Does SGD really happen in tiny subspaces?

Minhak Song[1] Kwangjun Ahn[2] Chulhee Yun[1]

[1]KAIST [2]Microsoft Research

ICLR 2025

Gradient descent happens in a tiny subspace

We revisit:

GRADIENT DESCENT HAPPENS IN A TINY SUBSPACE

Guy Gur-Ari*
School of Natural Sciences
Institute for Advanced Study
Princeton, NJ 08540, USA
guyg@ias.edu

Daniel A. Roberts*
Facebook AI Research
New York, NY 10003, USA
danr@fb.com

Ethan Dyer
Johns Hopkins University
Baltimore, MD 21218, USA
edyer4@jhu.edu

- During DNN training, gradients align with dominant subspace.
  (dominant subspace = top-k eigenspace of train loss Hessian)

Gradient descent happens in a tiny subspace

We revisit:

GRADIENT DESCENT HAPPENS IN A TINY SUBSPACE

Guy Gur-Ari*
School of Natural Sciences
Institute for Advanced Study
Princeton, NJ 08540, USA
guyg@ias.edu

Daniel A. Roberts*
Facebook AI Research
New York, NY 10003, USA
danr@fb.com

Ethan Dyer
Johns Hopkins University
Baltimore, MD 21218, USA
edyer4@jhu.edu

- During DNN training, gradients align with dominant subspace. (dominant subspace = top-k eigenspace of train loss Hessian)

We ask:

Q. *Can DNN be trained within the dominant subspace?*

Gradient descent happens in a tiny subspace

We revisit:

- During DNN training, gradients align with dominant subspace. (dominant subspace = top-k eigenspace of train loss Hessian)

We ask:

Q. *Can DNN be trained within the dominant subspace?*

Spoiler!

A. *No, dominant subspace is **not** where the learning happens!*

Problem setting

Task: k-class classification problem

Method: minimize the train loss $L(\theta)$ ($\theta \in \mathbb{R}^d, k \ll d$) with SGD

Definition (dominant/bulk subspace)

The **dominant subspace** $S_k(\theta)$ is a low-rank eigenspace of the top-k eigenvalues of $\nabla^2 L(\theta)$, and the **bulk subspace** $S_k^\perp(\theta)$ is its orthogonal complement.

Definition (projection onto dominant/bulk subspace)

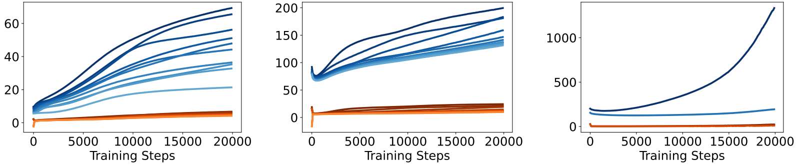
The projection matrix onto $S_k(\theta)$ ($S_k^\perp(\theta)$) is denoted by $P_k(\theta)$ ($P_k^\perp(\theta)$). The fraction of a given vector v in $S_k(\theta)$ is denoted by $\chi_k(v;\theta) := \|P_k(\theta)v\|/\|v\|$, or $\chi_k(v)$ in short.

Phenomenon 1: Gradient aligns with the dominant subspace

During DNN training with SGD,

1. Loss Hessian is approximately low-rank.

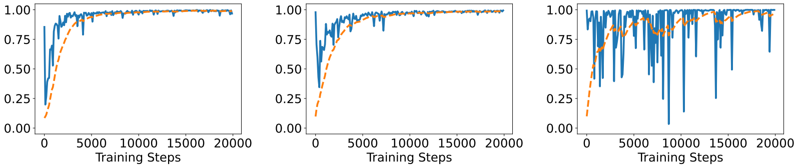
Top-k (blue) and next top-k (orange) eigenvalues:



(a) MLP on MNIST (b) CNN on CIFAR10 (c) Transformer on SST2

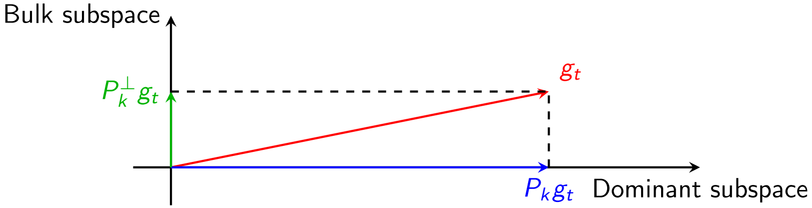
2. Gradients approximately align with the dominant subspace.

$\chi_k(\nabla L(\theta_t)) = \|P_k(\theta_t)v\|/\|v\|$ (orange dashed line denotes EMA):



(a) MLP on MNIST (b) CNN on CIFAR10 (c) Transformer on SST2

Phenomenon 2: Dominant subspace is NOT where the learning happens



Optimizers:

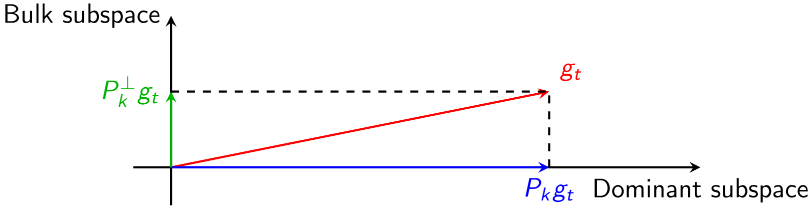
$$\theta_{t+1} \leftarrow \theta_t - \eta g_t \quad \text{(SGD)}$$

$$\theta_{t+1} \leftarrow \theta_t - \eta P_k(\theta_t) g_t \quad \text{(Dom-SGD)}$$

$$\theta_{t+1} \leftarrow \theta_t - \eta P_k^{\perp}(\theta_t) g_t \quad \text{(Bulk-SGD)}$$

where g_t denotes a stochastic gradient at t-th step.

