

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



May 2021



SGD vs. GD: Learning Rate Matters!

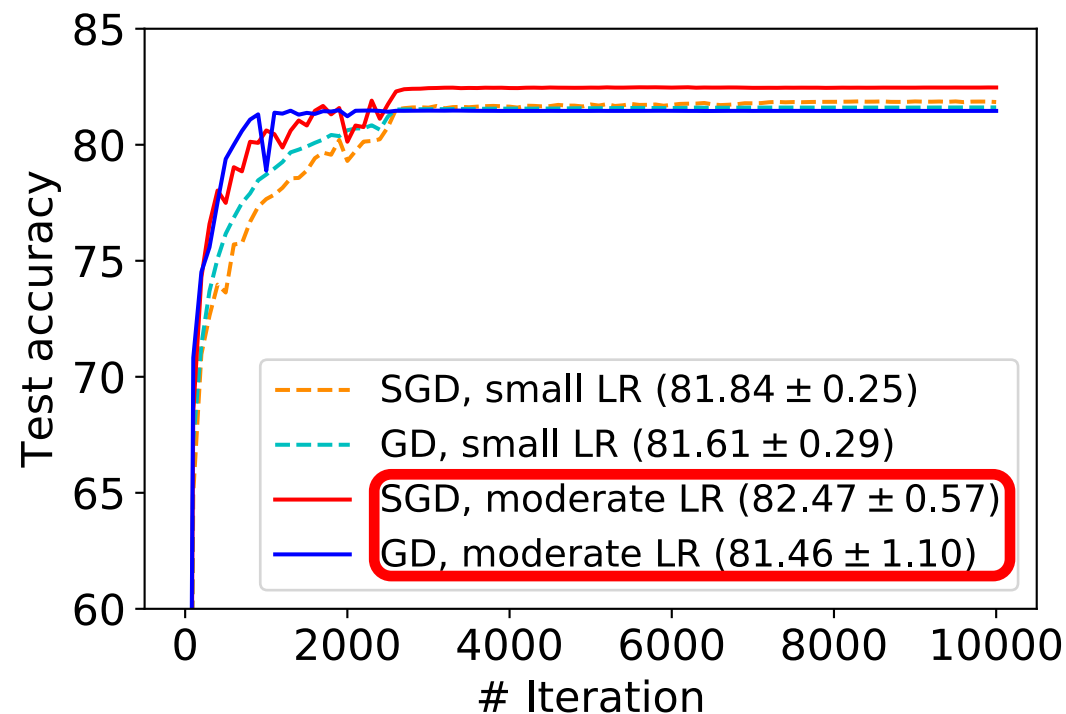
Loss $L_{\mathcal{S}}(w) = \frac{1}{n} \sum_{i=1}^n \ell_i(w)$

