# Review 111: [Short] DoG is SGD’s Best Friend: A Parameter-Free Dynamic Step Size Schedule, 30.07.23 https://arxiv.org/abs/2302.12022.pdf

**Paper: https://arxiv.org/abs/2302.12022v3**

נכון שכל פעם שאתם מאמנים רשתות אחד הפרמטרים שאתם צריכים להחליט עליו הוא קצב למידה (learning rate)? גם כאשר אתם משתמשים ב-adam אתם צריכים לבחור אותו. נכון שיש ערכים דיפולטים ש״אמורים לעבוד טוב עבור כל בעיה״ אבל עדיין בחירה לא טובה של קצב למידה עלולה להוביל לאימון ליעיל.

אז היום ב-shorthebrewpapereviews סוקרים מאמר המציע שיטה פשוטה שבה לא צריך לבחור קצב למידה ל-SGD והוא מחושב לפי נוסחה סגורה. קצב למידה(learning rate) תלוי במרחק המקסימלי בין הנקודה שאנו כרגע נמצאים בה מנקודת ההתחלה (= וקטור משקלים התחלתי) וגם בסכום של נורמות של גרדיאנטים עבור T הנקודות הקודמות. כלומר ככל שאנו ״הצלחנו להתרחק״ יותר מנקודת ההתחלה אנו מעלים את קצב הלמידה ובנוסף הוא ״מרוסן״ באמצעות סכום הגרדיאנטים שנצברו עד כה.

הבחירה של הנוסחה הזו נובעת מהתוצאה התיאורטית שאומרת אם עבור SGD (Stochastic Gradient Descent) עם קצב למידה ומספר איטרציות T קבועים, שקובעת שקצב למידה מקיים את הנוסחה המוצעת מוכפלת בקבוע c אז ניתן לחסום את אי האופטימליות שלו יחסית ל-SGD האופטימלי (עם הגורם התלוי בקבוע c). אבל כמובן שאנו לא רוצים להריץ T איטרציה בשביל לקבוע את קצב הלמידה. אז בוחרים אותו עם מקדם c=1 בכל איטרציה (לניתוח יותר מעמיק ועם כל ההוכחות ניתן למצוא במאמר עצמו).