Effective Theory of Deep Learning

Beyond the Infinite-Width Limit

Dan Roberts[a] and Sho Yaida[b]

[a]MIT, IAIFI, & Salesforce, [b]Facebook AI Research

Deep Learning Theory Summer School at Princeton
July 27, 2021 – August 8, 2021

Course Plan

~~Lecture 1~~ **~~Initialization~~, ~~Linear Models~~**

- ~~§0 + §7.1 + §10.4~~

~~Lecture 2~~ **~~Quadratic Models & Nearly-Kernel Methods~~**

- ~~§11.4 (+ §7.2) + §∞.2.2~~

~~Lecture 3~~ **~~The Principle of Sparsity (Recursing)~~**

- ~~§4, §8, §11.2, §∞.3~~

~~Lecture 4~~ **~~The Principle of Criticality~~**

- ~~§5, §9, §11.3, §∞.1, +§10.3.1~~

Lecture 5 **The End of Training, & More**

- §∞.2.3 + §10.3.2 + §A.3 + §ϵ

The End of Training

$$
\begin{aligned}
&z_{i;\dot\beta}(t=\infty)\\
=&z_{i;\dot\beta} - \sum_{j,k,\tilde\alpha_1,\tilde\alpha_2} \widehat{H}_{ij;\dot\beta\tilde\alpha_1}\left(\widehat{H}^{-1}\right)_{jk}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{k;\tilde\alpha_2}-y_{k;\tilde\alpha_2}\right)\\
&+\sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2}-\sum_{\tilde\alpha_5,\tilde\alpha_6}H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right]Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
&+\sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2}-\sum_{\tilde\alpha_5,\tilde\alpha_6}H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right]Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right]Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right]Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
&+O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{j,k,\tilde\alpha_1,\tilde\alpha_2} \widehat{H}_{ij;\dot\beta\tilde\alpha_1} \left(\widehat{H}^{-1}\right)_{jk}^{\tilde\alpha_1\tilde\alpha_2} \left(z_{k;\tilde\alpha_2} - y_{k;\tilde\alpha_2}\right) \\
&+ \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{j_1 ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{j_1 ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2} \right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
&+ \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2} \right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
&+ \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{j_1 ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{j_1 ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
&+ \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
&+ \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3} \right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
&+ \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
&+ O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

Complete Inverse of the Stochastic NTK

Here, we needed the complete inverse of the stochastic NTK:

$$\sum_{j,\tilde{\alpha}_2} \left(\widehat{H}^{-1}\right)_{ij}^{\tilde{\alpha}_1\tilde{\alpha}_2} \widehat{H}_{jk;\tilde{\alpha}_2\tilde{\alpha}_3} = \delta_{ik}\delta^{\tilde{\alpha}_1}_{\ \tilde{\alpha}_3} \, .$$

Complete Inverse of the Stochastic NTK

Here, we needed the complete inverse of the stochastic NTK:

$$\sum_{j,\tilde{\alpha}_2} \left(\widehat{H}^{-1}\right)_{ij}^{\tilde{\alpha}_1\tilde{\alpha}_2} \widehat{H}_{jk;\tilde{\alpha}_2\tilde{\alpha}_3} = \delta_{ik}\delta^{\tilde{\alpha}_1}_{\ \tilde{\alpha}_3} .$$

Defining a mean and fluctuation as

$$\widehat{H}_{i_1 i_2;\alpha_1\alpha_2} \equiv \delta_{i_1 i_2} H_{\alpha_1\alpha_2} + \widehat{\Delta H}_{i_1 i_2;\alpha_1\alpha_2} ,$$

we can then expand around the fluctuation to get:

$$\begin{aligned}\left(\widehat{H}^{-1}\right)_{ij}^{\tilde{\alpha}_1\tilde{\alpha}_2} =& \delta_{ij}\widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_3,\tilde{\alpha}_4\in\mathcal{A}} \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_3}\widehat{\Delta H}_{ij;\tilde{\alpha}_3\tilde{\alpha}_4}\widetilde{H}^{\tilde{\alpha}_4\tilde{\alpha}_2} \\ &+ \sum_{k=1}^{n_L}\sum_{\tilde{\alpha}_3,\ldots,\tilde{\alpha}_6\in\mathcal{A}} \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_3}\widehat{\Delta H}_{ik;\tilde{\alpha}_3\tilde{\alpha}_4}\widetilde{H}^{\tilde{\alpha}_4\tilde{\alpha}_5}\widehat{\Delta H}_{kj;\tilde{\alpha}_5\tilde{\alpha}_6}\widetilde{H}^{\tilde{\alpha}_6\tilde{\alpha}_2} \\ &+ O\!\left(\Delta^3\right) .\end{aligned}$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{j,k,\tilde\alpha_1,\tilde\alpha_2} \widehat{H}_{ij;\dot\beta\tilde\alpha_1} \left(\widehat{H}^{-1}\right)_{jk}^{\tilde\alpha_1\tilde\alpha_2} \left(z_{k;\tilde\alpha_2} - y_{k;\tilde\alpha_2}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{i j_1 j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{i j_1 j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{i;\tilde\alpha_2}-y_{i;\tilde\alpha_2}\right) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{j;\tilde\alpha_2}-y_{j;\tilde\alpha_2}\right) \\
& - \sum_{\substack{j,k\\ \tilde\alpha_1,\ldots,\tilde\alpha_4}}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\left(z_{k;\tilde\alpha_4}-y_{k;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
= & z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{i;\tilde\alpha_2}-y_{i;\tilde\alpha_2}) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{j;\tilde\alpha_2}-y_{j;\tilde\alpha_2}) \\
& - \sum_{\substack{j,k\\ \tilde\alpha_1,\ldots,\tilde\alpha_4}}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}(z_{k;\tilde\alpha_4}-y_{k;\tilde\alpha_4}) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

Two Extra Differentials??

$$
\widehat{\mathrm{dd}_{\mathrm{I}} H}^{(\ell)}_{i_0 i_1 i_2 i_3;\delta_0\delta_1\delta_2\delta_3}
\equiv \sum_{\ell_1,\ell_2,\ell_3=1}^{\ell} \sum_{\substack{\mu_1,\nu_1,\\ \mu_2,\nu_2,\\ \mu_3,\nu_3}} \lambda^{(\ell_1)}_{\mu_1\nu_1}\lambda^{(\ell_2)}_{\mu_2\nu_2}\lambda^{(\ell_3)}_{\mu_3\nu_3} \frac{d^3 z^{(\ell)}_{i_0;\delta_0}}{d\theta^{(\ell_1)}_{\mu_1} d\theta^{(\ell_2)}_{\mu_2} d\theta^{(\ell_3)}_{\mu_3}} \frac{dz^{(\ell)}_{i_1;\delta_1}}{d\theta^{(\ell_1)}_{\nu_1}} \frac{dz^{(\ell)}_{i_2;\delta_2}}{d\theta^{(\ell_2)}_{\nu_2}} \frac{dz^{(\ell)}_{i_3;\delta_3}}{d\theta^{(\ell_3)}_{\nu_3}}
$$