GD $w \leftarrow w - \eta \nabla L_{\mathcal{S}}(w)$

SGD $w \leftarrow w - \eta \nabla \ell_k(w)$

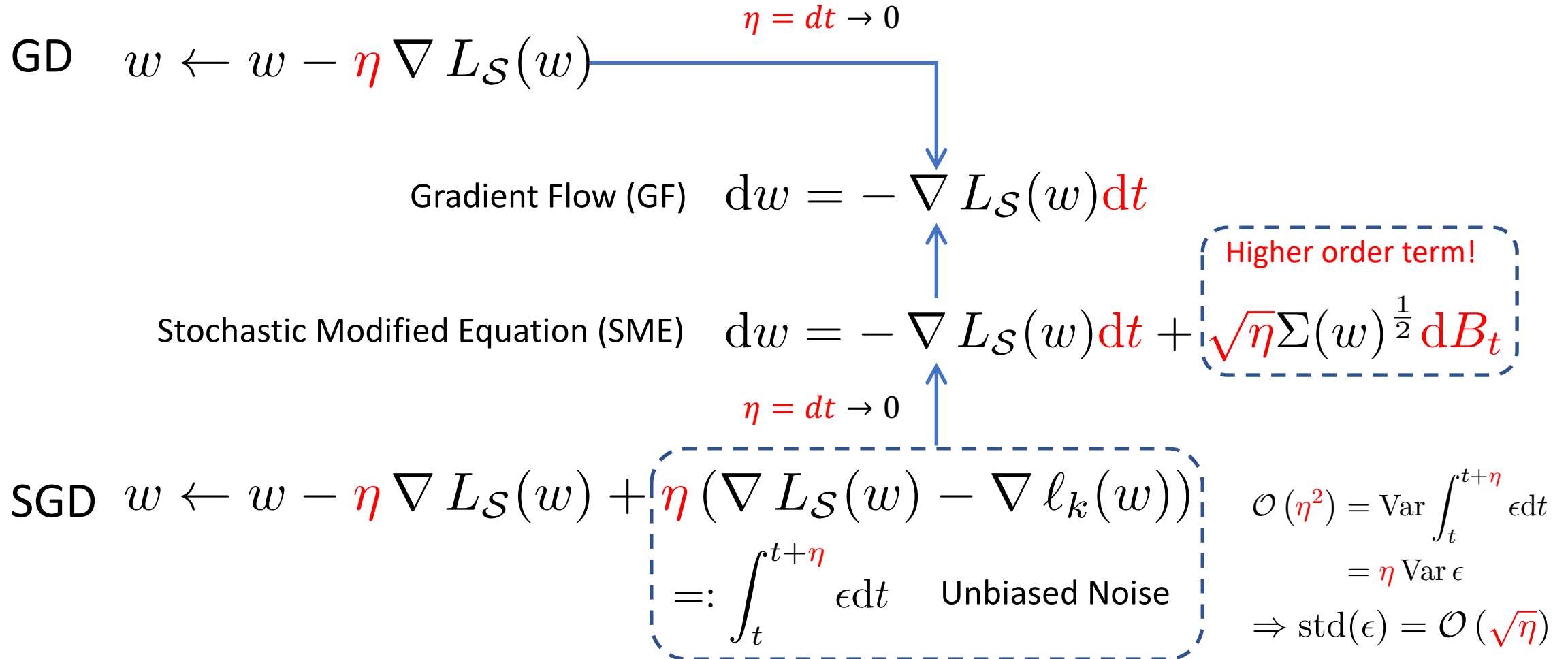
Questions

1. Small LR, SGD \approx GD?
2. Moderate LR, SGD \gg GD?
3. GD performs poorly anyhow?



A subset of FashionMNIST, a convolutional network

Small Learning Rate: SGD \approx GD



Effects of Moderate, Annealing Learning Rate

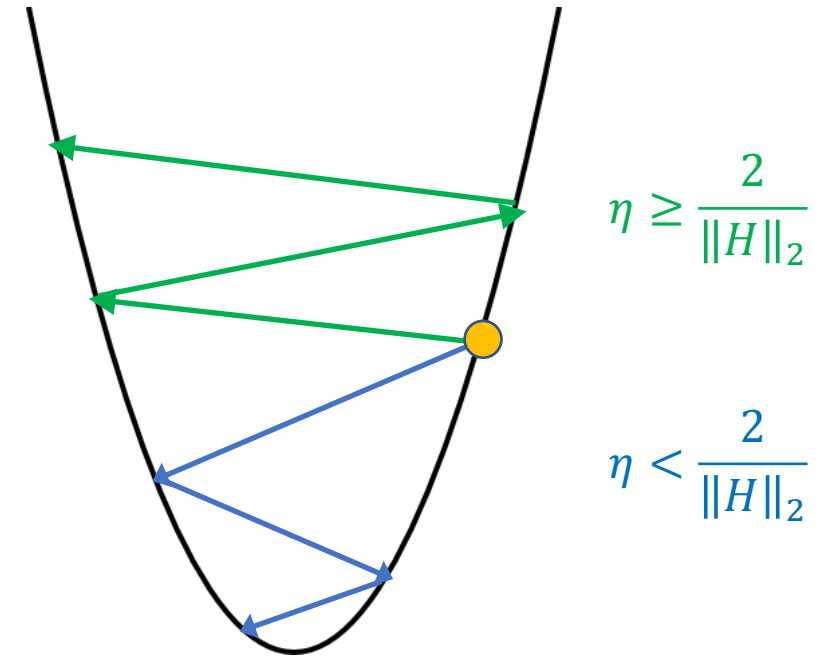
GD $w \leftarrow w - \eta \nabla L_{\mathcal{S}}(w)$

SGD $w \leftarrow w - \eta \nabla \ell_k(w)$

$$L_{\mathcal{S}}(w) = \frac{1}{n} \sum_{i=1}^n \ell_i(w)$$

Always flat!

