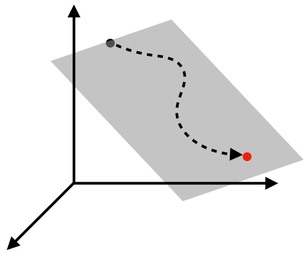
A geometric view of *intrinsic dimension* and *lottery subspaces*

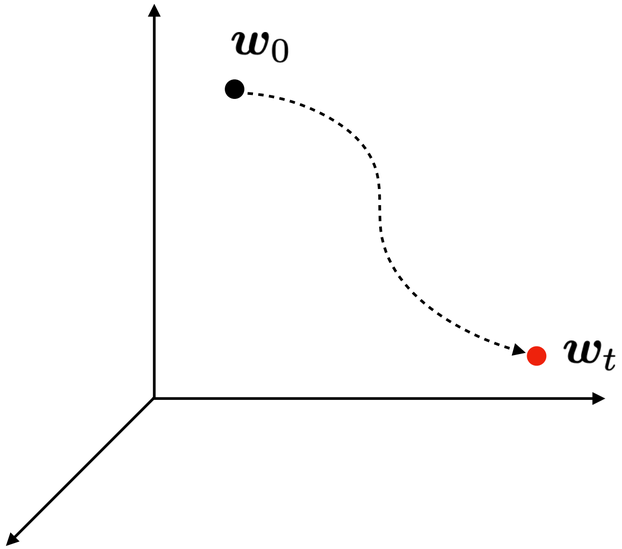


"Tony" Runzhe Yang

https://runzhe-yang.science

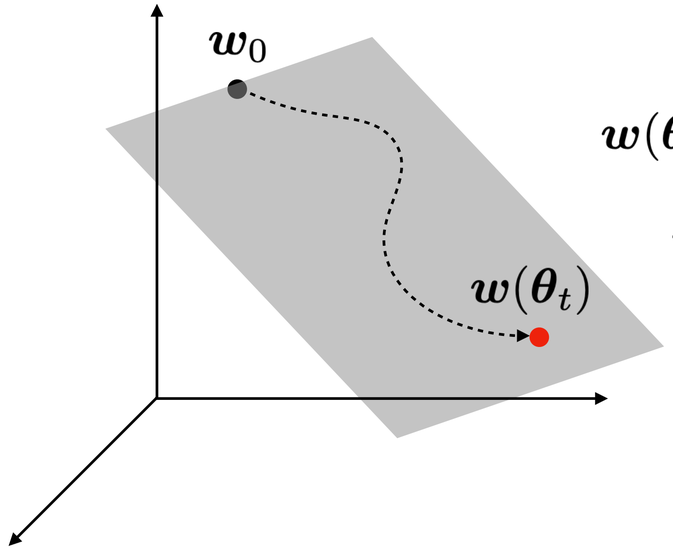
What happens if we limit the degrees of freedom of training a neural network?

Training a model with *D* parameters



$$\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \eta \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_{t-1})$$

Training a model with *D* parameters in a *d*-dimensional subspace

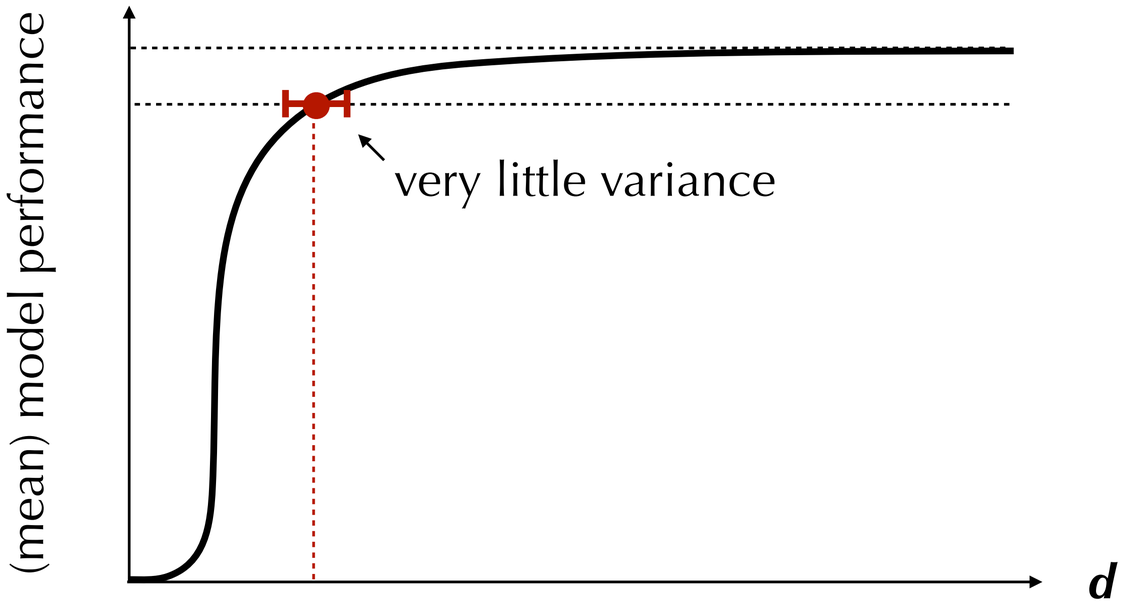


$$\boldsymbol{w}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{w}_0$$

$$\mathbf{A} \in \mathbb{R}^{D \times d}$$

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{w}(\boldsymbol{\theta}_{t-1}))$$

A small random subspace is enough to train a good model



$d_{int} << D$

"intrinsic dimension"

The intrinsic dimension measures invariance of objective landscape

Fully connected networks, on MNIST

The intrinsic dimension is invariant to model size.