$$
\widehat{\mathrm{dd}_{\mathrm{II}} H}^{(\ell)}_{i_1 i_2 i_3 i_4;\delta_1\delta_2\delta_3\delta_4}
\equiv \sum_{\ell_1,\ell_2,\ell_3=1}^{\ell} \sum_{\substack{\mu_1,\nu_1,\\ \mu_2,\nu_2,\\ \mu_3,\nu_3}} \lambda^{(\ell_1)}_{\mu_1\nu_1}\lambda^{(\ell_2)}_{\mu_2\nu_2}\lambda^{(\ell_3)}_{\mu_3\nu_3} \frac{d^2 z^{(\ell)}_{i_1;\delta_1}}{d\theta^{(\ell_1)}_{\mu_1} d\theta^{(\ell_3)}_{\mu_3}} \frac{d^2 z^{(\ell)}_{i_2;\delta_2}}{d\theta^{(\ell_2)}_{\mu_2} d\theta^{(\ell_3)}_{\nu_3}} \frac{dz^{(\ell)}_{i_3;\delta_3}}{d\theta^{(\ell_1)}_{\nu_1}} \frac{dz^{(\ell)}_{i_4;\delta_4}}{d\theta^{(\ell_2)}_{\nu_2}}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1} \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}\right) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2} \left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{j;\tilde\alpha_2} - y_{j;\tilde\alpha_2}\right) \\
& - \sum_{\substack{j,k \\ \tilde\alpha_1,\ldots,\tilde\alpha_4}} \left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\left(z_{k;\tilde\alpha_4} - y_{k;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{i j_1 j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{i j_1 j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1 i j_2 j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1 i j_2 j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{j;\tilde\alpha_2} - y_{j;\tilde\alpha_2}) \\
& - \sum_{\substack{j,k\\ \tilde\alpha_1,\ldots,\tilde\alpha_4}}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}(z_{k;\tilde\alpha_4} - y_{k;\tilde\alpha_4}) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3})(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3})(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{i;\tilde\alpha_2}-y_{i;\tilde\alpha_2}\right) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{j;\tilde\alpha_2}-y_{j;\tilde\alpha_2}\right) \\
& - \sum_{\substack{j,k\\ \tilde\alpha_1,\ldots,\tilde\alpha_4}}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\left(z_{k;\tilde\alpha_4}-y_{k;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot{\beta}}(t=\infty) \\
=& z_{i;\dot{\beta}} - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} H_{\dot{\beta}\tilde{\alpha}_1} \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_2} (z_{i;\tilde{\alpha}_2} - y_{i;\tilde{\alpha}_2}) \\
& + \sum_{j,\tilde{\alpha}_1,\tilde{\alpha}_2} \left[\widehat{\Delta H}_{ij;\dot{\beta}\tilde{\alpha}_1} - \sum_{\tilde{\alpha}_3,\tilde{\alpha}_4\in\mathcal{A}} H_{\dot{\beta}\tilde{\alpha}_3} \widetilde{H}^{\tilde{\alpha}_3\tilde{\alpha}_4} \widehat{\Delta H}_{ij;\tilde{\alpha}_4\tilde{\alpha}_1} \right] \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_2} (z_{j;\tilde{\alpha}_2} - y_{j;\tilde{\alpha}_2}) \\
& - \sum_{\substack{j,k \\ \tilde{\alpha}_1,\ldots,\tilde{\alpha}_4}} \left[\widehat{\Delta H}_{ij;\dot{\beta}\tilde{\alpha}_1} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6} H_{\dot{\beta}\tilde{\alpha}_5} \widetilde{H}^{\tilde{\alpha}_5\tilde{\alpha}_6} \widehat{\Delta H}_{ij;\tilde{\alpha}_6\tilde{\alpha}_1} \right] \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_2} \widehat{\Delta H}_{jk;\tilde{\alpha}_2\tilde{\alpha}_3} \widetilde{H}^{\tilde{\alpha}_3\tilde{\alpha}_4} (z_{k;\tilde{\alpha}_4} - y_{k;\tilde{\alpha}_4}) \\
& + \sum_{j_1,j_2,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4} \left[\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde{\alpha}_1\dot{\beta}\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6} H_{\dot{\beta}\tilde{\alpha}_5} \widetilde{H}^{\tilde{\alpha}_5\tilde{\alpha}_6} \widehat{\mathrm{d}H}_{j_1 i j_2;\tilde{\alpha}_1\tilde{\alpha}_6\tilde{\alpha}_2} \right] Z_{\mathrm{A}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} (z_{j_1;\tilde{\alpha}_3} - y_{j_1;\tilde{\alpha}_3})(z_{j_2;\tilde{\alpha}_4} - y_{j_2;\tilde{\alpha}_4}) \\
& + \sum_{j_1,j_2,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4} \left[\widehat{\mathrm{d}H}_{i j_1 j_2;\dot{\beta}\tilde{\alpha}_1\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6} H_{\dot{\beta}\tilde{\alpha}_5} \widetilde{H}^{\tilde{\alpha}_5\tilde{\alpha}_6} \widehat{\mathrm{d}H}_{i j_1 j_2;\tilde{\alpha}_6\tilde{\alpha}_1\tilde{\alpha}_2} \right] Z_{\mathrm{B}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} (z_{j_1;\tilde{\alpha}_3} - y_{j_1;\tilde{\alpha}_3})(z_{j_2;\tilde{\alpha}_4} - y_{j_2;\tilde{\alpha}_4}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6}} \left[\widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde{\alpha}_1\dot{\beta}\tilde{\alpha}_2\tilde{\alpha}_3} - \sum_{\tilde{\alpha}_7,\tilde{\alpha}_8} H_{\dot{\beta}\tilde{\alpha}_7} \widetilde{H}^{\tilde{\alpha}_7\tilde{\alpha}_8} \widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde{\alpha}_1\tilde{\alpha}_8\tilde{\alpha}_2\tilde{\alpha}_3} \right] Z_{\mathrm{IA}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} (z_{j_1;\tilde{\alpha}_4} - y_{j_1;\tilde{\alpha}_4})(z_{j_2;\tilde{\alpha}_5} - y_{j_2;\tilde{\alpha}_5})(z_{j_3;\tilde{\alpha}_6} - y