



ניסוי: הפצה לאחור של השגיאה

המימוש של המסמך "הפצה לאחור של השגיאה ברשת נוירונים",
כלשונו, על ספרות כתובות ביד

הורץ ב-10-08-2026

ד"ר ארז שיינר

מה הרצנו

הקוד `code/backprop.py` מממש את המסמך כלשונו: $n = 3$ שכבות עם סיגמואיד בכל שכבה $k \geq 2$, גם בפלט (הגדרה 2), מטריצות המשקלים וההטיות (הגדרה 4), השגיאה $E = \frac{1}{2} \|y^n - d\|^2$ (הגדרה 5), הרקורסיה של משפט 8 בדיוק בסדר שבו היא מנוסחת, ועדכון המשקלים אחרי כל קלט בנפרד (שיטה 7), במעבר אחד על קבוצת האימון בסדר שבו היא נתונה. אין מיני-`batch`, אין מומנטום, אין `softmax`, אין ערבוב ואין מעברים חוזרים: רק מה שבמסמך.

– **התמונות.** MNIST: 60000 ספרות לאימון ו-10000 לבדיקה.

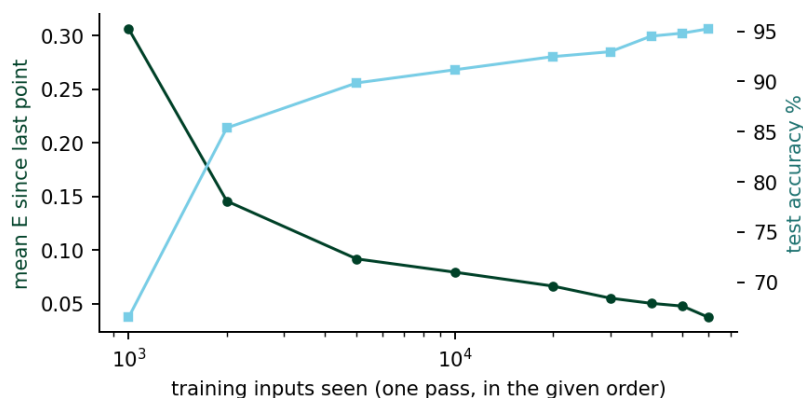
– **מה המסמך משאיר פתוח, ומה בחרנו.** הקלט x^1 הוא 784 הבהירויות מחולקות ב-255, כך שהוא ב- $[0, 1]^{784}$; הפלט הרצוי $d \in \mathbb{R}^{10}$ הוא וקטור עם 1 במקום הספרה הנכונה ו-0 בשאר; הסיווג הוא לפי הרכיב הגדול ביותר של y^3 . הערכים ההתחלתיים: W^k אחידים ב- $[-r, r]$ עם $r = 1/\sqrt{i_k}$, והטיות אפס. שתי הבחירות הראשונות אינן עניין של טעם, ובלעדיהן הרשת אינה לומדת כלל; סעיף 3 מראה זאת במספרים.

– **הרשת.** $i_1 = 784$, $i_2 = H = 64$, $i_3 = 10$; סך הכול 50,890 פרמטרים. $\eta = 0.5$. בניסויי ההשוואה שונה פעם H (16, 64, 128) ופעם η (0.1, 0.5, 2).

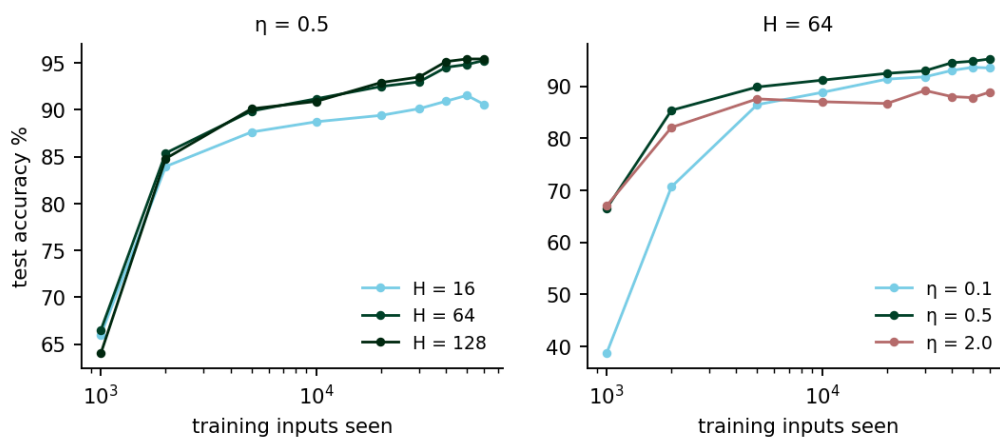
– **זמן.** 7 שניות למעבר של הרשת הסופית, כולל חישוב הדיוק בנקודות הביניים. הלולאה עוברת על הקלטים אחד-אחד, כמו במסמך, ולכן היא איטית פי כמה ממימוש שמעבד קבוצות.