One of them can be sharp



$$L(w) = 0.5 w^T H w$$
$$w_{k+1} = (I - \eta H) \cdot w_k$$

SGD + moderate and annealing LR

- Phase 1: moderate LR \Rightarrow fits flat losses first
- Phase 2: small LR \Rightarrow fits sharp losses then

A 2-D Example: Convergence Directions

$$\ell_1(w) = 0.5 w^\top H_1 w, \\ H_1 = \text{diag}(2\kappa, 0)$$

$$\ell_2(w) = 0.5 w^\top H_2 w, \\ H_2 = \text{diag}(0, 2)$$

$$L(w) = 0.5 w^\top H w, \\ H = \text{diag}(\kappa, 1)$$

$\kappa > 2$

Same limits

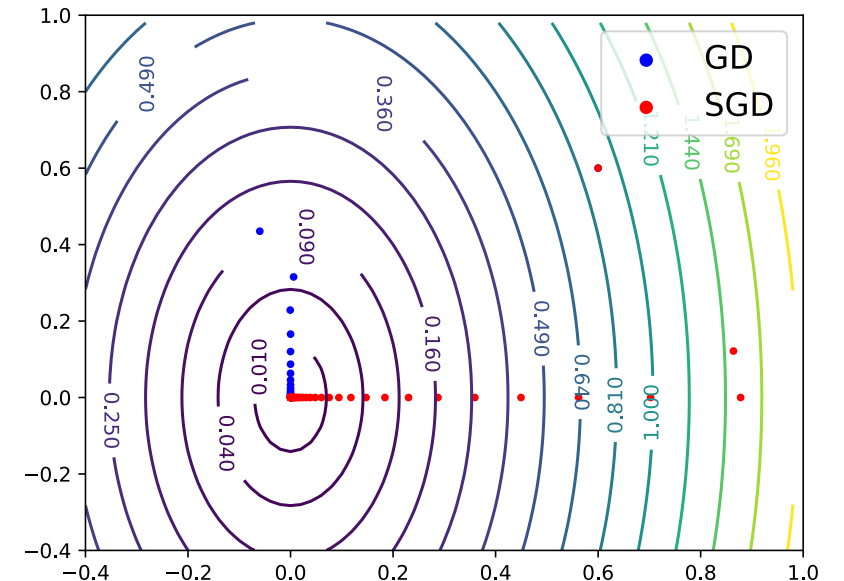
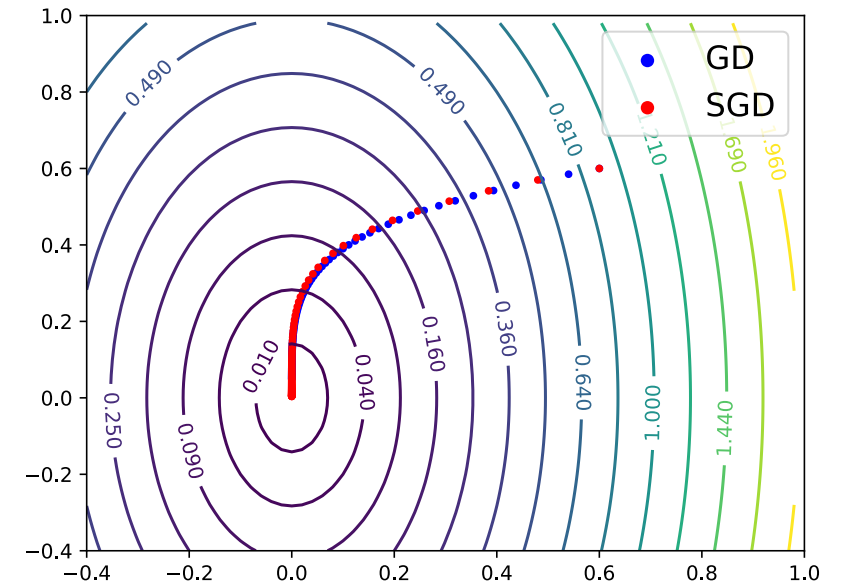
Different convergence directions

