אלגוריתם האימון – מורד הגרדיאנט המקרי

בסוף הפרק הקודם הגדרנו את פונקציית המחיר של מודל הסיווג ל10 מחלקות פריטי הלבוש, כך:

$$C(W,b)=\frac{1}{\#Data}\sum_{(X,Y_t)\in Data}H(X,Y_t)$$

$$.H(X,Y_t)=-\sum_{n=0}^{9}y_{tn}\log(y_n)$$

אם ברצוננו להשתמש באלגוריתם מורד הגרדיאנט בכדי לאמן את הרשת, **בכל איטרציה** עלינו לחשב את הגרדיאנט

$$.\nabla C(W,b)=\frac{1}{\#Data}\sum_{(X,Y_t)\in Data}\nabla H(X,Y_t)$$

יש לשים לב כי בגרדיאנט זה אנו מחשבים את הנגזרת של פונקציית המחיר לפי כל אחד מפרמטרי המודל, וכן שסכום זה רץ על כל דגימות פרטי הלבוש בסט האימון, מהן קיימות 60 אלף. כבר כעת, כאשר מדובר במודל הפשוט ביותר למשימת הסיווג, חישוב זה מהווה צוואר בקבוק בתהליך האימון. בהמשך, כאשר נעסוק במשימות ורשתות מורכבות יותר, וכאשר אוסף הנתונים יהיה גדול יותר, חישוב זה יהפוך בלתי אפשרי. מעבר לכך, קיימת **יתירות** רבה באוסף הנתונים: ישנן דגימות דומות מאוד (למשל זוג מגפיים דומים), אשר הגרדיאנט ∇H שלהן דומה גם הוא, ובעת חישוב ∇C כנ"ל אנו חוזרים על חישובים זהים שלא לצורך, תוך בזבוז משאבים.

בכדי להתמודד עם בעיות אלו, נבצע התאמות באלגוריתם מורד הגרדיאנט הקלאסי, הראשונה שבהן היא שימוש **בגרדיאנט המקרי** (Stochastic Gradient): במקום לחשב את הממוצע של ∇H על כל אוסף הנתונים, נדגום באקראי מהאוסף הגדול קבוצה קטנה (minibatch), ונחשב ממוצע זה רק על איבריה:

$$.\nabla C(W,b)=\frac{1}{\#minibatch}\sum_{(X,Y_t)\in minibatch}\nabla H(X,Y_t)$$

אם ה-minibatch אשר דגמנו מייצג נאמנה את אוסף הנתונים, אזי לא נפסיד הרבה במעבר לשימוש בקירוב, שכן $\nabla C\approx\nabla C$. לבד משינוי זה, אלגוריתם מורד הגרדיאנט המקרי (Stochastic Gradient Descent – SGD) זהה לאלגוריתם המקורי: מעדכנים את ערכי הפרמטרים בכיוון השלילי של הגרדיאנט המקרי, $(W,b)=(W,b)-\alpha\nabla C$ וחוזרים על החישוב באופן איטרטיבי.

הרנדומליות המובנית בתוך אלגוריתם זה אינה בהכרח דבר רע, שכן לפונקציות המחיר איתן נעבוד בהמשך יהיו מספר רב של נקודות מינימום מקומי, אשר אלגוריתם מורד הגרדיאנט הקלאסי יעצור בכל אחת מהן. אם לעתים האלגוריתם ייקח צעד שאינו בכיוון הירידה התלולה ביותר, תיווצר ההזדמנות לברוח ממינימום מקומי ולנוע בהמשך אל ערכי פרמטרים טובים יותר. למעשה קיימים אלגוריתמי אופטימיזציה אשר משלבים אלמנטים סטוכסטיים באופן מכוון, בדיוק למטרה זו.

באופן פרקטי, כדי להשתמש במלוא הנתונים בסט האימון, אנו מבצעים את הדגימה **ללא החזרה** עד אשר כל הסט הגיע למיצוי. במילים אחרות, נחלק **את כל** סט האימון ל-minibatches באקראי ועל סמך minibatches אלו נחשב איטרציות של SGD. לאחר השימוש בכולן – נחלק שוב את אוסף הנתונים באקראי (ובשונה מהחלוקה הקודמת) לקבוצות קטנות ונחזור על התהליך. מעבר אחד על כל הנתונים בסט האימון (על כל ה-minibatches) ייקרא **epoch**. בהנחה שהזמן הלוקח לעבור על כל סט נתוני האימון קבוע, חלוקתו ל-minibatches קטנים תאפשר לנו לבצע יותר צעדים ב-epoch יחיד, באותו מחיר חישובי, דבר אשר יוביל לרוב להתכנסות מהירה יותר.

שאלות לתרגול

1. טענו minibatch של 16 דגימות מסט האימון של אוסף הנתונים MNIST-Fashion לזכרון.

2. בצעו איטרציה אחת של SGD על הרשת אותה הגדרתם בפרק הקודם.

הערות:

- השתמשו במתודה `backward()` לחישוב הגרדיאנט.
- פרמטרי המודל נמצאים במאפיינים `z.weight` ו-`z.bias`.
- הגרדיאנט המחושב נמצא במאפיין `grad` של הפרמטרים.

3. בצעו איטרציה אחת של מורד הגרדיאנט (הגרדיאנט המלא).
4. השוו את זמני הריצה של איטרציה אחת בשני האלגוריתמים.
5. בדומה לבעיית הסיווג הבינארי, פונקציית הנראות המתאימה למודל הסיווג לעשר מחלקות היא

$$L(W,b)=\exp\left(-\#Data\cdot C(W,b)\right)=\prod_{(X,Y_t)\in Data} e^{H(X,Y_t)}=\prod_{(X,Y_t)\in Data}\prod_{n=0}^{9} y_n^{\,y_{tn}}$$

וערכי פרמטרים בהם מתקבל מינימום של $C(W,b)$ הם כאלו אשר בהם מתקבל מקסימום הנראות.

א. חשבו את $\frac{\partial L}{\partial w_{p,q}}$ עבור אחד הפרמטרים במודל, ונסו להסביר בעזרת הביטוי המתקבל מדוע עדיף לחפש את המינימום של $C(W,b)$ על פני המקסימום של $L(W,b)$.

רמז: בפרק הקודם חישבתם את $\frac{\partial H}{\partial w_{p,q}}$. השתמשו בחישוב זה ובכלל השרשרת.

ב. לעיל טענו שעבור הגרדיאנט המקרי, המחושב על בסיס minibatch, מתקיים $\nabla C \approx \nabla C$. הסבירו למה קשה יותר לחשב קירוב ל- ∇L על בסיס minibatch קטן.

רמז: השתמשו בתשובה לסעיף הקודם.