_{j_3;\tilde{\alpha}_6}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6}} \left[\widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\dot{\beta}\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} - \sum_{\tilde{\alpha}_7,\tilde{\alpha}_8} H_{\dot{\beta}\tilde{\alpha}_7} \widetilde{H}^{\tilde{\alpha}_7\tilde{\alpha}_8} \widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\tilde{\alpha}_8\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} \right] Z_{\mathrm{IB}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} (z_{j_1;\tilde{\alpha}_4} - y_{j_1;\tilde{\alpha}_4})(z_{j_2;\tilde{\alpha}_5} - y_{j_2;\tilde{\alpha}_5})(z_{j_3;\tilde{\alpha}_6} - y_{j_3;\tilde{\alpha}_6}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6}} \left[\widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde{\alpha}_1\tilde{\alpha}_2\dot{\beta}\tilde{\alpha}_3} - \sum_{\tilde{\alpha}_7,\tilde{\alpha}_8} H_{\dot{\beta}\tilde{\alpha}_7} \widetilde{H}^{\tilde{\alpha}_7\tilde{\alpha}_8} \widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_8\tilde{\alpha}_3} \right] Z_{\mathrm{IIA}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} (z_{j_1;\tilde{\alpha}_4} - y_{j_1;\tilde{\alpha}_4})(z_{j_2;\tilde{\alpha}_5} - y_{j_2;\tilde{\alpha}_5})(z_{j_3;\tilde{\alpha}_6} - y_{j_3;\tilde{\alpha}_6}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6}} \left[\widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\dot{\beta}\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} - \sum_{\tilde{\alpha}_7,\tilde{\alpha}_8} H_{\dot{\beta}\tilde{\alpha}_7} \widetilde{H}^{\tilde{\alpha}_7\tilde{\alpha}_8} \widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\tilde{\alpha}_8\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} \right] Z_{\mathrm{IIB}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} (z_{j_1;\tilde{\alpha}_4} - y_{j_1;\tilde{\alpha}_4})(z_{j_2;\tilde{\alpha}_5} - y_{j_2;\tilde{\alpha}_5})(z_{j_3;\tilde{\alpha}_6} - y_{j_3;\tilde{\alpha}_6}) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
=& z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1} \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2} \left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_4} \widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1} \right] \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{j;\tilde\alpha_2} - y_{j;\tilde\alpha_2}) \\
& - \sum_{\substack{j,k \\ \tilde\alpha_1,\ldots,\tilde\alpha_4}} \left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1} \right] \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2} \widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}(z_{k;\tilde\alpha_4} - y_{k;\tilde\alpha_4}) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2} \right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} (z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3})(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{i j_1 j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{i j_1 j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2} \right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} (z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3})(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} (z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} (z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3} \right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} (z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} (z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
&z_{i;\dot\beta}(t=\infty)\\
=&z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{i;\tilde\alpha_2}-y_{i;\tilde\alpha_2})\\
&+\sum_{j,\tilde\alpha_1,\tilde\alpha_2}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}(z_{j;\tilde\alpha_2}-y_{j;\tilde\alpha_2})\\
&-\sum_{\substack{j,k\\ \tilde\alpha_1,\dots,\tilde\alpha_4}}\left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1}\right]\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}(z_{k;\tilde\alpha_4}-y_{k;\tilde\alpha_4})\\
&+\sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right]Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3})(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4})\\
&+\sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right]Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3})(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4})\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right]Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6})\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{I}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6})\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right]Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6})\\
&+\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd}_{\mathrm{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4})(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5})(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6})\\
&+O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The End of Training