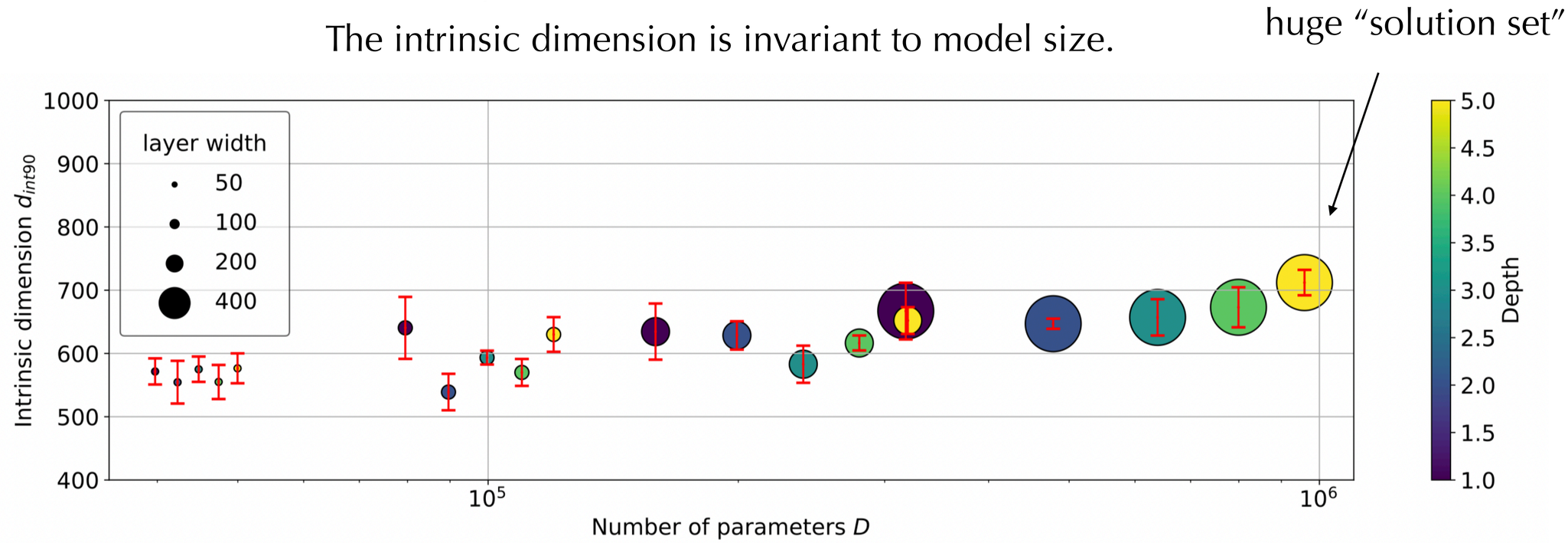
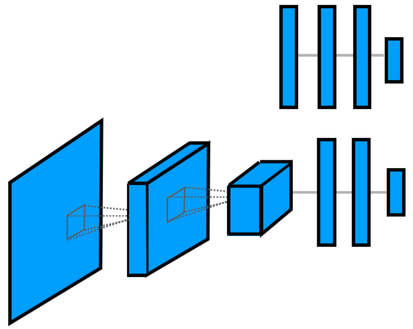
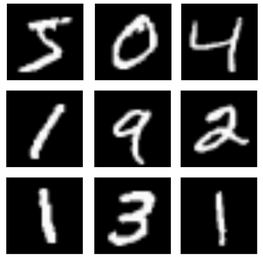
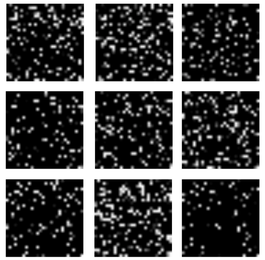


Fig S7 from Li et al. Measuring the intrinsic dimension of objective landscapes. ICLR’18

Intrinsic dimension approximates minimal description length (MDL)



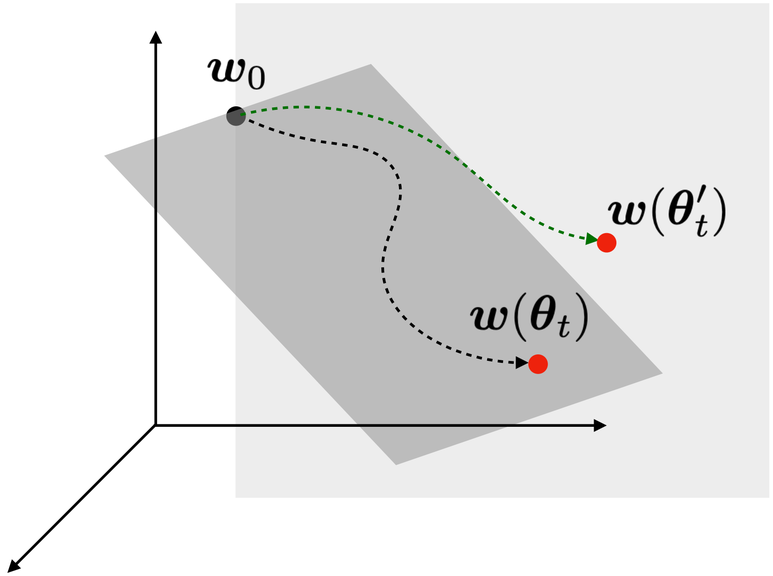
	MNIST	MNIST, shuffled pixels
		
FCN	d_{int}=750	d_{int}=750
CNN	d_{int}=290	d_{int}=1400

Modified from Blog: https://eng.uber.com/intrinsic-dimension/

Why do we always get a "lucky" subspaces when $d \geq d_{int}$?

