frac{(n - N + 1)\theta}{N\mu + (n - N + 1)\theta} , & B > n \geq N \\ 0 , & \text{otherwise} \end{cases}$$

נחזור לחישוב של $P\{Ab\}$, שמתקבל כעת על פי (1.10)

$$(1.15) \quad P\{Ab\} = \sum_{n=N}^{B-1} \frac{(n - N + 1)\theta}{N\mu + (n - N + 1)\theta} \pi_n .$$

דרך אלטרנטיבית, עבור המקרה בו $B = \infty$, היא שימוש במודל בו הנוטשים נדחים
 מראש מהמערכת (ראה §1.2). מכיוון שבמודל המקורי הנוטשים לא משפיעים על זמני
 ההמתנה של לקוחות אחרים, ולכן גם לא על האפשרות שהם נוטשים, הרי שדחיית
 הנוטשים מראש אינה משנה את ההסתברות שלקוח נוטש. מכאן שאפשר לחשב את
 ההסתברות לנטישה פשוט על ידי חישוב גודל זה במודל האלטרנטיבי. כאן ההסתברות

לנטישה היא למעשה ההסתברות שלקוח נדחה מהמערכת מראש.

התוצאה המופיעה אצל [8] Boxma and de Waal (עמוד 5 נוסחה (2.16):

$$\begin{aligned}
 P\{Ab\} &= \sum_{n=N}^{\infty} \pi_n^1 P\{Ab|Q^1 = n \text{ on arrival}\} \\
 &= 1 - \sum_{n=0}^{N-1} \pi_n^1 - \sum_{n=N}^{\infty} \left(\frac{N\mu}{N\mu + \theta} \right)^{n-N+1} \pi_n^1 \\
 (1.16) \quad &= 1 - \pi_0^1 \left[\sum_{n=0}^{N-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^N}{N!} \sum_{n=0}^{\infty} \left(\frac{\lambda}{N\mu} \right)^n \left(\frac{N\mu}{N\mu + \theta} \right)^{m(n)} \right], \\
 &\text{כאשר } m(n) = (n+1)(n+2)/2.
 \end{aligned}$$

הערה: על פי [8] Boxma and de Waal, הטור האינסופי בביטוי זה מתכנס מהר יותר מאשר ב-(1.15) (עם $B = \infty$). עובדה זו מתקבלת מהשוואת ההתנהגות של היחס בין האיבר ה- $n+1$ והאיבר ה- n . כאן מקבלים

$$(1.17) \quad \frac{a_{n+1}}{a_n} \sim \alpha^n, \quad n \uparrow \infty, \quad (0 < \alpha < 1)$$

ולעומת זאת ב-(1.15)

$$(1.18) \quad \frac{b_{n+1}}{b_n} \sim \frac{1}{n}, \quad n \uparrow \infty.$$

1.4 ההסתברות לנטישה - קשרים שימושיים

בחלק זה נציג מספר קשרים בין ההסתברות לנטישה למדדים אחרים. בהמשך נשתמש בקשרים אלה כאשר נבצע ניתוח אסימפטוטי עבור $N \uparrow \infty$.

נגדיר את העומס המוצע למוקדן על ידי $\rho = \frac{\lambda}{N\mu}$.

1. קשר להסתברות להמתנה (ראה [2], נוסחה (5.9) והוכחה שונה):

$$(1.19) \quad P\{Ab\} = (1 - 1/\rho)P\{Wait > 0\} + \pi_N/\rho$$

הוכחה:

נשים לב שעבור $B > n \geq N$ מתקיים

$$(1.20) \quad \pi_{n+1} = \frac{\lambda}{N\mu + (n - N + 1)\theta} \pi_n.$$

כעת יחד עם (1.9) ו-(1.15),

$$\begin{aligned}
 P\{Ab\} &= \sum_{n=N}^{B-1} \frac{(n-N+1)\theta}{N\mu + (n-N+1)\theta} \pi_n \\
 &= P\{Wait > 0\} - \sum_{n=N}^{B-1} \frac{N\mu}{N\mu + (n-N+1)\theta} \pi_n \\
 (1.21) \quad &= P\{Wait > 0\} - (1/\rho) \sum_{n=N}^{B-1} \pi_{n+1} = (1 - 1/\rho)P\{Wait > 0\} + \pi_N/\rho .
 \end{aligned}$$

□

2. קשר להסתברות לחסימה ($P\{Block\}$) במערכת $M/M/N/N$ (Loss System), עם אותם קצבי מופע ושירות:

$$(1.22) \quad P\{Ab\} = \frac{(\rho(1 - P\{Block\}) - 1)P\{Wait > 0\} + P\{Block\}}{\rho(1 - P\{Block\})}$$

הוכחה:

הקשר להסתברות לחסימה נובע מכך שכל עוד במערכת בה אנו דנים אין תור, היא מתנהגת כמו מערכת $M/M/N/N$, כלומר

$$(1.23) \quad \pi_N = P\{Block\} \cdot \sum_{n=0}^N \pi_n = P\{Block\}(1 - P\{Wait > 0\} + \pi_N) .$$

מכאן נקבל ביטוי עבור π_N :

$$(1.24) \quad \pi_N = \frac{(1 - P\{Wait > 0\})P\{Block\}}{1 - P\{Block\}} .$$

נציב ב-(1.19) ונקבל

$$(1.25) \quad P\{Ab\} = (1 - 1/\rho)P\{Wait > 0\} + \frac{(1 - P\{Wait > 0\})P\{Block\}}{\rho(1 - P\{Block\})} .$$

מכאן מתקבל המ.ש.ל. □

3. קשר לתוחלת זמן ההמתנה בתור ($E[Wait]$):

$$(1.26) \quad P\{Ab\} = \theta \cdot E[Wait]$$

הוכחה:

נחשב את תוחלת מספר הלקוחות בתור במצב יציב:

$$(1.27) \quad E[No. in queue] = \sum_{n=N}^{B-1} (n - N + 1)\pi_{n+1} .$$

מכאן על פי (1.15)

$$P\{Ab\} = \sum_{n=N}^{B-1} \frac{(n-N+1)\theta}{N\mu + (n-N+1)\theta} \pi_n = \frac{\theta}{\lambda} \sum_{n=N}^{B-1} (n-N+1) \pi_{n+1} = \frac{\theta}{\lambda} E[No. in queue]. \quad (1.28)$$

את המושל נקבל כאשר נציב את נוסחת Little ($E[No. in queue] = \lambda \cdot E[Wait]$).
הערות:

(א) מתקבל שהקשר בין אחוז הנוטשים לתוחלת זמן ההמתנה בתור הוא לינארי, ונחשפת דרך פשוטה יחסית לאמידת קצב הנטישה (θ): לאסוף במוקד זוגות נתונים של אחוז הנוטשים וממוצע זמן ההמתנה בתור, ולמצא את ישר הרגרסיה שעובר דרך נתונים אלה. השיפוע של ישר זה הוא אומדן לקצב הנטישה.

(ב) הקשר המתקבל ב-(1.28) מבטא למעשה שיווי משקל בין קצב הגעת "נוטשים" ($\lambda P\{Ab\}$) לקצב האפקטיבי של עזיבתם ($\theta E[No. in queue]$). הביטוי עבור קצב העזיבה מתבסס על כך שסבלנות הלקוחות מתוארת על ידי משתנים מקריים מעריכיים בלתי תלויים. אין פה שימוש בתכונות שהנחנו על תהליכי המופע והשירות (מלבד אי תלות), ולכן קשר זה יהיה נכון גם עבור מערכות $G/G/N+M$ במצב שיווי משקל.

(ג) על ידי חלוקת (1.26) ב- $P\{Wait > 0\}$ מתקבל קשר לא פחות מעניין:

$$P\{Ab|Wait > 0\} = \theta \cdot E[Wait|Wait > 0]. \quad (1.29)$$

4. קשר נוסף להסתברות להמתנה, עבור המקרה בו $B = \infty$ (מקום לא מוגבל במערכת):

$$P\{Ab\} = \left(1 - 1/\rho + \frac{(\lambda/\theta)^{N\mu/\theta-1} e^{-\lambda/\theta}}{\gamma(N\mu/\theta, \lambda/\theta)}\right) P\{Wait > 0\}. \quad (1.30)$$

(הגדרתה של הפונקציה $\gamma(x, y)$ מופיעה יחד עם (6.15))

הוכחה:

ברור שמתקיים

$$P\{Ab\} = P\{Ab|Wait > 0\} P\{Wait > 0\}, \quad (1.31)$$

לכן על פי (1.29)

$$(1.32) \quad P\{Ab\} = \theta \cdot E[Wait|Wait > 0]P\{Wait > 0\} .$$

מכאן לאחר הצבת תוצאה של Riordan [32] עבור $E[Wait|Wait > 0]$ (עמוד 111 נוסחה (84)) נקבל את המ.ש.ל. $\square$

פרק 2

סקר ספרות

בפרק זה נסקור באופן תמציתי את הספרות הרלוונטית לתחום המחקר בנושא מודלים של מערכות שירות עם תור יחיד, מספר שרתים זהים ולקוחות קצרי רוח. בנוסף למאמרים הרלוונטיים באופן ישיר (סעיף ראשון), נציין גם נסיונות שנעשו כדי להתמודד עם מודלים כלליים יותר (סעיף שני). אלה מודלים שאינם נכללים במסגרת של $G/G/N/B + G$, וקיימים בהם אלמנטים נוספים כמו חיוגים חוזרים, תהליכים לא הומוגניים בזמן ומשטרי שירות שונים. כמו כן, מכיוון שחלק נכבד מעבודה זו עוסק בניתוח התנהגות אסימפטוטית, כלולים גם מאמרים שונים בתחום זה (סעיף שלישי), חלקם כלליים וחלקם מטפלים במודלים ללא נטישות (ניתוחים אסימפטוטיים של מודלים עם נטישות מופיעים תחת הסעיף הראשון). בכל סעיף סדר ההצגה הוא כרונולוגי.

1. מודלים בהם ללקוחות יש סף סבלנות:

- על פי [32] Riordan, ההתייחסות הראשונה לתופעת הנטישות בספרות על תורים היא של [29] Palm, בהקשר של מטלפנים הממתנים לשירותי מרכזיה ידנית. Palm מניח שהתפלגות הזמן עד לנטישה היא מעריכית.
- בפרסום מאוחר יותר [31] Palm מבטא את מידת ה"עצבנות" (irritation) של אדם כפונקציה של זמן ההמתנה, ומצליח לקבל שהיא פרופורציונאלית לפונקציית הסיכון של התפלגות זמן הנטישה. הדיון מתחיל בהנחה שניתן לבטא את עצבנותו של אדם הממתין לשירות כפונקציה של זמן ההמתנה. המתנה מתמשכת תגרור במוקדם או במאוחר נטישה (פעולה שמטרתה להפיג את העצבנות). לכן סביר

להניח שהעצבנות היא פונקציה עולה בזמן ההמתנה, Palm-1 מציג אותה באופן כללי למדי כפונקציה מהצורה $c \cdot t^\lambda$, כאשר t מסמן את זמן ההמתנה ו- $0 < c, \lambda$ (עמוד 4, נוסחה (1)). בהמשך יש הנחה שמידת העצבנות שמצטברת באנטרוול זמן dt היא פרופורציונאלית להסתברות לנטישה בפרק זמן זה. מכיוון שההסתברות לנטישה בפרק זמן dt שווה לפונקציית הסיכון של התפלגות זמן הנטישה בזמן t כפול dt , הרי שפונקציית הסיכון והעצבנות בזמן t פרופורציונאלים. מכאן, על פי ההנחות לגבי העצבנות, Palm מקבל צורה כללית לפונקציית ההתפלגות של זמן הנטישה (עמוד 7, נוסחה (5)):

$$(2.1) \quad 1 - \exp\left(-\frac{c_0 c}{1 + \lambda} t^{1+\lambda}\right).$$

בהמשך יש ניתוחים ברוח דומה של מודל בו אדם שאינו מתקבל מיד, ממתין שיחזור אליו מהמרכזיה כאשר מתפנה קו, ומודל בו "לקוחות" צריכים לחייג בעצמם שוב ושוב עד שיתפנה קו. Palm משווה בין המודלים ומנסה לבחון את התוצאות האנליטיות על פי נתונים אמפיריים.

• אחד הניתוחים המוקדמים של מודל עם נטישות הוא של [3] Barrer אשר קובע את ההסתברות לנטישה ($P\{Ab\}$) במערכת $M/M/N + D$. הוא מניח שהסבלנות של לקוח היא קבוע τ_0 , ומנתח שני מקרים. במקרה הראשון לקוח נוטש אם הוא נמצא במערכת זמן τ_0 ושירותו טרם התחיל. עבור מקרה זה Barrer מקבל ביטוי מפורש להסתברות לנטישה (עמוד 654 נוסחאות (21),(22)):

$$(2.2) \quad P\{Ab\} = \begin{cases} \frac{(\rho-N)(\rho)^N e^{\alpha(\rho-N)}}{(\rho-N)N! \sum_{k=0}^N \frac{\rho^k}{k!} + \rho^{N+1}(e^{\alpha(\rho-N)} - 1)} & \rho \neq N \\ \left[(N!/N)^N \sum_{k=0}^N \frac{N^k}{k!} + \alpha N \right]^{-1} & \rho = N \end{cases}$$

כאשר $\rho = \lambda/\mu$, $\alpha = \mu\tau_0$.

במקרה השני לקוחות עלולים לנטוש גם לאחר ששירותם כבר התחיל. כאן הוא מצטמצם למערכת עם שרת יחיד (כלומר מערכת $M/M/1 + D$), ומקבל (עמוד 656 נוסחאות (29),(30)):

$$(2.3) \quad P\{Ab\} = \begin{cases} (1-\rho)e^{\alpha(\rho-1)}/[1-\rho e^{\alpha(\rho-1)}] & \rho \neq 1 \\ 1/(\alpha+1) & \rho = 1 \end{cases}$$

- בספרו של [32] Riordan קיים ניתוח של מערכת $M/M/N+M$. על ידי התניה באורך התור בעת הגעת לקוח הוא מקבל משוואה דיפרנציאלית רקורסיבית, שפתרונה נותן את התפלגות זמן ההמתנה של לקוח עם סבלנות בלתי מוגבלת, בהנתן שפגש n לקוחות בתור בעת הגעתו (F_{V_n}) :

$$(2.4) \quad F_{V_n}(t) = 1 - \frac{c(c+1) \cdots (c+n)}{n!} \sum_{k=0}^n (-1)^k \binom{n}{k} \frac{e^{-(c+k)\theta t}}{c+k}$$

כאן $c = N\mu/\theta$.

- טיפול במקרה הכללי של מערכת $G/G/1+G$ מופיע במאמר של [13] Daley. הוא מקבל משוואה אינטגרלית עבור ההתפלגות הסטציונרית של זמן ההמתנה וקובע תנאים לקיומה. בהמשך הוא מציג תוצאות גם עבור מערכות $G/G/1+D$, $M/G/1+M$, $M/G/1+D$.

- [17] Gnedenko and Kovalenko חוקרים את המערכות $M/M/N+D$ ו- $M/M/N+M$. עבור המקרה בו הסבלנות של לקוח היא קבוע, הם מפרידים לשני מקרים כמו בעבודתו של [3] Barrer. Gnedenko and Kovalenko מקבלים את אותם ביטויים מפורשים שקיבל Barrer (דהיינו (2.2) ו-(2.3)), אך הם טוענים שהניתוח ב-[3] אינו מדויק מבחינה מתמטית.

הערה: שים לב שיש טעות דפוס ב-[17] בביטוי המקביל ל-(2.2) (עמוד 38 נוסחה (13)) - בחזקה של האקספוננט נכתב $n\nu$ במקום $\tau\nu$. טעות זו תוקנה במהדורה השניה של הספר, משנת 1989.

- Yurkevich מנתח מערכת $M/M/N+G$. ב-[40] הוא מניח סף סבלנות שהוא המינימום בין משתנה מקרי מעריכי עם פרמטר a , וקבוע τ . כאשר $a \uparrow \infty$ מתקבלות התוצאות מ-[3] ו-[17], ועבור $\tau \uparrow \infty$ זה תואם מערכת $M/M/N+M$. ב-[41] הוא מרחיב את התוצאות לנטישות כלליות. בשני המקרים התוצאות מתקבלות על ידי פתרון משוואות דיפרנציאליות עבור הסתברויות המעבר.

- [20] Haugen and Skogan מטפלים במערכת $M/M/N+G$. הם מנתחים מערכת בה ההתפלגות הכללית של סף הסבלנות מקורבת על ידי התפלגות בדידה. הגישה היא שיש מספר סוגי לקוחות, כל אחד בעל סף סבלנות קבוע, ותהליך המופע של כל

סוג כנ"ל הוא תהליך פואסון (ב"ת בתהליכים האחרים). כאשר משאיפים את מספר ה"סוגים" לאינסוף, מקבלים תוצאות שמתאימות למודל עם סף סבלנות כללי.

- במאמרו של [35] Stanford יש ניתוח נוסף של מערכת $G/G/1+G$. הוא מסיק התפלגויות הקשורות לזמן ההמתנה ונוסחאות דומות לנוסחת Little בהן מופיעה ההסתברות לנטישה. כפי שהוא טוען במכתבו הקצר ב-[36], חלק מהתוצאות שקיבל מופיעות במאמר המאוחר יותר של [1] Baccelli, Boyer and Hebuterne, אם כי צורת הניתוח שונה.

- [2] Baccelli and Hebuterne דנים תחילה במערכת $G/G/1+G$. עבור סף סבלנות מהתפלגות F מתקבל תנאי הכרחי לקיום מצב יציב:

$$(2.5) \quad \lambda(1 - F(\infty)) \leq \mu .$$

להסבר על משמעות התנאי ראה (1.1) בעבודה זו. בהמשך הם מנתחים מערכות $M/G/1+G$ ו- $M/M/N+G$, ומוצגת מערכת משוואות למציאת התפלגות זמן ההמתנה וההסתברות ל- $0, \dots, m-1$ לקוחות במערכת. Baccelli and Hebuterne מטפלים בזמן ההמתנה של "test customer" - זהו לקוח בעל סבלנות בלתי מוגבלת (כלומר הוא לא נוטש). עבור מערכת $M/M/N+G$ הצפיפות $(v(x))$ של משך ההמתנה של "test customer" נתונה על ידי:

$$(2.6) \quad v(x) = \lambda \pi_{N-1} \exp \left[\lambda \int_0^x (1 - F(u)) du - N\mu x \right] , \quad x > 0$$

עם תנאי נירמול

$$(2.7) \quad \sum_{j=0}^{N-1} \pi_j + \int_0^\infty v(x) dx = 1 ,$$

כאשר נתון הקשר

$$(2.8) \quad \pi_j = \frac{(\lambda/\mu)^j}{j!} \pi_0 , \quad j = 0, \dots, N-1 .$$

עבור מערכת $M/M/N+D$ ו- $M/M/N+M$, ניתן לקבל ביטוי ל- $P\{Ab\}$ בצורה של טור אינסופי (נוסחאות (5.11) ו-(5.12)). המאמר של Baccelli, Boyer and Hebuterne [1] חופף במידה רבה למאמר זה, אך יש בו תוצאות נוספות עבור מערכת $G/G/1+G$, ואין בו התייחסות למודל עם מספר שרתים.

- במאמר של [8] Boxma and de Waal ישנה קודם כל סקירה של הספרות בנושא נטישות, ומובאות כמה מהתוצאות בנושא זה. בין השאר מוצג שם מודל אלטרנטיבי למערכת $M/M/N + M$, בו הלקוחות הנוטשים נדחים מראש מהמערכת (ראה פרוט בחלק §1.2 בעבודה זו). בהמשך הם עוסקים במציאת חסמים וקירובים להסתברות לנטישה, ובבדיקה נומרית של התוצאות. מסתבר שבמקרים רבים החסמים מהווים גם קירובים טובים. כמו כן, הם טוענים שניתן לעיתים לקרב בצורה טובה את ההסתברות לנטישה במודלים כלליים באמצעות אינטרפולציה של גודל זה במודלים פשוטים יותר (למשל, עבור $M/G/N + M$ כאשר מקדם ההשתנות של זמן השירות קטן מ-1, ניתן לקבל קירובים טובים מאינטרפולציה של $M/M/N + M$ ו- $M/M/N + D$).

- [16] Fleming, Stolyar and Simon מנתחים תחילה מערכת $M/M/N + M$. בדומה לנעשה בעבודה זו הם דנים בהתכנסות חלשה של התהליך הסטוכסטי של מספר הלקוחות במערכת, ושל ההתפלגות הסטציונרית של תהליך זה. ההנחה היא שקצב הנטישה אינו משתנה עם מספר השרתים. התהליך הגבולי שמתקבל הוא הרכבה של שני תהליכי Ornstein-Uhlenbeck. בהמשך יש הרחבה של תוצאות אלה למודל בו יש שני סוגי לקוחות בעלי סף סבלנות שונה. מאמר זה משחק תפקיד חשוב בעבודה זו, ופרוט התוצאות מופיע בהוכחות של למות 6.1.1, 6.1.2 ומשפט 6.2.

ישנן עבודות אשר עוסקות באפיון התנהגות מונוטונית במודל עם נטישות:

- [4] Bhattacharya and Ephremides מראים שמספר הלקוחות שנוטשים, בכל פרק זמן, הוא פונקציה מונוטונית (סטוכסטית) של קצבי ההגעה, השירות, והנטישה.
- שני מאמרים של [33], [34] Sadrolhefazi עוסקים בחקר תכונות מונוטוניות במערכת $G/M/N/B + M$. בפרט מתקבל שקצב המעבר דרך המערכת, קצב הנטישה האפקטיבי ופרופורציית הנוטשים בטווח הארוך, הם פונקציות מונוטוניות של קצב המופע, קצב השירות, מספר שרתים, מקום בתור וסף הסבלנות של הלקוחות. בהמשך, עבור מערכת בה המופעים הם פואסוניים, מקבלים מונוטוניות ממש של הפונקציות הנ"ל. במקרה זה, תחת הנחות מסוימות, ישנן תוצאות גם עבור מודל עם

חיוגים חוזרים.

2. מודלים כלליים:

- נסיון לטפל במודל כללי מופיע ב-[21] Harris, Hoffman and Saunders ובהמשך ב-[22] Hoffman and Harris. במודל שלהם תהליך המופע (ללא החיוגים החוזרים) הוא פואסון הומוגני, זמן השירות מעריכי, יש נטישות וחיוגים חוזרים, ויכולים להיות משטרי שירות שונים. תחת הנחה של מערכת העובדת בעומס גבוה הם מגיעים למערכת משוואות שווי משקל, שפתרון דורש הצבת חמישה גדלים הקשורים למערכת. מתוך הפתרון ניתן לחשב מדדי שירות שונים.

- גם Sze [38] מטפל במודל כללי. הבסיס של המודל שלו הוא $M/PH/N + PH$ (מסמן התפלגות Phase Type), בתוספת חיוגים חוזרים, משטר שירות עם עדיפויות ותהליך מופע לא סטציונרי. המודל אינו ניתן לניתוח אנליטי ולכן רמת האיוש נקבעת על פי קירובים. חלק מהעבודה מתמקד בקבלת תוצאות משיטות קירוב שונות, והשוואתן לתוצאות של סימולציה.

3. ניתוחים אסימפטוטיים:

- במאמר של Stone [37] מופיע קריטריון פשוט להתכנסות חלשה של תהליך לידה ומוות לתהליך דיפוזיה רציף שמקבל ערכים על כל הישר הממשי. בחלק §6.3 בעבודה זו אנו עושים שימוש בקריטריון לשם הוכחת התכנסות חלשה.

- במאמר של Iglehart [23] הוא מטפל במערכת $M/M/N$ ובתהליך לידה ומוות המתאים ל-Repairman Problem. בשני המקרים מוכח שכאשר $N \uparrow \infty$, ותחת התנהגות מתאימה של הפרמטרים, מתקבלת התכנסות חלשה לתהליך Ornstein-Uhlenbeck. הכלי להוכחת ההתכנסות הוא הקריטריון של Stone [37].

- ניתוח יסודי ומקיף של ההסתברות לחסימה במערכת $M/M/N/N$ (Erlang Loss Function) מופיע במאמר של Jagerman [24]. בין השאר מוצעות דרכים שונות לחישוב (חלקן בעזרת פונקציות מרוכבות) וניתוחים אסימפטוטיים (גם של נגזרות).

- שני מאמרים אשר היוו מוטיבציה לעבודה זו הם [19] Halfin and Whitt ו-[39] Whitt. בראשון יש ניתוח של סדרת מערכות $M/M/N$ בעומס גבוה, כאשר

$N \uparrow \infty$. מתקבל שתנאי הכרחי ומספיק לכך שהגבול של ההסתברות להמתנה (α) אינו מנוון (כלומר $0 < \alpha < 1$) הוא שהגודל $\beta \equiv \lim_{N \rightarrow \infty} \sqrt{N}(1 - \frac{\lambda_N}{N\mu})$ יקיים $0 < \beta < \infty$ במקרה זה מתקיים:

$$(2.9) \quad \alpha = \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1}.$$

כמו כן אם $\beta = 0$ אז $\alpha = 1$, ואם $\beta = \infty$ אז $\alpha = 0$. בהמשך מרחיבים את התוצאות הנ"ל גם עבור מערכת $G/M/N$. מציאת גבול דיפוזיה ל- $G/G/N$ קשה יותר מכיוון שהתהליך הגבולי שמתקבל אינו מרקובי, אך נטען שזה אפשרי עבור $G/PH/N$, ומובאת דוגמא למקרה $G/H_2/N$ מסמן התפלגות היפראקספוננציאלית עם שני רכיבים). הגבול שמתקבל בדוגמא זו הוא תהליך דיפוזיה תלת-מימדי. קיימת הכללה ל- $G/PH/N$ בעבודה של M. Rieman and T. Puhalskii, שטרם פורסמה. בין השאר הם טוענים ש-Halfin and Whitt טועים ביחס למקרה של $G/H_2/N$, והתהליך הגבולי שם הוא דו-מימדי.