Phenomenon 2: Dominant subspace is NOT where the learning happens



Optimizers:

$$\theta_{t+1} \leftarrow \theta_t - \eta g_t \qquad \text{(SGD)}$$
$$\theta_{t+1} \leftarrow \theta_t - \eta P_k(\theta_t) g_t \qquad \text{(Dom-SGD)}$$
$$\theta_{t+1} \leftarrow \theta_t - \eta P_k^{\perp}(\theta_t) g_t \qquad \text{(Bulk-SGD)}$$

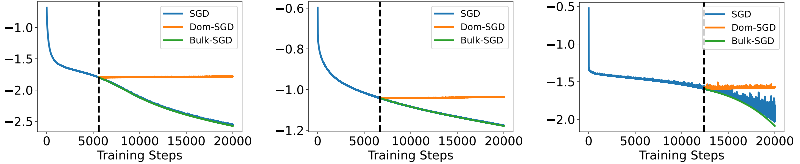
where g_t denotes a stochastic gradient at t-th step.

- Since gradient aligns with dominant subspace, we may expect: *Dom-SGD is as effective as SGD, but Bulk-SGD isn't.*

Phenomenon 2: Dominant subspace is NOT where the learning happens

Experiment: We switch from SGD to Dom-SGD/Bulk-SGD after gradient aligns with the dominant subspace.

Training loss curves (log-scale):



(a) MLP on MNIST (b) CNN on CIFAR10 (c) Transformer on SST2

- Surprisingly, Dom-SGD fails to further decrease the loss.
- In contrast, Bulk-SGD is as effective as SGD.

The "spurious" alignment between gradient and dominant subspace.

Phenomenon 3: The "spurious" alignment is due to the stochastic noise

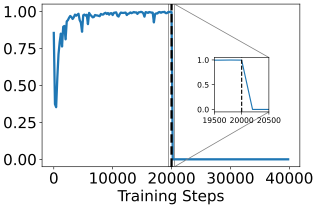
Q. *What causes the "spurious" alignment?*

Phenomenon 3: The "spurious" alignment is due to the stochastic noise

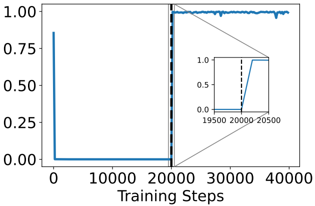
Q. *What causes the "spurious" alignment?*

Experiment: We switch from (a) SGD to GD, and (b) GD to SGD.

Fraction of (full-batch) gradient in the dominant subspace $\chi_k(\nabla L)$:



(a) SGD to GD



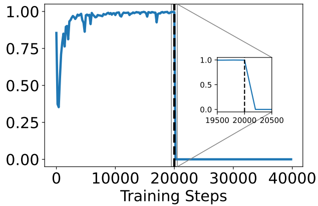
(b) GD to SGD

Phenomenon 3: The "spurious" alignment is due to the stochastic noise

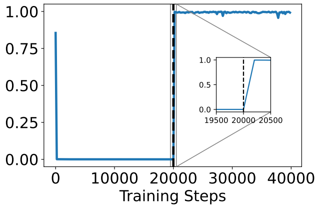
Q. *What causes the "spurious" alignment?*

Experiment: We switch from (a) SGD to GD, and (b) GD to SGD.

Fraction of (full-batch) gradient in the dominant subspace $\chi_k(\nabla L)$:



(a) SGD to GD

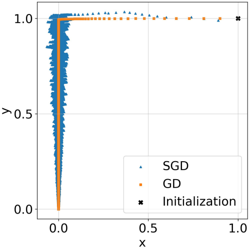


(b) GD to SGD

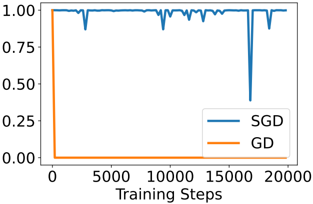
A. *The "spurious" alignment is caused by stochastic noise of SGD.*

Toy model experiment

Ill-conditioned quadratic loss $L(x, y) = \frac{1}{2}(1000x^2 + y^2)$:



(a) (S)GD trajectory



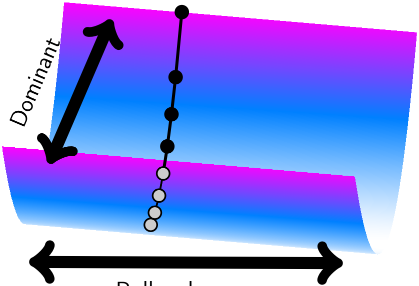
(b) $\chi_1(\nabla L(\theta_t))$

Toy model recovers all the observed phenomena (Phenomena 1–3)

- SGD oscillates in the high-curvature (dominant) direction, resulting in gradient alignment, but training progresses in the flat (bulk) direction.

Our mental model of loss landscape in DNN training

Our mental model: Ill-conditioned valley loss landscape



- SGD's noise bumps parameters up the steep walls (dominant direction), but "true" training progress happens along the bottom of a narrow and steep valley (bulk direction).

Key takeaway

DNN cannot be trained within the dominant subspace, and bulk subspace plays an essential role during training.

Key takeaway

DNN cannot be trained within the dominant subspace, and bulk subspace plays an essential role during training.

- We extend our observations to practical settings, including the large learning rate regime (Edge of Stability), Sharpness-Aware Minimization (SAM), momentum, and adaptive optimizers.
- For more details, see our paper or visit our poster session!

Contact: `minhaksong@kaist.ac.kr`