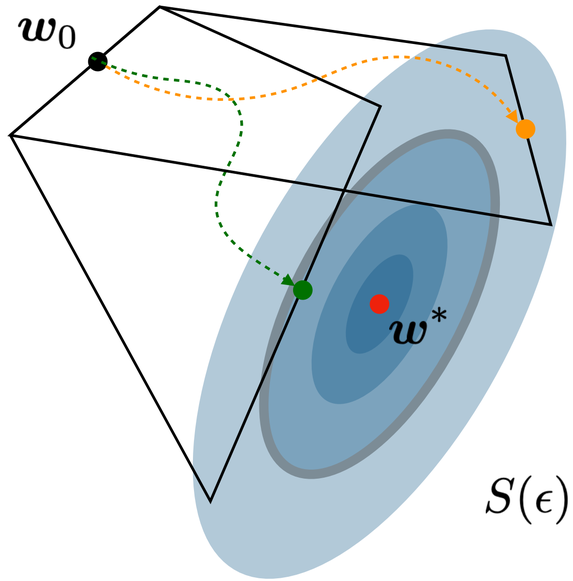
(and why always unlucky subspaces when $d < d_{int}$)

Training a model with *D* parameters in a *d*-dimensional subspace



$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{w}(\boldsymbol{\theta}_{t-1}))$$

Success probability in hitting a sub-level set

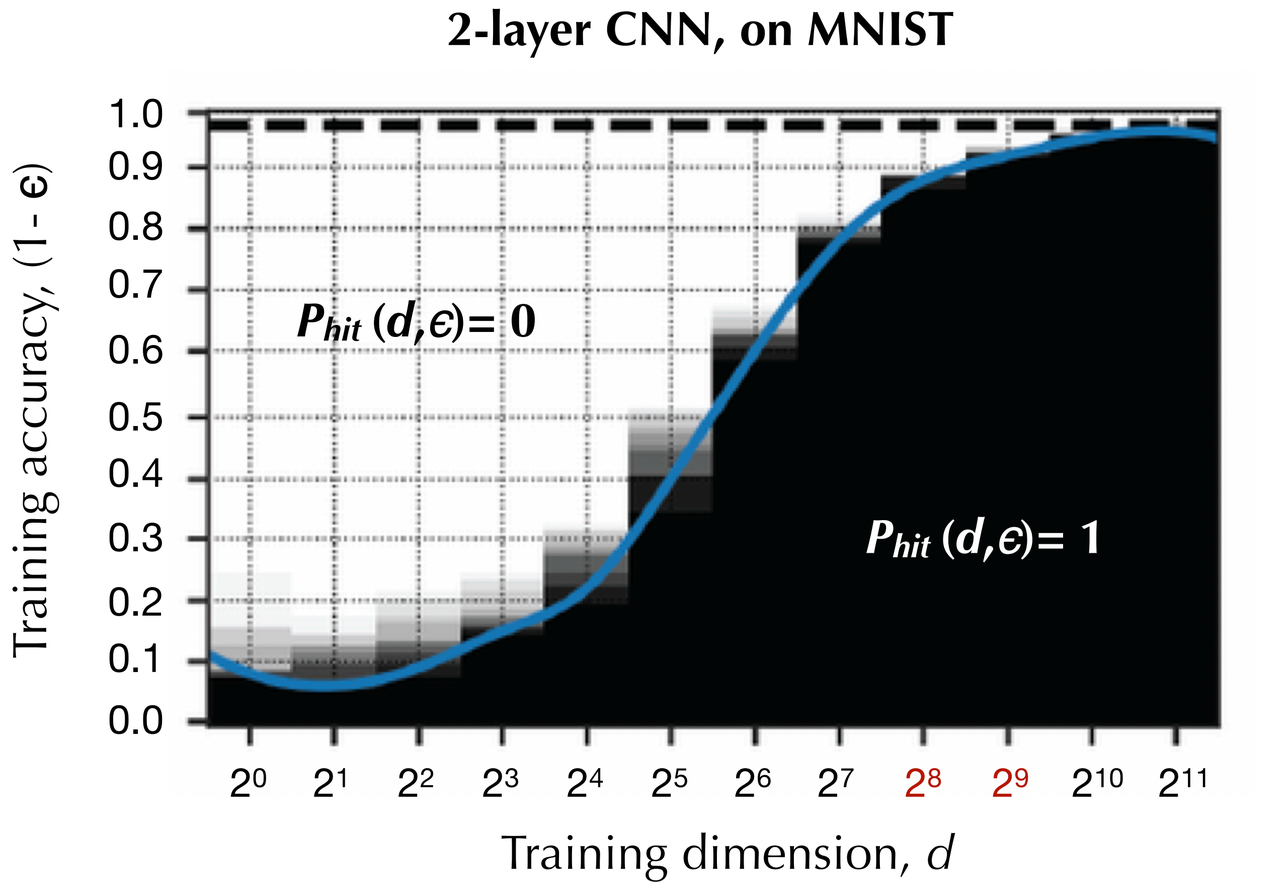


sub-level set

$$S(\epsilon) := \{\boldsymbol{w} \in \mathbb{R}^D : \mathcal{L}(\boldsymbol{w}) \leq \epsilon\}$$