במאמר השני [39] יש שימוש בתוצאות המאמר הראשון לקבלת מדד מקורב המוגדר על ידי

$$(2.10) \quad \gamma = \sqrt{N}(1 - \rho).$$

מדד זה מהווה קנה מידה לרמת השירות במערכת, והוא נקשר באופן ישיר, על פי תוצאות המאמר הראשון [19], להסתברות שלקוח נאלץ להמתין בתור לפני תחילת שירותו. מכאן נגזר כלל אצבע פשוט הקובע את היחס בין מספר השרתים (N) לעומס (ρ), ומידת השפעתם על רמת השירות.

• [26] Mandelbaum and Pats מציגים משפטי קיום ויחידות של גבול נוזלים וגבול דיפוזיה המתאימים למשפחה רחבה של מערכות תורים שהם state dependent, וניתנים הכלים למציאת הגבולות הנ"ל. במחקר הנוכחי נעשה שימוש בתוצאות אלה למציאת גבול דיפוזיה במקרה בו התהליך הגבולי מצטמצם לחלק מהישר הממשי (ראה חלק §6.3).

• דרך לקבלת ההתפלגות הסטציונרית של תהליך דיפוזיה מופיעה במאמר של [10] Browne and Whitt. כאשר הממוצע האינפיניטסימלי הוא לינארי למקוטעין

והשונות האינפנייטסימלית קבועה למקוטעין, מתקבלת התפלגות סטציונרית בעלת צפיפות נורמלית ו/או מעריכית למקוטעין (פרוט התוצאות ודוגמא ליישומן מופיעים בחלק §6.4 כאן בהמשך). בהמשך המאמר גם דנים בטיבו של תהליך דיפוזיה כקירוב לתהליך לידה ומוות.

פרק 3

חישוב מדויק של מדדים במערכת $M/M/N/B+M$ (קצב נטישה θ)

כפי שצוין במבוא, בעת עיצוב מוקד שירות טלפוני קיימות מטרות מסוימות אשר ניתן להגדירן באופן מדויק בעזרת מדדים שונים. למשל בעזרת הממד "ההסתברות, במצב יציב, שלקוח יצטרך להמתין בתור לשירות" ניתן להגדיר מטרה שהיא "במצב יציב, לא יותר מ-10% מהלקוחות ימתינו בתור לשירות". אנו רוצים למצא את הקשר בין מדדים שונים לפרמטרים של המערכת, כך שניתן יהיה לקבוע פרמטרים אלה על פי המטרות שהוגדרו (או לקבוע שלא ניתן להשיג את כל המטרות יחד).

יש לשים לב שהמדדים בהם נדון מניחים שהמערכת נמצאת במצב יציב. לכאורה יש כאן בעיה - ברוב המוקדים של שירות טלפוני קצב המופע משתנה משמעותית עם הזמן (על פי שעות היום, היום בשבוע, חגים, עונות וכו'). הפתרון מבוסס על ההנחות שבפרקי זמן קצרים יחסית (שעה למשל), קצב המופע קבוע בקירוב, והמערכת מגיעה למצב יציב במהירות. לכן בפועל, קובעים את החלוקה של ציר הזמן לאינטרוולים, ובכל אינטרוול משתמשים בתוצאות הניתוח של מערכת במצב יציב, עם קצב מופע קבוע השווה לקצב המופע הממוצע באינטרוול.

בפרק זה, אם כן, נעסוק בשיטות לחישוב מדדים שונים הרלוונטיים למערכות $M/M/N/B+M$ במצב יציב. נציג סכימה כללית לחישוב מדדים. לשם כך נשתמש בסימונים הבאים:

• V - זמן ההמתנה הוירטואלי במצב יציב. זהו הזמן אותו יבלה בתור לקוח בעל סבלנות אינסופית.

טבלה 3.1: מדדים מהצורה $E[f(V, R)]$
 Performance measures of the form $E[f(V, R)]$

$f(v, r)$	$E[f(V, R)]$
$1_{\{v > r\}}$	$P\{V > R\} = P\{Ab\}$
$1_{(t, \infty)}(v \wedge r)$	$P\{Wait > t\}$
$1_{(t, \infty)}(v \wedge r) 1_{\{v > r\}}$	$P\{Wait > t; Ab\}$
$(v \wedge r) 1_{\{v > r\}}$	$E[Wait; Ab]$
$(v \wedge r) 1_{(t, \infty)}(v \wedge r) 1_{\{v > r\}}$	$E[Wait; Wait > t; Ab]$
$g(v \wedge r)$	$E[g(Wait)]$

• V_n - זמן ההמתנה הוירטואלי בתנאי שבעת הגעת הלקוח למערכת היו בתור n לקוחות אחרים, $n = 0, 1, \dots$. פונקציית ההתפלגות של מ"מ זה תסומן F_{V_n} .

• R - סבלנותו של לקוח. כאמור $R \sim \exp(\theta)$.

• $Wait \equiv V \wedge R$ - הזמן שלקוח מבלה בתור בפועל.

הערה: שים לב שהתפלגות המ"מ V אינה נתונה מראש, והיא מתקבלת כתוצאה מניתוח המודל. לעומת זאת, משיקולים של חוסר זכרון ותחרות בין משתנים מקריים מעריכיים ב"ת (ראה דוגמא ב-§1.2), המ"מ V_n הוא סכום של $n + 1$ משתנים מקריים מעריכיים ב"ת עם פרמטרים $N\mu, N\mu + \theta, \dots, N\mu + n\theta$. כמו כן נדגיש ש- R ו- V ב"ת.

המדדים אותם נחשב הם מהצורה $E[f(V, R)]$, כאשר f פונקציה כלשהי. בטבלה 3.1 מופיעות דוגמאות שונות ל- f , והמדד שמתקבל מתוך מדדים אלה ניתן לחשב את המדדים ה"טבעיים" יותר שהוצגו בסוף המבוא, למשל:

$$(3.1) \quad P\{Ab | Wait > t\} = \frac{P\{Wait > t; Ab\}}{P\{Wait > t\}}.$$

כעת נפתח באופן כללי ביטויים עבור $E[f(V, R)]$. נשתמש בפירוק הבא:

$$\begin{aligned} E[f(V, R)] &= E[f(V, R) \cdot 1_{\{V>0\}}] + E[f(V, R) \cdot 1_{\{V=0\}}] \\ (3.2) \quad &= E[f(V, R) \cdot 1_{\{V>0\}}] + E[f(0, R)] \cdot (1 - P\{Wait > 0\} + \pi_B) . \end{aligned}$$

נמשיך את הפיתוח עבור המחובר הראשון. על פי התניה במספר הלקוחות במערכת בעת הגעת לקוח

$$\begin{aligned} E[f(V, R) \cdot 1_{\{V>0\}}] &= \sum_{n=0}^{B-N-1} \pi_{N+n} E[f(V_n, R)] \\ (3.3) \quad &= - \sum_{n=0}^{B-N-1} \pi_{N+n} \int_0^\infty \int_0^\infty f(t, r) \theta e^{-r\theta} dr d\bar{F}_{V_n}(t) , \\ &\text{כאשר } \bar{F}_{V_n}(t) = 1 - F_{V_n}(t) . \end{aligned}$$

[32] Riordan מקבל את ההתפלגות של V_n ובפרט (עמוד 111 נוסחה (83))

$$(3.4) \quad \bar{F}_{V_n}(t) = \frac{c(c+1) \cdots (c+n)}{n!} \sum_{k=0}^n (-1)^k \binom{n}{k} \frac{e^{-(c+k)\theta t}}{c+k} .$$

כאן $c = N\mu/\theta$

נציב את התוצאה הנ"ל ב-(3.3) ונקבל

$$\begin{aligned} E[f(V, R) \cdot 1_{\{V>0\}}] &= \\ (3.5) \quad &\sum_{n=0}^{B-N-1} \pi_{N+n} \int_0^\infty \int_0^\infty f(t, r) \theta^2 e^{-r\theta} \frac{c(c+1) \cdots (c+n)}{n!} \sum_{k=0}^n (-1)^k \binom{n}{k} e^{-(c+k)\theta t} dr dt . \end{aligned}$$

מכאן נמשיך את הפיתוח בשתי דרכים.

דרך א': נרשום את (3.5) באופן הבא

$$(3.6) \quad E[f(V, R) \cdot 1_{\{V>0\}}] = \sum_{n=0}^{B-N-1} \pi_{N+n} \frac{c(c+1) \cdots (c+n)}{n!} \sum_{k=0}^n (-1)^k \binom{n}{k} I(k) ,$$

כאשר

$$\begin{aligned} I(k) &= \frac{1}{c+k} \int_0^\infty \int_0^\infty f(t, r) (c+k) \theta e^{-(c+k)\theta t} \theta e^{-\theta r} dt dr \\ (3.7) \quad &= \frac{1}{c+k} E[f(X_k, R)] , \end{aligned}$$

ו- X_k הוא משתנה מקרי בלתי תלוי ב- R , מפולג $\exp(N\mu + k\theta)$.

תוך שימוש בקשר

$$(3.8) \quad \pi_{N+n} = \frac{(\lambda/\theta)^n}{(c+1) \cdots (c+n)} \pi_N, \quad n = 1, 2, \dots$$

ולאחר שינוי סדר הסכימה ב-(3.6), נקבל את הביטוי הסופי

$$(3.9) \quad E[f(V, R) \cdot 1_{\{V>0\}}] = c\pi_N \sum_{k=0}^{B-N-1} (-1)^k \frac{(\lambda/\theta)^k}{k!} I(k) \sum_{n=0}^{B-N-1-k} \frac{(\lambda/\theta)^n}{n!}.$$

דרך ב': נתחיל בהצגת (3.5) בצורה שונה במעט:

$$(3.10) \quad E[f(V, R) \cdot 1_{\{V>0\}}] = \sum_{n=0}^{B-N-1} \pi_{N+n} \int_0^\infty \int_0^\infty f(t, r) \theta^2 e^{-(r+ct)\theta} \frac{c(c+1) \cdots (c+n)}{n!} \sum_{k=0}^n \binom{n}{k} (-e^{-\theta t})^k dr dt.$$

נשים לב שכעת הסכום השני הרשום הוא בדיוק בפורמט של הבינום של ניוטון, ויחד עם (3.8)

נקבל

$$(3.11) \quad E[f(V, R) \cdot 1_{\{V>0\}}] = \theta^2 c \pi_N \sum_{n=0}^{B-N-1} \frac{(\lambda/\theta)^n}{n!} J(n),$$

כאשר

$$(3.12) \quad J(n) = \int_0^\infty \int_0^\infty f(t, r) e^{-(r+ct)\theta} (1 - e^{-\theta t})^n dr dt.$$

בשתי הדרכים חישוב מדדים מצריך חישוב של ההתפלגות הסטציונרית π . יתרונה של הדרך הראשונה הוא הקלות היחסית בה ניתן בדרך כלל לחשב את $I(k)$. החסרון הוא שהביטוי שמחשבים הוא בצורה של טור עם סימנים מתחלפים, ולעיתים האיברים הם גדולים מאד בערכם המוחלט, דבר שהופך אותו ללא יציב נומרית. בביטויים שמתקבלים על פי הדרך השנייה נפתרת בעיה זו, אך החישוב של $J(n)$ מסובך יותר - בדרך כלל לא ניתן לקבל עבורו ביטוי מפורש, אך במקרים אלה ניתן לקבל נוסחאות רקורסיביות.

הערה: כאשר $B = \infty$, בחישוב ההתפלגות הסטציונרית ובשתי הסכימות שהצגנו (בפרט (3.7) ו-(3.12)) מתקבלים טורים אינסופיים. טורים אלה מתכנסים, לכן ניתן לקרב אותם על ידי טורים קטומים, אך עדיף להשתמש בתוצאות הבאות:

1. [30] Palm מקבל קשר בין ההתפלגות הסטציונרית של מערכת $M/M/N+M$ להסתברות

לחסימה במערכת $M/M/N/N$:

$$\pi_n = \begin{cases} \frac{P\{Block\}}{1 + (A(\frac{\lambda}{N\mu}, \frac{N\mu}{\theta}) - 1)P\{Block\}} \cdot \frac{N!}{n! \left(\frac{\lambda}{\mu}\right)^{N-n}}, & n < N \\ \frac{P\{Block\}}{1 + (A(\frac{\lambda}{N\mu}, \frac{N\mu}{\theta}) - 1)P\{Block\}} \cdot \frac{\left(\frac{\lambda}{\theta}\right)^{n-N}}{\left(\frac{N\mu}{\theta} + 1\right) \cdots \left(\frac{N\mu}{\theta} + (n - N)\right)}, & n \geq N \end{cases} \quad (3.13)$$

כאשר

$$A(x, y) = \frac{ye^{xy}}{(xy)^y} \cdot \gamma(xy, y). \quad (3.14)$$

2. ביטוי לצפיפות של V מתקבל כמקרה פרטי של התוצאה של [2] Baccelli and Hebuterne

(ראה ניסוח כללי ב-(2.6)) :

$$f_V(t) = \lambda \pi_{N-1} \exp \left\{ \frac{\lambda}{\theta} (1 - e^{-\theta t}) - N\mu t \right\}. \quad (3.15)$$

בעזרת ביטוי זה אפשר לנסות לחשב מדדים באופן ישיר מתוך

$$E[f(V, R) \cdot 1_{\{V>0\}}] = \int_0^\infty \int_0^\infty f(t, r) \theta e^{-r\theta} f_V(t) dr dt. \quad (3.16)$$

ניתן להראות שמדדים המתקבלים מפונקציה $f(t, r)$ שנתונה על ידי סכומים ומכפלות של פונקציות מהצורה

$$1_{\{v>r\}}, 1_{\{v<r\}}, 1_A(v), 1_B(r), 1_{(t,\infty)}(v \wedge r), 1_{[0,t]}(v \vee r), P(r),$$

ניתנים לתאור בעזרת Gamma Functions ו/או Incomplete Gamma Functions. (כאן

A ו- B הם אינטרוולים ב- $\mathbb{R}_+$, ו- $P(r)$ הוא פולינום ב- r). דוגמא למדדים מסוג זה ניתן

לראות ב-(7.14) ו-(7.15).

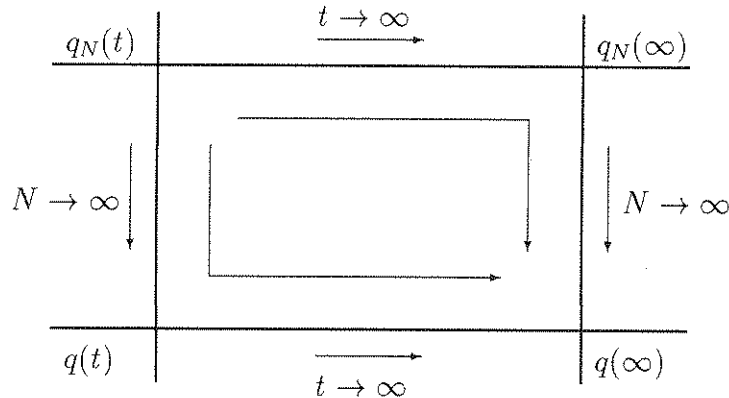
פרק 4

קירובים

החל מפרק זה נבחן את ההתנהגות האסימפטוטית של גדלים שונים כדי לנסות ולקבל קירובים למדדים כדוגמת אלו שהוצגו בפרק הקודם. למרות שקיבלנו סכימות מפורשות על פיהן ניתן לחשב מדדים באופן מדויק, נרצה למצא גם קירובים. המוטיבציה היא מציאת קשרים פשוטים בין המדדים לפרמטרים של המערכת, מהם ניתן לקבל "insight" על התנהגות המערכת. ברור שמהביטויים הכלליים שקיבלנו (3.9) ו-(3.11) בפרק הקודם) לא ניתן להבין את אופי הקשר בין המדדים לפרמטרים.

באופן הדרגתי נסיף מספר הנחות על הפרמטרים של המודל, כשהראשונה מביניהן, שתיושם כבר עכשיו, היא שאין מגבלה על אורך התור ($B = \infty$). כלומר המערכת שמנותחת היא $M/M/N + M$. הנחה זו מקלה על הניתוח, אך מדוע היא סבירה? ההצדקה כאן נעוצה בכך שאנו דנים דווקא במוקדי שירות טלפוניים, בהם מספר הקוים עשוי להיות משמעותית גדול יותר ממספר המוקדנים (אורך התור המקסימלי הוא ההפרש בין גדלים אלה). הציוד הנדרש לטיפול בתור של ממתנינים במוקד שירות טלפוני הוא יחסית זול כיום, לכן שכח לאמץ מדיניות של יצירת מקומות רבים לממתנינים כדי להקטין את הסיכוי שלקוח נתקל ב"תפוס". כלל לא ברור אם מדיניות שכזאת תמיד תואמת את העדפות הלקוחות. דיון בסוגיה זו מופיע בפרק 7. בניתוח האסימפטוטי נטפל בסדרת מערכות $M/M/N+M$, עם פרמטרים התלויים ב- N (מספר המוקדנים). באופן יותר ספציפי, נטפל בסדרת התהליכים הסטוכסטים $\{Q_N, N = 1, 2, \dots\}$, כאשר $\{Q_N(t), t \geq 0\}$ הוא התהליך הסטוכסטי של מספר הלקוחות במערכת עם N מוקדנים (ראה בחלק §1.1 תיאור כתהליך לידה ומוות).

איור 4.1: קירובים לתהליך $\{q_N(t), t \geq 0\}$
 Approximating the process $\{q_N(t), t \geq 0\}$



שים לב שמעטה יופיע האינדקס N על הפרמטרים שמתאימים למערכת ה- N . כמו כן, כדי להקל על הסימונים נניח שהזמן נמדד ביחידות של ממוצע זמן שירות. באופן שקול המשמעות היא חלוקת קצבי המעבר של תהליך הלידה ומוות Q , ב- μ (כלומר אופרטיבית $\mu = 1$). הניתוח נעשה עבור $N \uparrow \infty$, כאשר גם כאן העובדה שמדובר במוקדי שירות טלפוני, בהם יש לעיתים מאות מוקדנים, מצדיקה את הנסיון לקבל קירובים בצורה זו. בהמשך העבודה נראה שגם כאשר מספר המוקדנים הוא בסדר גודל של עשרות הקירובים המתקבלים עדין סבירים.

הניתוח האסימפטוטי יעשה בשתי רמות. הרמה הבסיסית היא לבחון מדדים ספציפיים. ברמה השניה נבחן את ההתנהגות האסימפטוטית של ההתפלגות הסטציונרית של התהליך. כפי שצויין, המדדים שמעניינים אותנו מתייחסים למערכת במצב יציב, לכן ה"מידע" שדרוש לחישובם מצוי בהתפלגות הסטציונרית. ההתנהגות האסימפטוטית במקרה זה היא במובן של התכנסות לגבולות חלשים, ובפועל נטפל בסדרת תהליכים $\{q_N, N = 1, 2, \dots\}$ שמתקבלים מה-rescaling הבא:

$$(4.1) \quad q_N = \frac{Q_N - N}{\sqrt{N}}.$$

כאן ניתן לנקוט בשתי גישות שונות, אותן נתאר באופן סכמטי בעזרת איור 4.1. דרך אחת היא למצא את ההתפלגות הסטציונרית, אם קיימת (צלע עליונה), ואחר כך לבחון את ההתכנסות

של התפלגות זו (צלע ימין). דרך שניה היא לבחון את ההתכנסות של התהליך כולו לתהליך סטוכסטי אחר (צלע שמאל), ולאחר קבלת התהליך הסטוכסטי הגבולי, למצא את ההתפלגות הסטוציונרית שלו (צלע תחתונה). אנו נפעל בשתי הדרכים. אחת המטרות תהיה להראות שההתפלגות $q(\infty)$ שמתקבלת בסוף (פינה ימנית תחתונה) היא אותה עבור שתי הדרכים. כלומר בכך נראה שניתן לשנות את סדר הגבולות, דהיינו:

$$(4.2) \quad \lim_{t \rightarrow \infty} \lim_{N \rightarrow \infty} q_N(t) \stackrel{d}{=} \lim_{N \rightarrow \infty} \lim_{t \rightarrow \infty} q_N(t).$$

כאן " $\stackrel{d}{=}$ " מסמן שיוויון בהתפלגות.

פרוצדורה דומה מופיעה בעבודה של [19] Halfin and Whitt עבור סדרת מערכות $M/M/N$.

כפי שנראה בפרקים הבאים, התוצאות שנקבל תהיינה תלויות בערכי פרמטרים שונים של המודל. ננסה להתמקד בתחומי ערכים אשר מתאימים למודל של מוקד שירות טלפוני, ושעשויים לספק קירובים טובים למציאות.

פרק 5

התנהגות אסימפטוטית של ההסתברות לנטישה

היחוד של המודל הוא קיום הנטישות, לכן הגודל הראשון שנתמקד בניתוחו הוא ההסתברות שלקוח נוטש את המערכת לפני תחילת שירותו. בפרק זה נתחיל לבחון את ההתנהגות האסימפטוטית של ההסתברות לנטישה עבור ערכים שונים של הפרמטרים, וזאת כדי לזהות תחומים בהם ההתנהגות שונה, ולקבוע באיזה תחומים כדאי להתמקד בהמשך. ההפרדה הראשונה לתחומים היא על פי הקשר בין מספר המוקדנים (N) והעומס המוצע למוקדן ($\rho_N = \frac{\lambda_N}{N}$).

כפי שראינו קל להגיע לביטויים אנליטיים עבור ההסתברות לנטישה, אך את הניתוח האסימפטוטי הבסיסי (קרי מציאת $\lim_{N \rightarrow \infty} P_N\{Ab\}$) נוח יותר לבצע על ידי השוואה למודלים אחרים עבורם קיימות תוצאות ידועות. בפרט נתייחס לניתוח האסימפטוטי של מערכת $M/M/N/N$ (Loss System) כפי שמופיע ב-[24] Jagerman.

סימונים:

- $P_N\{Ab\}$ - ההסתברות לנטישה במערכת $M(\lambda_N)/M(1)/N + M(\theta_N)$
- $P_N\{Wait > 0\}$ - ההסתברות שלקוח ימתין במערכת $M(\lambda_N)/M(1)/N + M(\theta_N)$
- $P_N\{Block\}$ - ההסתברות לחסימה במערכת $M(\lambda_N)/M(1)/N/N$ (loss function)
- $\rho_\infty \equiv \lim_{N \rightarrow \infty} \rho_N$ (אנו נניח שהפרמטרים הם כאלו כך שגבול זה תמיד קיים)
- הסימן " $\sim$ " יתייחס להתנהגות אסימפטוטית כאשר $N \uparrow \infty$, אלא אם יצויין אחרת.

הערה: קצבי המופע, השירות והנטישה מצוינים כאן במפורש כדי להדגיש את היותם תלויים/בלתי תלויים במספר המוקדנים.

משפט 5.1 נניח שמתקיים $0 \leq \rho_\infty < \infty$, $\lim_{N \rightarrow \infty} \rho_N = \rho_\infty$, אזי ההתנהגות של ההסתברות לנטישה, כאשר $N \uparrow \infty$, נתונה על ידי:

$$\lim_{N \rightarrow \infty} P_N\{Ab\} = \begin{cases} 0 & \rho_\infty \leq 1 \\ 1 - \frac{1}{\rho_\infty} & \rho_\infty > 1 \end{cases}$$

אינטואיציה לתוצאה קיימת עבור המצב בו $\rho_\infty \geq 1$. במצב זה המערכת היא בעומס גבוה ולכן במצב יציב נקבל שיוויון בין קצב העבודה המקסימלי של המוקדנים (N) לקצב הגעת לקוחות לשירות ($\lambda_N(1 - P_N\{Ab\})$). כאשר $N \uparrow \infty$ מתקבלת התוצאה שבלמה. ניתן לצפות להתנהגות מונוטונית בעומס המוצע למוקדן - במערכת פחות עמוסה נצפה לפחות נוטים - וכך מתקבלת התוצאה עבור $\rho_\infty \leq 1$.

כעת באופן פורמלי:

הוכחה:

בהסתמך על [4] Bhattacharya and Ephremides, הסתברות לנטישה מונוטונית עולה בקצב המופע. עבור N נתון, קצב השירות קבוע, לכן עליה ב- ρ_N נובעת מעליה ב- λ_N . כלומר, במצב זה ההסתברות לנטישה מונוטונית עולה בעומס המוצע למוקדן. לכן מספיק לטפל במקרה של $\rho_\infty \geq 1$, והתוצאה עבור שאר התחום תנבע מהמונוטוניות ומכך שנקבל שההסתברות לנטישה היא אפס עבור $\rho_\infty = 1$.