$$
\begin{aligned}
& z_{i;\dot\beta}(t=\infty) \\
={}& z_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2} H_{\dot\beta\tilde\alpha_1} \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}\right) \\
& + \sum_{j,\tilde\alpha_1,\tilde\alpha_2} \left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} H_{\dot\beta\tilde\alpha_3} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_4} \widehat{\Delta H}_{ij;\tilde\alpha_4\tilde\alpha_1} \right] \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}\left(z_{j;\tilde\alpha_2} - y_{j;\tilde\alpha_2}\right) \\
& - \sum_{\substack{j,k \\ \tilde\alpha_1,\ldots,\tilde\alpha_4}} \left[\widehat{\Delta H}_{ij;\dot\beta\tilde\alpha_1} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\Delta H}_{ij;\tilde\alpha_6\tilde\alpha_1} \right] \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2} \widehat{\Delta H}_{jk;\tilde\alpha_2\tilde\alpha_3} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_4}\left(z_{k;\tilde\alpha_4} - y_{k;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\dot\beta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2} \right] Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4} \left[\widehat{\mathrm{d}H}_{i j_1 j_2;\dot\beta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\dot\beta\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_5\tilde\alpha_6} \widehat{\mathrm{d}H}_{i j_1 j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2} \right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3} - y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4} - y_{j_2;\tilde\alpha_4}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\dot\beta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\dot\beta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3} \right] Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + \sum_{\substack{j_1,j_2,j_3, \\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}} \left[\widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\dot\beta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\dot\beta\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_7\tilde\alpha_8} \widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} \right] Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4} - y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5} - y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6} - y_{j_3;\tilde\alpha_6}\right) \\
& + O\left(\frac{1}{n^2}\right)
\end{aligned}
$$