Small LR

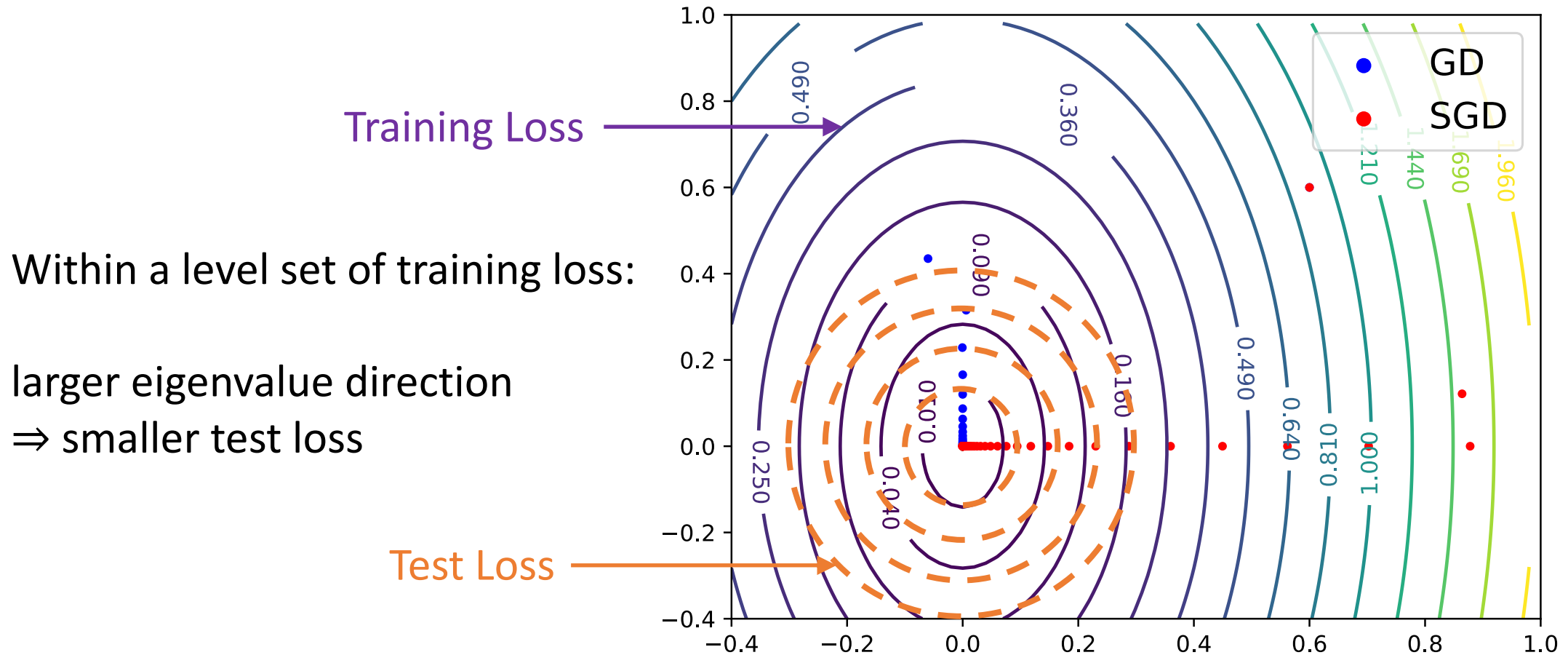
$$\eta_t = 0.1/\kappa, t = 1, \dots, T$$

Moderate LR

$$\eta_t = \begin{cases} 1.1/\kappa, t = 1, \dots, T_1 \\ 0.1/\kappa, t = T_1 + 1, \dots, T_2 \end{cases}$$



A 2-D Example: Effects on Generalization



A High Dimensional Linear Regression

Setups

- Test data $x = \zeta \cdot \xi \in \mathbb{R}^d$, where $\begin{cases} \zeta \in (0, 1] \\ \xi \sim \mathcal{U}(S^{d-1}) \end{cases}$
- $\ell(x; w) = (w - w_*)^\top x x^\top (w - w_*)$
- Training data $X = (x_1, \dots, x_n)$, *i.i.d.*, $d \gg n$

WOLG

- Let $\lambda_i = \|x_i\|_2^2 \in (0, 1]$
- Assume $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$
- Let P be the projection onto the column space of X
- $P_\perp = I - P$

Theorem 0:

- There are multiple minima for $L_S(w) = \frac{1}{n} (w - w_*)^\top X X^\top (w - w_*)$
- The iterates of gradient methods belong to a hypothesis class $\mathcal{H}_S = \{w: P_\perp w = P_\perp w_0\}$
- If gradient methods find a global minima, then it is the one closest to initialization

Remark: this is also known as “minimal-norm solution” since the initialization is usually zero