על פי [4] Bhattacharya and Ephremides ההסתברות לנטישה מונוטונית עולה גם בקצב הנטישה, ולכן ההסתברות לחסימה במערכת $M(\lambda_N)/M(1)/N/N$ (מתקבלת כגבול של מערכות $M(\lambda_N)/M(1)/N + M(\theta)$ כאשר $\theta \uparrow \infty$) מהווה חסם עליון להסתברות לנטישה במערכת $M(\lambda_N)/M(1)/N + M(\theta_N)$ עם קצב נטישה כלשהו. לקבלת ביטוי מפורש לחסם העליון נשתמש בתוצאות של [24] Jagerman:

אם $\lambda_N = N\rho_\infty$, $\rho_\infty > 1$ אזי (עמוד 538 נוסחה (63))

$$(5.1) \quad P_N\{Block\} \sim \left[\frac{\rho_\infty}{\rho_\infty - 1} - \frac{\rho_\infty}{(\rho_\infty - 1)^3} \frac{1}{N} + \frac{2\rho_\infty^2 + \rho_\infty}{(\rho_\infty - 1)^5} \frac{1}{N^2} \right]^{-1},$$

ואם $\lambda_N \equiv N$ אזי (עמוד 540 אחרי נוסחה (70))

$$(5.2) \quad P_N\{Block\} \sim \left[\sqrt{\frac{\pi N}{2}} + \frac{2}{3} + \frac{1}{12} \sqrt{\frac{\pi}{2N}} \right]^{-1}.$$

מכאן, כאשר $\rho_\infty > 1$, $N \uparrow \infty$, נקבל

$$(5.3) \quad \limsup_{N \rightarrow \infty} P_N\{Ab\} \leq \lim_{N \rightarrow \infty} P_N\{Block\} = \frac{\rho_\infty - 1}{\rho_\infty}.$$

באופן כללי, אם $\lambda_N = N\rho_\infty + o(N)$, ולכל $\rho_\infty > \varepsilon > 0$ נגדיר $\lambda_N^+ = N(\rho_\infty + \varepsilon)$, נקבל בגבול $N \uparrow \infty$ שההסתברות לחסימה במערכת $M(\lambda_N)/M(1)/N/N$ חסומה מלמעלה על ידי ההסתברות לחסימה במערכת $M(\lambda_N^+)/M(1)/N/N$. באותו אופן נקבל חסם תחתון כאשר נגדיר $\lambda_N^- = N(\rho_\infty - \varepsilon)$. כלומר שלכל $\rho_\infty > \varepsilon > 0$ מתקיים

$$(5.4) \quad \frac{\rho_\infty - \varepsilon - 1}{\rho_\infty - \varepsilon} \leq \liminf_{N \rightarrow \infty} P_N\{Block\} \leq \limsup_{N \rightarrow \infty} P_N\{Block\} \leq \frac{\rho_\infty + \varepsilon - 1}{\rho_\infty + \varepsilon},$$

כך שכאשר $\varepsilon \downarrow 0$ נקבל את (5.3).

לסיום ההוכחה דרוש לנו החסם התחתון. בכל מערכת $M/M/N + M$ קיים מצב יציב ובו קצב מופע העבודה אל המוקדנים לא עולה על קצב עבודתם המקסימלי. מטיעון זה מתקבל האי שיויון הבא:

$$(5.5) \quad \lambda_N(1 - P_N\{Ab\}) \leq N.$$

לאחר העברת אגפים מתקבל החסם:

$$(5.6) \quad P_N\{Ab\} \geq 1 - \frac{1}{\rho_N},$$

ובגבול

$$(5.7) \quad \liminf_{N \rightarrow \infty} P_N\{Ab\} \geq 1 - \frac{1}{\rho_\infty}.$$

□

יש לשים לב שלא מופיעה בתוצאה זו כל תלות בקצב הנטישה - אותה התנהגות גבולית מתקבלת עבור לקוחות מאד קצרי רוח ($\theta_N \uparrow \infty$) ועבור לקוחות בעלי סבלנות רבה ($\theta_N \downarrow 0$), כל עוד קיימים הגבולות הרלוונטיים $\rho_N \rightarrow \rho_\infty$ כאשר $N \uparrow \infty$.

פרק 6

התנהגות אסימפטוטית בעומס גבוה (Heavy Traffic)

לאחר שבפרק הקודם עשינו מיפוי גס של ההתנהגות האסימפטוטית של ההסתברות לנטישה כאשר $N \uparrow \infty$, נבצע עתה מיפוי עדין יותר תוך התמקדות בתחום פרמטרים מסוים. כפי שצויין במבוא, האתגר בעיצוב מוקדי שירות טלפוניים הוא השגת איזון בין יעילות ורמת שירות, ושאיפה למצב בו שניהם גבוהים. מסיבה זו נתמקד מעתה בתחום הפרמטרים בו באופן אסימפטוטי העומס האפקטיבי למוקדן הוא 1.

הערה: יש לשים לב שבמודל עם נטיות קיים הבדל בין העומס המוצע למוקדן (ρ_N) לעומס האפקטיבי למוקדן. העומס האפקטיבי למוקדן נמוך יותר כי חלק מהלקוחות שמגיעים למערכת נוטים לפני קבלת שירות, והוא מחושב על ידי הכפלת העומס המוצע למוקדן בפרופורציית מקבלי השירות ($1 - P_N\{Ab\}$). נזכיר גם שבמודל עם נטיות תמיד קיים מצב יציב, בפרט כאשר $\rho_N \geq 1$.

בשלב זה נתחיל להתעניין גם בהתנהגות האסימפטוטית של ההסתברות להמתנה. על פי Halfin and Whitt [19] במודל ללא נטיות, הגבול

$$(6.1) \quad \alpha = \lim_{N \rightarrow \infty} P_N\{Wait > 0\}$$

אינו מנוון (כלומר $0 < \alpha < 1$) אם ורק אם (אמ"מ) הגודל

$$(6.2) \quad \beta = \lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N),$$

מקיים $0 < \beta < \infty$.

כעת ניתן לקבוע תחום פרמטרים להמשך הניתוח. על פי האמור בתחילת הפרק אנחנו רוצים להתמקד בתחום בו מתקיים

$$(6.3) \quad \lim_{N \rightarrow \infty} \rho_N (1 - P_N\{Ab\}) = 1.$$

ממשפט 5.1 נובע ש-(6.3) מתקיים אם $\rho_\infty \geq 1$.

מצד שני על פי (1.22)

$$(6.4) \quad \lim_{N \rightarrow \infty} P_N\{Wait > 0\} = \lim_{N \rightarrow \infty} \frac{\rho_N(1 - P_N\{Block\})P_N\{Ab\} - P_N\{Block\}}{\rho_N(1 - P_N\{Block\}) - 1},$$

וכאשר $\rho_\infty > 1$ נקבל בעזרת (5.1)

$$(6.5) \quad \begin{aligned} \lim_{N \rightarrow \infty} P_N\{Wait > 0\} &= \lim_{N \rightarrow \infty} \frac{[N(\rho_\infty - 1)^2 - \rho_\infty](\rho_\infty - 1) - N(\rho_\infty - 1)^3}{[N(\rho_\infty - 1)^2 - \rho_\infty]\rho_\infty - [N(\rho_\infty - 1)^2 - 1]\rho_\infty} \\ &= \lim_{N \rightarrow \infty} \frac{\rho_\infty(\rho_\infty - 1)}{\rho_\infty(\rho_\infty - 1)} = 1. \end{aligned}$$

בדרך כלל, בהגדרת רמת שירות גבוהה דורשים שאחוז הלקוחות שנאלצים להמתין בתור לא יהיה גבוה. מכאן שהתחום בו $\rho_\infty > 1$ פחות מתאים למערכות עם רמת שירות גבוהה (על פי הגדרה זו), ולכן נבחר להתמקד בתחום $\rho_\infty = 1$. זהו עדיין תחום "נרחב" ונרצה התמקדות צרה יותר. כאן נכניס שיקולים יותר אינטואיטיביים, ויש מקום לעבודה נוספת שתבחן את הנושא בצורה יותר מדויקת. כאשר קצב הנטישה שואף לאפס נצפה לקבל בגבול מערכת $M/M/N$ (כלומר מערכת ללא נטישות). על פי העבודה של Halfin and Whitt התחום ה"מעניין" (במובן שאין ניוון ב-(6.1)) במקרה זה הוא כאשר $\rho_N \sim 1 - \beta/\sqrt{N}$. כאן מתקיים $\rho_\infty = 1$, ולכן סביר שגם אנו נתמקד בתחום זה. מעתה נקרא לתחום זה "תחום HW ". התייחסות נוספת להתמקדות בתחום HW מופיעה בפרק 7.

שים לב לכך שאצל Halfin and Whitt [19] הדיון מוגבל ל- $\beta > 0$ עקב הדרישה לקיום מצב יציב. במערכת $M/M/N + M$ תמיד קיים מצב יציב, לכן אנו נתייחס לכל התחום $-\infty < \beta < \infty$ (קיימת כאן הנחה שהגבול בהגדרה (6.2) של β תמיד קיים).

בניתוח של מערכת $M/M/N + M$ בתחום HW נסתמך על התוצאה של משפט 5.1 שבתחום

זה

$$(6.6) \quad \lim_{N \rightarrow \infty} P_N\{Ab\} = 0 ,$$

על ארבעת הקשרים הבאים (מופיעים בחלק §2.4, ללא האינדקס N):

$$(6.7) \quad P_N\{Ab\} = P_N\{Ab|Wait > 0\} P_N\{Wait > 0\}$$

$$(6.8) \quad P_N\{Ab\} = (1 - 1/\rho_N) P_N\{Wait > 0\} + \pi_N/\rho_N$$

$$(6.9) \quad P_N\{Ab\} = \frac{[\rho_N(1 - P_N\{Block\}) - 1] P_N\{Wait > 0\} + P_N\{Block\}}{\rho_N(1 - P_N\{Block\})}$$

$$(6.10) \quad P_N\{Ab|Wait > 0\} = 1 - 1/\rho_N + \frac{(\lambda_N/\theta_N)^{N\mu/\theta_N-1} \exp(-\lambda_N/\theta_N)}{\gamma(N\mu/\theta_N, \lambda_N/\theta_N)}$$

ועל התוצאה שמופיעה אצל Jagerman [24] (משפט 14) לגבי ההסתברות לחסימה במערכת $M/M/N/N$ בתחום HW :

$$(6.11) \quad \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Block\} = h(-\beta) ,$$

(הגדרתה של הפונקציה $h(x)$ מופיעה ב-(6.13)).

כמו כן נניח שקצב הנטישה הוא מהצורה $\theta_N = \delta \cdot N^\gamma$, כאשר $-\infty < \gamma < \infty$, $\delta > 0$. כפי שנראה הנחה זו תקל עלינו את הניתוח, ותאפשר חלוקה נוחה לתחומים עם מאפיינים שונים של התנהגות לקוחות, על פי ערך הפרמטר γ .

בחלקו הראשון של פרק זה נתמקד בניתוח אסימפטוטי עדין יותר של ההסתברות לנטישה, ונטפל גם בהסתברות להמתנה. בחלק השני הטיפול כללי יותר, ובמקום למצא קירובים למדדים מסוימים, נמצא קירובים להתפלגות הסטציונרית של מספר הלקוחות במערכת, לאחר rescaling מתאים. החלק השלישי הוא הכללי ביותר במובן ששם נמצא קירובי דיפוזיה לכל התהליך הסטוכסטי המתאר את מצב המערכת. בחלק האחרון נקשר בין קירובי הדיפוזיה ובין הקירובים שמצאנו להתפלגות הסטציונרית.

6.1 ההסתברות לנטישה וההסתברות להמתנה ($\theta_N \propto N^\gamma$)

כפי שנראה בחלק זה, מתקבל מגוון רחב של תוצאות אשר תלויות בפרמטרים של המודל. התלות העיקרית היא בפרמטר γ , אך בתחומים מסוימים מופיעה גם תלות ב- β . התוצאות מרוכזות בטבלה 6.1. בהמשך חלק זה נביא סדרה של למות בהן נוכיח את התוצאות, בהפרדה לשלושה תחומים על פי הערך של הפרמטר γ : $\gamma < 0$, $\gamma > 0$, $\gamma = 0$. בטבלה 6.1 מופיעה בכל "משבצת" הפניה ללמות הרלוונטיות. יש לשים לב לכך שאין אחידות ברמת הדיוק שהושגה עבור ההתנהגות האסימפטוטית, אם כי ברוב המקרים נתונים קצבי ההתכנסות.

6.1.1 $\gamma = 0$

עבור המקרה בו $\gamma = 0$ (כלומר, קצב הנטישה אינו תלוי במספר המוקדנים) נשתמש בתוצאות של Fleming, Stolyar and Simon [16].

למה 6.1.1 נניח $\gamma = 0$. אזי מתקיים:

$$\lim_{N \rightarrow \infty} P_N\{\text{Wait} > 0\} = w(-\beta, \sqrt{\frac{\mu}{\delta}})$$

הוכחה:

Fleming, Stolyar and Simon [16] מטפלים בהתפלגות הסטציונרית של התהליך $\frac{Q_N(t) - \lambda_N}{\sqrt{\lambda_N}}$. הם מוכיחים התכנסות חלשה של $\frac{Q_N(\infty) - \lambda_N}{\sqrt{\lambda_N}} \xrightarrow{d} \bar{q}(\infty)$, ומקבלים שלהתפלגות של $\bar{q}(\infty)$ יש את הצפיפות הבאה

$$(6.16) \quad f(x) = \begin{cases} C \frac{\phi(x)}{\phi(\beta)}, & x \leq \beta \\ C \frac{\phi((x-\beta)\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}})}{\phi(\beta\sqrt{\frac{\mu}{\delta}})}, & x > \beta \end{cases}$$

כאשר

$$(6.17) \quad C = h(-\beta) \left[1 + \frac{h(-\beta)\sqrt{\frac{\mu}{\delta}}}{h(\beta\sqrt{\frac{\mu}{\delta}})} \right]^{-1}.$$

טבלה 6.1: ריכוז תוצאות עבור התנהגות אסימפטוטית בתחום HW
Results for asymptotic behavior in HW regime

	$P_N\{Wait > 0\}$	$P_N\{Ab\}$
$\gamma < 0, \beta < 0$	~ 1 < 6.1.10 >	$\sim -\beta N^{-0.5}$ < 6.1.13 >
$\gamma < 0, \beta = 0$	$\sim 1 - \sqrt{\frac{\delta}{\mu}} N^{\gamma/2}$ < 6.1.10 , 6.1.14 >	$\sim \sqrt{\frac{2\delta}{\pi\mu}} N^{(\gamma-1)/2}$ < 6.1.12 >
$\gamma < 0, \beta > 0$	$\sim \alpha(\beta)$ < 6.1.9 >	$o(N^{-0.5})$ < 6.1.11 >
$\gamma = 0$	$\sim w(-\beta, \sqrt{\frac{\mu}{\delta}})$ < 6.1.1 >	$\sim [\sqrt{\frac{\delta}{\mu}} h(\beta \sqrt{\frac{\mu}{\delta}}) - \beta] \cdot w(-\beta, \sqrt{\frac{\mu}{\delta}}) N^{-0.5}$ < 6.1.2 >
$0 < \gamma < 1$	$\sim \sqrt{\frac{\pi\mu}{2\delta}} h(-\beta) N^{-\gamma/2}$ (?) < 6.1.3 , 6.1.4 , 6.1.8 >	$\sim h(-\beta) N^{-0.5}$ < 6.1.5 >
$\gamma = 1$	$\sim c(\frac{\mu}{\delta}) h(-\beta) N^{-0.5}$ < 6.1.4 , 6.1.7 >	$\sim h(-\beta) N^{-0.5}$ < 6.1.5 >
$\gamma > 1$	$\sim h(-\beta) N^{-0.5}$ < 6.1.4 , 6.1.6 >	$\sim h(-\beta) N^{-0.5}$ < 6.1.5 >

(?) - הוכח רק עבור המקרה $\beta = 0$. ראה הערה בלמה 6.1.8.

כאשר

$$(6.12) \quad w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)} \right]^{-1}$$

$$(6.13) \quad h(x) = \frac{\phi(x)}{1 - \Phi(x)}$$

$$(6.14) \quad \alpha(x) = \left[1 + \frac{x}{h(-x)} \right]^{-1}$$

$$(6.15) \quad c(x) = \frac{\gamma(x, x)}{x^{x-1} e^{-x}}$$

כאן ϕ ו- Φ הן פונקציית הצפיפות ופונקציית ההתפלגות המצטברת של מ"מ נורמלי סטנדרטי בהתאמה. $\gamma(x, y)$ זו ה-Incomplete Gamma Function המוגדרת על ידי האנטגרל $\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt$.

כעת

$$(6.18) \quad \frac{Q_N(\infty) - N}{\sqrt{N}} = \sqrt{\frac{\lambda_N}{N}} \left[\frac{Q_N(\infty) - \lambda_N}{\sqrt{\lambda_N}} + \frac{\lambda_N - N}{\sqrt{\lambda_N}} \right] \xrightarrow{d} \bar{q}(\infty) - \beta = q(\infty) .$$

מכאן בחישוב ישיר נקבל את המ.ש.ל:

$$(6.19) \quad \begin{aligned} \lim_{N \rightarrow \infty} P_N\{Wait > 0\} &= \lim_{N \rightarrow \infty} P\{Q_N(\infty) \geq N\} = \lim_{N \rightarrow \infty} P\left\{\frac{Q_N(\infty) - N}{\sqrt{N}} \geq 0\right\} \\ &= P\{\bar{q}(\infty) - \beta \geq 0\} = \int_{\beta}^{\infty} f(x)dx = w(-\beta, \sqrt{\frac{\mu}{\delta}}) . \end{aligned}$$

□

למה 6.1.2 נניח $\gamma = 0$. אזי מתקיים:

$$\lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} = \left[\sqrt{\frac{\delta}{\mu}} h(\beta \sqrt{\frac{\mu}{\delta}}) - \beta \right] \cdot w(-\beta, \sqrt{\frac{\mu}{\delta}})$$

הוכחה:

נשתמש ב-(6.9) ובתוצאת למה 6.1.1, ונקבל

$$(6.20) \quad \begin{aligned} \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} &= \lim_{N \rightarrow \infty} \frac{[\rho_N(1 - P_N\{Block\}) - 1] P_N\{Wait > 0\}}{\rho_N(1 - P_N\{Block\})} \\ &\quad + \lim_{N \rightarrow \infty} \frac{P_N\{Block\}}{\rho_N(1 - P_N\{Block\})} \\ &= -(\beta + h(-\beta)) \cdot w(-\beta, \sqrt{\frac{\mu}{\delta}}) + h(-\beta) \\ &= \left[\sqrt{\frac{\delta}{\mu}} h(\beta \sqrt{\frac{\mu}{\delta}}) - \beta \right] \cdot w(-\beta, \sqrt{\frac{\mu}{\delta}}) , \end{aligned}$$

כאשר המעבר האחרון מתקבל לאחר מספר מניפולציות אלגבריות פשוטות ושימוש בהגדרה של $w(x, y)$. □

הערה: התוצאה שהוכחה בלמה זו מופיעה בעבודה של Fleming, Stolyar and Simon עם טעות

(עמוד 10 נוסחה (5.3) - צריך להיות $\dots - nb(1 - \Phi(nb))$ במקום $\dots - b(1 - \Phi(nb))$).

נשתמש עתה בתוצאות שקיבלנו כדי להוכיח תוצאה אנליטית טהורה שתשמש אותנו בהמשך.

למה 6.1.3

$$\lim_{N \rightarrow \infty} \sqrt{N} \frac{N^{N-1} e^{-N}}{\gamma(N, N)} = \sqrt{\frac{2}{\pi}} ,$$

הוכחה:

נניח $\gamma = 0$, $\lambda_N \equiv N$, $\delta = 1$, $\mu = 1$. אזי על פי התוצאות של למות 6.1.1 ו-6.1.2 מתקיים

$$(6.21) \quad \begin{cases} \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} = h(0)w(0, 1) , \\ \lim_{N \rightarrow \infty} P_N\{Wait > 0\} = w(0, 1) , \end{cases}$$

מכאן על פי (6.7)

$$(6.22) \quad \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab|Wait > 0\} = \lim_{N \rightarrow \infty} \frac{\sqrt{N} P_N\{Ab\}}{P_N\{Wait > 0\}} = h(0) .$$

מצד שני כאשר נציב את ערכי הפרמטרים ב-(6.10) נקבל

$$(6.23) \quad P_N\{Ab|Wait > 0\} = \frac{N^{N-1} e^{-N}}{\gamma(N, N)} .$$

□

6.1.2 $\gamma > 0$

נמשיך את הניתוח עבור המקרים בהם קצב הנטישה הולך ומתגבר ($\gamma > 0$).

למה 6.1.4 נניח $\gamma > 0$. אזי

$$\lim_{N \rightarrow \infty} P_N\{Wait > 0\} = 0$$

הוכחה:

נשתמש בתוצאה שהוכחנו בלמה 6.1.1 לגבי ההסתברות להמתנה במקרה של $\gamma = 0$. ההסתברות להמתנה מונוטונית יורדת בקצב הנטישה, לכן לכל $\delta > 0$ מתקיים

$$(6.24) \quad \lim_{N \rightarrow \infty} P_N\{Wait > 0\} \leq w(-\beta, \sqrt{\frac{\mu}{\delta}}) .$$

כעת כאשר נשאיף $\delta \uparrow \infty$ נקבל על פי הגדרת $w(x, y)$ (ראה (6.12))

$$(6.25) \quad \lim_{N \rightarrow \infty} P_N\{Wait > 0\} \leq \lim_{\delta \rightarrow \infty} w(-\beta, \sqrt{\frac{\mu}{\delta}}) = 0 .$$

□

למה 6.1.5 נניח $\gamma > 0$ אזי

$$\lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} = h(-\beta)$$

הוכחה:

מתקבל באופן מיידי על פי (6.9) ולמה 6.1.4

$$(6.26) \quad \begin{aligned} \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} &= - \lim_{N \rightarrow \infty} \left(\sqrt{N}(1 - \rho_N) + \sqrt{N} P_N\{Block\} \right) P_N\{Wait > 0\} \\ &+ \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Block\} = h(-\beta) . \end{aligned}$$

□

הערה: שים לב שתוצאה זו היא לכל $\gamma > 0$, כלומר בתחום זה קצב ההתכנסות של ההסתברות לנטישה כאשר $N \uparrow \infty$ אינו תלוי בערך הפרמטר γ .

לסיום הטיפול במקרים בהם $\gamma > 0$, נבחן את קצב ההתכנסות של $P_N\{Wait > 0\}$.

למה 6.1.6 נניח $\gamma > 1$ אזי

$$\lim_{N \rightarrow \infty} \sqrt{N} P_N\{Wait > 0\} = h(-\beta)$$

הוכחה:

נקבל חסם תחתון עבור ההסתברות לנטישה בהנתן המתנה. נסמן

$$(6.27) \quad P\{Ab|k\} \equiv P\{Ab|kth \text{ in queue on arrival}\} .$$

אזי

$$(6.28) \quad P_N\{Ab|Wait > 0\} = \frac{\sum_{k=1}^{\infty} P\{Ab|k\}\pi_{N+k-1}}{P_N\{Wait > 0\}} .$$

כעת, ברור שלכל $k \geq 1$ מתקיים

$$(6.29) \quad P\{Ab|k\} \geq P\{Ab|1\} ,$$

ומכאן החסם

$$(6.30) \quad P_N\{Ab|Wait > 0\} \geq \frac{\sum_{k=1}^{\infty} P\{Ab|1\}\pi_{N+k-1}}{P_N\{Wait > 0\}} = P\{Ab|1\} .$$

ובגבול

$$(6.31) \quad \lim_{N \rightarrow \infty} P_N\{Ab|Wait > 0\} \geq \lim_{N \rightarrow \infty} P\{Ab|1\} = \lim_{N \rightarrow \infty} \frac{\theta_N}{\theta_N + N} = 1 .$$

כלומר במקרה זה $\lim_{N \rightarrow \infty} P_N\{Ab|Wait > 0\} = 1$, לכן על פי (6.7) ולמה 6.1.5

$$(6.32) \quad \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Wait > 0\} = \lim_{N \rightarrow \infty} \frac{\sqrt{N} P_N\{Ab\}}{P_N\{Ab|Wait > 0\}} = h(-\beta) .$$

□

למה 6.1.7 נניח $\gamma = 1$, אזי

$$\lim_{N \rightarrow \infty} \sqrt{N} P_N\{Wait > 0\} = c\left(\frac{\mu}{\delta}\right)h(-\beta)$$

הוכחה:

מתקבל באופן מיידי מ-(6.7) ו-(6.10). □

למה 6.1.8 נניח $0 < \gamma < 1$, $\beta = 0$, אזי

$$\lim_{N \rightarrow \infty} N^{\gamma/2} P_N\{Wait > 0\} = \sqrt{\frac{\mu}{\delta}}$$

הוכחה:

על פי (6.7)

$$(6.33) \quad \lim_{N \rightarrow \infty} N^{\gamma/2} P_N\{Wait > 0\} = \lim_{N \rightarrow \infty} \frac{\sqrt{N} P_N\{Ab\}}{N^{(1-\gamma)/2} P_N\{Ab|Wait > 0\}}.$$

נשתמש ב-(6.10) ונציב את התוצאות של למה 6.1.5

$$(6.34) \quad \lim_{N \rightarrow \infty} N^{\gamma/2} P_N\{Wait > 0\} = h(0) \lim_{N \rightarrow \infty} N^{(\gamma-1)/2} \frac{\gamma(N/\theta_N, N/\theta_N)}{(N/\theta_N)^{N/\theta_N-1} e^{-N/\theta_N}}.$$

לסיום ההוכחה נציב $M_N = N/\theta_N = \frac{\mu}{\delta} N^{1-\gamma}$ מכיון ש- $\lim_{N \rightarrow \infty} M_N = \infty$, נקבל על פי

למה 6.1.3

$$(6.35) \quad \lim_{N \rightarrow \infty} N^{\gamma/2} P_N\{Wait > 0\} = h(0) \lim_{M_N \rightarrow \infty} \frac{\sqrt{\frac{\mu}{\delta}}}{\sqrt{M_N}} \frac{\gamma(M_N, M_N)}{(M_N)^{M_N-1} e^{-M_N}} = \sqrt{\frac{\mu}{\delta}}.$$

□

הערה: מתוך (6.34) ניתן לשער שבמקרה הכללי, עבור β כלשהו, יתקיים

$$(6.36) \quad \lim_{N \rightarrow \infty} N^{\gamma/2} P_N\{Wait > 0\} = \frac{h(-\beta)}{\sqrt{2/\pi}} \sqrt{\frac{\mu}{\delta}}.$$

השערה זו עולה בקנה אחד עם תוצאות נומריות.

6.1.3 $\gamma < 0$

לסיום נעבור לטפל באופן דומה במקרים בהם קצב הנטישה הולך ומאט, כלומר $\gamma < 0$. כאן לראשונה משפיעים ערכים של β על ההתנהגות האסימפטוטית מסדר ראשון.

