# Review 159: [Short] Stochastic Re-weighted Gradient Descent via Distributionally Robust Optimization, 02.10.2023

**Paper: https://arxiv.org/abs/2306.09222v5**

https://openreview.net/forum?id=EhK6wBBJNS

האם כל הדוגמאות בדאטהסט שלכם שווים? כשאתם מאמנים מודל שלכם (נגיד רשת נוירונים) אתם עושים זאת עם באצ'ים כאשר כל דוגמא בבאץ' תורמת לעדכון המשקלים באותה מידה (לפי ערך הגרדיאנט בה). אבל האם זה אופטימלי? היום ב-#shorthebrewpapereviews אנו סוקרים מאמר שמציע שיטה למשקול של תרומת הדוגמאות לעדכון משקלי המודל פרופורציאונלית לאקספוננט של ערך של פונקציית בה.

כלומר ככל שדוגמא קשה יותר היא תתרום יותר לשינוי משקלי המודל. זה נשמע די הגיוני - כבר ראינו את הגישה הזו ב-AdaBoost לפני שנים. אבל איך המחברים הגיעו לכך? אוקיי, אז קודם כל אנו מעוניינים לאמן מודל הממזער את השגיאה על הדאטה שלנו.

אבל אין לנו את כל הדאטה אלא רק דאטהסט אימון. אחת הדרכים להתחשב בכך יש אין לנו רק מדגם ולא כל הדאטה היא לקנוס את המודל על שונות של המודל (על הדאטה - זה קשור ל bias variance tradeoff). מכיוון שחישוב שונות המודל הוא מורכב מאוד (ושערוכו רועש למדי) אז משתמשים בסוג של קירוב שממזער את הערך המקסימלי של פונקציית לוס על כל התפלגויות הדאטה הקרובות להתפלגות של דאטהסט האימון.

הקרבה מחושבת באמצעות f-divergences שזו משפחה של פונקציות מרחק בין התפלגויות (כמו KL, total variation distance או לא מעט אחרים). גישה של נקראת DRO - distribution robust optimization. המחברים מצאו קשר בין DRO לבין משקול תרומה של דוגמאות בעדכון של הגרדיאנט. מאמר מאוד מעניין - ממליץ לצלול לעומק לפרטים המתמטיים.