# Review 85: [Short] Anticorrelated Noise Injection for Improved Generalization

**Paper: https://arxiv.org/abs/2202.02831v3**

עבודה מעניינת המציע שיטה מאוד פשוטה לשיפור ביצועים של (stochastic gradient descent (SGD

הוספת רעש אקראי לעדכוני משקלים (perturbed GD) המתבצעים במהלך SGD נחקרה בכמה עבודות בשנים האחרונות. אחת המסקנות של עבודות אלו היא שהזרקה של רעש (בדרך כלל גאוסי) בעל שונות נמוכה יחסית ל- SGD עשויה לשפר את יכולת ההכללה של הרשת המאומנת.

אבל למה זה עוזר בעצם? מקובל להסביר את ההשפעה החיובית של הזרקת רעש ל-SGD בכך שזה עוזר לפונקציית לוס להתכנס לנקודת מינימום ״רחבה״ כלומר כזו שברוב הנקודות בסביבה הקרובה שלה ערך של פונקצית לוס נותר נמוך. מינימום רחב של פונקציית לוס נחשב לטוב ליכולת הכללה של המודל לעומת מינימום חד כלומר כזה שאפילו בסביבתו הקרובה יש עלייה ניכרת בערכי פונקציית לוס. הסיבה לכך היא טמונה בהנהה (יש גם תוצאות תיאורטיות חלקיות המוכיחות זאת עבור מודלים פשוטים יחסית) שנקודת מינימום רחבה ״מבטאת את הסיגנל האמיתי״ מהדאטה ולא נוצרת כתוצאה של הרעש שנמצא בדאטה.

אוקיי, אז מה מציע המאמר? בדרך כלל מוספים רעש גאוסי בלתי תלוי לעדכוני משקלים של SGD והמאמר שואל שאלה לגיטימית: האם זה אופטימלי לביצועים. מתברר שלא כל כך….

המאמר מציע להוסיף רעש בעל קורלציה שווה ל -1. איך עושים זאת? דוגמים סדרה ארוכה של וקטורים גאוסית x\_i ויוצרים סדרת הפרשים בין איבר סדרה הסמוכים (x\_2 -x\_1, x\_3 - x\_2…, x\_n+1 - x\_n...) ומוסיפים אותה לעדכוני משקלים של SGD באיטרציה n. מתברר שזה עוזר בלא מקרים ליכולת ההכללה של הרשת. הם גם מראים שהערכים העצמיים של ההסיאן בנקודת מינימום עם השיטה שלהן (השיטה נקראת Anti-PGD) יותר נמוכים מאשר ל-PGD עם רעש חסר קורלציה.

המאמר גם דן (לא התעמקתי) בקשר בין הטכניקה המוצעת להוספה של רעש אקראי ללייבלים וגם לשיטות החלקה שונות (smoothing).

המאמר כתוב היטב ודי קל להבנה.

מאמר: https://arxiv.org/abs/2202.02831