למה 6.1.9 נניח $\gamma < 0$, $\beta > 0$. אזי

$$\lim_{N \rightarrow \infty} P_N\{Wait > 0\} = \alpha(\beta)$$

הוכחה:

ההסתברות להמתנה במקרה זה חסומה מלעיל על ידי ההסתברות להמתנה במערכת $M(\lambda_N)/M(1)/N$ (מערכת זו מתקבלת כגבול של מערכות $M/M/N + M$ כאשר $\theta \downarrow 0$). זו המערכת שמנותחת במאמר של [19] Halfin and Whitt, שם הם מקבלים (עבור $\beta > 0$)

$$(6.37) \quad \lim_{N \rightarrow \infty} P_N\{Wait > 0\} = \alpha(\beta) .$$

כמו כן, עבור N מספיק גדול ההסתברות להמתין במערכת $M(\lambda_N)/M(1)/N + M(\theta_N)$ חסומה מלרע על ידי ההסתברות לאותו מאורע במערכת $M(\lambda_N)/M(1)/N + M(\delta)$ (בה $\gamma = 0$), כאשר $\delta > 0$ כלשהו.

מסקנה: לכל $\delta > 0$ מתקיים

$$(6.38) \quad w(-\beta, \sqrt{\frac{\mu}{\delta}}) \leq \liminf_{N \rightarrow \infty} P_N\{Wait > 0\} \leq \limsup_{N \rightarrow \infty} P_N\{Wait > 0\} \leq \alpha(\beta) ,$$

ולכן גם בגבול כאשר $\delta \downarrow 0$ הנ"ל נכון. נסתמך על האסימפטוטיקה המוכרת של זנב ההתפלגות הנורמלית הסטנדרטית, קרי

$$(6.39) \quad \frac{1}{h(x)} = \sqrt{2\pi} e^{x^2/2} (1 - \Phi(x)) \sim \frac{1}{x} , \quad x \uparrow \infty ,$$

ובגבול נקבל

$$(6.40) \quad \begin{aligned} \lim_{\delta \rightarrow 0} w(-\beta, \sqrt{\frac{\mu}{\delta}}) &= \lim_{\delta \rightarrow 0} \left[1 + \frac{h(\beta \sqrt{\frac{\mu}{\delta}})}{\sqrt{\frac{\mu}{\delta}} h(-\beta)} \right]^{-1} \\ &= \lim_{\delta \rightarrow 0} \left[1 + \frac{\beta \sqrt{\frac{\mu}{\delta}}}{\sqrt{\frac{\mu}{\delta}} h(-\beta)} \right]^{-1} = \alpha(\beta) . \end{aligned}$$

ומכאן המ.ש.ל. $\square$

למה 6.1.10 נניח $\beta \leq 0, \gamma < 0$ אזי

$$\lim_{N \rightarrow \infty} P_N\{Wait > 0\} = 1$$

הוכחה:

□ כמו בלמה הקודמת, מתקבל בעזרת החסם התחתון $w(-\beta, \sqrt{\frac{\mu}{\delta}})$ והשאפת δ לאפס.

כעת נעבור ללמות שמטפלות בהתנהגות האסימפטוטית של ההסתברות לנטישה.

למה 6.1.11 נניח $\beta > 0, \gamma < 0$ אזי

$$\lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} = 0$$

הוכחה:

על פי (6.9) ולמה 6.1.9 נקבל

$$(6.41) \quad \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} = h(-\beta) - [\beta + h(-\beta)]\alpha(\beta) \equiv 0.$$

□

למה 6.1.12 נניח $\beta = 0, \gamma < 0$ אזי

$$\lim_{N \rightarrow \infty} N^{(1-\gamma)/2} P_N\{Ab\} = \sqrt{\frac{2\delta}{\pi\mu}}$$

הוכחה:

על פי (6.7) ו-(6.10) נקבל

$$(6.42) \quad N^{(1-\gamma)/2} P_N\{Ab\} = N^{(1-\gamma)/2} \frac{(N/\theta_N)^{N/\theta_N-1} e^{-N/\theta_N}}{\gamma(N/\theta_N, N/\theta_N)} P_N\{Wait > 0\}.$$

□ לאחר שנציב $M_N = N/\theta_N$, על פי למה 6.1.3 ולמה 6.1.10, נקבל את המ.ש.ל.

למה 6.1.13 נניח $\gamma < 0, \beta < 0$. אזי

$$\lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} = -\beta$$

הוכחה:

נשתמש בפיתוח שמופיע ב-(6.9). על פי למה 6.1.10 נקבל

$$\begin{aligned} \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Ab\} &= \lim_{N \rightarrow \infty} \sqrt{N} P_N\{Block\} (1 - \rho_N P_N\{Wait > 0\}) \\ &\quad + \lim_{N \rightarrow \infty} \sqrt{N} (\rho_N - 1) P_N\{Wait > 0\} \\ (6.43) \qquad &= \lim_{N \rightarrow \infty} \sqrt{N} (\rho_N - 1) = -\beta . \end{aligned}$$

□

כעת נוכל לחזור ולבחון בצורה יותר עדינה את ההתכנסות של ההסתברות להמתנה כאשר $\beta = 0$

למה 6.1.14 נניח $\gamma < 0, \beta = 0$. אזי

$$\lim_{N \rightarrow \infty} N^{-\gamma/2} (1 - P_N\{Wait > 0\}) = \sqrt{\frac{\delta}{\mu}}$$

הוכחה:

מתקבל מהצבת (6.4) ומשימוש בתוצאות הלמות הקודמות. □

6.2 התכנסות חלשה של ההתפלגויות הסטציונריות $\{q_N(\infty)\}$

נטפל בסדרת התהליכים הסטוכסטיים $q_N = \{q_N(t), t \geq 0\}$ אשר הוגדרו ב-(4.1). בחלק זה נוכיח התכנסות חלשה של ההתפלגויות הסטציונריות של סדרת התהליכים $\{q_N, N = 0, 1, \dots\}$. לתהליכים אלה תמיד קיימות התפלגויות סטציונריות, שנשמך F_N . נתחיל מהתנאי ההכרחי להתכנסות חלשה - tightness.

משפט 6.1 סדרת ההתפלגויות F_N היא tight אם ורק אם $\gamma \geq 0$ או $\beta > 0$.

הוכחה:

להוכחת tightness מספיק למצא סדרות של התפלגויות F_N^+ ו- F_N^- אשר מתכנסות חלש, וחוסמות את הסדרה F_N מלעיל ומלרע בהתאמה. הוכחה בשיטה דומה, עבור המקרה של $\gamma = 0$, מופיעה אצל [16] Fleming, Stolyar and Simon.

תהי F_N^+ ההתפלגות הסטציונרית המתאימה לתהליך המתקבל מה-rescaling של תהליך לידה ומוות עם קצבי לידה (λ_k) ומוות (μ_k) הבאים:

$$(6.44) \quad \lambda_k = \begin{cases} \lambda_N, & k < N \\ 0, & \text{otherwise} \end{cases} ; \quad \mu_k = \begin{cases} k, & k \leq N \\ 0, & \text{otherwise} \end{cases}$$

(התהליך הנתון הוא בדיוק תהליך לידה ומוות המתאים למערכת $(M(\lambda_N)/M(1)/N/N)$.

אזי לכל N מתקיים $F_N^+ \geq F_N$, והסדרה F_N^+ מתכנסת חלש להתפלגות

$$(6.45) \quad F^+(x) = \begin{cases} \frac{\Phi(x+\beta)}{\Phi(\beta)}, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

את בחירת החסם התחתון נפריד לשני מקרים:

1. $\beta > 0$:

תהי F_N^- ההתפלגות הסטציונרית המתאימה ל-rescaling של תהליך לידה ומוות עם קצבים

$$(6.46) \quad \lambda_k = \begin{cases} \lambda_N, & k \geq N \\ 0, & \text{otherwise} \end{cases} ; \quad \mu_k = \begin{cases} N, & k > N \\ 0, & \text{otherwise} \end{cases}$$

(תהליך זה מתאים למערכת $M(\lambda_N)/M(N)/1$ "מוזן" ב- N).

לכל N מתקיים $F_N^- \leq F_N$, ועל פי [19] Halfin and Whitt הסדרה F_N^- מתכנסת חלש להתפלגות

$$(6.47) \quad F^-(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\beta x}, & x \geq 0 \end{cases}$$

2. $\gamma \geq 0$:

במקרה זה נגדיר את הסדרה F_N^- על סמך rescaling של תהליך לידה ומוות אחר, עם

קצבים

$$(6.48) \quad \lambda_k = \begin{cases} \lambda_N, & k \geq N \\ 0, & \text{otherwise} \end{cases}; \quad \mu_k = \begin{cases} N + (k - N)\frac{\delta}{\mu}, & k > N \\ 0, & \text{otherwise} \end{cases}$$

לכל N מתקיים $F_N^- \leq F_N$.

למה 6.2.1 הסדרה F_N^- מתכנסת חלש להתפלגות

$$F^-(x) = \begin{cases} 0, & x < 0 \\ 1 - \frac{\Phi(x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}})}{\Phi(-\beta\sqrt{\frac{\mu}{\delta}})}, & x \geq 0 \end{cases}$$

הוכחה:

תוצאה זו מופיעה למעשה אצל Fleming, Stolyar and Simon [16] (עם rescaling שממרכז סביב β במקום סביב 0). בהוכחה ב-[16] מראים tightness, ואחר כך מוצאים את הצפיפות של הגבול החלש. כאן נביא הוכחה אחרת, בה נחשב בצורה ישירה את הגבול של סדרת הפונקציות $\{F_N^-\}$.

לכל $x \geq 0$

$$(6.49) \quad \begin{aligned} F_N^-(x) &= P\{q_N(\infty) \leq x | q_N(\infty) \geq 0\} \\ &= P\{Q_N(\infty) \leq N + x\sqrt{N} | Q_N(\infty) \geq N\} = \frac{\sum_{n=0}^{x\sqrt{N}} \pi_{n+N}}{\sum_{n=0}^{\infty} \pi_{n+N}}. \end{aligned}$$

נציב את ההתפלגות הסטציונרית על פי (1.3) ונקבל בעזרת קירוב Stirling

$$(6.50) \quad F_N^-(x) = \frac{\sum_{n=0}^{x\sqrt{N}} (\lambda_N \frac{\mu}{\delta})^n \frac{\Gamma(N\frac{\mu}{\delta}+1)}{\Gamma(N\frac{\mu}{\delta}+n+1)}}{\sum_{n=0}^{\infty} (\lambda_N \frac{\mu}{\delta})^n \frac{\Gamma(N\frac{\mu}{\delta}+1)}{\Gamma(N\frac{\mu}{\delta}+n+1)}} \sim \frac{\sum_{n=N\frac{\mu}{\delta}}^{N\frac{\mu}{\delta}+x\sqrt{N}} e^{-\lambda_N \frac{\mu}{\delta}} \frac{(\lambda_N \frac{\mu}{\delta})^n}{n!}}{\sum_{n=N\frac{\mu}{\delta}}^{\infty} e^{-\lambda_N \frac{\mu}{\delta}} \frac{(\lambda_N \frac{\mu}{\delta})^n}{n!}}.$$

נזכור שאנו דנים בתחום HW לכן

$$(6.51) \quad N\frac{\mu}{\delta} \sim \lambda_N \frac{\mu}{\delta} + (\beta\sqrt{\frac{\mu}{\delta}})\sqrt{\lambda_N \frac{\mu}{\delta}},$$

$$(6.52) \quad N\frac{\mu}{\delta} + x\sqrt{N} \sim \lambda_N \frac{\mu}{\delta} + (\beta\sqrt{\frac{\mu}{\delta}} + x\sqrt{\frac{\delta}{\mu}})\sqrt{\lambda_N \frac{\mu}{\delta}}.$$

עכשיו נשתמש במשפט גבול מרכזי למ"מ פואסוני (ראה [15] Feller עמודים 245 – 244, 194, 190) שקובע

$$(6.53) \quad \lim_{\lambda \rightarrow \infty} \sum_{n=0}^{\lambda + \alpha \sqrt{\lambda}} e^{-\lambda} \frac{\lambda^n}{n!} = \Phi(\alpha) .$$

נחזור ונציב ב-(6.50), ונקבל כאשר $N \uparrow \infty$:

$$(6.54) \quad \begin{aligned} \lim_{N \rightarrow \infty} F_N^-(x) &= \lim_{N \rightarrow \infty} \frac{\sum_{n=0}^{N \frac{\mu}{\delta} + x \sqrt{N}} e^{-\lambda_N \frac{\mu}{\delta}} \frac{(\lambda_N \frac{\mu}{\delta})^n}{n!} - \sum_{n=0}^{N \frac{\mu}{\delta} - 1} e^{-\lambda_N \frac{\mu}{\delta}} \frac{(\lambda_N \frac{\mu}{\delta})^n}{n!}}{1 - \sum_{n=0}^{N \frac{\mu}{\delta} - 1} e^{-\lambda_N \frac{\mu}{\delta}} \frac{(\lambda_N \frac{\mu}{\delta})^n}{n!}} \\ &= \frac{\Phi(\beta \sqrt{\frac{\mu}{\delta}} + x \sqrt{\frac{\delta}{\mu}}) - \Phi(\beta \sqrt{\frac{\mu}{\delta}})}{1 - \Phi(\beta \sqrt{\frac{\mu}{\delta}})} , \end{aligned}$$

ומכאן נקבל את המ.ש.ל. $\square$

לקבלת ההכרחיות נראה שכאשר $\gamma < 0$ וגם $\beta \leq 0$ אין tightness. לשם כך נוכיח את הלמה הבאה.

למה 6.2.2 אם $\gamma < 0$ וגם $\beta \leq 0$ אזי לכל x סופי מתקיים

$$\lim_{N \rightarrow \infty} F_N(x) = 0$$

הוכחה:

עבור $x \leq 0$ התוצאה נובעת מלמה 6.1.10, כי הרי במקרה זה

$$(6.55) \quad \lim_{N \rightarrow \infty} F_N(x) \leq \lim_{N \rightarrow \infty} F_N(0) = 1 - \lim_{N \rightarrow \infty} P_N\{\text{Wait} > 0\} = 0 .$$

לכן נניח $x > 0$. כעת

$$(6.56) \quad \begin{aligned} F_N(x) &= P\{q_N(\infty) \leq x\} = P\{Q_N(\infty) \leq N + x \sqrt{N}\} \\ &= 1 - P_N\{\text{Wait} > 0\} + \sum_{n=N}^{N+x\sqrt{N}} \pi_n . \end{aligned}$$

ראשית נבחן את הטור. לאחר שנציב את ההתפלגות הסטציונרית מ-(1.3) ונכפול מונה ומכנה ב- $e^{-\lambda_N}$ נקבל

$$(6.57) \quad \sum_{n=N}^{N+x\sqrt{N}} \pi_n = \frac{\sum_{n=0}^{x\sqrt{N}} \frac{(\lambda_N)^n e^{-\lambda_N}}{(N+\theta_N) \cdots (N+n\theta_N)}}{\sum_{n=0}^N \frac{(\lambda_N)^n e^{-\lambda_N}}{n!} + \sum_{n=1}^{\infty} \frac{(\lambda_N)^n e^{-\lambda_N}}{(N+\theta_N) \cdots (N+n\theta_N)}}$$

נשים לב לשתי נקודות:

1. $\lim_{N \rightarrow \infty} \sum_{n=0}^N \frac{(\lambda_N)^n e^{-\lambda_N}}{n!} = \Phi(\beta)$, ולכן המכנה חיובי ממש.

2. לגבי המונה מתקיים

$$(6.58) \quad \begin{aligned} \sum_{n=0}^{x\sqrt{N}} \frac{(\lambda_N)^n e^{-\lambda_N}}{(N+\theta_N) \cdots (N+n\theta_N)} &\leq e^{-\lambda_N} \sum_{n=0}^{x\sqrt{N}} (\lambda_N/N)^n \\ &= e^{-\lambda_N} \frac{((\lambda_N/N)^{x\sqrt{N}+1} - 1)}{1 - \lambda_N/N} \\ &\sim \frac{\sqrt{N} e^{-\lambda_N}}{\beta} \left((1 - \beta/\sqrt{N})^{x\sqrt{N}+1} - 1 \right). \end{aligned}$$

כלומר כאשר $N \uparrow \infty$ המונה מתכנס ל-0.

המסקנה היא

$$(6.59) \quad \lim_{N \rightarrow \infty} \sum_{n=N}^{N+x\sqrt{N}} \pi_n = 0,$$

ולכן

$$(6.60) \quad \lim_{N \rightarrow \infty} P\{q_N(\infty) \leq x\} = 1 - \lim_{N \rightarrow \infty} P_N\{Wait > 0\}.$$

בכך סיימנו כי על פי למה 6.1.10, כאשר $\gamma < 0$ ו- $\beta \leq 0$, מתקיים

$$(6.61) \quad \lim_{N \rightarrow \infty} P_N\{Wait > 0\} = 1.$$

□ □

לאחר הוכחת התנאים ל-tightness של הסדרה F_N , נמצא את ההתפלגות שמתקבלת כאשר יש התכנסות חלשה. נפריד לתחומי ערכים שונים של γ . בכל המקרים נטפל בנפרד בתחום בו במערכת המקורית יש תור (כלומר $x > 0$), ובתחום בו אין תור ($x \leq 0$). בהנתן $x \leq 0$, אין נטישות, ולכן ההתפלגות היא כמו F^+ (ראה (6.45)), כלומר התפלגות נורמלית קטומה. ההסתברות שנהיה במצב זה ($x \leq 0$) היא בדיוק ההסתברות שלקוח יגיע למערכת ולא יצטרך להמתין בתור לשירות, כלומר $1 - \lim_{N \rightarrow \infty} P\{Wait > 0\}$. כעת נפריד למקרים על פי ערך הפרמטר γ :

1. $\gamma > 0$: כפי שהוכח בלמה 6.1.4, במקרה זה $\lim_{N \rightarrow \infty} P_N\{Wait > 0\} = 0$ המסקנה היא שכל ההתפלגות מרוכזת בתחום $x \leq 0$, כלומר ש- F היא התפלגות נורמלית קטומה, עם פונקציית צפיפות שנתונה על ידי

$$(6.62) \quad f(x) = \begin{cases} \frac{\phi(x+\beta)}{\Phi(\beta)}, & x \leq 0 \\ 0, & x > 0 \end{cases}$$

2. $\gamma = 0$: מקרה זה (עד כדי הזזה ב- β) טופל במאמר של [16] Fleming, Stolyar and Simon, והתוצאות פורטו בהוכחה של למה 6.1.1. פונקציית הצפיפות כאן היא

$$(6.63) \quad f(x) = \begin{cases} C \frac{\phi(x+\beta)}{\phi(\beta)}, & x \leq 0 \\ C \frac{\phi(x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}})}{\phi(\beta\sqrt{\frac{\mu}{\delta}})}, & x > 0 \end{cases}$$

3. $\gamma < 0$: (כאן $\beta > 0$, כפי שהוכח כאשר $\beta \leq 0$ אין התכנסות חלשה).

כאמור בתחום $x \leq 0$ ההתפלגות היא נורמלית קטומה, ומתקיים

$$(6.64) \quad F(0) = 1 - \lim_{N \rightarrow \infty} P_N\{Wait > 0\} = 1 - \alpha(\beta),$$

כאשר השיויון האחרון הוא על פי למה 6.1.9.

ההתנהגות בתחום בו $x > 0$ תוכח בלמה הבאה:

למה 6.2.3 נניח $\gamma < 0$, $\beta > 0$, ונסמן $\hat{F}(x) = \lim_{N \rightarrow \infty} P\{q_N(\infty) \leq x | q_N(\infty) > 0\}$,

אזי

$$\hat{F}(x) = 1 - e^{-\beta x}, \quad x \geq 0$$

הוכחה:

תחילה נראה שעבור ההתפלגות הסטציונרית (המותנה) במקרה של Fleming, Stolyar and Simon [16], שתסומן $\bar{F}(x)$, מתקבל הגבול הנ"ל כאשר $\delta \downarrow 0$. נציב את הביטוי

המתאים כפי שמופיע בלמה 6.1.1

$$(6.65) \quad \bar{F}(x) = \int_0^x C \frac{\phi(t\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}})}{w(-\beta, \sqrt{\frac{\mu}{\delta}})\phi(\beta\sqrt{\frac{\mu}{\delta}})} dt = C \frac{\Phi(x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})}{\sqrt{\frac{\delta}{\mu}} w(-\beta, \sqrt{\frac{\mu}{\delta}})\phi(\beta\sqrt{\frac{\mu}{\delta}})}.$$

כעת על פי ההגדרות של C , $w(x, y)$, $h(x)$ נקבל

$$\begin{aligned}\bar{F}(x) &= \frac{h(\beta\sqrt{\frac{\mu}{\delta}}) [\Phi(x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})]}{\phi(\beta\sqrt{\frac{\mu}{\delta}})} \\ (6.66) \quad &= 1 - \frac{h(\beta\sqrt{\frac{\mu}{\delta}})\phi(x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}})}{h(x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}})\phi(\beta\sqrt{\frac{\mu}{\delta}})}.\end{aligned}$$

לסיום נשתמש בהתנהגות האסימפטוטית של $h(x)$ כאשר $x \uparrow \infty$, כפי שמופיע ב-(6.39)

$$(6.67) \quad \lim_{\delta \rightarrow 0} \bar{F}(x) = 1 - \lim_{\delta \rightarrow 0} \frac{x\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}}{\beta\sqrt{\frac{\mu}{\delta}}} e^{-(x^2\frac{\delta}{\mu} + 2x\beta)/2} = 1 - e^{-\beta x} = F^-(x).$$

בכך למעשה סיימנו את ההוכחה מכיוון שלכל $\delta > 0$ מתקיים $\bar{F} \geq \hat{F}$, לכן גם הגבול מהווה חסם עליון, וכפי שצויין בתחילת חלק זה F^- מהווה חסם תחתון. הצפיפות שמתקבלת במקרה זה:

$$(6.68) \quad f(x) = \begin{cases} \frac{[1-\alpha(\beta)]\phi(x+\beta)}{\Phi(\beta)}, & x \leq 0 \\ \alpha(\beta)\beta e^{-\beta x}, & x > 0 \end{cases}$$

□

6.3 התכנסות חלשה של התהליכים הסטוכסטיים $\{q_N\}$

חלק זה מהווה תקיפה יסודית יותר של נושא ההתכנסות. במקום לבחון באופן ספציפי את ההתנהגות האסימפטוטית של מדד זה או אחר, או של ההתפלגות הסטוציונרית, נמצא את ההתנהגות האסימפטוטית של התהליך הסטוכסטי כולו. לאחר מציאת הגבול החלש (עם rescaling מתאים של התהליך המקורי), נוכל לקבל קירובים למדדים על ידי חישוב המדדים בתהליך הגבולי (בהסתמך על משפט ההעתקה הרציפה).

נטפל בתהליכים הסטוכסטיים $q_N = \{q_N(t), t \geq 0\}$, המוגדרים (כמו ב-(4.1)) על ידי

$$(6.69) \quad q_N = \frac{Q_N - N}{\sqrt{N}},$$

ונוכיח התכנסות חלשה של סדרת התהליכים $\{q_N, N = 1, 2, \dots\}$. ההתכנסות פה היא במרחב הפונקציות $D[0, \infty)$. גם כאן נקבל תוצאות שונות, כלומר תהליכים גבוליים שונים, בהתאם לערך הפרמטר γ .

משפט 6.2 נניח $\theta_N = \delta N^\gamma$, $\lambda_N \sim N\mu - \beta\sqrt{N}\mu$

אם $q_N(0) \xrightarrow{d} V_0$, אזי

$$q_N \xrightarrow{d} V$$

כאשר V מתקבל כפתרון של אחת מהמשוואות הדיפרנציאליות הסטוכסטיות הבאות, בהתאם לערך של γ :

• נטישות זניחות ($\gamma < 0$):

$$\begin{cases} dV_t = f(V_t)dt + \sqrt{2\mu}dW_t \\ f(x) = \begin{cases} -(\beta + x)\mu, & x \leq 0 \\ -\beta\mu, & x > 0 \end{cases} \end{cases}$$

• נטישות "מאוזנות" ($\gamma = 0$):

$$\begin{cases} dV_t = f(V_t)dt + \sqrt{2\mu}dW_t \\ f(x) = \begin{cases} -(\beta + x)\mu, & x \leq 0 \\ -(\beta + x\delta/\mu)\mu, & x > 0 \end{cases} \end{cases}$$

• תור זניח ($\gamma > 0$):

$$\begin{cases} dV_t = -(\beta + V_t)\mu dt + \sqrt{2\mu}dW_t - dY_t \\ Y_0 = 0, Y \uparrow 0, VdY = 0 \\ V_t \leq 0, \forall t. \end{cases}$$

כאן W מסמן תנועה בראונית סטנדרטית.

הערות:

1. שים לב שאת הנוסחאות הנ"ל הבענו ביחידות זמן ולא ביחידות של ממוצע זמן שירות.

כלומר קצב השירות הוא μ ולא אחד.

2. תהליכי הדיפוזיה המתקבלים כאן הם "הרכבה" של תהליכי דיפוזיה מסוג Ornstein-

Uhlenbeck (OU) ו/או Brownian Motion (BM), מוגדרים כל אחד על $\mathbb{R}_+$ או $\mathbb{R}_-$.

בפרט:

- $OU + BM : \gamma < 0$.
- $OU + OU : \gamma = 0$.
- $\gamma > 0$: מתקבל תהליך OU עם החזרה (reflection) שיסומן ROU .

הוכחה:

נסמן

$$(6.70) \quad X_N = Q_N - N .$$

כעת נטפל בנפרד בכל אחד משלושת המקרים.

