# Review 91: Gradients Are Not All You Need

**Paper: https://arxiv.org/abs/2111.05803v2**

עוד מאמר אחד מהסדרה של ""X is/are .../not all you need" שכמובן לא יכולתי לפספס. זה המאמר הכי עמוק מהמאמרים נושאים את השם הזה לפחות בשנה האחרונה (למרות שבחירת השם לא אידיאלית ולא משקפת את מסקנות המאמר).

אז הנה לכם הסקירה הקצרה מבית shortdeepnightlearners#.

כמו שאתם יודעים הרוב המוחלט של שיטות אופטימיזציה לרשתות נוירונים היום מבוססת על גרדיאנטים של פונקציית לוס. בדרך כלל גרדיאנטים אלו מחושבים באמצעות כלים של גזירה אוטומטית. כאשר שיטות אלו מיושמות למערכות "איטרטיביות" הן עלולות "להתברר" ולגרום למצב כאוטי.

עכשיו השאלה מה זה מערכת איטרטיבית ומה זה מצב כאוטי בהקשר הזה? המאמר מספק כמה דוגמאות למערכות איטרטיביות במאמר:

אימון של רשתות נוירונים עמוקות: פונקציה שמופעלת באופן איטרטיבי היא טרנספורמציה של שכבה

למידה באמצעות חיזוקים: פונקצית מדרגה המתארת מעבר בין מצבי מערכת שונים

למידה של אופטימייזר (נגיד לומדים פרמטרים אופטימליים של אופטימייזר): פונקציה איטרטיבית כאן היא הפעלת אופטימייזר

מה זה מצב כאוטי במערכות דינמיות? זה מצב שבו שינוי מאוד קטן של קלט גורם לשינוי גדול מאוד של פלט. כלומר אם חישוב של גרדיאנט הוא כאוטי אנו עלולים לקבל גרדיאנטים לא יציבים (exploding)

בנוסף לפעמים חישוב גרדיאנטים יכול להתבסס על reparameterization trick המאפשר לחשב את נגזרת גם במקרים שפונקציית לוס כוללת דגימה מהתפלגות בעלת פרמטרים נלמדים. המאמר טוען שבמערות איטרטיביות חישוב כזה עלול להוביל למצב כאוטי גם כן.

אבל למה זה קורה בעצם? המאמר טוען שהסיבה לכך היא הנוכחות של "אקראיות" בחישוב של גרדיאנט. צריך להבין שיש לא מעט "אקראיות" במנגנוני חישוב נגזרת שונים. אקראיות עלול לצוץ מכמה כיוונים:

מיני-באטצ'ים באימון של רשתות

"רעש" הנוצר מחישובי floating point

דגימות של "מצב מערכת" בלמידה באמצעות חיזוקים.

שלא נדבר על המקרים בהם משתמשים ב- reparameterization trick לגזירה.

המאמר טוען שניתן לזהות מצב כאוטי באמצעות ניתוח של ערכים עצמיים של יעקוביאן של המערכת האיטרטיבית.

המאמר מביא כמה דוגמאות מעשיות לכך שחישוב של גרדיאנט במערכות איטרטיביות אכן גורם למצב כאוטי:

Rigid body simulation

מטה למידה

סימולציה של דינמיקה מולקולרית

לבסוף המחברים מציעים כמה דרכים להתמודדות עם התופעה הזו לדומיינים ומשימות שונות.

מסקנה: אם אתם מחליטים לעבוד עם כלים אוטומטיים לחישוב של גרדיאנטים למערכות איטרטיביות - תדאגו ״להביא״ את המערכת (את היעקוביאן שלה) למצב יציב. כך ששם המאמר לא מוצלח במיוחד - גרדיאנטים זה כן כל מה שאנחנו רוצים (עד שנצליח לחשב הסיאנים למודלים עם טריליון פרמטרים) אבל אנו צריכים גרדיאנטים שמתנהגים יפה.

לינקים: ארקיב, Reddit