Theory: Convergence Directions

Theorem 1 (informal): Consider SGD with moderate LR,

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

then

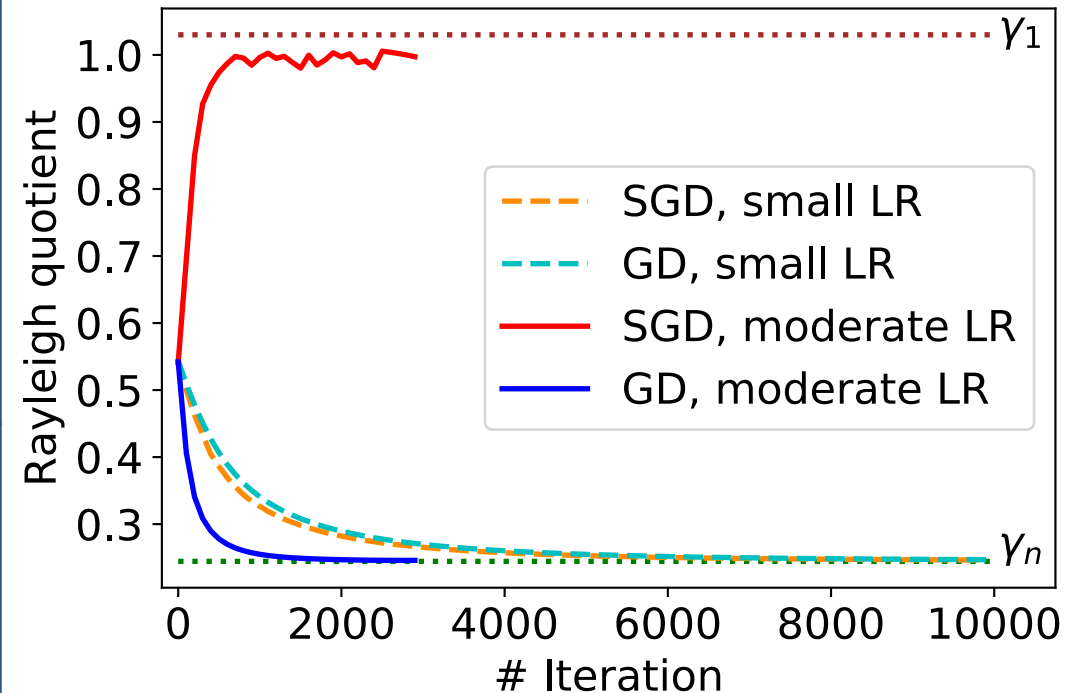
$$\frac{P(w - w_*)}{\|P(w - w_*)\|_2} \rightarrow v_1 \pm o(1)$$

Theorem 2 (informal): Consider GD with moderate or small LR,

$$\eta_t \in \left(0, n/2\lambda_2 - o(1)\right), \quad t = 1, \dots, T_2$$

then

$$\frac{P(w - w_*)}{\|P(w - w_*)\|_2} \rightarrow v_n \pm o(1)$$



$$\text{Rayleigh quotient: } R(XX^T, u) = \frac{u^T XX^T u}{u^T u}$$

Remark: v_1 (v_n) is the largest (smallest) eigen vector of XX^T

Continued Theory: Generalization Separation

$$L_{\mathcal{D}}(w^{\text{alg}}) - \inf_w L_{\mathcal{D}}(w) = \underbrace{L_{\mathcal{D}}(w^{\text{alg}}) - \inf_{w' \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{D}}(w')}_{\Delta(w^{\text{alg}}), \text{ estimation error}} + \underbrace{\inf_{w' \in \mathcal{H}_{\mathcal{S}}} L_{\mathcal{D}}(w') - \inf_w L_{\mathcal{D}}(w)}_{\text{approximation error}}$$

determined by the algorithms and hyperparameters intrinsic error, not improvable

- α -level set: $\mathcal{W}_{\alpha} = \{w \in \mathcal{H}_{\mathcal{S}} : L_{\mathcal{S}}(w) = \alpha\}$
- Optimal estimation error within a level set: $\Delta_{\alpha}^* = \min_{w \in \mathcal{W}_{\alpha}} \Delta(w)$

Theorem 3:

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

Remark: γ_1 (γ_n) is the largest (smallest) eigenvalue of XX^{\top}

Take Home

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

Get the paper ->