$\gamma < 0$

כאן קצב הנטישה הולך לאפס. ככל ש- N גדל נקבל תהליך בו תופעת הנטישות פחות משמעותית, ואכן כפי שמיד יוכח, בגבול מתקבל אותו תהליך סטוכסטי שמופיע במאמר של [19] Halfin and Whitt, כגבול חלש של סדרת מערכות $M/M/N$ ללא נטישות. ההוכחה עצמה גם כן מתקבלת כמו ב-[19] Halfin and Whitt, קרי על ידי שימוש בקריטריון של [37] Stone (ראה הוכחת משפט 2 ב-[19]). עקב ה-rescaling מתקבל שמרחב המצב של התהליך q_N נהיה צפוף בישר הממשי כאשר $N \uparrow \infty$. בנוסף צריך לבחון את ההתכנסות של התוחלת והשונות האינפניטסימליים:

$$(6.71) \quad \mu_N(x) = \begin{cases} -\frac{|N+\sqrt{N}x|\mu}{\sqrt{N}} + \frac{\lambda_N}{\sqrt{N}}, & x \leq 0 \\ -\frac{N\mu+|\sqrt{N}x|\theta_N}{\sqrt{N}} + \frac{\lambda_N}{\sqrt{N}}, & x > 0 \end{cases}$$

$$(6.72) \quad \sigma_N^2(x) = \begin{cases} \frac{|N+\sqrt{N}x|\mu}{N} + \frac{\lambda_N}{N}, & x \leq 0 \\ \frac{N\mu+|\sqrt{N}x|\theta_N}{N} + \frac{\lambda_N}{N}, & x > 0 \end{cases}$$

ובגבול מתקבל כנדרש

$$(6.73) \quad \lim_{N \rightarrow \infty} \mu_N(x) = \begin{cases} -(\beta+x)\mu, & x \leq 0 \\ -\beta\mu, & x > 0 \end{cases}$$

$$(6.74) \quad \lim_{N \rightarrow \infty} \sigma_N^2(x) = 2\mu .$$

□

הערה: הכלים ב-Mandelbaum and Pats [26] אינם מאפשרים טיפול בכל התחום הנ"ל ($\gamma < 0$).
 כאשר $-0.5 < \gamma < 0$ מקבלים (ראה הגדרות במשפט 4.2 ב-[26], או ב-(6.83) כאן בהמשך)

$$\begin{aligned} f_\mu(\xi) &\equiv \lim_{N \rightarrow \infty} \sqrt{N} \left[\frac{1}{N} ((N\xi \wedge N) + \theta_N[N\xi - N]^+) - (\xi \wedge 1) \right] \\ (6.75) \quad &= \lim_{N \rightarrow \infty} \sqrt{N} \theta_N[\xi - 1]^+ . \end{aligned}$$

כלומר, הפונקציה $f_\mu(\xi)$ אינה מוגדרת עבור $\xi > 1$.

: $\gamma = 0$

מקרה זה מופיע ב-Fleming, Stolyar and Simon [16], כהשערה ללא הוכחה. ההוכחה מתקבלת בדיוק כמו במקרה $\gamma < 0$, כאשר כאן התוחלת והשונות האינפניטיסימליים נתונים על ידי

$$(6.76) \quad \mu_N(x) = \begin{cases} -\frac{|N+\sqrt{N}x|\mu}{\sqrt{N}} + \frac{\lambda_N}{\sqrt{N}}, & x \leq 0 \\ -\frac{N\mu+|\sqrt{N}x|\delta}{\sqrt{N}} + \frac{\lambda_N}{\sqrt{N}}, & x > 0 \end{cases}$$

$$(6.77) \quad \sigma_N^2(x) = \begin{cases} \frac{|N+\sqrt{N}x|\mu}{N} + \frac{\lambda_N}{N}, & x \leq 0 \\ \frac{N\mu+|\sqrt{N}x|\delta}{N} + \frac{\lambda_N}{N}, & x > 0 \end{cases}$$

ובגבול

$$(6.78) \quad \lim_{N \rightarrow \infty} \mu_N(x) = \begin{cases} -(\beta + x)\mu, & x \leq 0 \\ -(\beta + x\delta/\mu)\mu, & x > 0 \end{cases}$$

$$(6.79) \quad \lim_{N \rightarrow \infty} \sigma_N^2(x) = 2\mu .$$

□

הערה: אותה התוצאה מתקבלת גם על ידי שימוש בכלים שב-Mandelbaum and Pats [26].

: $\gamma > 0$

כאן ההוכחה יותר מורכבת. במקרה זה לא ניתן להשתמש בקריטריון של Stone כי התהליך הגבולי אינו "חי" על כל הישר הממשי (בהסתברות אחד הוא נמצא ב- $(-\infty, 0]$). הכלים שמופיעים ב-Mandelbaum and Pats [26] מאפשרים לטפל במקרים בהם התהליך הגבולי מוגבל

לחלק מהישר הממשי, אבל קיימת דרישה שקצבי הלידה והמוות יתכנסו (תחת rescaling מתאים) לפונקציות שהן locally Lipschitz. תנאי זה לא מתקיים במקרה שלנו, בו קצב הנטישה (ומכאן גם המוות) שואף לאינסוף כאשר $N \uparrow \infty$.

נתאר באופן כללי את שלבי ההוכחה. הרעיון הוא לחלק את $\mathbb{R}_+$, המתאים לציר הזמן, לשתי קבוצות זרות ומשלימות, בהתאם לזמנים בהם התהליך X_N נמצא ב- $(-\infty, 0]$ או ב- $(0, \infty)$. מכאן על ידי time change ניתן לקבל מ- X_N שני תהליכים, כל אחד "חי" בחלק אחר של הישר הממשי.

השלב הראשון הוא להראות שהתהליך ש"חי" ב- $(-\infty, 0]$ מתכנס לגבול הרצוי. תוצאה זו מתקבלת על ידי הפעלת הכלים הנתונים ב-Mandelbaum and Pats [26]. בשלב השני נראה שהזמן בו התהליך X_N נמצא ב- $(0, \infty)$ נהיה "זניח" כאשר $N \uparrow \infty$. שני השלבים האלה, יחד עם משפט ה-Inverse Random Time Change (ראה Appendix ב-Nguyen [28]), נותנים את המ.ש.ל.

כעת נעבור לפרטי ההוכחה. התהליך

$$(6.80) \quad \tau_N(t) \equiv \int_0^t 1_{\{q_N(s) \leq 0\}} ds$$

מתאר את משך הזמן שהתהליך q_N שהה ב- $(-\infty, 0]$ עד זמן t . מתוך ההגדרה, זהו גם הזמן שהתהליך X_N שהה שם. נגדיר תהליך סטוכסטי חדש,

$$(6.81) \quad X_N^{(-)}(t) = X_N \circ \tau_N^{-1}.$$

תוצאה מוכרת (ראה למשל [6] Blumenthal and Gettoor, עמוד 212) היא שהתהליך $X_N^{(-)}$ הוא מרקובי. בפרט, $X_N^{(-)}$ הוא תהליך לידה ומוות על קבוצת המצבים $\{-N, -N+1, \dots, 0\}$, עם קצבי לידה ומוות הנתונים על ידי

$$(6.82) \quad \lambda_k = \begin{cases} \lambda_N, & k < 0 \\ 0, & \text{otherwise} \end{cases}; \quad \mu_k = \begin{cases} (k+N)\mu, & k \leq 0 \\ 0, & \text{otherwise} \end{cases}$$

שים לב שתהליך זה מתאים למצב בו אין תור במערכת המקורית, ולכן גם אין נטישות. נפעיל את הכלים מ-Mandelbaum and Pats [26] על התהליך $X_N^{(-)}$ (בפועל מפעילים על התהליך $-X_N^{(-)}$, ומטעמי סימטריה התוצאה זהה פרט לסימן). נחשב בהתאם למופיע שם (נוסחאות

(4.4) ו-(4.8):

$$(6.83) \quad \lambda(\xi) \equiv \lim_{N \rightarrow \infty} \frac{1}{N} [N - N\xi]^+ \mu = [1 - \xi]^+ \mu ,$$

$$(6.84) \quad \mu(\xi) \equiv \lim_{N \rightarrow \infty} \frac{1}{N} \lambda_N = \mu ,$$

$$(6.85) \quad f_\lambda(\xi) \equiv \lim_{N \rightarrow \infty} \sqrt{N} ([1 - \xi]^+ \mu - \lambda(\xi)) = 0 ,$$

$$(6.86) \quad f_\mu(\xi) \equiv \lim_{N \rightarrow \infty} \sqrt{N} (\lambda_N/N - \mu(\xi)) = -\beta \mu .$$

כעת מכיוון ש- $V_0 \xrightarrow{d} q_N(0)$, הרי שבהכרח

$$(6.87) \quad \lim_{N \rightarrow \infty} \frac{1}{N} X_N^{(-)}(0) = 0 , \text{ a.s. .}$$

כמו כן מתקיים $\mu(0) = \lambda(0) = \mu$, לכן על פי המופיע ב-[26] אנו נמצאים במקרה של "critically loaded", ובגבול נקבל תהליך עם החזרה באפס, ותוחלת ושונות אינפנייטסימליים הנתונים על ידי

$$(6.88) \quad \mu(x) = f_\lambda(0) - f_\mu(0) + (\lambda'(0) - \mu'(0))x = (\beta - x)\mu ,$$

$$(6.89) \quad \sigma^2(x) = 2\lambda(0) = 2\mu .$$

הערה: ב-[26] יש טעות בנוסחאות שמתייחסות מפורשות למקרה של "critically loaded" (עמוד 251). במעבר מ-(4.9) הושמט המחובר עם הפונקציות f_λ ו- f_μ . מאידך, הנוסחה הכללית (4.9) מדויקת.

לסיכום, ההתכנסות שהתקבלה היא

$$(6.90) \quad q_N^{(-)} \equiv \frac{X_N^{(-)}}{\sqrt{N}} = q_N \circ \tau_N^{-1} \xrightarrow{d} ROU .$$

בכך סיימנו את השלב הראשון.

כעת מכיוון שקיבלנו התכנסות חלשה של $q_N \circ \tau_N^{-1}$, על פי משפט ה-Inverse Random Time Change [28], להשלמת ההוכחה מספיק להראות שמתקיים $e \xrightarrow{d} \tau_N^{-1} e$, כאשר $e(t) = t$ (פונקציית הזהות).

שלב שני:

נתבונן בתהליך המקורי Q_N . התהליך עובר לסרוגין בין מצב בו יש תור ($Q_N > N$) למצב בו אין תור ($Q_N \leq N$). נסמן:

A_N^i - משך התקופה ה- i בה אין תור.

B_N^i - משך התקופה ה- i בה יש תור.

ויהיו $C_N^i = A_N^i + B_N^i$, $i = 1, 2, \dots$ אלה "מחזורים" המורכבים מתקופה אחת ללא תור, ואחריה תקופה אחת עם תור. המרקוביות של התהליך Q_N מובילה לכך ש- C_N^i , $i = 1, 2, \dots$ הם משתנים מקריים ב"ת. C_N^i , $i = 2, 3, \dots$ הם גם שווי התפלגות. נגדיר

$$(6.91) \quad \sigma_N(T) = \min\{n : \sum_{i=1}^n C_N^i > T\},$$

אזי $\sigma_N(T)$ הוא זמן עצירה (stopping time) ביחס ל- $\{C_N^i\}$. גודל זה הוא מספר ה"מחזורים" שהתחילו עד זמן T .

בהתאם להגדרות הנ"ל אנו צריכים להוכיח שלכל $\varepsilon, T > 0$ מתקיים

$$(6.92) \quad \lim_{N \rightarrow \infty} P\left\{\sum_{i=1}^{\sigma_N(T)} B_N^i > \varepsilon\right\} = 0.$$

כלומר שהגודל $\sum_{i=1}^{\sigma_N(T)} B_N^i$ מתכנס בהסתברות לאפס לכל $T > 0$. אנחנו נוכיח שיש התכנסות של $\sum_{i=1}^{\sigma_N(T)} B_N^i$ לאפס ב- $\mathcal{L}^1$, והתכנסות זו גוררת התכנסות בהסתברות.

למה 6.3.1 לכל $T > 0$ מתקיים

$$\lim_{N \rightarrow \infty} E\left[\sum_{i=1}^{\sigma_N(T)} B_N^i\right] = 0$$

הוכחה:

נניח בשלב ראשון $Q_N(0) = N$, כך של- C_N^1 אותה התפלגות כמו לכל C_N^i . בהמשך נשתחרר מהנחה זו, ונראה שתוחלת הזמן עד שהתהליך Q_N מגיע לראשונה ל- N (כלומר $E[B_N^1]$) נעשית זניחה כאשר $N \uparrow \infty$.

במהלך ההוכחה של השלב הראשון נראה את הדברים הבאים:

$$E[A_N^1] \sim \Omega(1/\sqrt{N}) \bullet$$

$$E[B_N^1] \sim o(1/\sqrt{N}) \bullet$$

ומכאן נקבל $E[C_N^1] \sim \Omega(1/\sqrt{N})$.

אחר כך נשתמש בתוצאה זו למציאת ההתנהגות האסימפטוטית של $E[\sigma_N(T)]$. כפי שניתן

לשער מתוך ההגדרה, נקבל $E[\sigma_N(T)] \sim \Omega(\sqrt{N})$.

זה מסיים את ההוכחה מאחר ונראה $E[\sum_{i=1}^{\sigma_N(T)} B_N^i] = E[\sigma_N(T)]E[B_N^1]$.

נתחיל מחישוב התוחלות של A_N^1 ו- B_N^1 .

נסמן:

$$(6.93) \quad z_N^A = \min\{t : Q_N(t) > N\} \quad , \quad U_N^n = E[z_N^A | Q_N(0) = n]$$

$$(6.94) \quad z_N^B = \min\{t : Q_N(t) \leq N\} \quad , \quad V_N^n = E[z_N^B | Q_N(0) = n]$$

$$E[B_N^1] = V_N^{N+1}, E[A_N^1] = U_N^N$$

על ידי התניה בצעד ראשון נקבל משוואת הפרשים עבור U_N^n :

$$(6.95) \quad \begin{cases} U_N^n = U_N^{n+1} + \frac{\sum_{k=0}^n (\lambda_N/\mu)^k / k!}{(\lambda_N/\mu)^{n+1} / n!} , & 0 \leq n \leq N \\ U_N^{N+1} = 0 , \end{cases}$$

ומכאן באופן מידי מקבלים

$$(6.96) \quad U_N^N = \frac{\sum_{k=0}^N (\lambda_N/\mu)^k / k!}{(\lambda_N/\mu)^{N+1} / N!} = \frac{1}{(\lambda_N/\mu) \cdot P_N\{Block\}} \sim \frac{1}{\sqrt{N}h(-\beta)} ,$$

כאשר ההתנהגות האסימפטוטית נגזרת מהתוצאה של Jagerman [24] (ראה (6.11)).

לחישוב V_N^{N+1} נשתמש בעובדה שבתהליך לידה ומוות תוחלת הזמן בין ביקורים עוקבים במצב i נתונה על ידי $[\pi_i \cdot q_i]^{-1}$, כאשר q_i קצב היציאה ממצב i . מצד שני את תוחלת הזמן בין ביקורים עוקבים במצב N ניתן לחשב על ידי התניה בצעד ראשון. מהשוואת שני הביטויים נקבל:

$$(6.97) \quad \frac{1}{(\lambda_N + N\mu)\pi_N} = \frac{1}{\lambda_N + N\mu} + \frac{N\mu}{\lambda_N + N\mu} U_{N-1}^{N-1} + \frac{\lambda_N}{\lambda_N + N\mu} V_N^{N+1} .$$

עבור U_{N-1}^n ניתן לקבל אותן משוואות הפרשים כמו ל- U_N^n , רק תנאי השפה יהיה $U_{N-1}^N = 0$,

ומכאן ש

$$(6.98) \quad U_{N-1}^{N-1} = \frac{\sum_{k=0}^{N-1} (\lambda_N/\mu)^k / k!}{(\lambda_N/\mu)^N / (N-1)!} .$$

כעת ניתן להציב ב-(6.97), ולחלץ את V_N^{N+1} :

$$(6.99) \quad V_N^{N+1} = \frac{1}{\lambda_N \cdot \pi_N} - \frac{1}{\lambda_N} - \frac{N\mu}{\lambda_N} U_{N-1}^{N-1} = \frac{1}{\lambda_N \cdot \pi_N} - U_N^N.$$

על פי (1.24)

$$(6.100) \quad \lambda_N \cdot \pi_N = \frac{\lambda_N(1 - P_N\{Wait > 0\})P_N\{Block\}}{1 - P_N\{Block\}},$$

ומלמות 6.1.4 ו-6.1.5 נקבל

$$(6.101) \quad \lambda_N \cdot \pi_N \sim \sqrt{N}h(-\beta) + o(\sqrt{N}).$$

מכאן

$$(6.102) \quad V_N^{N+1} \sim \frac{1}{\sqrt{N}h(-\beta) + o(\sqrt{N})} - \frac{1}{\sqrt{N}h(-\beta)} = o(1/\sqrt{N}).$$

הערה: שים לב שבתוצאה זו נבדל המקרה $\gamma > 0$ מהמקרים בהם $\gamma \leq 0$, עבורם אפשר לקבל בדרך דומה $V_N^{N+1} = \Omega(1/\sqrt{N})$. כפי שנראה בהמשך, התנהגות זו היא ה"לב" של ההוכחה.

כלומר קיבלנו

$$(6.103) \quad E[A_N^1] \sim \frac{1}{\sqrt{N}h(-\beta)},$$

$$(6.104) \quad E[B_N^1] \sim o(1/\sqrt{N}),$$

ומתוצאות אלה נובע

$$(6.105) \quad E[C_N^1] \sim \frac{1}{\sqrt{N}h(-\beta)}.$$

נשתמש בתוצאה זו למציאת ההתנהגות האסימפטוטית של $E[\sigma_N(T)]$:

למה 6.3.2 לכל $T > 0$ מתקיים

$$E[\sigma_N(T)] \sim \sqrt{N}T \cdot h(-\beta)$$

הוכחה:

באופן אינטואיטיבי מכיוון שב-(6.105) קיבלנו שבממוצע ה"מתזורים" קטנים כמו $1/\sqrt{N}$, נצפה שמספר המתזורים שהתחילו עד זמן T יגדל כמו $\sqrt{N}$. הרעיון בהוכחה הוא לשנות את סקלת

הזמן בפקטור $\sqrt{N}$ כך שנקבל מחזורים שתוחלתם קבועה (אסימפטוטית). בניסוח החדש התוצאה שאנו רוצים לקבל דומה למשפט החידוש הבסיסי (Elementary renewal theorem), וההוכחה תהיה דומה להוכחת משפט זה כפי שמופיע ב-[18] Grimmett and Stirzaker (עמוד 394).

נגדיר

$$(6.106) \quad \hat{\sigma}_N(T) = \sigma_N(T/\sqrt{N}) = \min\{n : \sum_{i=1}^n \sqrt{N}C_N^i > T\}$$

על פי הגדרה זו

$$(6.107) \quad T < \sum_{i=1}^{\hat{\sigma}_N(T)} \sqrt{N}C_N^i .$$

עבור התוחלת, על פי משפט Wald (ראה למשל [18] Grimmett and Stirzaker עמוד 396) נקבל

$$(6.108) \quad T < E\left[\sum_{i=1}^{\hat{\sigma}_N(T)} \sqrt{N}C_N^i\right] = E[\hat{\sigma}_N(T)] \cdot E[\sqrt{N}C_N^i] ,$$

ולאחר העברת אגפים

$$(6.109) \quad \frac{E[\hat{\sigma}_N(T)]}{T} > \frac{1}{\sqrt{N}E[C_N^i]} .$$

הנ"ל לכל $T > 0$ ולכן

$$(6.110) \quad \frac{E[\hat{\sigma}_N(\sqrt{N}T)]}{\sqrt{N}T} > \frac{1}{\sqrt{N}E[C_N^i]} ,$$

ובגבול כאשר $N \uparrow \infty$

$$(6.111) \quad \liminf_{N \rightarrow \infty} \frac{E[\hat{\sigma}_N(\sqrt{N}T)]}{\sqrt{N}T} \geq \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N}E[C_N^i]} = h(-\beta) .$$

מצאנו חסם תחתון, נעבור למציאת חסם עליון. יהי $a > 0$ ונגדיר

$$(6.112) \quad \bar{C}_N^i = C_N^i \wedge (a/\sqrt{N}) ,$$

כלומר $\sqrt{N}\bar{C}_N^i = \sqrt{N}C_N^i \wedge a$ כמו כן נגדיר

$$(6.113) \quad \bar{\sigma}_N(T) = \min\{n : \sum_{i=1}^n \sqrt{N}\bar{C}_N^i > T\} .$$

על פי הגדרת $\bar{C}_N^i$ מתקיים $\sqrt{N}\bar{C}_N^i \leq \sqrt{N}C_N^i$ ומכאן

$$(6.114) \quad \bar{\sigma}_N(T) \geq \hat{\sigma}_N(T) ,$$

וכמובן שעבור התוחלות

$$(6.115) \quad E[\bar{\sigma}_N(T)] \geq E[\hat{\sigma}_N(T)] .$$

כעת על פי הגדרת $\bar{\sigma}_N(T)$

$$(6.116) \quad T \geq \sum_{i=1}^{\bar{\sigma}_N(T)} \sqrt{N} \bar{C}_N^i - \sqrt{N} \bar{C}_N^{\bar{\sigma}_N(T)} .$$

שוב עבור התוחלת ועל פי משפט Wald נקבל

$$(6.117) \quad T \geq E[\bar{\sigma}_N(T)] \cdot E[\sqrt{N} \bar{C}_N^i] - E[\sqrt{N} \bar{C}_N^{\bar{\sigma}_N(T)}] \geq \sqrt{N} E[\hat{\sigma}_N(T)] \cdot E[\bar{C}_N^i] - a .$$

לאחר העברת אגפים

$$(6.118) \quad \frac{E[\hat{\sigma}_N(T)]}{T} \leq \frac{1}{\sqrt{N} E[\bar{C}_N^i]} + \frac{a}{\sqrt{N} E[\bar{C}_N^i] \cdot T} .$$

הנ"ל לכל $T > 0$ ולכן

$$(6.119) \quad \frac{E[\hat{\sigma}_N(\sqrt{N}T)]}{\sqrt{N}T} \leq \frac{1}{\sqrt{N} E[\bar{C}_N^i]} + \frac{a/(\sqrt{N}T)}{\sqrt{N} E[\bar{C}_N^i]} ,$$

ובגבול $N \uparrow \infty$ מתקבל

$$(6.120) \quad \limsup_{N \rightarrow \infty} \frac{E[\hat{\sigma}_N(\sqrt{N}T)]}{\sqrt{N}T} \leq \lim_{N \rightarrow \infty} \frac{1}{\sqrt{N} E[\bar{C}_N^i]} \equiv \frac{1}{f(a)} .$$

כאן f זו הפונקציה המוגדרת על ידי הגבול $\lim_{N \rightarrow \infty} \sqrt{N} E[\bar{C}_N^i]$.

כעת נשאוף את $a \uparrow \infty$. על פי משפט ההתכנסות המונוטונית נקבל

$$(6.121) \quad \limsup_{N \rightarrow \infty} \frac{E[\hat{\sigma}_N(\sqrt{N}T)]}{\sqrt{N}T} \leq \frac{1}{f(\infty)} = h(-\beta) .$$

כלומר קיבלנו גם חסם עליון, ולכן קיים הגבול

$$(6.122) \quad \lim_{N \rightarrow \infty} \frac{E[\hat{\sigma}_N(\sqrt{N}T)]}{\sqrt{N}T} = h(-\beta) ,$$

ובמונחי $\sigma_N(T)$

$$(6.123) \quad \lim_{N \rightarrow \infty} \frac{E[\sigma_N(T)]}{\sqrt{N}T} = h(-\beta) .$$

בכך סיימנו את הוכחת הלמה. □

כעת נחשב את $E[\sum_{i=1}^{\sigma_N(T)} B_N^i]$:

$$(6.124) \quad E\left[\sum_{i=1}^{\sigma_N(T)} B_N^i\right] = E\left[\sum_{i=1}^{\infty} B_N^i 1_{\{i \leq \sigma_N(T)\}}\right] = \sum_{i=1}^{\infty} E[B_N^i 1_{\{i \leq \sigma_N(T)\}}],$$

כאשר המעבר האחרון מסתמך על משפט ההתכנסות המונוטונית. מכיוון שהמאורע $\{i \leq \sigma_N(T)\}$ תלוי רק במשתנים המקריים $C_N^1, \dots, C_N^{i-1}$, הרי שהוא ב"ת ב- C_N^i , ולכן גם ב"ת ב- B_N^i . מהאי תלות נקבל

$$(6.125) \quad \begin{aligned} E\left[\sum_{i=1}^{\sigma_N(T)} B_N^i\right] &= \sum_{i=1}^{\infty} E[B_N^i] P\{\sigma_N(T) \geq i\} = E[B_N^1] \sum_{i=1}^{\infty} P\{\sigma_N(T) \geq i\} \\ &= E[B_N^1] E[\sigma_N(T)], \end{aligned}$$

ומכאן שבגבול

$$(6.126) \quad \lim_{N \rightarrow \infty} E\left[\sum_{i=1}^{\sigma_N(T)} B_N^i\right] = \lim_{N \rightarrow \infty} \sqrt{N} \cdot o(1/\sqrt{N}) = 0.$$

לסיום ההוכחה נותר להשתחרר מההנחה לגבי $Q_N(0)$. ברור שאם $Q_N(0) < N$ (כלומר מתחילים ממצב בו יש שרתים בטלים) כמות הזמן שהתהליך ימצא במצב בו יש תור, עד זמן T , קטנה יותר (סטוכסטית) ולכן נקבל את התוצאה הרצויה. לכן נניח $Q_N(0) > N$. עבור מקרה זה נגדיר $A_N^1 = 0$ ו- B_N^1 הוא הזמן עד הפעם הראשונה שאין תור. מספיק להראות

$$(6.127) \quad \lim_{N \rightarrow \infty} E[B_N^1] = 0,$$

כי מרגע שהתהליך Q_N יגיע לראשונה ל- N , כמות הזמן שהוא ימצא במצב עם תור, עד זמן T , חסומה (סטוכסטית) מלמעלה על ידי גודל זה במקרה בו $Q_N(0) = N$ (שטופל כבר). על פי הסימונים שהוגדרו (ראה (6.94)) $E[B_N^1] = V_N^{Q_N(0)}$. באופן כללי מתקיים

$$(6.128) \quad V_N^{N+k} = \sum_{n=1}^k (V_N^{N+n} - V_N^{N+n-1}) = \sum_{n=1}^k W_N(N+n, -1),$$

כאשר $W_N(a, b)$ מסמן את תוחלת הזמן שלוקח לתהליך Q_N להגיע לראשונה ל- $a+b$ בהנתן שהוא מתחיל ב- a . שים לב שעל פי הגדרה זו מתקיים $V_N^{N+1} = W_N(N+1, -1)$.