– **חומרה.** Intel(R) Ultra Core(TM) 9, 16 ליבות, 31.5 GB זיכרון. הכול ב-Numpy על המעבד; הכרטיס הגרפי NVIDIA GeForce RTX 5060 Laptop GPU (MiB) 8151 לא היה בשימוש. Windows 11, Python 3.12.10, Numpy 2.4.4.

1 האימון

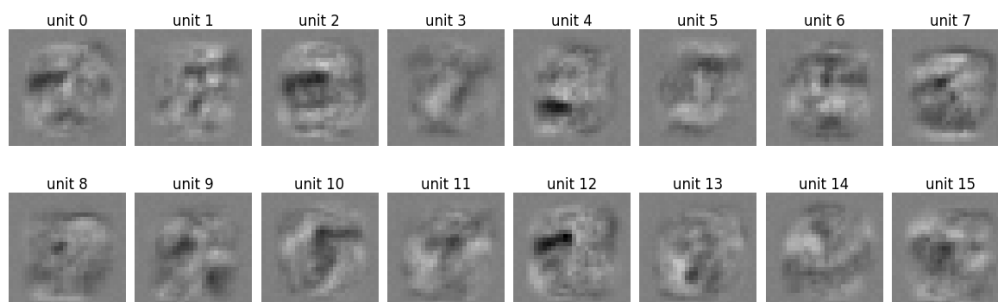


הרשת הסופית ($\eta = 0.5$, $H = 64$) לאורך המעבר היחיד: השגיאה הממוצעת E של הקלטים מאז נקודת המדידה הקודמת (ירוק, ציר שמאל) ודיוק הסיווג על תמונות הבדיקה (תכלת, ציר ימין), אחרי כך וכך קלטים. אחרי 1,000 קלטים הדיוק 66.5%, אחרי 10,000 91.2%, אחרי 30,000 93.0%, ובסוף 95.24%; השגיאה הממוצעת יורדת מ-0.306 באלף הראשונים ל-0.037 בעשרת האלפים האחרונים.



משמאל: שלושה גדלים של השכבה החבוייה עם $\eta = 0.5$: $H = 16$ מגיע ל-90.6%, $H = 64$ ל-95.2%, $H = 128$ ל-95.4% (20 שניות במקום 7). מימין: שלושה ערכים של η עם $H = 64$: $\eta = 0.1$ לומד לאט יותר (38.7% אחרי 1,000 קלטים לעומת 66.5%) ומגיע ל-93.5%; $\eta = 2$ גדול מדי, הצעדים מדלגים מעל המינימום והדיוק נתקע ב-88.9%.

2 הייצוג שנלמד



16 מתוך 64 שורות המטריצה W^1 , כל אחת מסודרת בחזרה כתמונה 28×28 (לבן חיובי, שחור שלילי). כל שורה היא "מבחן" שהרשת למדה להפעיל על התמונה: במקום הקשרים האקראיים של הפרספטון, כאן המשקלים של השכבה הראשונה נקבעו ע"י ההפצה לאחור, ורואים בהם קווים וקשתות של ספרות.

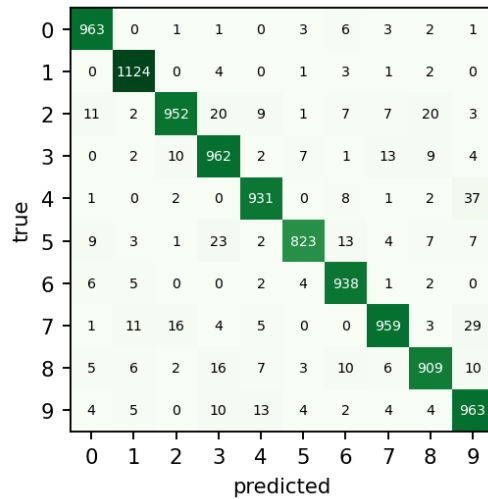
3 מה נדרש כדי שהפשוט יעבוד

המסמך אינו אומר מה הם המשקלים בתחילת הלמידה ובאיזה קנה מידה נתון הקלט. שתי הבחירות האלה אינן שיפור של האלגוריתם, אבל בלעדיהן הוא אינו עובד, ולכן הן חלק מהקורס. בדקנו כל אחת מהן בנפרד, עם $H = 64$ ו- $\eta = 0.5$:

