Decomposing SGD Dynamics in Neural Networks: Teacher-Induced Spikes and Variance Inflation

Samet Demir[1,*] Zafer Doğan[1,2] {sdemir20, zdogan}@ku.edu.tr *now at Oden Institute, UT Austin
[1]MLIP Research Group, KUIS AI Center, Koç University [2]Department of EEE, Koç University

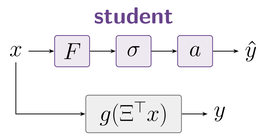
paper code

Motivation

- Two-layer network; the target depends on a few **teacher directions**
- Feature learning is usually tracked through alignment with these directions
- Most first-layer parameters lie in the **teacher-orthogonal complement**, which carries no information about the label
- Common simplification: a few summary quantities, or isotropic noise

Question: Does SGD keep the complement isotropic? And does the complement matter for generalization?

Setting



- **Teacher:** $x \sim \mathcal{N}(0, I_n)$ (input dimension n), $y = g(\Xi^\top x)$
  $\Xi \in \mathbb{R}^{n\times k_\star}$: orthonormal teacher directions, $k_\star = O(1)$; $z_\star = \Xi^\top x$: teacher input
- **Student:** $\hat{y}(x) = \frac{1}{k}\sum_{i=1}^{k} a_i\, \sigma(f_i^\top x)$
  k neurons; first layer $F \in \mathbb{R}^{k\times n}$ with rows f_i; second layer $a \in \mathbb{R}^k$; activation σ (theory: smooth and bounded; simulations: mostly ReLU)
- **Training:** online SGD on the squared loss, one fresh sample (x_t, y_t) per step
  step sizes η_f (first layer), η_a (second layer); error $e_t = \hat{y}(x_t) - y_t$
- **Proportional limit:** $n, k \to \infty$ with $k/n \to \alpha$; $t = \tau n$ steps
- **Test error:** $\mathcal{L}_t = \frac{1}{2}\mathbb{E}\big[(\hat{y}(x) - y)^2\big]$

Decomposition of the first layer

$$F_t = \underbrace{M_t \Xi^\top}_{\text{signal}} + \underbrace{F_t^\perp}_{\text{complement}}, \qquad F_t^\perp = \operatorname{diag}(r_t)\, U_t$$

$M_t = F_t\Xi \in \mathbb{R}^{k\times k_\star}$: signal matrix; $F_t^\perp = F_t P_\perp$ with $P_\perp = I_n - \Xi\Xi^\top$: complement matrix; $r_t \in \mathbb{R}^k$: row norms of $F_t^\perp$; U_t: its unit-norm row directions

teacher subspace: $k_\star$ directions

k neurons { **complement:** $n - k_\star$ directions, no information about y

- Each SGD step changes the complement by a **rank-one** term:

$$F_{t+1}^\perp = F_t^\perp - \frac{\eta_f}{k}\, v_t\, g_t^\top$$

$v_{i,t} = e_t\, a_{i,t}\, \sigma'(f_{i,t}^\top x_t)$: per-neuron driver; $g_t = P_\perp x_t$: isotropic in the complement

- **Complement Gram matrix:** $G_t = \frac{1}{n_\perp} F_t^\perp (F_t^\perp)^\top \in \mathbb{R}^{k\times k}$, $n_\perp = n - k_\star$. Given a_t and M_t, the test error sees the complement only through G_t.

1. SGD creates spikes in the complement

- Split each driver into a teacher-driven mean and a centered part:

$$v_s = \mu_s + \tilde{v}_s, \qquad \mu_s = \mathbb{E}\big[v_s \mid z_{\star,s}, \text{past}\big]$$

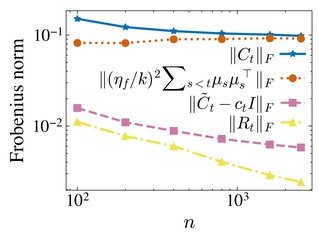
$z_{\star,s} = \Xi^\top x_s$; μ_s averages over the complement input g_s only
- **Teacher-induced structure:** $S_s = \mathbb{E}\big[\mu_s \mu_s^\top \mid \text{past}\big] \succeq 0$

Lemma 3.1 (complement geometry)

Accumulated driver covariance:

$$C_t = \tfrac{1}{n} I_k + \left(\tfrac{\eta_f}{k}\right)^2 \sum_{s<t} v_s v_s^\top = \underbrace{\tilde{C}_t}_{\text{isotropic bulk: } \approx c_t I_k} + \underbrace{\left(\tfrac{\eta_f}{k}\right)^2 \sum_{s<t} S_s}_{\text{teacher-induced: adds up}} + \underbrace{R_t}_{\text{vanishes}}$$