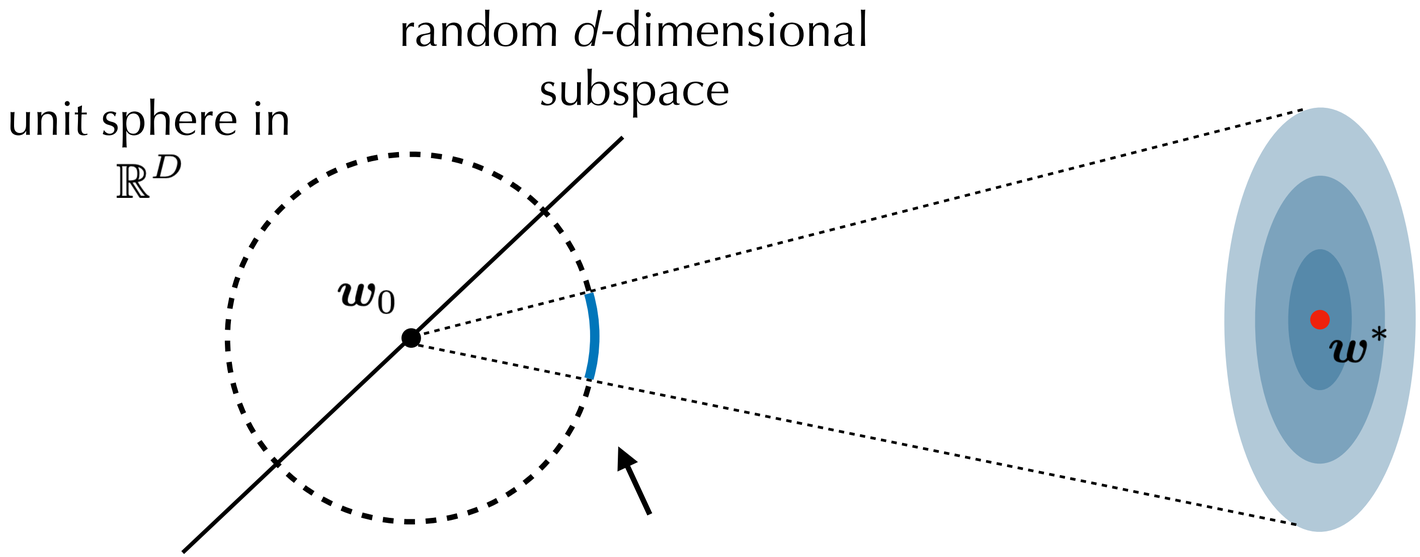
$$P_{hit}(d, \epsilon) = \mathbb{P}\left[S(\epsilon) \cap \{\boldsymbol{w}_0 + \mathrm{span}(\mathbf{A})\} \neq \emptyset\right]$$

A sharp phase transition in ϵ by *d* plane



Modified from Fig.2 in Larsen et al. How many degrees of freedom do we need to train deep networks? A loss landscape perspective, ICLR'22

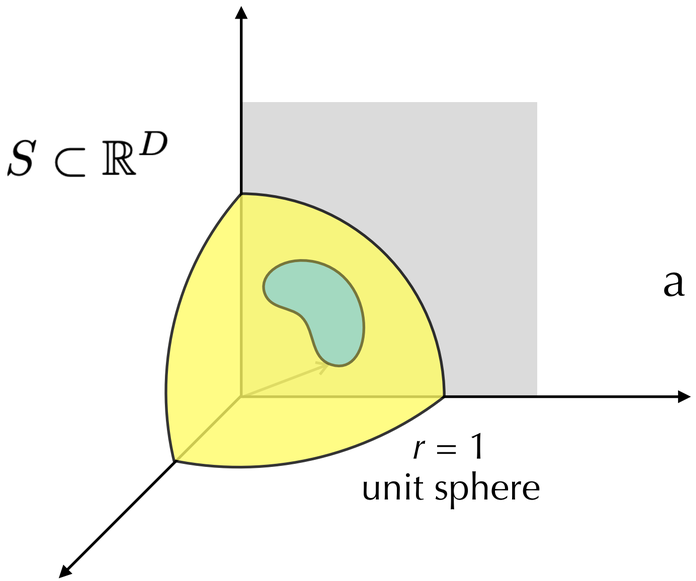
Equivalence to hitting the sub-level projection on a unit sphere



$$\text{proj}_{\boldsymbol{w}_0}(S(\epsilon)) = \left\{ \frac{\boldsymbol{x} - \boldsymbol{w}_0}{\|\boldsymbol{x} - \boldsymbol{w}_0\|_2} : \boldsymbol{x} \in S(\epsilon) \right\}$$

$$\Rightarrow P_{hit}(d, \epsilon) = \mathbb{P}\left[\text{proj}_{\boldsymbol{w}_0}(S(\epsilon)) \cap \text{span}(\mathbf{A}) \neq \emptyset\right]$$

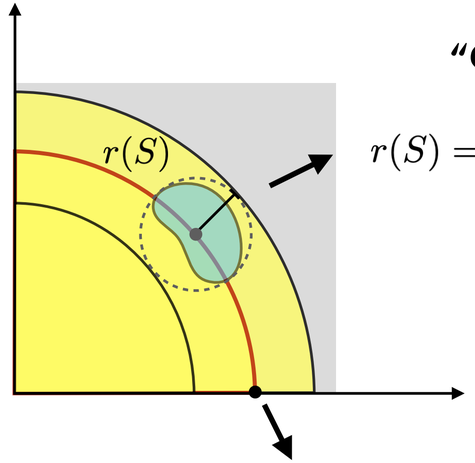
Picturing random projections in a high dimension space



$$\mathbf{G} \in \mathbb{R}^{k \times D}$$
$$G_{ij} \sim \mathcal{N}(0, 1)$$

random projection to
a *k-dimensional subspaces*

a.s.