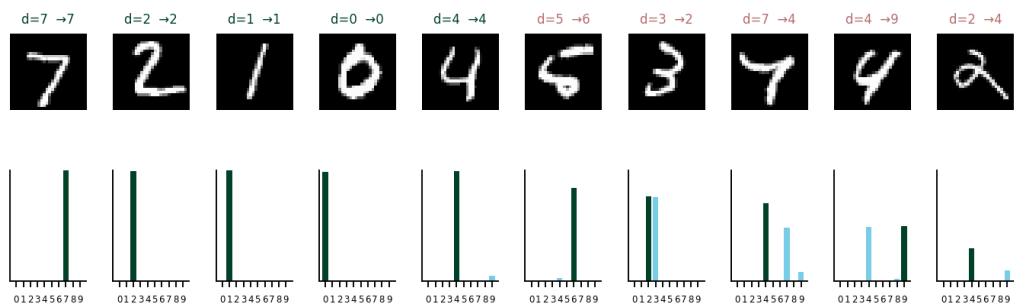
- **משקלים התחלתיים אפס.** הדיוק אחרי מעבר שלם: 30.8%. הסיבה סימטריה: כאשר כל השורות של W^1 שוות (כולן אפס), כל ניווני השכבה החבויה מחשבים אותו y_i^2 , לפי משפט 8 מקבלים אותו $\frac{\partial E}{\partial x_i^2}$, ולכן אותו עדכון; הם נשארים זהים לנצח, והרשת כולה שקולה לרשת עם ניוון חבוי אחד. אתחול אקראי קטן שובר את השוויון, וזה כל תפקידו.
- **הקלט בבירוריות 0 עד 255 במקום ב-[0,1].** הדיוק: 11.7%, כמו ניחוש. $x^2 = W^1 x^1 + b^1$ יוצא בגודל מאות, הסיגמואיד רווי, y_i^2 הוא 0 או 1 כמעט בדיוק, והגורם $y_i^2(1 - y_i^2)$ שבמשפט 8 (ב) מאפס את הגרדיאנט: הרשת אינה יכולה ללמוד כי הנגזרת שהיא מחשבת היא אפס.
- η . פשרה, כפי שהאיור של סעיף 1 מראה: קטן מדי ואיטי, גדול מדי ומדלג.

4 התוצאות

על תמונות הבדיקה: 95.24% סיווג נכון (476 טעויות מתוך 10000), לעומת 95.34% על תמונות האימון. הספרה הקלה ביותר 1 (99.0%), הקשה ביותר 2 (92.2%).



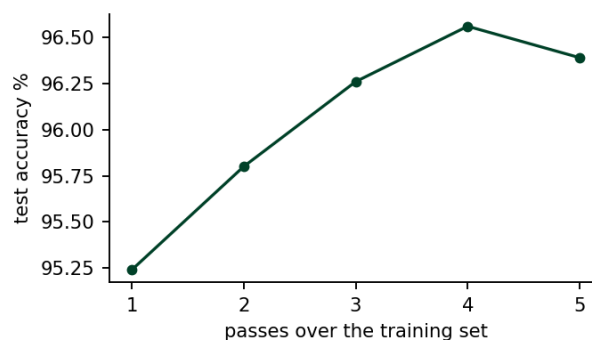
מטריצת הבלבול על תמונות הבדיקה: שורה, הספרה הנכונה; עמודה, הספרה שהרשת בחרה.



עשר תמונות בדיקה והפלט $y^3 \in [0, 1]^{10}$ של הרשת: חמש נכונות וחמש שגויות. הרשת מסווגת לפי הרכיב הגדול ביותר. בטעויות הפלט בדרך כלל מפוצל בין שתי ספרות, והרכיב הנכון קרוב לגדול.

5 מה לא עשינו: מעברים חוזרים

המסמך מתאר עדכון לכל קלט ואינו אומר כמה פעמים עוברים על הנתונים; המימוש עובר פעם אחת. בניסוי צד חזרנו על קבוצת האימון, באותו סדר, עוד ארבע פעמים.



הדיוק על תמונות הבדיקה אחרי כל מעבר. אחרי 5 מעברים 96.39%, כלומר 1.1 נקודות יותר ממעבר אחד, בזמן פי 5. השיפור קיים אבל קטן, ולכן המימוש נשאר מעבר אחד; מי שרוצה את הנקודה הנוספת יודע עכשיו מה מחירה.

6 מה למדנו

- הרקורסיה של משפט 8, מיושמת מילה במילה עם סיגמואיד בפלט ושגיאה ריבועית, במעבר אחד על הנתונים, מגיעה ל-95.24% על ספרות כתובות ביד: כ-4 נקודות יותר מהפרספטרון עם 3000 נזירונים אקראיים, עם 50,890 פרמטרים בלבד.
- הרשת לומדת גם את הייצוג: שורות W^1 אינן אקראיות אלא תבניות של ספרות. זה ההבדל בין יחידה 2 ליחידה 4.
- שני דברים שהמסמך שותק עליהם הכרחיים: אתחול אקראי (בלעדיו 30.8%) וקלט בקנה מידה קטן (בלעדיו 11.7%). שניהם נובעים ישירות ממשפט 8, לא מניסיון.
- η הוא פשרה: קטן מדי ואיטי, גדול מדי ומדלג. H גדול יותר משפר מעט ועולה הרבה בזמן. מעברים חוזרים מוסיפים נקודה.
- העדכון אחרי כל קלט הוא הגרסה של 1986. המימוש המודרני מעבד קבוצות של קלטים בבת אחת ומשתמש ב-softmax ובאנטרופיה צולבת בפלט; זה מהיר פי עשרות ונותן אחוז או שניים יותר, אבל המתמטיקה של ההפצה לאחור זהה.