The Algorithm Projectors

Gradient Descent:

$$
\begin{aligned}
&Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \equiv Y_2^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\,,\\
&Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \equiv Y_2^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} + \frac{\eta}{2} X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\,,\\
&Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - \frac{\eta}{2} Y_4^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,\\
&Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - \frac{\eta}{2} Y_4^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - \frac{\eta}{2} Y_1^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - \frac{\eta^2}{6} X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,\\
&Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,\\
&Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - Y_3^{\tilde\alpha_2\tilde\alpha_1\tilde\alpha_3\tilde\alpha_5\tilde\alpha_4\tilde\alpha_6} - Y_3^{\tilde\alpha_1\tilde\alpha_3\tilde\alpha_2\tilde\alpha_4\tilde\alpha_6\tilde\alpha_5} - \eta Y_4^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - \eta Y_1^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,.
\end{aligned}
$$

Various Shorthands:

$$
\begin{aligned}
&Y_1^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_4\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_6} - \sum_{\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_7} X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_7\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,\\
&Y_2^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \equiv \widetilde{H}^{\tilde\alpha_1\tilde\alpha_3} \widetilde{H}^{\tilde\alpha_2\tilde\alpha_4} - \sum_{\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_2\tilde\alpha_5} X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_5\tilde\alpha_3\tilde\alpha_4},\\
&Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv \widetilde{H}^{\tilde\alpha_1\tilde\alpha_4} \widetilde{H}^{\tilde\alpha_2\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_6} - \sum_{\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_2\tilde\alpha_7} X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_7\tilde\alpha_4\tilde\alpha_5} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_6} - \sum_{\tilde\alpha_7} \widetilde{H}^{\tilde\alpha_1\tilde\alpha_4} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_7} X_{\mathrm{II}}^{\tilde\alpha_2\tilde\alpha_7\tilde\alpha_5\tilde\alpha_6} + \sum_{\tilde\alpha_7,\tilde\alpha_8,\tilde\alpha_9} \widetilde{H}^{\tilde\alpha_3\tilde\alpha_9} X_{\mathrm{II}}^{\tilde\alpha_2\tilde\alpha_9\tilde\alpha_7\tilde\alpha_8} X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_7\tilde\alpha_8\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,\\
&Y_4^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv \widetilde{H}^{\tilde\alpha_1\tilde\alpha_4} X_{\mathrm{II}}^{\tilde\alpha_2\tilde\alpha_3\tilde\alpha_5\tilde\alpha_6} - \sum_{\tilde\alpha_7,\tilde\alpha_8} X_{\mathrm{II}}^{\tilde\alpha_2\tilde\alpha_3\tilde\alpha_7\tilde\alpha_8} X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_7\tilde\alpha_8\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,.
\end{aligned}
$$

Inverting Tensors:

$$
\begin{aligned}
\delta^{\tilde\alpha_1}_{\tilde\alpha_5}\delta^{\tilde\alpha_2}_{\tilde\alpha_6} &= \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \left(\widetilde{H}_{\tilde\alpha_3\tilde\alpha_5}\delta_{\tilde\alpha_4\tilde\alpha_6} + \delta_{\tilde\alpha_3\tilde\alpha_5}\widetilde{H}_{\tilde\alpha_4\tilde\alpha_6} - \eta \widetilde{H}_{\tilde\alpha_3\tilde\alpha_5}\widetilde{H}_{\tilde\alpha_4\tilde\alpha_6}\right)\,,\\
\delta^{\tilde\alpha_1}_{\tilde\alpha_7}\delta^{\tilde\alpha_2}_{\tilde\alpha_8}\delta^{\tilde\alpha_3}_{\tilde\alpha_9} &= \sum_{\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6} X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\Big[\Big(\widetilde{H}_{\tilde\alpha_4\tilde\alpha_7}\delta_{\tilde\alpha_5\tilde\alpha_8}\delta_{\tilde\alpha_6\tilde\alpha_9} + \delta_{\tilde\alpha_4\tilde\alpha_7}\widetilde{H}_{\tilde\alpha_5\tilde\alpha_8}\delta_{\tilde\alpha_6\tilde\alpha_9} + \delta_{\tilde\alpha_4\tilde\alpha_7}\delta_{\tilde\alpha_5\tilde\alpha_8}\widetilde{H}_{\tilde\alpha_6\tilde\alpha_9}\Big)\\
&\quad - \eta\Big(\widetilde{H}_{\tilde\alpha_4\tilde\alpha_7}\widetilde{H}_{\tilde\alpha_5\tilde\alpha_8}\delta_{\tilde\alpha_6\tilde\alpha_9} + \widetilde{H}_{\tilde\alpha_4\tilde\alpha_7}\delta_{\tilde\alpha_5\tilde\alpha_8}\widetilde{H}_{\tilde\alpha_6\tilde\alpha_9} + \delta_{\tilde\alpha_4\tilde\alpha_7}\widetilde{H}_{\tilde\alpha_5\tilde\alpha_8}\widetilde{H}_{\tilde\alpha_6\tilde\alpha_9}\Big) + \eta^2\Big(\widetilde{H}_{\tilde\alpha_4\tilde\alpha_7}\widetilde{H}_{\tilde\alpha_5\tilde\alpha_8}\widetilde{H}_{\tilde\alpha_6\tilde\alpha_9}\Big)\Big]\,.
\end{aligned}
$$