"Gaussian width"

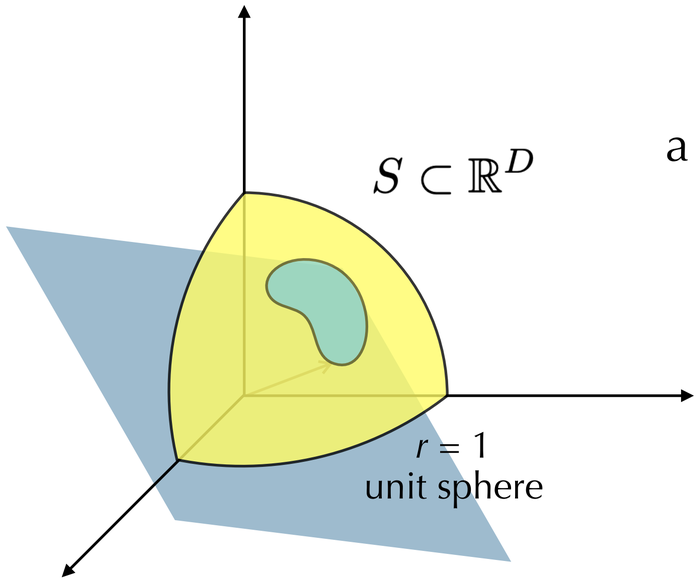
$$r(S) = \frac{1}{2}\mathbb{E} \sup_{\boldsymbol{x},\boldsymbol{y} \in S} \langle \boldsymbol{g}_D, \boldsymbol{x} - \boldsymbol{y} \rangle$$

$$\alpha_k = \mathbb{E}[\|\boldsymbol{g}_k\|]$$
$$= \sqrt{2}\frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \approx \sqrt{k - \frac{1}{2}}$$

mean of Chi distribution

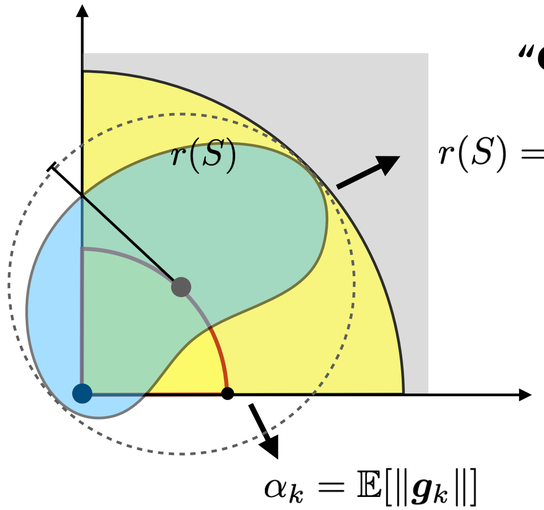
$$\mathbb{E}\left[\min_{\boldsymbol{x} \in S} \|\mathbf{G}\boldsymbol{x}\|_2\right] \geq \alpha_k - r(S) \qquad \mathbb{E}\left[\max_{\boldsymbol{x} \in S} \|\mathbf{G}\boldsymbol{x}\|_2\right] \leq \alpha_k + r(S)$$

High probability of intersection with the null space (when $\alpha_k \leq r(S)$)



$$\mathbf{G} \in \mathbb{R}^{k \times D}$$
$$G_{ij} \sim \mathcal{N}(0,1)$$

random projection to
a *k-dimensional subspaces*



"Gaussian width"

$$r(S) = \frac{1}{2}\mathbb{E} \sup_{\boldsymbol{x},\boldsymbol{y} \in S} \langle \boldsymbol{g}_D, \boldsymbol{x} - \boldsymbol{y} \rangle$$

$$\mathrm{span}(\mathbf{A}) = \mathrm{null}(\mathbf{G})$$
$$d + k = D$$

When $\alpha_k \leq r(S)$, or approximately $k \leq r^2(S)$,
w.h.p. the null space of $\mathbf{G}$ intersect with ball(S).

$$\Rightarrow$$

$$d \geq D - \underbrace{r^2(\mathrm{proj}_{\boldsymbol{w}_0}(S(\epsilon)))}_{\text{"local angular dimension"}}$$

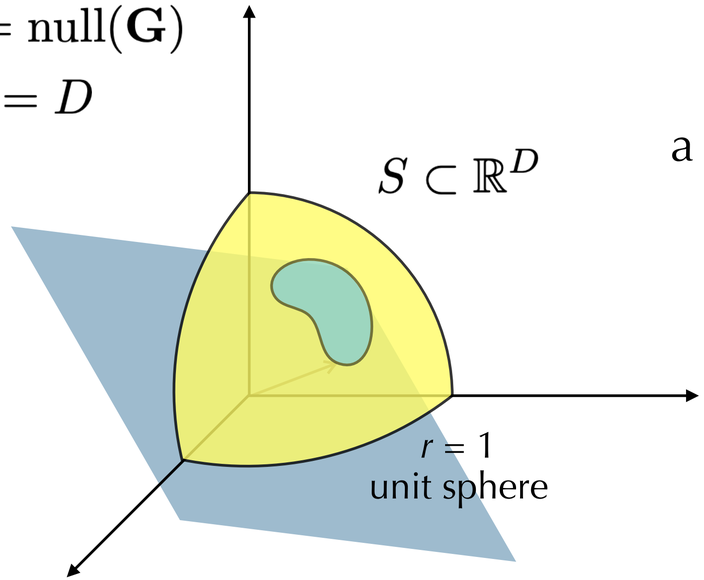
$$P_{hit}(d, \epsilon) \geq 1 - \exp\{-m_{d,\epsilon}\}$$

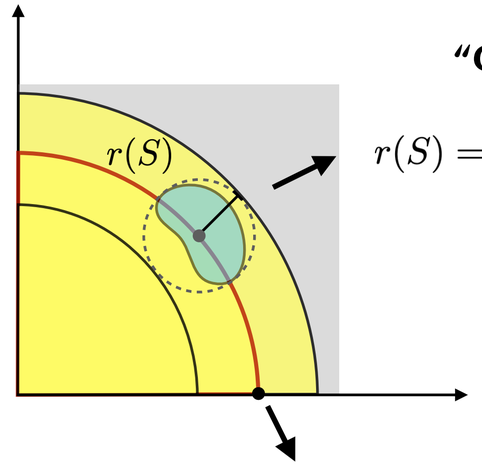
Gordon's escape from a mesh theorem (when $\alpha_k > r(S)$)