מהתניה בצעד ראשון מקבלים

$$(6.129) \quad W_N(N+n, -1) = \frac{1}{\lambda_N + N\mu + n\theta_N} + \frac{\lambda_N}{\lambda_N + N\mu + n\theta_N} (W_N(N+n+1, -1) + W_N(N+n, -1)) , \quad n = 1, 2, \dots ,$$

ומכאן מקבלים

$$(6.130) \quad W_N(N+n+1, -1) = \frac{N\mu + n\theta_N}{\lambda_N} W_N(N+n, -1) - \frac{1}{\lambda_N} , \quad n = 1, 2, \dots .$$

באופן טריוויאלי מתקיים

$$(6.131) \quad W_N(N+n+1, -1) \leq \frac{N\mu + n\theta_N}{\lambda_N} W_N(N+n, -1) .$$

הצבה בחזרה ב-(6.128) נותנת

$$(6.132) \quad V_N^{N+k} \leq V_N^{N+1} + \sum_{n=1}^{k-1} \frac{N\mu + n\theta_N}{\lambda_N} W_N(N+n, -1) ,$$

ומכאן באופן רקורסיבי נקבל

$$(6.133) \quad V_N^{N+k} \leq V_N^{N+1} \sum_{n=0}^{k-1} \left(\prod_{j=1}^n \frac{N\mu + j\theta_N}{\lambda_N} \right) .$$

נקבל חסם לסכום על ידי

$$(6.134) \quad \sum_{n=0}^{k-1} \left(\prod_{j=1}^n \frac{N\mu + j\theta_N}{\lambda_N} \right) \leq \sum_{n=0}^{k-1} \left(\frac{N\mu + (k-1)\theta_N}{\lambda_N} \right)^n = \frac{1 - \left(\frac{N\mu + (k-1)\theta_N}{\lambda_N} \right)^k}{1 - \frac{N\mu + (k-1)\theta_N}{\lambda_N}} .$$

כעת נשים לב לתכונת מונוטוניות של הגודל V_N^{N+k} . עבור תהליך לידה ומוות ארגודי, אם מגדילים את קצבי המעבר כך שהיחס בין קצבי הלידה לקצבי המוות בכל מצב נשמר, נקבל ש- V_N^{N+k} קטן. זאת מכיוון שהגדלת קצבי המעבר בצורה זו שקולה באופן מעשי להאצת הזמן (לא בהכרח בצורה הומוגנית). אפשרות אחרת היא להגדיל את קצבי המוות ובמקביל להקטין את קצבי הלידה, באופן שבו עבור כל מצב סכומם ישאר קבוע. כאן אנו לא משנים את סקלת הזמן, אבל מגדילים את הסיכויים למיתות. לכן גם במקרה זה V_N^{N+k} קטן. כמוכן שגם שילוב של שתי הפעולות הנ"ל יגרום ל- V_N^{N+k} לקטון, ומכאן נובע שגודל זה מונוטוני יורד בקצב הנטישה. נשתמש במונוטוניות זו לקבלת חסם נוסף.

יש לשים לב שיתכנו שני מקרים אפשריים בהתאם ל- V_0 . אם $V_0 \leq 0$ התהליך הגבולי יהיה רציף על $[0, \infty)$, אחרת הוא יהיה רציף רק על $(0, \infty)$. מבחינה איכותית, אם $V_0 > 0$ תהיה קפיצה מיידית בזמן $0+$ לנקודה 0. אנו לא נטפל כאן במקרה זה. דוגמא לטיפול במקרים דומים ניתן למצא ב- [11] Chen and Mandelbaum.

כעת אם $V_0 \leq 0$, הרי שאם מתקיים $Q_N(0) > N$ אז

$$(6.135) \quad Q_N(0) = N + o(\sqrt{N}) .$$

כלומר אם נרשום $Q_N(0) = N + k_N$, אז בהכרח $k_N = o(\sqrt{N})$. לכן אם נגדיר קצב נטישה חדש על ידי

$$(6.136) \quad \hat{\theta}_N = \left(\frac{N}{k_N^2} \right) \wedge \theta_N ,$$

נקבל שמתקיימים שלושת הדברים הבאים:

$$(6.137) \quad \hat{\theta}_N \leq \theta_N ,$$

$$(6.138) \quad \lim_{N \rightarrow \infty} \hat{\theta}_N = \infty ,$$

$$(6.139) \quad \lim_{N \rightarrow \infty} \frac{\hat{\theta}_N k_N^2}{\lambda_N} \leq \mu .$$

נחזור ל-(6.133), ונשתמש בחסמים שמתקבלים על סמך המונוטוניות, ממערכת עם קצב נטישה $\hat{\theta}_N$. מעל לגדלים המתאימים למערכת זו יתווסף הסימן " $\wedge$ ". נקבל:

$$(6.140) \quad V_N^{N+k_N} \leq \hat{V}_N^{N+k_N} \leq \hat{V}_N^{N+1} \frac{1 - \left(\frac{N\mu + (k_N-1)\hat{\theta}_N}{\lambda_N} \right)^{k_N}}{1 - \frac{N\mu + (k_N-1)\hat{\theta}_N}{\lambda_N}} .$$

על פי (6.138), בסדרת המערכות " $\wedge$ " ההתנהגות האסימפטוטית של קצב הנטישה היא כמו במקרה $\gamma > 0$. לכן ניתן להשתמש בתוצאה שקיבלנו ב-(6.102) עבור ההתנהגות האסימפטוטית של $\hat{V}_N^{N+1}$, לפיה $\hat{V}_N^{N+1} \sim o(1/\sqrt{N})$.

הערה: יתכן מצב שקצב השאיפה של $\hat{\theta}_N$ לאינסוף איטי כל כך שלא ניתן להציגו בצורה δN^γ , $\gamma > 0$ (למשל אם $\hat{\theta}_N \sim \log N$). ניתן להראות שגם במקרה זה (6.101) מתקיים (ומכאן שגם (6.102) מתקיים), למשל על ידי חסימה מלמטה עם מערכות בהן $\gamma = 0$, והשאפת $\delta \uparrow \infty$ (בדיוק כמו בהוכחת למה 6.1.4).

בכך סיימנו כי מכאן

$$(6.141) \quad \lim_{N \rightarrow \infty} E[B_N^1] = \lim_{N \rightarrow \infty} V_N^{Q_N(0)} \leq \lim_{N \rightarrow \infty} \hat{V}_N^{N+1} (1 - e) O(\sqrt{N}) = 0 .$$

□ □

6.4 קירובים להתפלגות הסטציונרית - סיכום

לאחר הוכחת ההתכנסות החלשה של סדרת התהליכים $\{q_N\}$, נודא שההתפלגויות הסטציונריות של התהליכים הגבוליים שהתקבלו (כל אחד בהתאם לפרמטרים של המודל) זהות להתפלגויות שהתקבלו כגבול החלש של ההתפלגות הסטציונרית (ראה חלק 6.2). דרך למציאת ההתפלגות הסטציונרית של תהליך דיפוזיה מהסוג הנידון כאן מופיעה אצל Browne and Whitt [10] (חלק 18.3). הם מטפלים בתהליכי דיפוזיה רציפים למקוטעין, כלומר תהליכי דיפוזיה בהם התוחלת והשונות האינפניטסימליים $\mu(x)$ ו- $\sigma^2(x)$ (בהתאמה) הם פונקציות רציפות למקוטעין, כאשר $\sigma^2(x) > 0$. מניחים שקיימות נקודות $s_0 < s_1 < \dots < s_k$ כך שהפונקציות $\mu(x)$ ו- $\sigma^2(x)$ רציפות ב- (s_{i-1}, s_i) (יתכן $s_0 = -\infty$ או $s_k = \infty$). כמו כן מניחים שנקודות השפה s_0 ו- s_k הן מחזירות (reflecting), אם סופיות, או לא נגישות (inaccessible), אם אינסופיות. תחת הנחות אלה ההתפלגות הסטציונרית היא בעלת צפיפות $f(x)$ הנתונה על ידי:

$$(6.142) \quad f(x) = I \cdot m(x) , \quad s_0 \leq x \leq s_k ,$$

כאשר

$$(6.143) \quad m(x) = \frac{2}{\sigma^2(x)s(x)} ,$$

$$(6.144) \quad s(x) = \exp \left\{ - \int_{\theta}^x \frac{2\mu(y)}{\sigma^2(y)} dy \right\}$$

עבור $s_0 < \theta < s_k$ כלשהו (בחירה שרירותית),

$$(6.145) \quad I = \left[\int_{s_0}^{s_k} m(y) dy \right]^{-1} .$$

נשתמש בפרוצדורה הנ"ל למציאת ההתפלגויות הסטציונריות של תהליכי הדיפוזיה שהתקבלו בחלק הקודם. נפריד למקרים על פי ערך הפרמטר γ :

1. $\gamma < 0$:

נקודות החלוקה:

$$(6.146) \quad k = 2 : (s_0, s_1, s_2) = (-\infty, 0, \infty)$$

תוחלת ושונות אינפיניטסימליים:

$$(6.147) \quad \mu(x) = \begin{cases} -(\beta + x), & x \leq 0 \\ -\beta, & x > 0 \end{cases}, \quad \sigma^2(x) = 2$$

כעת נבחר $\theta = 0$ ונקבל

$$(6.148) \quad s(x) = \begin{cases} \exp\{-\int_0^x -(\beta + y)dy\} = \exp(x\beta + x^2/2), & x \leq 0 \\ \exp\{-\int_0^x -\beta dy\} = \exp(x\beta), & x > 0 \end{cases}$$

ומכאן

$$(6.149) \quad m(x) = \begin{cases} \exp(-x\beta - x^2/2), & x \leq 0 \\ \exp(-x\beta), & x > 0 \end{cases}$$

והצפיפות (לאחר ביצוע הנרמול):

$$(6.150) \quad f(x) = \begin{cases} \frac{1-\alpha(\beta)}{\sqrt{2\pi}\Phi(\beta)} \exp(-(x+\beta)^2/2), & x \leq 0 \\ \alpha(\beta)\beta \exp(-x\beta), & x > 0 \end{cases}$$

שים לב שהאנטגרל בהגדרת I (ראה (6.145) מתבדר כאשר $\beta \leq 0$. כלומר הצפיפות הנתונה היא רק עבור $\beta > 0$, ורק במקרה זה קיימת לתהליך הדיפוזיה התפלגות סטציונרית.

2. $\gamma = 0$:

נקודות החלוקה:

$$(6.151) \quad k = 2 : (s_0, s_1, s_2) = (-\infty, 0, \infty)$$

תוחלת ושונות אינפיניטסימליים:

$$(6.152) \quad \mu(x) = \begin{cases} -(\beta + x), & x \leq 0 \\ -(\beta + x\delta/\mu), & x > 0 \end{cases}, \quad \sigma^2(x) = 2$$

כעת נבחר $\theta = 0$ ונקבל

$$(6.153) \quad s(x) = \begin{cases} \exp\{-\int_0^x -(\beta + y)dy\} = \exp(x\beta + x^2/2), & x \leq 0 \\ \exp\{-\int_0^x -(\beta + y\delta/\mu)dy\} = \exp(x\beta + x^2\delta/(2\mu)), & x > 0 \end{cases}$$

ומכאן

$$(6.154) \quad m(x) = \begin{cases} \exp(-x\beta - x^2/2), & x \leq 0 \\ \exp(-x\beta - x^2\delta/(2\mu)), & x > 0 \end{cases}$$

והצפיפות (לאחר ביצוע הנירמול):

$$(6.155) \quad f(x) = \begin{cases} h(-\beta) \cdot w(\beta\sqrt{\frac{\mu}{\delta}}, \sqrt{\frac{\delta}{\mu}}) \exp(-x\beta - x^2/2), & x \leq 0 \\ h(-\beta) \cdot w(\beta\sqrt{\frac{\mu}{\delta}}, \sqrt{\frac{\delta}{\mu}}) \exp(-x\beta - x^2\delta/(2\mu)), & x > 0 \end{cases}$$

3. $\gamma > 0$:

נקודות החלוקה:

$$(6.156) \quad k = 1 : (s_0, s_1) = (-\infty, 0)$$

תוחלת ושונות אינפניטיסימליים:

$$(6.157) \quad \mu(x) = -(\beta + x) \quad , \quad \sigma^2(x) = 2$$

כעת נבחר $\theta = 0$ ונקבל

$$(6.158) \quad s(x) = \exp\left\{-\int_0^x -(\beta + y)dy\right\} = \exp(x\beta + x^2/2)$$

ומכאן

$$(6.159) \quad m(x) = \exp(-x\beta - x^2/2)$$

והצפיפות (לאחר ביצוע הנירמול):

$$(6.160) \quad f(x) = h(-\beta) \exp(-x\beta - x^2/2)$$

פרק 7

עיצוב מוקד שירות - דיון

לאחר הצגת הניתוחים האסימפטוטיים של מערכת $M(\lambda_N)/M(\mu)/N + M(\theta_N)$, ננסה בפרק זה להשליך ממגוון התוצאות שקיבלנו על סוגיית עיצוב מוקד שירות, כפי שתוארה במבוא. בחלק הראשון נבחן את טיב הקירובים שקיבלנו, כדי לזהות האם ומתי סביר להשתמש בהם. בחלק השני נשתמש במסקנות החלק הראשון, וננסה למצא כללי אצבע לעיצוב מוקד שירות טלפוני עם לקוחות קצרי רוח. לסיום הדיון, נציע בחלק השלישי מדד לרמת השירות במוקד טלפוני.

תחילה נחזור לדון בחשיבות השימוש במודל עם נטישות. כעת יש בידינו כלים לחישוב מדדים במודל עם נטישות, ונביא דוגמא אשר תמחיש את המשמעות האופרטיבית של נטישות בעת עיצוב מוקד שירות טלפוני.

נניח מוקד שירות טלפוני עם הפרמטרים הבאים:

- מספר מוקדנים $(N): 50$

- קצב מופע $(\lambda): 48$ לקוחות לדקה

- קצב שירות $(\mu): 1$ לדקה

- קצב נטישה $(\theta): 0.5$ לדקה

נשווה בין תוצאות המתקבלות מניתוח על פי מודל ללא נטישות $(M/M/N)$ עם פרמטרים אלה (והתעלמות מ- θ), לתוצאות מניתוח מודל עם נטישות $(M/M/N + M)$. למרות שעל פי המודל עם נטישות רק כ-3% מהלקוחות נוטשים לפני תחילת שירותם, נראה שיש הבדלים משמעותיים (לעיתים בסדרי גודל) בין התוצאות המתקבלות משני הניתוחים.

ללא נטישות, זמן ההמתנה הממוצע בתור הוא כ-21 שניות. לעומת זאת, על פי מודל עם נטישות זמן ההמתנה הממוצע בתור הוא רק כ-4 שניות. כדי להגיע למצב בו זמן ההמתנה הממוצע במערכת ללא גטישות הוא 4 שניות, יש צורך ב-54 – 53 מוקדנים. כלומר תוספת של 8% – 6% (בגלל שהמערכת בעומס מאד גבוה, יש רגישות גבוהה לשינוי במספר המוקדנים). גם ה"זנב" של התפלגות זמן ההמתנה בתור שונה: ללא נטישות מתקבל שכ-9.5% מהלקוחות ממתינים בתור יותר מדקה. עם נטישות, אותו אחוזון נמצא בערך ב-13.5 שניות. עד כאן ההשוואות מתייחסות לנקודת המבט של הלקוח, אך יש היבטים נוספים. מבחינת המוקדנים גודל רלוונטי הוא הנצילות (אחוז הזמן שעסוקים): ללא נטישות 96%, עם נטישות כ-93%. כלומר אין הבדל משמעותי במיוחד. ישנו גם היבט של התשתית הדרושה - למשל מספר הקוים התפוסים בממוצע: ללא נטישות כ-65, עם נטישות כ-51. כאן ההבדל משמעותי - תוספת של יותר מרבע במודל ללא נטישות.

אנו רואים ששני המודלים נותנים תמונה שונה למדי של המתרחש במערכת. הפעלת מוקד כך שלא יהיו בו נוטשים כלל, דורשת תוספת משמעותית של משאבים (מוקדנים). מאידך, אם מאפשרים אחוז קטן של נוטשים, העומס על המערכת יורד משמעותית. צריך לזכור שנטישה של לקוחות אמנם פוגעת בנגישות, אבל לא בהכרח ברמת השירות - לקוח שנוטש "משפר" את רמת השירות של הלקוחות האחרים. המסקנה היא שלעיתים נטישה של חלק מהלקוחות דווקא תשפר את רמת השירות במוקד, וצריך למצא את האיזון המתאים להעדפות הלקוחות. מבחינת מתכנן/מנהל המוקד, הוספת נטישות למערכת $M/M/N$ מוסיפה "מימד" למדדים של רמת השירות - "מימד נטישה". במקום להעריך את רמת השירות על סמך מדדים המבוססים על זמן המתנה בלבד, נוספים מדדים המתייחסים גם לנטישות. תוספת המימד מאפשרת למתכנן/מנהל לנתח את המערכת ולעצב את המוקד כך שיתקבל האיזון הרצוי בין שני האספקטים של רמת השירות - המתנה ונטישה. חשוב לפתח מודלים אנליטיים של מוקדים בהם קיימות נטישות, לקבלת ניתוחים מדויקים. כפי שראינו בדוגמא, לעיתים נדרש ניתוח עדין ולא ניתן להסתפק בניתוח "אינטואיטיבי".

המסקנה אליה הגענו, שלעיתים "מניעת" שירות מחלק מהלקוחות יכולה בסך הכל להעלות את רמת השירות, רלוונטית גם למרכיב אחר של המודל - קיבולת המערכת. בהתאם למקובל

היום במוקדי שירות טלפוניים, הנחנו בפרק 4 שקיבולת המערכת אינה מוגבלת ($B = \infty$). כפי שצויין שם, כלל לא ברור שמצב זה תואם את העדפות הלקוחות. במיוחד במערכות שפועלות בעומס גבוה, קיבולת אינסופית יכולה להיות בעייתית. במקרה זה יכול להיווצר תור מאד גדול, הלקוחות יאלצו להמתין זמן רב, או שהם יחליטו לנטוש. ברור שעבור לקוחות אלה השירות אינו טוב, ודווקא במערכת עם קיבולת סופית היה נמנע מצב זה - ברגע שהמערכת "מלאה", לקוחות חדשים נתקלים ב"תפוס" וכניסתם למערכת נמנעת. אמנם גם קבלת "תפוס" אינה נחשבת לשירות טוב, אך יתכן וזה פחות גרוע מהמתנה ממושכת בתור, שעלולה להסתיים בנטישה לפני קבלת השירות. כמו כן, קבלת "תפוס" זו העברה מיידית ואוטומטית של אינפורמציה (חלקית) על מצב המערכת ללקוח, ואינפורמציה זו מאפשרת לו להחליט כיצד לפעול, בהתאם להעדפותיו. אחוז הלקוחות הנתקלים ב"תפוס" הינו מימד נוסף של רמת השירות, ויש מקום לעבודה נוספת שתרחיב את הנעשה בעבודה זו למודלים עם קיבולת סופית.

על סמך ניתוח זה אפשר לנסות ולאפיין מוקד "אידיאלי": לקוחות ממתנינים יקבלו מידע על מצב המערכת (למשל, אומדנים לזמן עד תחילת שירותם או מקומם בתור). בכל רגע נתון לקוח יוכל לעזוב את המערכת, להשאיר מספר טלפון ושעה בה הוא מעוניין שיצרו עימו קשר. כעת, בנוסף למתן שירות ללקוחות במערכת, יש לנסות ולעמוד בלוח הזמנים שמוכתב על ידי בקשות הלקוחות שעזבו. ניתוח מערכת כזו הוא מורכב מאד, וברור שלא נכלל בעבודה זו.

7.1 בדיקת טיב הקירובים

מטרתנו היא להשתמש בתוצאות שקיבלנו כקירובים עבור מוקדי שירות טלפוניים המתאימים לתיאור הסכמטי באיור 0.1. תוצאות הניתוח הראו שאת התנהגות המודל בתחום HW ניתן לסווג לשלושה תחומים עיקריים על פי ערך הפרמטר γ . כלומר, נתונים לנו למעשה שלושה תת-מודלים: לקוחות מאד קצרי רוח ($\gamma > 0$), לקוחות מאד סבלניים ($\gamma < 0$) ותחום ביניים ($\gamma = 0$). ננסה לקבוע עבור איזה תחומי פרמטרים הקירובים שלנו "טובים", וכיצד לבחור תת-מודל "המתאים למציאות".

טבלה 7.1: התנהגות אסימפטוטית של ההסתברות להמתנה בתחום HW
Asymptotic behavior of the waiting probability in HW regime

	$\lim_{N \rightarrow \infty} P_N\{Wait > 0\}$
$\gamma < 0, \beta \leq 0$	1
$\gamma < 0, \beta > 0$	$\alpha(\beta)$
$\gamma = 0$	$w(-\beta, \sqrt{\frac{\mu}{\delta}})$
$\gamma > 0$	0

7.1.1 תלות במספר המוקדנים

נזכור שהקירובים שקיבלנו הם עבור $N \uparrow \infty$, לכן בשלב ראשון נבדוק האם סביר להשתמש בהם (בכולם או בחלקם) עבור מוקדי שירות טלפוניים בהם מספר המוקדנים הוא בסדר גודל של עשרות או מאות. נשתמש בתוצאות המופיעות בטבלה 6.1, אך נסתפק בהתנהגות האסימפטוטית מסדר ראשון. התוצאות הרלוונטיות להסתברות להמתנה משוחזרות כאן בטבלה 7.1. אין כרגע התייחסות להסתברות לנטישה היות ובתחום HW מתקיים לכל γ $\lim_{N \rightarrow \infty} P_N\{Ab\} = 0$, ולכן אין טעם להשוות בין תת-המודלים על סמך ההתנהגות מסדר ראשון של גודל זה. בהמשך נתייחס גם לקירובים שמתקבלים תוך שימוש בהתנהגות מסדר שני (קצב ההתכנסות) של ההסתברות לנטישה.

בטבלה 7.2 אנו בוחנים את מידת הדיוק של הקירובים מסדר ראשון להסתברות להמתנה (כפי שמופיעים בטבלה 7.1), עבור ערכים שונים של הפרמטרים β ו- γ . מטעמי נוחות נקבע $\delta = \mu/2$, לקבלת $\sqrt{\frac{\delta}{\mu}} = \sqrt{2}$. הדיוק נמדד על פי הסטיה המוחלטת של הקירוב מהערך המדויק, ובטבלה מצויין המספר המינימלי של מוקדנים כך שסטיה זו תהיה קטנה מ-0.05. למשל, כאשר $\beta = \gamma = -0.5$, רק עבור מוקדים עם לפחות 100 מוקדנים נקבל דיוק של 0.05. ככל שמספר המוקדנים המינימלי שנדרש קטן יותר, הקירוב יחשב טוב יותר. הערה: מספר המוקדנים נלקח מתוך הערכים $\{20, 50, 100, 200, 500, 1000, 2000, 5000, 10000\}$. במקרים בהם גם עבור 10,000 מוקדנים לא הושג הדיוק הנדרש מסומן בטבלה '+'.

טבלה 7.2: מספר מוקדנים מינימלי לקבלת דיוק של 0.05
Minimum number of servers to achieve 0.05 accuracy

$\beta \setminus \gamma$	-1	-0.5	-0.1	0	0.1	0.5	1	2	5
-2	20	20	20	20	+	+	+	5000	5000
-1	20	20	20	20	+	+	5000	1000	1000
-0.5	20	20	+	20	+	+	5000	500	500
-0.1	50	2000	+	20	+	+	2000	500	500
0	200	+	+	20	+	+	2000	500	500
0.1	50	5000	+	20	+	+	1000	500	500
0.5	20	50	+	20	+	+	500	100	100
1	20	20	20	20	+	10000	200	50	50
2	20	20	20	20	20	20	20	20	20

מעיון בטבלה 7.2 ניתן לראות שיש מקרים שמתאימים למוקדים בגודל הנדון, ומקרים אחרים שאינם מתאימים. באופן בולט כאשר $\gamma = 0$ מתקבלים הקירובים הטובים ביותר, וכאשר $\gamma \neq 0$ וקטן בערכו המוחלט, מתקבלים הקירובים הגרועים ביותר. לכאורה יש כאן "אי רציפות" של טיב הקירוב בערך הפרמטר γ , אך זו נובעת מהעובדה שעבור המקרים $\gamma = 0$ ו- $\gamma \neq 0$ אנו משתמשים בתת-מודלים שונים לקבלת הקירוב.

כעת ננתח את המופיע בטבלה 7.2 באופן יסודי יותר: המקרים שבהם לא התקבל קירוב טוב עבור מספר "סביר" של מוקדנים הם בעיקר כאשר ערכו המוחלט של הפרמטר γ קטן, אך שונה מאפס. נזכור שהתוצאות שקיבלנו עבור ההתכנסות החלשה של ההתפלגות הסטציונרית ושל התהליך הסטוכסטי של מספר הלקוחות במערכת הראו שכאשר $\gamma < 0$ מקבלים בגבול התפלגות/תהליך שמתאימים למערכת ללא נטישות (בה $\theta = 0$), ועבור $\gamma > 0$ ההתפלגות/תהליך הגבוליים מתאימים למערכת ללא תור (בה $\theta = \infty$). מכאן שכל שההתכנסות של θ איטית

יותר (ל-0 או ∞), נצפה שגם ההתכנסות להתפלגות/תהליך הגבוליים תהיה איטית יותר, וכנ"ל לגבי המדדים. המסקנה היא שהשימוש בתת-מודל עם $\gamma \neq 0$ אינו "מוצלח" מבחינתנו כאשר הפרמטר γ קטן בערכו-המוחלט.

