A Derandomization Framework for Structure Discovery: Applications in Neural Networks and Beyond

Nikos Tsikouras[1,2] *Yorgos Pantis*[1,2] *Ioannis Mitliagkas*[2,3] *Christos Tzamos*[1,2]

[1]National and Kapodistrian University of Athens, Greece [2]Archimedes, Athena Research Center, Greece [3]Mila & Université de Montréal, Canada

ICLR

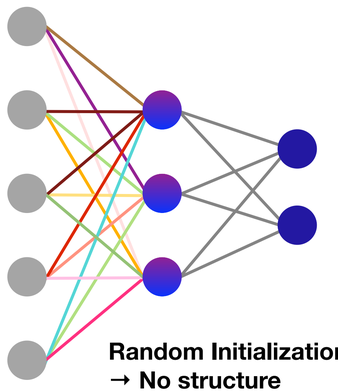
ARCHIMEDES Mila

Université de Montréal

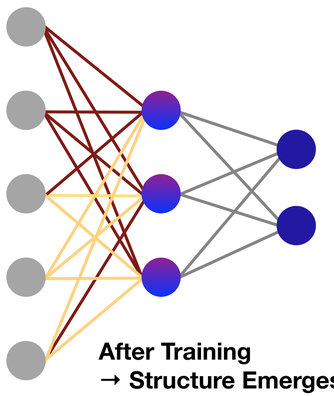
HELLENIC REPUBLIC
National and Kapodistrian University of Athens

It is empirically known that the first-layer weights **collapse** to a low-rank subspace.
Prior work needed strong assumptions, i.e. strong regularization, simple architecture to explain this.

Can low-rank structure emerge under more natural assumptions?



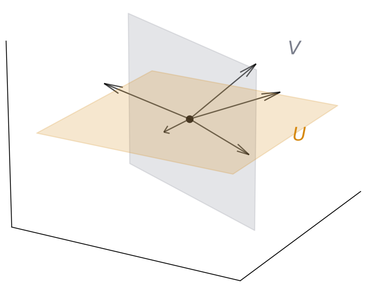
Random Initialization
→ No structure



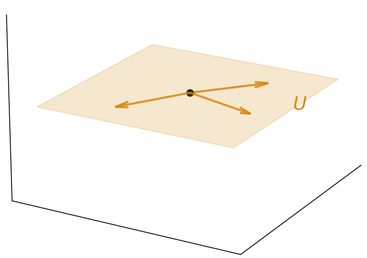
After Training
→ Structure Emerges

More specifically:

At initialization, first layer weight spans the full space.



After training, first layer weights converge to hidden structure. **Why?**



The solution comes from a natural solution concept, ρ-second-order stationary points (ρ-SOSPs).

- Standard GD/SGD almost surely avoid strict saddles.
- GD with noise efficiently converges to ρ-SOSP.

Definition (ρ-SOSP): A point $\mathbf{x}^*$ satisfying:
$\|\nabla f(\mathbf{x}^*)\|_2 \le \rho. \quad \|\lambda_{\min}(\nabla^2 f(\mathbf{x}^*))\|_2 \ge -\sqrt{K\rho}.$

Main Contribution

Consider a general family of the form:

$$f(\mathbf{W}, \mathbf{b}; \theta) = \mathbb{E}_x \left[g_\theta(\mathbf{W}\mathbf{x} + \mathbf{b}) \right] + \lambda \|\mathbf{W}\|_F^2,$$

where $\lambda > 0$ is an **arbitrarily small regularization** parameter. Our main result is presented below.

Key Derandomization Lemma

Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ be a standard multivariate Gaussian random variable. For the objective function defined above, with $\lambda > \sqrt{K\rho}/2$ where $g_\theta(\,\cdot\,)$ is both L-smooth and K-Hessian Lipschitz, any ρ-SOSP satisfies: $\|\mathbf{W}\|_F \le \rho/(2\lambda - \sqrt{K\rho})$.

We use this to get results in:
- Structure Discovery in NN
- Derandomizing JL
- Derandomizing GW MAXCUT

How does this translate to NNs?

Essentially, this shows why optimization drives the first-layer weights to lie within any signal direction, eliminating any component perpendicular to it.

Let $\mathbf{x} \in \mathbb{R}^d$ be a standard Gaussian distribution. The target labels are generated by a ***teacher*** multiple-index teacher model:

$$y = h(\langle \mathbf{u}_1, \mathbf{x} \rangle, \ldots, \langle \mathbf{u}_k, \mathbf{x} \rangle; \epsilon) \equiv h(U\mathbf{x}; \epsilon).$$

Let $\mathbf{W} \in \mathbb{R}^{k \times d}$ and $\mathbf{b} \in \mathbb{R}^k$, and consider the first layer of a NN given by $\mathbf{W}\mathbf{x} + \mathbf{b}$. We decompose this into components that are parallel and perpendicular to the signal subspace U, as follows:

$$\mathbf{W}\mathbf{x} + \mathbf{b} = \mathbf{W}_{\|}\mathbf{x}_{\|} + \mathbf{W}_{\perp}\mathbf{x}_{\perp} + \mathbf{b}.$$

Using this decomposition, we can write the risk population risk as:

$$R(\mathbf{W}_{\perp}, \mathbf{b}; \theta') = \mathbb{E}_{\mathbf{x}_{\perp}} \left[\ell'_{\theta'} \left(\mathbf{W}_{\perp}\mathbf{x}_{\perp} + \mathbf{b} \right) \right] + \lambda \|\mathbf{W}_{\perp}\|_F^2,$$

where $\ell'_{\theta'}(\,\cdot\,)$ is any loss function of your choice in which we have suppressed the label.

Our main theorem says that:

Theorem 4.1 (Informal)

Under mild assumptions, any $\rho-$SOSP for $R(\mathbf{W}, \mathbf{b}; \theta')$, with respect to $(\mathbf{W}, \mathbf{b})$ yields a weight matrix that satisfies: $\|\mathbf{W}_{\perp}\| \le \dfrac{\rho}{2\lambda - \sqrt{K\rho}}$.

We also show that you can reach such a point in polynomial time version (Theorem 4.2).