$$\mathbf{G} \in \mathbb{R}^{k \times D}$$
$$G_{ij} \sim \mathcal{N}(0,1)$$

random projection to
a *k-dimensional subspaces*

$$\text{span}(\mathbf{A}) = \text{null}(\mathbf{G})$$
$$d + k = D$$





"Gaussian width"

$$r(S) = \frac{1}{2}\mathbb{E} \sup_{\boldsymbol{x},\boldsymbol{y} \in S} \langle \boldsymbol{g}_D, \boldsymbol{x} - \boldsymbol{y} \rangle$$

$$\alpha_k = \mathbb{E}[\|\boldsymbol{g}_k\|]$$

$$\mathbb{P}\left[S \cap \{\text{span}(\mathbf{A})\} = \emptyset\right] = \mathbb{P}\left[\min_{\substack{\boldsymbol{x} \in S \\ \boldsymbol{y} \in \text{null}(\mathbf{G})}} \|\boldsymbol{x} - \boldsymbol{y}\|_2 > 0\right]$$

$$\geq \lim_{\epsilon \to 0} \mathbb{P}\left[\min_{\substack{\boldsymbol{x} \in S \\ \boldsymbol{y} \in \text{null}(\mathbf{G})}} \|\mathbf{G}(\boldsymbol{x} - \boldsymbol{y})\|_2 \geq \epsilon(1+\delta)\mathbb{E}\|\mathbf{G}\|_2\right]$$
$$- \exp\{-\delta^2 \alpha_k^2 / 2\}$$
$$\geq 1 - \frac{7}{2}\exp\left\{-\frac{1}{18}\left(\alpha_k - r(S)\right)^2\right\}$$

$$\Rightarrow \quad d < D - \underbrace{r^2(\text{proj}_{\boldsymbol{w}_0}(S(\epsilon)))}_{\text{"local angular dimension"}}$$
$$P_{hit}(d, \epsilon) \leq \exp\{-m_{1-d,1-\epsilon}\}$$

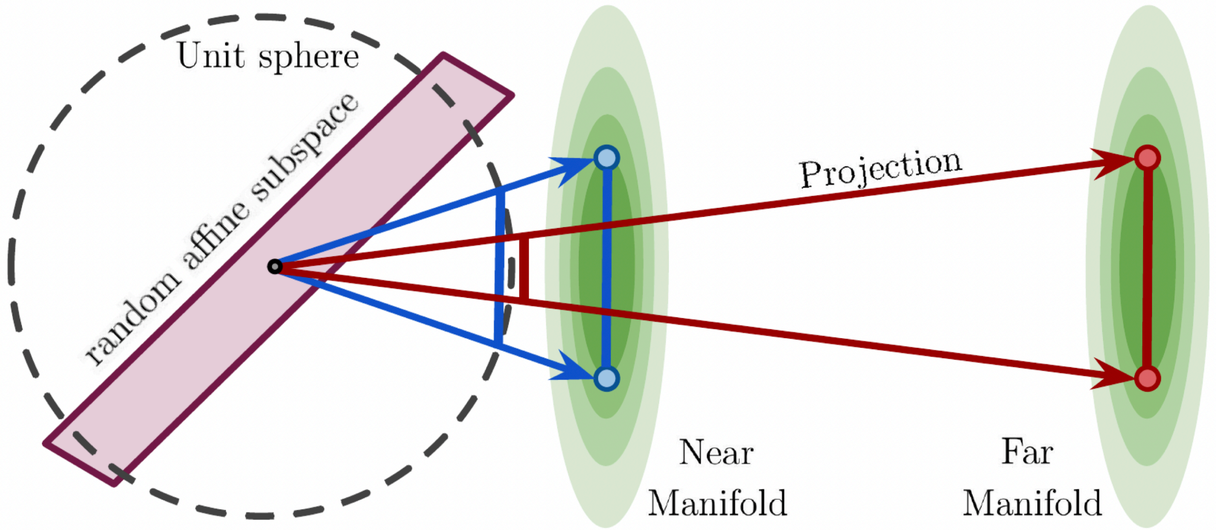
Intrinsic dimension as a function of initial point and sub-level set