תופעה נוספת שניתן לזהות בטבלה 7.2 היא תלות בערך הפרמטר β . כאן צריך להפריד בין המקרים בהם $\gamma > 0$ לאלה בהם $\gamma < 0$:

- כאשר $\gamma > 0$ ההתנהגות היא מונוטונית ב- β . התנהגות זו ניתן להסביר על פי אופי המערכת המתאימה לתהליך הגבולי שמתקבל במקרה זה: בגבול מקבלים תהליך דיפוזיה שמתאים למערכת ללא תור. ככל שהערך של β גדול יותר, העומס המוצע למוקדן קטן יותר, והמערכת המקורית תמצא פחות במצב בו יש תור, כלומר היא דומה יותר בהתנהגותה למערכת $M/M/N/N$. לכן הקירוב טוב יותר עבור ערכי β גדולים. שים לב לכך שבתחום זה הרגישות לערך הפרמטר β גבוהה מהרגישות לערך של γ .

- במקרה בו $\gamma < 0$, המערכת שמתאימה לתהליך הגבולי היא מערכת ללא נטישות. הקירוב שנלקח תלוי בסימן של β (ראה בטבלה 7.1). כאשר $\beta \leq 0$ הקירוב שנלקח הוא של מערכת ש"מתפוצצת" - זו מערכת ללא נטישות עם עומס מוצע למוקדן גדול או שווה 1. כאן ככל שהערך של β קטן יותר העומס גדול יותר, וה"פיצוץ" מהיר יותר. לכן בתחום זה הקירוב משתפר ככל שהערך של β קטן. לעומת זאת כאשר $\beta > 0$, הקירוב הוא של מערכת שלא "מתפוצצת". בדומה לטיעון שניתן עבור $\gamma > 0$, כאשר העומס המוצע הוא נמוך, המערכת המקורית דומה יותר למערכת ללא נטישות, אז רק חלק קטן מהלקוחות נאלצים להמתין, ועוד פחות נוטשים. לכן בתחום זה טיב הקירוב מונוטוני עולה ב- β .

נחזור לנקודה הראשונה שצינו, והיא שהקירובים שמתקבלים עבור $\gamma = 0$, טובים במיוחד. הסיבה העיקרית לכך היא שהקירוב שנלקח במקרה זה תלוי גם ב- β וגם ב- δ (ראה טבלה 7.1). כלומר יש תלות בשני פרמטרים, בעוד ערכי הקירובים האחרים נקבעים לכל היותר על ידי פרמטר אחד. התרשמות נוספת מטיב הקירוב שמתקבל מתת-מודל $\gamma = 0$ ניתן לקבל מעיון בטבלה 7.3. בטבלה זו מצוין מספר המוקדנים המינימלי לקבלת סטיה מוחלטת שקטנה מ-0.01. ניתן לראות שגם דיוק כזה בר השגה במוקדים בגודל הנדון. הערה: כאן מספר המוקדנים נלקח מתוך הערכים {10, 20, 30, 50, 75, 100, 150, 200, 250}.

טבלה 7.3: $\gamma = 0$ - מספר מוקדנים מינמלי לקבלת דיוק של 0.01
 $\gamma = 0$ - Minimum number of servers to achieve 0.01 accuracy

β	
-2	10
-1	10
-0.5	30
-0.1	150
0	150
0.1	150
0.5	75
1	20
2	50

שים לב שבטבלה 7.3 אין מונוטוניות בפרמטר β . אין לנו הסבר לכך - כנראה שתופעה זו מתבטאת רק בהתנהגות מסדר שני (קצב התכנסות).

7.1.2 יישום

בבואנו לנתח מוקד שירות טלפוני, ניתן לאמוד מתוך נתונים סטטיסטיים את קצבי השירות, המופע, והנטישה. אנו מניחים שקצב השירות אינו משתנה, לכן סביר לאמוד אותו על פי קצב השירות הממוצע. כפי שנדון בתחילת פרק 3, בדרך כלל קצב המופע אינו קבוע, לכן נצטמצם לפרק זמן בו הקצב בקירוב קבוע, ונאמוד אותו על פי הממוצע בפרק זמן זה. את קצב הנטישה אפשר לאמוד במספר דרכים. דרך אחת היא על פי הקשר $P\{Ab\} = \theta \cdot E[Wait]$, כפי שמוסבר בהערות שאחרי (1.26). דרך שניה היא בעזרת אמד (נראות מירבית) מקובל לסבלנות הממוצעת של לקוח:

$$(7.1) \quad \frac{\hat{1}}{\theta} = \frac{\left(\begin{smallmatrix} No. \\ served \end{smallmatrix} \right) \times \left(\begin{smallmatrix} Average\ wait \\ of\ served \end{smallmatrix} \right) + \left(\begin{smallmatrix} No. \\ abandoning \end{smallmatrix} \right) \times \left(\begin{smallmatrix} Average\ time \\ to\ abandon \end{smallmatrix} \right)}{No. \text{ abandoning}}$$

כדי להשתמש במודל האנליטי שנותח בפרק 6, יש לקבוע גם את ערכי הפרמטרים β , γ , δ , ויש להניח שהמוקד פועל בתחום HW . מתי סביר להניח זאת? באופן איכותי מוקד ש"פועל

בתחום HW הוא קודם כל מוקד שעובד בעומס גבוה (שקול ליעילות גבוהה). בנוסף, במוקד כזה אחוז הלקוחות הממתנינים בתור לפני שירותם אינו זניח, אך גם אינו "גורף", ואחוז הנוטשים קטן. אם אכן ניתן להניח שהמוקד פועל בתחום HW , ניתן לקבוע ערך לפרמטר β על פי

$$(7.2) \quad \beta = \frac{N\mu - \lambda}{\sqrt{N\mu}}.$$

נותר למצא את ערך הפרמטר γ (או את δ , שהרי בהנתן N ו- θ הם קובעים זה את זה באופן חד-חד-ערכי דרך $\theta = \delta N^\gamma$).

נטפל בבעיה קלה יותר - במקום לנסות לקבוע את הערך המדויק של γ , ננסה לזהות באיזה משלושת התחומים ($\gamma > 0$, $\gamma = 0$, $\gamma < 0$) הוא נמצא. אם קצב הנטישה מאד גבוה, סביר לקבוע $\gamma > 0$; אם הנטישה מאד איטית נקבע $\gamma < 0$, ובמקרי הביניים $\gamma = 0$. הבעיה היא שלא ברור כמה קיצוני צריך להיות הערך של קצב הנטישה כדי להצדיק את השימוש בתת-המודלים בהם $\gamma \neq 0$. לבדיקת הנושא בחרנו גדלים אופייניים של מוקד שירות טלפוני גדול בעומס גבוה:

• מספר מוקדנים (N): 100

• קצב מופע (λ): 95 לקוחות לדקה

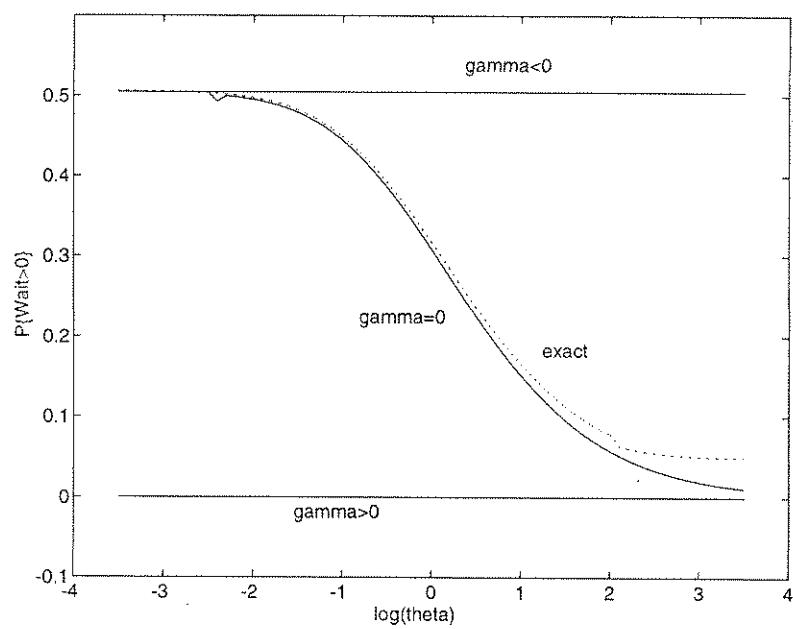
• קצב שירות (μ): 1 לדקה

מכאן על פי (7.2) נקבל $\beta = (100 - 95)/10 = 0.5$.

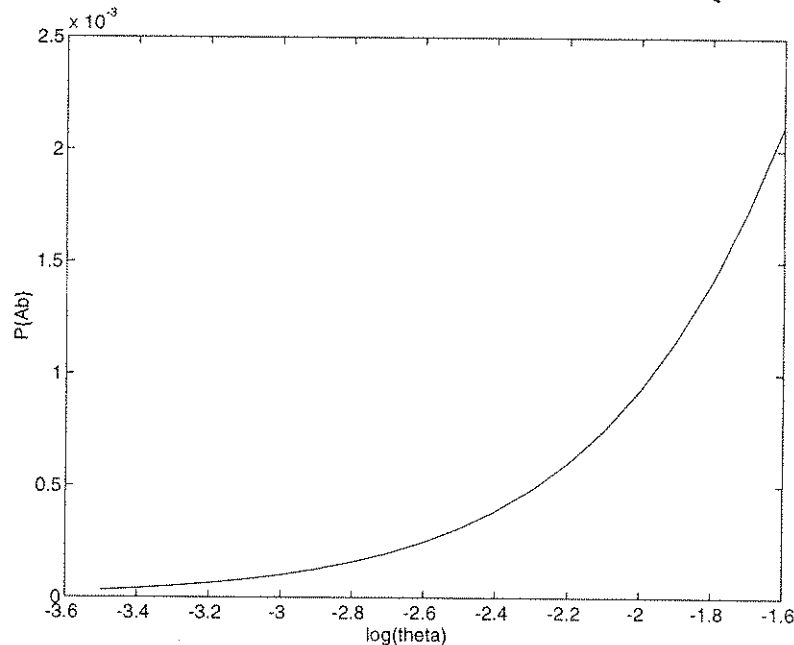
כעת, עבור טווח ערכים גדול של קצב הנטישה בדקנו את ההתאמה של הקירובים להסתברות להתנתה (על פי טבלה 7.1), לערך המדויק. את התוצאות ניתן לראות באיור 7.1. הערה: שים לב שבסביבות הערך $\log(\theta) = -2.5$ מופיעה "בליטה" בגרף של תת-המודל $\gamma = 0$. בליטה זו היא כתוצאה מאי דיוק נומרי הנובע מאופן החישוב - בתחום $\log(\theta) < -2.4$ נעשה שימוש בחישוב מקורב.

המסקנות הן חד משמעיות, וממשיכות את הנאמר בעיקבות הבדיקות ב-7.1.1: תת-המודל $\gamma = 0$ הוא המוצלח ביותר. ברוב התחום שנבדק, הקירוב שמתקבל מתת-מודל זה, טוב בהרבה משני הקירובים האחרים. גם בערכים הקיצוניים, בהם מתקבל שהקירובים מתת-המודלים האחרים טובים גם כן (כל אחד בתחום המתאים), לא נראה שיש סיבה להעדיף אותם על פני תת-המודל $\gamma = 0$. סיבה נוספת לאמץ את תת-המודל $\gamma = 0$ היא שעבורו ניתן לקבוע את

איור 7.1: $P\{Wait > 0\}$ במערכת $M(95)/M(1)/100 + M$ כתלות ב- θ
 Dependence of $P\{Wait > 0\}$ on θ in an $M(95)/M(1)/100 + M$ system



איור 7.2: $P\{Ab\}$ במערכת $M(95)/M(1)/100 + M$ עם קצבי נטישה "איטיים"
 $P\{Ab\}$ in an $M(95)/M(1)/100 + M$ system with "slow" abandonment rates



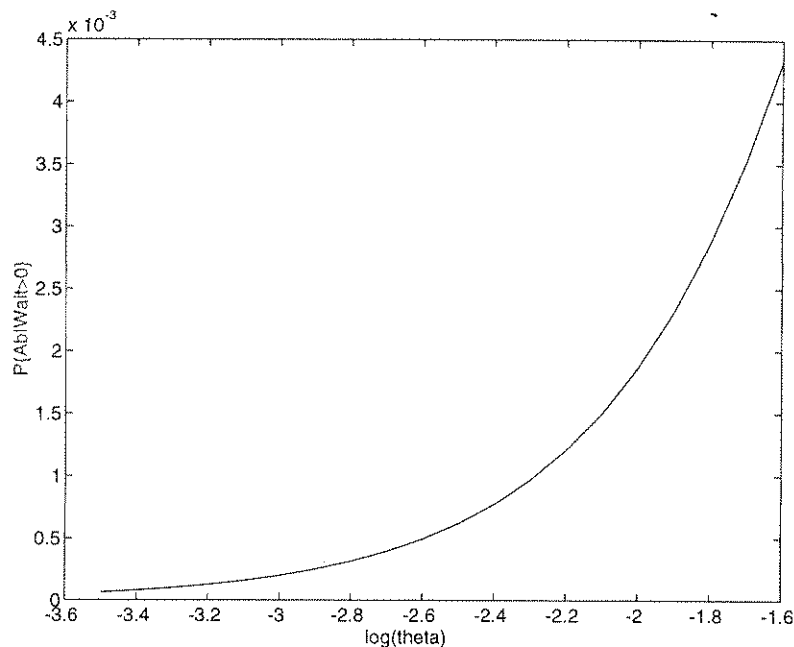
כל הפרמטרים. נזכור שבמקרים האחרים עדין נותר לקבוע את ערכו המדויק של הפרמטר γ , בלעדיו לא ניתן להשתמש בקירובים העדינים יותר שמופיעים בטבלה 6.1.

כפי שניתן לראות מהגבולות החלשים שהתקבלו בפרק 6, במקרים בהם $\gamma \neq 0$ "מאבדים" בעת לקיחת הגבול $N \uparrow \infty$ חלק ממאפייני המערכת - כאשר $\gamma < 0$ מאבדים את הנטישות, וכאשר $\gamma > 0$ מאבדים את התור. כתוצאה מכך הקירובים (מסדר ראשון) שמתקבלים במקרים אלה פחות טובים בדרך כלל.

7.1.3 קצבי נטישה "קיצוניים"

לפני שנתמקד בתת-המודל בו $\gamma = 0$, נבחן בצורה מעמיקה יותר את תת-המודלים האחרים, במקרים בהם קצב הנטישה הוא "קיצוני" מספיק כדי להצדיק את השימוש בקירובים שהם נותנים. נדון בכל מקרה בנפרד:

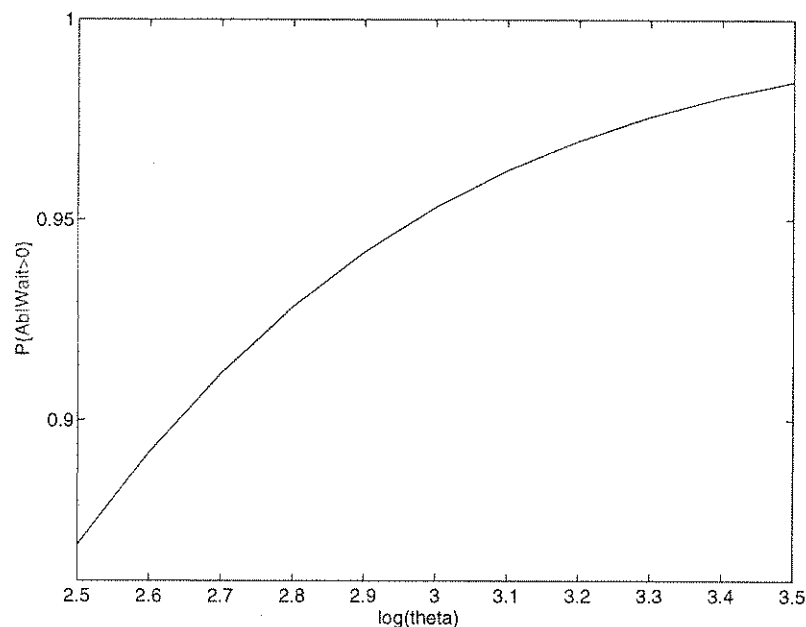
איור 7.3: $P\{Ab|Wait > 0\}$ במערכת $M(95)/M(1)/100 + M$ עם קצבי נטישה "איטיים"
 $P\{Ab|Wait > 0\}$ in an $M(95)/M(1)/100 + M$ system with "slow" abandonment rates



• $\gamma < 0$: באיור 7.2 מופיעה ההסתברות לנטישה עבור קצבי נטישה "איטיים". מתוך טווח הערכים של קצב הנטישה שהופיע באיור 7.1, אנחנו מתמקדים בתחום בו $\log(\theta) < -1.5$. בתחום זה הקירוב להסתברות להמתנה שמתקבל מתת-המודל $\gamma < 0$ הוא סביר. הערכים שמתקבלים מאד קטנים (פחות מ- 3×10^{-3} - שלושה מאלף נוטשים), כלומר, דומה מאד למתקבל במודל שמזניח את תופעת הנטישות ($M/M/N$).

את הדמיון למערכת $M/M/N$ אפשר לבחון בצורה יותר טובה בעזרת הגודל $P\{Ab|Wait > 0\}$ (אחוז הנוטשים מבין אלה שנאלצו להמתין בתור). אחוז נמוך של נוטשים יכול להתקבל גם במוקד בו הסיכוי של לקוח להמתין בתור הוא נמוך. לכן, כדי לאשש את ההנחה שללקוחות יש בקירוב סבלנות בלתי מוגבלת, עדיף לבחון רק את הלקוחות שהגיעו למצב של המתנה בתור, ולראות איזה אחוז מהם נוטשים. במודל ללא נטישות גודל זה הוא זהותית אפס. באיור 7.3 רואים שעבור תת-המודל $\gamma < 0$ גודל זה עדין קטן מאד - כ-4 נוטשים מ-1000 בתחום הרלוונטי.

איור 7.4: $P\{Ab|Wait > 0\}$ במערכת $M(95)/M(1)/100 + M$ עם קצבי נטישה "מהירים"
 $P\{Ab|Wait > 0\}$ in an $M(95)/M(1)/100 + M$ system with "fast" abandonment rates



• $\gamma > 0$: מתוך איור 7.1 ניתן לראות שהקירוב להסתברות להמתנה שמתקבל מתת-מודל זה הוא סביר בתחום בו $\log(\theta) > 2.5$, באיור 7.4 ניתן לראות שבתחום זה רוב הלקוחות שנאלצים להמתין בתור, נוטשים. כלומר, מצב דומה למתקבל במודל ללא תור $(M/M/N/N)$ - כאשר אין תור, כל לקוח שנאלץ "להמתין בתור" נוטש מיד.

7.2 כללי אצבע

בחלק זה ננסה לנסח כללי אצבע לעיצוב מוקד שירות טלפוני עם לקוחות קצרי רוח. על פי תוצאות ומסקנות החלק הקודם, תת-המודל $\gamma = 0$ הוא המוצלח ביותר, ולא ברור באם קיימים מקרים בהם תת-המודלים האחרים עדיפים עליו. כמו כן, במצבים בהם קצבי הנטישה "קיצוניים", אם מוותרים על השימוש בתת-המודל $\gamma = 0$, נראה סביר להשתמש במודלים ללא נטישות כלל - $M/M/N$ או $M/M/N/N$ - כל אחד בתחום המתאים. עבור מודלים אלה קיימות תוצאות רבות (למשל [19] ו-[24]). לכן תת-המודל עליו נתבסס לקביעת כללי אצבע,

יהיה אותו מודל שהוצג על ידי [16] Fleming, Stolyar and Simon, בו קצב הנטישה אינו תלוי במספר המוקדנים.

האם סביר להניח שקצב הנטישה אינו תלוי במספר המוקדנים? נראה סביר שבמוקדי שירות קיים קשר בין קצב המופע למספר המוקדנים - אם קצב המופע גדל נצפה שמספר המוקדנים יוגדל כדי שהמערכת תצליח לעמוד בעומס הנוסף. לעומת זאת הנחנו שקצב השירות של מוקדן אינו תלוי במספר המוקדנים שעובדים איתו במקביל. על פי תכונת "התורים הנסתרים" של מוקד שירות טלפוני, הגיוני לטעון שהלקוח אינו יודע כמה מוקדנים עובדים במוקד בזמן שהוא מתקשר, ולכן סבלנותו לא יכולה להיות תלויה בגודל זה. מאידך, במקרים רבים הלקוח יודע באופן "איכותי" האם המוקד אליו הוא מתקשר הוא "גדול" (למשל בזק 144) או "קטן" (למשל סניף מקומי של חברה לאספקת גז), ויתכן שסבלנותו תושפע מכך. אין לנו תשובה מוחלטת לסוגיה זו, אך נראה שההנחה שהסבלנות לא תלויה במספר המוקדים היא טבעית ואינה בלתי סבירה.

למעשה היינו מצפים שקצב הנטישה יהיה תלוי בקצב השירות - משך הזמן שלקוח מוכן להמתין עד לתחילת שירותו קשור במשך השירות הצפוי לו. מכיוון שאנו מניחים קצב שירות קבוע, גם מבחינה זו סביר שנניח קצב נטישה קבוע.

ננסה לנסח כללי אצבע למתכנן/מנהל המוקד. תחת מודל $M/M/N+M$, אנו מניחים שהפרמטר היחיד עליו שולט המתכנן/מנהל הוא מספר המוקדנים שמאיישים את המוקד. תחילה נראה שתחת תת-המודל $\gamma = 0$, ההנחה שהמוקד פועל בתחום HW (כלומר $\lambda_N \sim N\mu + O(\sqrt{N})$, μ קבוע) מקבלת תוקף של "כלל אצבע" במובן שזוהו התחום שסביר לנסות לפעול בו לקבלת מוקד ביעילות גבוהה ורמת שירות שאינה קיצונית (לטוב או לרע). על פי המונטוניות של ההסתברות להמתנה בקצב המופע (ניתנת להסקה באופן ישיר מתוך (1.9), על פי ההתנהגות של ההתפלגות הסטציונרית), מתקבלות התוצאות הבאות:

1. כאשר $\lambda_N \sim N\mu + bN^a$, $b > 0$, $a > 0.5$,

$$(7.3) \quad \lim_{N \rightarrow \infty} P_N\{\text{Wait} > 0\} \geq \lim_{\beta \rightarrow \infty} w(-\beta, \sqrt{\frac{\mu}{\delta}}) = 1.$$

2. כאשר $\lambda_N \sim N\mu + bN^a$, $b < 0$, $a > 0.5$

$$(7.4) \quad \lim_{N \rightarrow \infty} P_N\{Wait > 0\} \leq \lim_{\beta \rightarrow -\infty} w(-\beta, \sqrt{\frac{\mu}{\delta}}) = 0.$$

נזכיר את ההגדרה של הפונקציה w מ-(6.12): $w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}$. המסקנה בשני המקרים $\rho_\infty = 1$, לכן על פי משפט 5.1 מתקיים $\lim_{N \rightarrow \infty} P_N\{Ab\} = 0$. המסקנה היא שבתחומים "מחוץ" לתחום HW , בגבול $N \uparrow \infty$ ההסתברות להמתנה מנוונת (כלומר אפס או אחד), וכמו שמופיע במאמר של [19] Halfin and Whitt ביחס למערכות ללא נטישות, התחום ה"מעניין" הוא תחום HW . תחום זה מתאים למצב בו רוצים שפרופורציית הממתנים תהיה ממש בין 0 ל-1, כלומר נסיון למצא איזון בין היעילות לרמת השירות, שאפשר לכנותו תכנון/ניהול "רציונלי". מספר המוקדנים במקרה זה, נקבע על ידי

$$(7.5) \quad N \approx \frac{\lambda}{\mu} + \hat{\beta} \sqrt{\frac{\lambda}{\mu}}.$$

הערה: הצגה זו של הקשר בין מספר המוקדנים לקצב המופע בתחום HW שקולה ל-(7.2) כאשר $\hat{\beta} = \beta \sqrt{\mu}$.

יישום אפשרי, כפי שמסביר [39] Whitt, מתייחס למשל לתרחיש שקצב המופע למערכת נתונה צפוי להשתנות, ורוצים לקבוע את רמת האיוש הנחוצה כדי לשמור על אותה רמת שירות. הפרמטר β מהווה אינדקס לרמת השירות במובן שהוא מבטא את העודף/חוסר במוקדנים ביחס לכמות העבודה אשר מגיעה למערכת. עבור קצב מופע קבוע, ככל שהערך של β גדול יותר המשמעות היא שיש יותר מוקדנים ורמת השירות עולה ולא ניתן לתת אינטרפרטציה יותר מדויקת מזו. לכן כדי לשמור על β קבוע, האיוש הדרוש מתקבל מ-(7.5).

