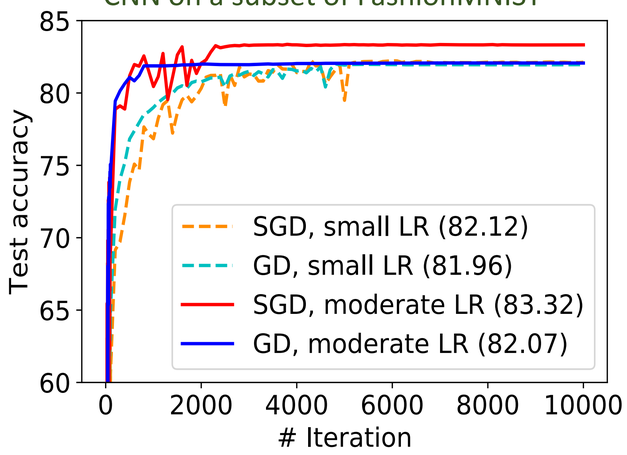
Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu
Johns Hopkins University & UCLA

1. Background

- SGD: $w_t = w_{t-1} - \eta_t \frac{1}{\mathrm{b}} \sum_{i \in B_t} \nabla \ell(x_i; w_{t-1})$
- GD: $w_t = w_{t-1} - \eta_t \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \nabla \ell(x_i; w_{t-1})$

CNN on a subset of FashionMNIST



	Small LR	Moderate LR
GD	☹	☹
SGD	☹	☺

Questions

- Small LR, SGD ≈ GD?
- Moderate LR, SGD >> GD?
- GD performs poorly anyhow?

2. Theory

Setups

- Test data $x = \zeta \cdot \xi \in \mathbb{R}^d$, where $\zeta \in (0, 1]$, $\xi \sim \mathcal{U}(S^{d-1})$.
- Linear model $\ell(x; w) = (w - w_*)^\top x x^\top (w - w_*)$
- Training data $\mathrm{X} = (x_1, \dots, x_n)$, $i.i.d.$, $d \gg n$
- Let $\lambda_i = \|x_i\|_2^2$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
- Let P be the projection onto the column space of X
- Let γ_1 (γ_n) be the largest (smallest non-zero) eigenvalue of $XX^\top$, and v_1 (v_n) be the corresponding eigenvector

Theorem 1: Consider SGD with batch size b and moderate LR,

$$\eta_t = \begin{cases} \eta \in \left({}^{b}/_{\lambda_1} + o(1), {}^{b}/_{\lambda_2} - o(1) \right), t = 1, \dots, T_1 \\ o(1), \qquad\qquad t = T_1 + 1, \dots, T_2 \end{cases}$$

then

$$\frac{P\left(w^{sgd} - w_*\right)}{\|P(w^{sgd} - w_*)\|_2} \to v_1 \pm o(1)$$

Theorem 2: Consider GD with moderate or small LR,

$$\eta_t \in \left(0, {}^{n}/_{2\lambda_2} - o(1)\right), \qquad t = 1, \dots, T_2$$

then

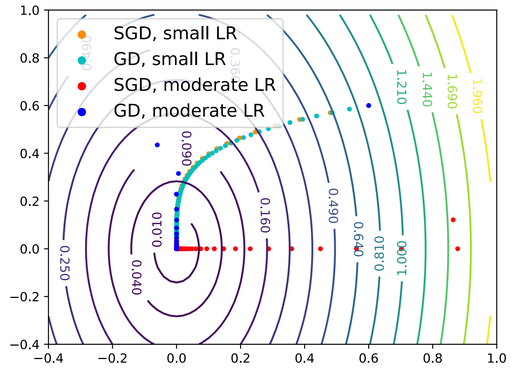
$$\frac{P\left(w^{gd} - w_*\right)}{\|P(w^{gd} - w_*)\|_2} \to v_n \pm o(1)$$

Theorem 3: Let $\Delta(w)$ be the estimation error and Δ_α^* be the optimal estimation error within an α-level set where the algorithms stop.

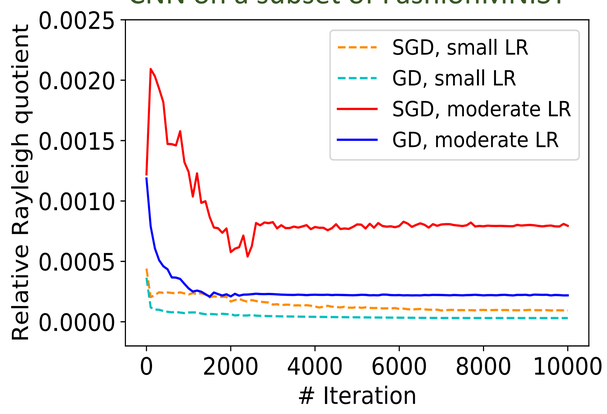
- For SGD with moderate LR, $\Delta\left(w^{sgd}\right) < \left(1 + o(1)\right) \cdot \Delta_\alpha^*$
- For GD with moderate/small LR, $\Delta\left(w^{gd}\right) > \left({}^{\gamma_1}/_{\gamma_n} - o(1)\right) \cdot \Delta_\alpha^*$

3. Verifications

2D numerical simulation



CNN on a subset of FashionMNIST



4. Conclusions

- SGD + moderate LR: converges along large eigenvalue directions
- GD + moderate/small LR: converge along small eigenvalue directions
- The former directional bias benefits generalization