$\epsilon \uparrow$, $d_{int} \downarrow$

$$d_{int} = D - r^2(\mathrm{proj}_{\boldsymbol{w}_0}(S(\epsilon)))$$

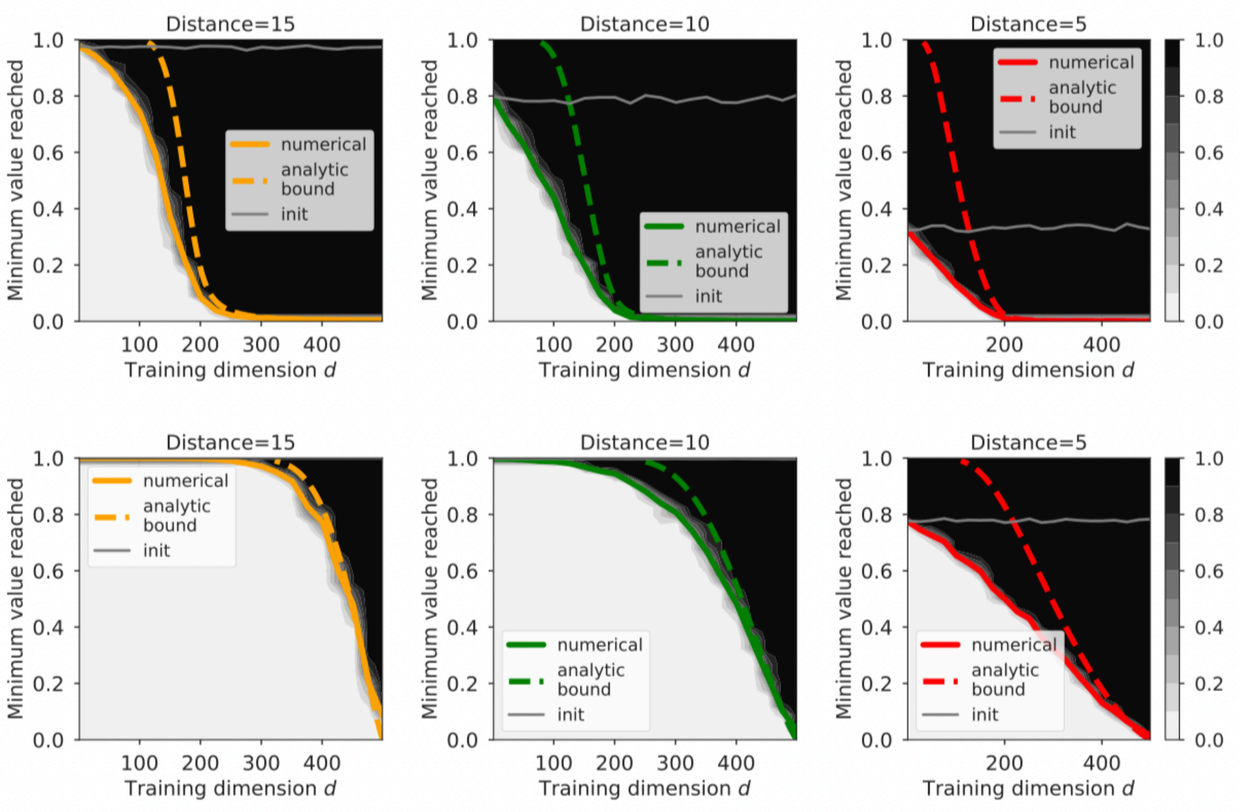
random subspace training after $\boldsymbol{w}_t$

if $\boldsymbol{w}_t$ is closer to $S(\boldsymbol{\epsilon})$, $d_{int} \downarrow$

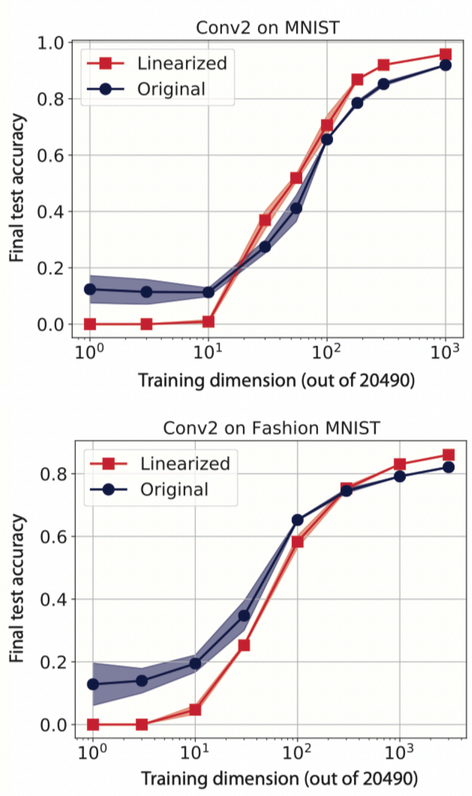


Modified from Fig.4 in Larsen et al. How many degrees of freedom do we need to train deep networks? A loss landscape perspective, ICLR'22

Deriving an analytical upper bound on intrinsic dimension



Optimizing $\mathcal{L}(w) = w^{\intercal}\mathbf{H}w$, w/ different init. An analytical upper bound on $d_{int.}$



Linearized networks using NTK.

Modified from Fig.5 and Fig.9 in Larsen et al. How many degrees of freedom do we need to train deep networks? A loss landscape perspective, ICLR'22

An analogy to lottery tickets (Frankle & Carbin, 2019)

	Training not used	Training used for init. only	Training used for pruning only	Training used for init. and pruning
Axis-aligned subspaces	Random weight pruning	Random weight pruning at step t	Lottery tickets, rewound to init.	Lottery tickets, rewound to step t
General subspaces	Random affine subspaces	Burn-in affine at step t	Lottery subspaces	Lottery subspaces at step t

smaller intrinsic dimension required

"Lottery subspace": $w(\theta) = \mathrm{U}_d\boldsymbol{\theta} + \boldsymbol{w}_t$

top *d* principal component of a training trajectory

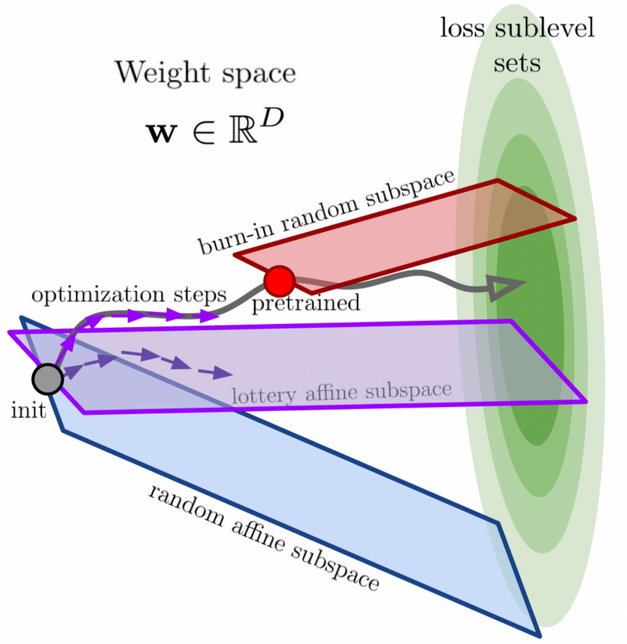


Fig.1 and Table 1 from Larsen et al. How many degrees of freedom do we need to train deep networks? A loss landscape perspective, ICLR'22