$\tilde{C}_t$: the same sum with $\tilde{v}_s$ in place of v_s; $c_t = \operatorname{Tr}(\tilde{C}_t)/k$; R_t: cross terms



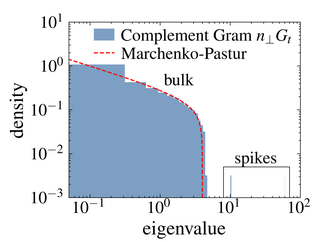
Growing n: the bulk deviation $\|\tilde{C}_t - c_t I\|_F$ and the remainder R_t shrink, while the teacher-induced term stays and dominates C_t.
(Simulation: $\mu_s\mu_s^\top$ in place of S_s.)

Lemma 3.2 (one-step update of the complement Gram)

$$\mathbb{E}\big[G_{t+1} \mid \text{past}\big] = G_t - \underbrace{\tfrac{\eta_f}{k}\big(J_t G_t + G_t J_t\big)}_{\text{damping}} + \underbrace{\left(\tfrac{\eta_f}{k}\right)^2 K_t}_{\text{injected covariance} \succeq 0}$$

Up to $O(n^{-3/2})$; $J_t = \mathbb{E}[\partial v_t/\partial h_t^\perp \mid \text{past}]$ (Stein operator), with complement preactivations $h_t^\perp = F_t^\perp g_t$;
$K_t = \mathbb{E}[v_t v_t^\top \mid \text{past}] = S_t + \mathbb{E}[\tilde{v}_t \tilde{v}_t^\top \mid \text{past}]$

Mechanism (Remark 3.3): the coherent part S_t is approximately low rank, so it creates **outliers (spikes)** above a Marchenko–Pastur bulk in the spectrum of G_t.



Spectrum of $n_\perp G_t$ in a simulation ($n = k = 1000$, $\tau = 2$): Marchenko–Pastur bulk plus spikes.

How large are the spikes?

- Top eigenvalue of $n_\perp G_t$: $\lambda_1 \approx 63$; initial bulk edge: 4
- λ_1/k is 40% of the mean signal variance $\frac{1}{k}\sum_i \|m_{i,t}\|^2$ and 5.5% of the mean complement variance $\frac{1}{k}\sum_i r_{i,t}^2$ ($m_{i,t}$: rows of M_t)
  $\Rightarrow$ small next to the complement, comparable to the signal
- Halving n roughly halves λ_1, so λ_1/k stays constant: not a finite-size effect

2. Decomposed Dynamics (DD)

- Goal: the same test error as SGD, with the drift separated from the injected covariance (which carries the spikes)
- Keep the SGD updates of the second layer a_t and the signal M_t; replace the complement update by

$$F_{t+1}^\perp = (I - \tilde{\Lambda}_t)\, F_t^\perp + \zeta_t$$

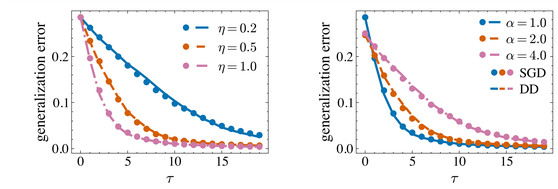
$\tilde{\Lambda}_t = \frac{\eta_f}{k} J_t$: the mean step $-\tilde{\Lambda}_t F_t^\perp$ is that of SGD (same drift)
ζ_t: Gaussian, isotropic in the complement, with neuron covariance $\Gamma_t = \left(\frac{\eta_f}{k}\right)^2 K_t$, the second moment of the SGD step (contains S_t)

Theorem 3.4 (equivalence of SGD and DD)

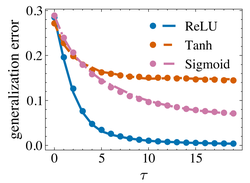
For every fixed τ,

$$\sup_{t \le \tau n} \big|\mathcal{L}_t^{\text{SGD}} - \mathcal{L}_t^{\text{DD}}\big| \longrightarrow 0 \quad \text{in probability.}$$

$\sigma, \sigma', \sigma'', \sigma'''$ bounded; g bounded and Lipschitz; bounded initial second layer



(a) step size $\eta = \eta_f = \eta_a$ (b) width ratio $\alpha = k/n$



(c) activation σ

- Markers: SGD; lines: DD ($n = 500$)
- The match holds even for ReLU, beyond the assumptions