בתחום HW , בהנתן $\delta \equiv \mu, \theta$, אפשר לקבל קירובים למדדים שונים בהתבסס על התוצאות מפרק 6.1 (ראה טבלה 6.1) וקשרים שהופיעו בפרק 1, למשל:

$$(7.6) \quad P\{Wait > 0\} \approx w(-\beta, \sqrt{\frac{\mu}{\theta}})$$

$$(7.7) \quad P\{Ab\} \approx \left[\sqrt{\frac{\theta}{\mu}} \cdot h\left(\beta \sqrt{\frac{\mu}{\theta}}\right) - \beta\right] \cdot w(-\beta, \sqrt{\frac{\mu}{\theta}}) \cdot \lambda^{-0.5}$$

$$(7.8) \quad P\{Ab|Wait > 0\} \approx \left[\sqrt{\frac{\theta}{\mu}} \cdot h\left(\beta \sqrt{\frac{\mu}{\theta}}\right) - \beta\right] \cdot \lambda^{-0.5}$$

$$(7.9) \quad E[Wait] \approx \left[\sqrt{\frac{\theta}{\mu}} \cdot h\left(\beta \sqrt{\frac{\mu}{\theta}}\right) - \beta \right] \cdot w\left(-\beta, \sqrt{\frac{\mu}{\theta}}\right) \cdot \lambda^{-0.5} / \theta$$

$$(7.10) \quad E[Wait|Wait > 0] \approx \left[\sqrt{\frac{\theta}{\mu}} \cdot h\left(\beta \sqrt{\frac{\mu}{\theta}}\right) - \beta \right] \cdot \lambda^{-0.5} / \theta$$

בדומה למדיניות תכנון/ניהול "רצינות" המתאימה לתחום HW , ניתן להגדיר תחומים נוספים המתאימים למוקדים בעומס גבוה, בהם ה"איוון" בין רמת השירות ליעילות שונה:

• מכון "שירות": כאן יש "העדפה" של רמת השירות על פני היעילות.

$$(7.11) \quad N \approx \frac{\lambda}{\mu} + \beta_1 \frac{\lambda}{\mu}, \quad \beta_1 > 0.$$

• מכון "יעילות": במקרה זה מצמצמים את כוח האדם על חשבון רמת השירות.

$$(7.12) \quad N \approx \frac{\lambda}{\mu} + \beta_2.$$

כאשר הפרמטרים β_1 ו- β_2 הם אינדקסים של רמת השירות בשני המקרים. בעבודה של Borst, [7] Mandelbaum and Reiman ניתן ביסוס נוסף להפרדה זו לתחומים, ומראים שם כיצד בחירת תחום נובעת משיקולי עלויות.

הערה: שים לב שמתוך הקשר $N \approx \frac{\lambda}{\mu} + b\lambda^a$ בחרנו את המקרים $a = 0$ ו- $a = 1$ עבור מדיניות מכוונת "שירות" ו-"יעילות" בהתאמה (בנוסף ל- $a = 0.5$ עבור "רצינות"). למעשה בהתאם לניתוח שלנו כל בחירה של $0.5 < a \leq 1$ תתאים למכון "שירות", ו- $0 \leq a < 0.5$ למכון "יעילות".

7.3 מדד לרמת שירות

ראינו שהוספת נטישות למערכת $M/M/N$, מוסיפה "מימד" לרמת השירות, ומתקבל עושר של מדדים. אך כדי שהשימוש בכללי האצבע יהיה פשוט ונוח, רצוי להגדיר את רמת השירות על ידי מדד יחיד. ננסה להגדיר מדד, אשר ישקף בצורה טובה את רמת השירות במוקד. ישנן שתי אוכלוסיות אשר "מדאיגות" את מתכנן/מנהל המוקד, וגודלן היחסי (ביחס לכלל הלקוחות) יכול להיות מדד לרמת השירות:

1. לקוחות שקיבלו שירות, אך המתינו זמן רב בתור.

2. לקוחות שנטשו ללא קבלת שירות, אך רק כאלה שהמתינו לפחות פרק זמן מינימלי בתור, ולא נטשו מיד כשהבינו שיהיה עליהם להמתין.

מתכנן/מנהל המוקד שואף למצב בו שתי האוכלוסיות הנ"ל קטנות.

מכאן אפשר להגדיר מדד לרמת השירות, SL , על ידי

$$(7.13) \quad SL = aP\{Wait > T_1 ; \neg Ab\} + (1 - a)P\{Wait > T_2 ; Ab\}, \quad 0 \leq a \leq 1.$$

כלומר ממוצע משוקלל של פרופורציות שתי האוכלוסיות הנ"ל, כאשר המשקולות נקבעות על ידי מתכנן/מנהל המוקד, בהתאם לחשיבות שהוא מייחס לכל אוכלוסיה. בעזרת הפרמטרים T_1 ו- T_2 מוגדרות במדויק שתי האוכלוסיות שהוגדרו לעיל באופן איכותי. על פי המופיע בסוף פרק 3 (בהערות על המקרה $B = \infty$) ניתן לקבל ביטויים מדויקים עבור המחברים ב-(7.13):

$$(7.14) \quad P\{Wait > T_1 ; \neg Ab\} = c\pi_N \cdot e^{\frac{\lambda}{\theta}} \left(\frac{\theta}{\lambda}\right)^{c+1} \gamma(c+1, \frac{\lambda}{\theta} e^{-\theta T_1}),$$

$$(7.15) \quad P\{Wait > T_2 ; Ab\} = c\pi_N \cdot e^{\frac{\lambda}{\theta}} \left(\frac{\theta}{\lambda}\right)^c \left[e^{-\theta T_2} \gamma(c, \frac{\lambda}{\theta} e^{-\theta T_2}) - \frac{\theta}{\lambda} \gamma(c+1, \frac{\lambda}{\theta} e^{-\theta T_2}) \right],$$

כאן $c = \frac{N\mu}{\theta}$.

בעבודה זו לא הצלחנו למצא קירובים טובים למדדים דוגמת (7.13), במובן שלא הושג השילוב בין פשטות ודיוק. את ההתנהגות האסימפטוטית של ה-Incomplete Gamma Function ניתן למצא בספרות (ראה למשל [14]), אך נדרשת עבודה נוספת מעבר לכך.

אם נסתפק במדד פחות "עדין", בו נקבע $T_1 = T_2 = 0$ (כלומר ההתייחסות היא לכל הלקוחות שהמתינו בתור לפני קבלת שירות, ולכל הלקוחות שנטשו), נקבל על פי הקירובים (7.6)-(7.10)

$$\begin{aligned} SL &= a(P\{Wait > 0\} - P\{Wait > 0 ; Ab\}) + (1 - a)P\{Wait > 0 ; Ab\} \\ &= aP\{Wait > 0\} + (1 - 2a)P\{Ab\} \\ (7.16) \quad &\approx w(-\beta, \sqrt{\frac{\mu}{\theta}}) \cdot \left(a + (1 - 2a) \left[\sqrt{\frac{\theta}{\mu}} \cdot h(\beta \sqrt{\frac{\mu}{\theta}}) - \beta \right] \cdot \lambda^{-0.5} \right). \end{aligned}$$

יתרוננו של מדד זה בכך שהוא פשוט למדידה/מעקב.

סיכום / מחקר עתידי

בעבודה זו עסקנו בניתוח של מוקד שירות טלפוני עם לקוחות קצרי רוח. המטרה היתה לספק כלים מדויקים וכללים (חלקם מבוססים על קירובים), שיסייעו למתכנן/מנהל המוקד להשיג את האיזון הרצוי בין יעילות לרמת שירות. התוצאות שקיבלנו ממחישות את הצורך להשתמש במודלים הכוללים את תופעת הנטישות בעת עיצוב מוקדי שירות טלפוניים. יש מקום לעבודות נוספות בתחום זה, כפי שיפורט בהמשך.

מהמודל הסכמטי שהוצג במבוא (איור 0.1) קיבלנו מודל אנליטי בסיסי, לאחר הזנחת תופעת החיגים החוזרים, ובתוספת מספר הנחות מקובלות בתורת התורים ורחבות מספיק להיות עדן שימושיות. מודל זה סומן ב- $M/M/N/B + M$. הצגנו דרכים לחישוב מדויק של מגוון מדדים במודל $M/M/N/B + M$ במצב יציב, עם דגש על המדד "פרופורציית הלקוחות שנוטשים לפני תחילת שירותם" (= ההסתברות לנטישה).

חלק מרכזי בעבודה עסק במציאת קירובים. בהנחת התוצאות המדויקות, מטרת הקירובים היתה לקבל קשרים פשוטים יותר בין הפרמטרים של המודל למדדים, כדי לספק "insight". לשם כך נדרשנו למספר הנחות נוספות שהובילו למודל אנליטי פשוט יותר. הנחות אלה התבססו בעיקר על מאפיינים כיום של מוקדי שירות טלפוני גדולים:

- מספר הקווים בדרך כלל גדול $B = \infty$.
- מספר המוקדנים גדול יחסית (עשרות-מאות) $\Leftarrow$ ניתוח עבור $N \uparrow \infty$.
- ניתן ורצוי לעבוד בעומס גבוה $\Leftarrow 1 \uparrow \frac{\lambda N}{N\mu}$ כאשר $N \uparrow \infty$.
- מעוניינים בתחום בו העומס גבוה, ואחוז הממתנינים/נוטשים קטן אך לא זניח $\Leftarrow HW$ תחום.

בצענו ניתוחים אסימפטוטיים של מודל זה בכמה רמות - מדדים ספציפיים, התהליך הסטוכסטי של מספר הלקוחות במערכת, וההתפלגות הסטציונרית של תהליך זה. ניתוחים אלה הניבו קירובים, ונמצא שהקירובים הטובים ביותר מתקבלים כאשר מניחים קצב נטישה קבוע. קירובים אלה הם בעלי דיוק טוב כבר עבור מוקדים עם עשרים מוקדנים ויותר. כללי האצבע שקיבלנו דומים לאלה שקיבל [39] Whitt עבור מערכות $M/M/N$. לעומת זאת, הוספת תופעת הנטישות מוסיפה "מימד" לרמת השירות, ומתקבל מגוון רחב של מדדים לרמת השירות. ניסינו להגדיר מדד אחד שיכלול את שתי התופעות העיקריות שמשפיעות על רמת השירות - המתנה ונטישה.

יש מקום להמשך עבודה במספר כיוונים. ראשית אפשר להעמיק ולחקור בתחומים בהם התמקדנו:

- מציאת סכימת קירוב כללית למדדים מדויקים דוגמת אלו שהוצגו בפרק 3.
 - שימוש בגבולות הדיפוזיה שהתקבלו עבור קירובים להתנהגויות טרנזיאנטיות (כלומר למצבים בהם המערכת לא נמצאת במצב יציב).
 - "שיכלול" כללי האצבע (על סמך הקירובים החדשים).
- חלק מהקשיים בהם נתקלנו נבעו מכך שניתחנו את התהליך הסטוכסטי של מספר הלקוחות במערכת, בעוד אנו מעוניינים למצא קירובים למדדים שקשורים בזמני המתנה בתור. נסיון לנתח את התהליך הסטוכסטי של כמות העבודה במערכת (מציאת התפלגות סטציונרית, התכנסות חלשה וכדומה) יכול להיות מועיל.

באופן טבעי, יש עניין רב לבצע מחקר ברוח דומה עבור מודלים בהם מוסרות חלק מההנחות המקלות שהנחנו כאן:

• מספר קווים סופי ($B < \infty$).

• מודל עם חיוגים חוזרים.

ניתוח מודלים אלה עשוי להיות מורכב למדי. יתכן וכדאי להתחיל מניתוח מודלים המתקבלים מהוספת כל אחת מתכונות אלו, בנפרד, למערכת $M/M/N$. אין ספק שניתוח של מודל עם נטישות, קיבולת סופית וחיוגים חוזרים הוא בעל ערך רב. בעבודה של Mandelbaum, Massey

[25] and Reiman יש טיפול במודל מסוג זה (בתוספת מדיניות שירות עם עדיפויות ומוקד במבנה של רשת), תחת הנחות מסוימות.

כמו כן, אפשר לנתח מודלים בהם מוסרות חלק מההנחות שעמדו בבסיס המודל האנליטי כגון:

- מופע הלקוחות לא פואסוני/לא הומוגני בזמן.
- זמני שירות לא מעריכיים.
- סבלנות הלקוחות לא מעריכית.
- מבנה המוקד הוא רשת של מוקדנים.
- מדיניות שירות עם עדיפויות.

השירות הטלפוני נמצא בהתפתחות מתמדת, ולסיום נציין שני מודלים חדשים יחסית בהם יש עניין לשלב נטישות:

- Skill-based routing - מודל בו כל מוקדן מאופיין על פי סדרת כישורים, ויש למצא "שידוך אופטימלי" בין לקוח למוקדן, בהתאם לשירות הנדרש ומצב התעסוקה של המוקדנים. שלב ה"קליטה" של לקוח וזיהוי מבוקשו יכול להיות ממומש במענה קולי.
- מודלים בהם המוקד "יוזם" שיחות ללקוחות (למשל לביצוע מכירות). במודלים אלה יכול גם להיות שילוב - טיפול בשיחות "נכנסות" (inbound), ומנגנון אוטומטי שיוזם שיחות "יוצאות" (outbound) בהתאם לעומס במוקד.

רשימת מקורות

- [1] Baccelli, F., Boyer, P., Hebuterne, G. (1984). Single-server queues with impatient customers. *Adv. Appl. Prob.* **16**, 887-905.
- [2] Baccelli, F., Hebuterne, G. (1981). On queues with impatient customers. In: *Performance '81*, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam) pp. 159-179.
- [3] Barrer, D.Y. (1957). Queueing with impatient customers and ordered service. *Operations Research* **5**, 650-656.
- [4] Bhattacharya, P.P., Ephremides, A. (1991). Stochastic monotonicity properties of multiserver queues with impatient customers. *J. Appl. Probab.* **28**, 673-682.
- [5] Bitran, G.R., Leong, T.Y., (1995). *The Service Sector in the United States*. MIT, Sloan School.
- [6] Blumenthal, R.M., Gettoor, R.K., (1968). *Markov Processes and Potential Theory*. Academic Press.
- [7] Borst, S., Mandelbaum, A., Reiman, M., (1998). Dimensioning of large call centers. *Preprint*.
- [8] Boxma, O.J., de Waal, P.R. (1994). Multiserver queues with impatient customers. *ITC* **14**,
- [9] Brigandi, A.J., Dargon, D.R., Sheehan, M.J., Spencer, T. III, (1994). AT&T's call processing simulator (CAPS) operational design for inbound call centers. *INTERFACES* **24**, 6-28.
- [10] Browne, S., Whitt, W., (1995). Piecewise-linear diffusion processes. *Probability and Stochastic Series: Advances in Queueing. Theory, Methods, and Open Problems*. 463-480. Ed. J.H. Dshalalow, CRC Press.

- [11] Chen, H., Mandelbaum, A., (1991). Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *The Annals of Probability*, Vol. **19** , No. **4** , 1463-1519.
- [12] Coffman, E.G. Jr., Puhalskii, A.A., Reiman, M.I., Wright, P.E., (1994). Processor-shared buffers with reneging. *Performance Evaluation* **19** 25-46.
- [13] Daley, D.J., (1965). General customer impatience in the queue $GI/G/1$. *J. Appl. Probab.* **2**, 186-205.
- [14] Erdelyi, A., (1953). *Higher Transcendental Functions, Vol. I*, McGraw-Hill.
- [15] Feller, W., (1968). *An Introduction to Probability Theory and Its Applications, Vol. I*, Ed. 3. John Wiley & Sons, New York.
- [16] Fleming, P.J., Stolyar, A., Simon, B., (1996). Heave traffic limit for a mobile phone system loss model. *Preprint* .
- [17] Gnedenko, B.V., Kovalenko, I.N., (1968). *Introduction to Queueing Theory* (Israel Program for Scientific Translations, Jerusalem).
- [18] Grimmett, G.R., Stirzaker, D.R., (1992). *Probability and Random Processes*. Ed. 2. Clarendon Press.
- [19] Halfin, S., Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**, 567-587.
- [20] Haugen, R.B., Skogan, E. (1980). Queueing systems with stochastic time out. *IEEE Trans. Commun.* **COM-28**, 1984-1989.
- [21] Harris, C.M., Hoffman, K.L., Saunders, P.B., (1987). Modeling the IRS telephone taxpayer information system. *Operations Research* **35**, 504-523.
- [22] Hoffman, K.L., Harris, C.M., (1986). Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research* **27**, 207-214.
- [23] Iglehart, D.L., (1965). Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* **2**, 429-441.

- [24] Jagerman, D.L., (1974). Some properties of the Erlang loss function. *The Bell System Technical Journal* Vol. **53**, No. **3**, 525-551.
- [25] Mandelbaum, A., Massey, W.A., Reiman, M., (1998). Strong approximations for Markovian service networks. *Preprint*.
- [26] Mandelbaum, A., Pats, G., (1995). State-dependent queues: approximations and applications. *Stochastic Networks*, 239-282. Ed. F.P. Kelly, R.J. Williams, Springer-Verlag.
- [27] Newell, G.F., (1982). *Applications to Queueing Theory*. 2nd. ed., Chapman and Hall, London.
- [28] Nguyen, V., (1993). Processing networks with parallel and sequential tasks: heavy traffic analysis and Brownian limits. *The Annals of Applied Probability*. **3**, 28-55.
- [29] Palm, C., (1937). Etude des delais d'attente. *Ericsson Technics* **5**, 37-56.
- [30] Palm, C., (1943). Intensitatsschwankungen im fernsprechverkehr, *Ericsson Technics* **44(1)**, 1-189.
- [31] Palm, C., (1953). Methods of judging the annoyance caused by congestion. *Tele.* **4**, 189-208.
- [32] Riordan, J., (1962). *Stochastic Service Systems*, Wiley.
- [33] Sadrolhefazi, A., (1993). Monotonicity results for $M/M/c/k$ queues with abandonments and retries. AT&T tech report.
- [34] Sadrolhefazi, A., (1993). Stochastic monotonicity in multi-server, finite-buffer queues with abandonment. AT&T tech report.
- [35] Stanford, R.E., (1979). Reneging phenomena in single channel queues. *Math. of Oper. Res.* **4**, 162-178.
- [36] Stanford, R.E., (1990). On queues with impatience. *Adv. Appl. Prob.* **22**, 768-769.
- [37] Stone, C., (1963). Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois Journal of Mathematics* **7**, 638-660.

- [38] Sze, D.Y., (1984). A queueing model for telephone operator staffing. *Operations Research* **32**, 229-249.
- [39] Whitt, W., (1992). Understanding the efficiency of multi-server service systems. *Management Science* Vol. **38**, No. **5**, 708-723.
- [40] Yurkevich, O.M., (1970). Many-line queueing systems with limited waiting time. *Engineering Cybernetics* No. **5**, 868-875.
- [41] Yurkevich, O.M., (1970). Multilinear systems with random constraints on waiting time.

נספח: קוד MATLAB לחישוב מדדים מדויקים

בנספח זה נתון קוד ליישום החישוב של מדדים מדויקים על פי הסכימה הראשונה, כפי שמופיע בפרק 3 (נוסחה (3.9)). הקוד מתאים לתכנת MATLAB. התכנית הנתונה מהווה מסגרת חיצונית, ועבור כל מדד יש צורך להגדיר פונקציה $I(k)$ מתאימה (ראה הגדרה ב-(3.7)). הסברים נוספים מופיעים בגוף התכנית.

```
function [out] = msrs(B,theta,lambda,mu,N,t)
% B - system capacity
% theta - abandonment rate
% lambda - arrival rate
% mu - service rate
% N - number of servers
% t - time parameter (for P{Wait>t} etc.)

%-- Part 1
%-- calculating the normalizing constant of the stationary distribution
nrml = 1;
add = lambda/mu;
for i = 1:N-1
    nrml = nrml + add;
    add = add * lambda/mu/(i+1);
```

end

imdt = nrml;

prd = (lambda/mu)^N/gamma(N+1);

nrml = nrml + prd;

for j = N+1:B

 prd = prd * lambda/(N*mu+(j-N)*theta);

 nrml = nrml + prd;

end

%-- End Part 1

imdt = imdt/nrml; %imdt - imediate service = P{Wait=0}

busy = prd/nrml; %busy - buffer is full = P{Block}

pi_N = (lambda/mu)^N/gamma(N+1)/nrml;

c = N*mu/theta ;

lt = lambda/theta ;

scnd_sum = 1;

coeff = (-lt)^(B-N-1)/gamma(B-N) ;

% The function ik is to be defined according to the performance measure
% beeing calculated.

Ik = ik(B-N-1,B,theta,lambda,mu,N,t) ;

Ef = coeff*Ik ;

for n = B-N-2:-1:0

```

    scnd_sum = scnd_sum + (1t)^(B-N-1-n)/gamma(B-N-n);
    coeff = coeff*(n+1)/(-1t);
    Ik = ik(n,B,theta,lambda,mu,N,t);
    Ef = Ef + coeff*Ik*scnd_sum;
end

Ef = Ef*pi_N*c;

output = [B theta lambda mu N t Ef];

[out] = output;

```

Designing a Telephone Call Center with Impatient Customers

Research Thesis
Submitted in Partial Fulfillment of the Requirements
for the degree of Master of Science
in Operations Research and Systems Analysis

OFER GARNETT

Submitted to the Senate of the Technion - Israel Institute of Technology

Adar 5758

Haifa

March 1998

The research was carried out under the supervision of Prof. Avishai Mandelbaum in the faculty of Industrial Engineering & Management.

I am grateful to Prof. Mandelbaum for his guidance during this research, his patience and assistance. I would also like to thank Dr. Marty Reiman of Bell Labs for his help.

The generous financial help of the Technion Graduate School is gratefully acknowledged.

Table of Contents

Introduction	4
1 The analytic model	9
1.1 Description as a Birth and Death process	11
1.2 Alternative description as a two-node network	12
1.3 Elementary performance measures - waiting and abandonment probabilities	13
1.4 Abandonment probability - usefull results	15
2 Literature review	19
3 Exact calculation of performance measures in an $M/M/N/B+M$ system (abandonment rate θ)	27
4 Approximations	32
5 Asymptotic behavior of the abandonment probability	35
6 Asymptotic behavior in Heavy Traffic	38
6.1 Abandonment and waiting probabilities ($\theta_N \propto N^\gamma$)	41
6.1.1 $\gamma = 0$	41
6.1.2 $\gamma > 0$	44
6.1.3 $\gamma < 0$	47
6.2 Weak convergence of the stationary distributions $\{q_N(\infty)\}$	50
6.3 Weak convergence of the processes $\{q_N\}$	56

Table of contents continued:

6.4	Approximating the stationary distribution - conclusion	70
7	Designing a service center - discussion	73
7.1	Examining quality of the approximations	75
7.1.1	Dependence on the number of servers	76
7.1.2	Application	79
7.1.3	"Extreme" abandonment rates	82
7.2	Rules of thumb	84
7.3	Service-level measures	87
	Conclusion / Future research	89
	Appendix: MATLAB code for exact calculations of performance measures	96

List of Figures

0.1	Schematic model of a telephone call center	5
1.1	$\{Q(t), t \geq 0\}$ - Transition diagram	11
4.1	Approximating the process $\{q_N(t), t \geq 0\}$	33
7.1	Dependence of $P\{Wait > 0\}$ on θ in an $M(95)/M(1)/100 + M$ system . .	81
7.2	$P\{Ab\}$ in an $M(95)/M(1)/100 + M$ system with "slow" abandonment rates	82
7.3	$P\{Ab Wait > 0\}$ in an $M(95)/M(1)/100 + M$ system with "slow" abandonment rates	83
7.4	$P\{Ab Wait > 0\}$ in an $M(95)/M(1)/100 + M$ system with "fast" abandonment rates	84

List of Tables

3.1	Performance measures of the form $E[f(V, R)]$	28
6.1	Results for asymptotic behavior in HW regime	42
7.1	Asymptotic behavior of the waiting probability in HW regime	76
7.2	Minimum number of servers to achieve 0.05 accuracy	77
7.3	$\gamma = 0$ - Minimum number of servers to achieve 0.01 accuracy	79

Abstract

During the recent decades there has been an explosive growth in the number of companies providing services via telephone, and in the variety of telephone services provided. In this work we focus on telephone call centers that provide information or business services, customer help desks, and telemarketing centers. Telephone call centers of this type are large (number-of-agents-wise), relative to other service centers.

The overall challenge in designing and managing a service center is to achieve a balance between efficiency and service level. Service via the telephone requires a relatively short response time (seconds), while the number of daily calls to a large telephone call center can reach tens of thousands. In these circumstances, one must rely on analytical models to achieve the sought-after balance. A common model for the analysis of telephone call centers with many servers is the $M/M/N/B$ model, but it lacks an important feature: in service centers, particularly service via telephone, customers tend to be impatient, therefore customers waiting in queue might decide to leave (abandon) before their service begins.

It turns out that the design of a call center using results from the analysis of an $M/M/N/B$ model (without abandonment) may cause a significant excess in server allocation. The abandonment of a client hurts the level of service provided, but by abandoning he is also relieving the workload in the system and improving the service level of the other customers. Therefore, it is important to equip the call center's designer/manager with tools which will enable the assesment of the impact of abandonment on service level and efficiency. Abandonment in the $M/M/N/B$ model also adds a "dimension" to performance measures, and the call center's designer/manager can account for "abandoning" customers in addition to "waiting" customers.

We have analyzed the simplest case, in which the customers' patience is modeled by an exponential random variable, independent of the arrival and service processes, and the state of the system. Assuming independence on the state of the system is appropriate for

service via the telephone, where the queues are "invisible" in the sense that customers typically have no information about the state of the system (in contrast, for example, to a bank, in which the customer sees the queue which typically effects behavior). For an $M/M/N/B$ model with "exponential" abandonment we use the notation $M/M/N/B+M$. We introduce two schemes for exact calculation of performance measures of the $M/M/N/B+M$ model in steady state. We continue by focusing on the asymptotic analysis of an $M/M/N+M$ system, in a regime appropriate for large service centers (many servers, high efficiency, high service level). Here we assume that the system's waiting capacity is unlimited. (This assumption coincides with the prevalent policy of avoiding busy-signals by allocating a large number of trunks for customers awaiting service). Our asymptotic analysis was performed at two levels:

1. For two elementary performance measures - fraction of customers required to wait in queue and fraction of customers who abandon before their service begins.
2. The stationary distribution of the stochastic process of the number of customers in the system. Here we used two different approaches - finding the weak limit of a series of stationary distributions, and finding the stationary distribution of the weak limit of a series of stochastic processes. The results in both cases coincide.

In both cases the asymptotic behavior depends on the values of the system's parameters, and their interrelations. Our primary goal is to use the asymptotic results as approximations. It is possible to determine which values of the parameters lead to the most "successful" approximations, and this defines a "sub-model" which is most suitable for the analysis of telephone call centers with impatient customers. Following the asymptotic behavior of this sub-model, we derive approximations for a variety of performance measures and propose "rules of thumb" for the design of a telephone call center.