High “compression ratio” (*D/d*) compared with pruning

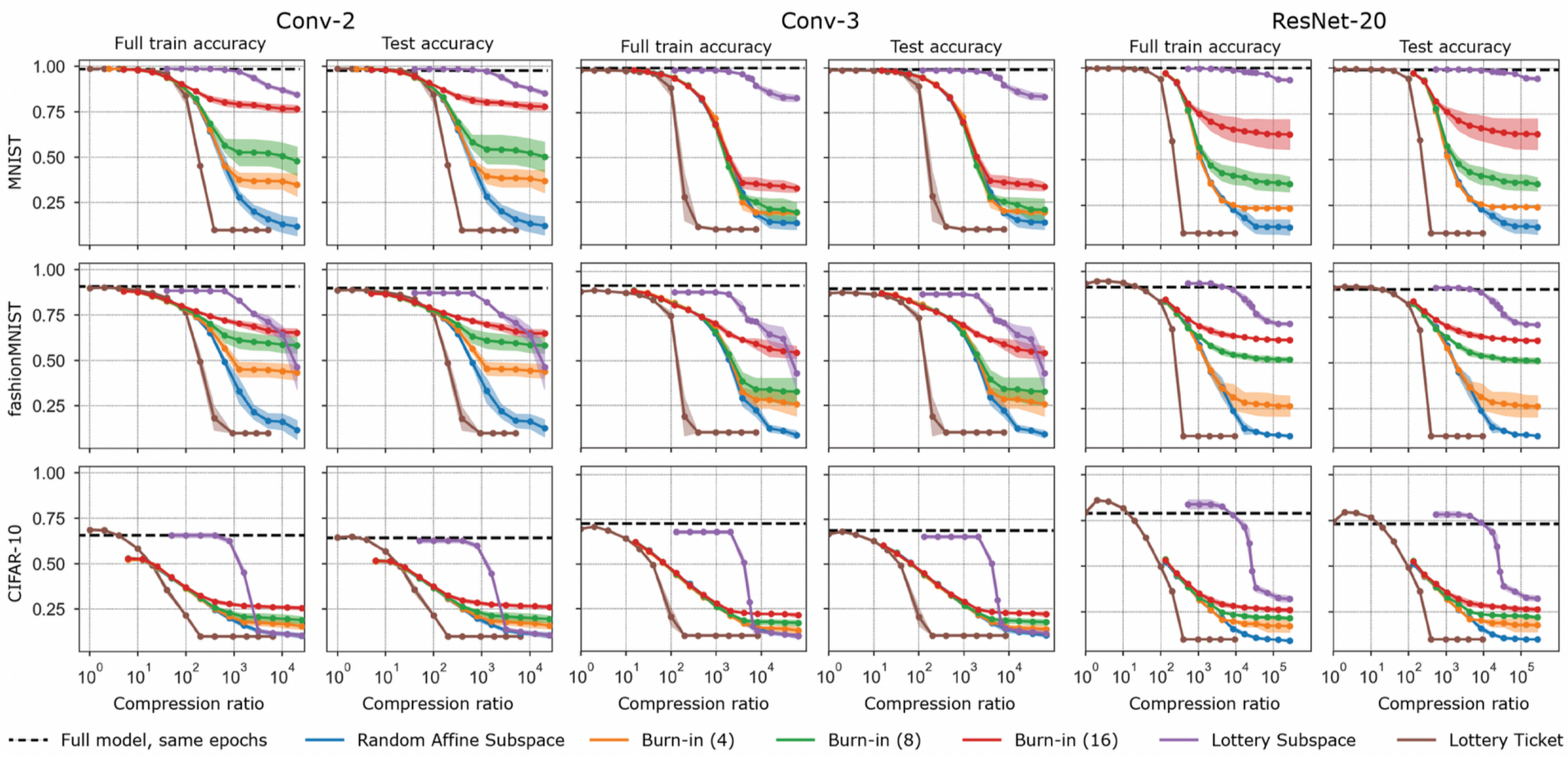
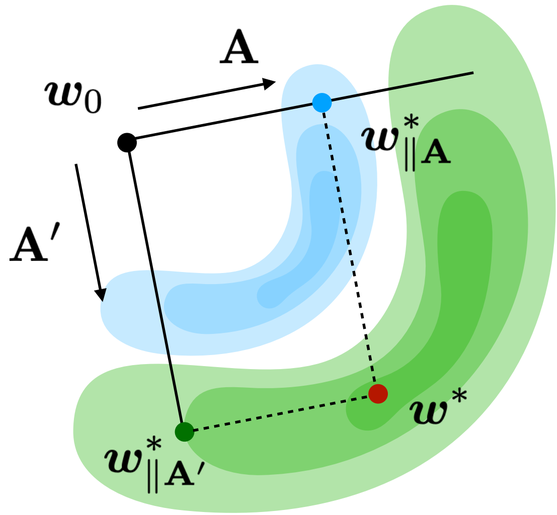


Fig.6 from Larsen et al. How many degrees of freedom do we need to train deep networks? A loss landscape perspective, ICLR’22

Discussion

- When $d \geq d_{int}$ and d << D, w.h.p. two random affine subspace span{A} and span{A′} hit the solution
  - A and A′ are nearly orthogonal: Pr[span{A} $\perp$ span{A′}] ≃ 0
  - The training trajectory in A has zero projection in A′, vice versa
  - Can we make use of the redundant degrees of freedom?

$$w(\{\theta^{(i)}\}_{i=1}^{K}) = \sum_{i=1}^{K} \mathbf{A}^{(i)}\theta^{(i)} + w_0$$



- ~4 x subnets: Havasi et al. MIMO. ICLR'21
  ~40 x inputs: Murahari et al. DataMux. 2022

- Larger D/d_{int} for (transformers, text)?
  Warm-up allows more subnets?
  Disentangled readout from all layers / via (soft) weight masking?