3. Spikes inflate the variance

- Bulk plus spikes: $n_\perp G_t \approx \hat{c}_t\, I_k + B_t^{\text{acc}} (B_t^{\text{acc}})^\top$
  $\hat{c}_t$: bulk level; B_t^{acc}: accumulated teacher-induced loadings (spikes)
- Effective predictor $\hat{y}_t^{\text{eff}}$: the same network, with complement preactivations of exactly this covariance; $\mathcal{L}_t^{\text{eff}}$: its test error

Proposition 3.5 (variance inflation)

$$2\,\mathcal{L}_t^{\text{eff}} = \underbrace{\mathcal{L}_{\text{bias},t}}_{\text{bias}} + \underbrace{\mathcal{L}_{\text{res},t}}_{\text{variance from the bulk}} + \underbrace{\mathcal{L}_{\text{str},t}}_{\text{variance along the spikes}}$$

- $\mathcal{L}_{\text{str},t} \ge 0$, and $\mathcal{L}_{\text{str},t} = 0$ if $B_t^{\text{acc}} = 0$ (no spikes)
- $\mathcal{L}_{\text{str},t} > 0$ when the spikes overlap with the slope-weighted readout ω_t
  $\omega_{i,t}$: $a_{i,t}$ times the average slope σ' of neuron i



Bulk: independent across neurons $\Rightarrow$ averages out

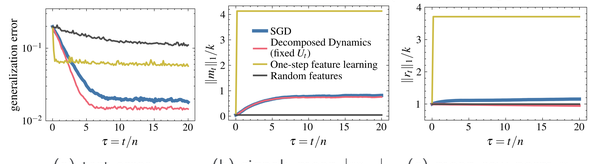


Spikes: shared across neurons $\Rightarrow$ add up

4. Freezing the complement directions

First layer trained in four ways (second layer: SGD):

Random features: fixed
SGD: full training
One-step feature learning: one large step, then fixed
DD with fixed U_t: complement directions frozen; M_t, r_t trained by SGD steps (not the surrogate)



(a) test error (b) signal: mean $|m_{i,t}|$ ($k_\star = 1$) (c) mean row norm $r_{i,t}$

- Fixed U_t: **lower test error** than SGD (ReLU student)
- Similar signal growth: the gain is not from faster signal learning

Freezing also stops the growth of r_t (c), so the comparison does not isolate the spikes; it is consistent with variance inflation.

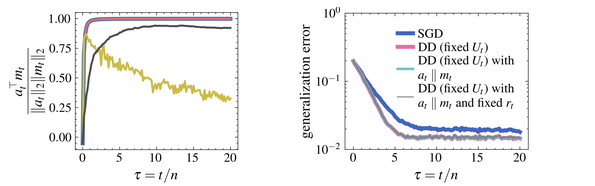
5. The two layers couple quickly

Proposition 3.6 (fast alignment of m_t and a_t)

Single-index teacher ($k_\star = 1$), oracle no-spike dynamics (complement channel replaced by independent Gaussians): for every $\varepsilon > 0$ there is $\tau_0(\varepsilon) > 0$ with

$$1 - \cos_t^2 \le \varepsilon \quad \text{for } k^{2/3} \le t \le \tau_0(\varepsilon)\, k,$$

so m_t aligns with a_t within $o(k)$ steps.
$\cos_t = a_t^\top m_t / (\|a_t\|_2 \|m_t\|_2)$; odd g, σ with nonzero linear parts; bounded a_0



(a) alignment $\cos_t$ (b) test error with $m_t \parallel a_t$

- Alignment is fast under SGD too (a; colors as in block 4)
- Fixed U_t plus fixed r_t and $M_t = a_t \nu_t^\top$ ($\nu_t \in \mathbb{R}^{k_\star}$ trained): the same test error as fixed U_t with $k + k_\star$ instead of $k(n+1)$ trainable parameters (b); a diagnostic (uses Ξ)

Takeaways

- **Isotropic?** No: SGD creates teacher-induced **spikes**
- **Does it matter?** Yes: a variance term; freezing the directions lowered the test error (simulations)
- DD asymptotically matches the test error of SGD
- Suggests far fewer effective degrees of freedom than the raw first layer

Supported in part by TÜBİTAK ARDEB 1001 (No. 124E063); S.D. also by a KUIS AI Center AI Fellowship and a TÜBİTAK BİDEB 2211 PhD Scholarship.