The Algorithm Projectors

Gradient Flow (ODE limit):

$$Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \equiv Y_2^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\,,$$
$$Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \equiv Y_2^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4},$$
$$Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,$$
$$Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,$$
$$Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,$$
$$Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv -Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} - Y_3^{\tilde\alpha_2\tilde\alpha_1\tilde\alpha_3\tilde\alpha_5\tilde\alpha_4\tilde\alpha_6} - Y_3^{\tilde\alpha_1\tilde\alpha_3\tilde\alpha_2\tilde\alpha_4\tilde\alpha_6\tilde\alpha_5}\,.$$

Various Shorthands:

$$Y_2^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \equiv \widetilde{H}^{\tilde\alpha_1\tilde\alpha_3}\widetilde{H}^{\tilde\alpha_2\tilde\alpha_4} - \sum_{\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_2\tilde\alpha_5}X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_5\tilde\alpha_3\tilde\alpha_4},$$
$$Y_3^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} \equiv \widetilde{H}^{\tilde\alpha_1\tilde\alpha_4}\widetilde{H}^{\tilde\alpha_2\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_6} - \sum_{\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_2\tilde\alpha_7}X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_7\tilde\alpha_4\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_6} - \sum_{\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_4}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_7}X_{\mathrm{II}}^{\tilde\alpha_2\tilde\alpha_7\tilde\alpha_5\tilde\alpha_6} + \sum_{\tilde\alpha_7,\tilde\alpha_8,\tilde\alpha_9}\widetilde{H}^{\tilde\alpha_3\tilde\alpha_9}X_{\mathrm{II}}^{\tilde\alpha_2\tilde\alpha_9\tilde\alpha_7\tilde\alpha_8}X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_7\tilde\alpha_8\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\,,$$

Inverting Tensors:

$$\delta^{\tilde\alpha_1}_{\tilde\alpha_5}\delta^{\tilde\alpha_2}_{\tilde\alpha_6} = \sum_{\tilde\alpha_3,\tilde\alpha_4\in\mathcal{A}} X_{\mathrm{II}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(\widetilde{H}_{\tilde\alpha_3\tilde\alpha_5}\delta_{\tilde\alpha_4\tilde\alpha_6} + \delta_{\tilde\alpha_3\tilde\alpha_5}\widetilde{H}_{\tilde\alpha_4\tilde\alpha_6}\right)\,,$$
$$\delta^{\tilde\alpha_1}_{\tilde\alpha_7}\delta^{\tilde\alpha_2}_{\tilde\alpha_8}\delta^{\tilde\alpha_3}_{\tilde\alpha_9} = \sum_{\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6} X_{\mathrm{III}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(\widetilde{H}_{\tilde\alpha_4\tilde\alpha_7}\delta_{\tilde\alpha_5\tilde\alpha_8}\delta_{\tilde\alpha_6\tilde\alpha_9} + \delta_{\tilde\alpha_4\tilde\alpha_7}\widetilde{H}_{\tilde\alpha_5\tilde\alpha_8}\delta_{\tilde\alpha_6\tilde\alpha_9} + \delta_{\tilde\alpha_4\tilde\alpha_7}\delta_{\tilde\alpha_5\tilde\alpha_8}\widetilde{H}_{\tilde\alpha_6\tilde\alpha_9}\right)$$

The Algorithm Projectors

Direct Optimization:

$$
\begin{aligned}
&Z_{\mathrm{A}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \equiv 0\,,\\
&Z_{\mathrm{B}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \equiv \frac{1}{2}\widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_3}\widetilde{H}^{\tilde{\alpha}_2\tilde{\alpha}_4},\\
&Z_{\mathrm{IA}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} \equiv 0\,,\\
&Z_{\mathrm{IB}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} \equiv -\frac{1}{6}\widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_4}\widetilde{H}^{\tilde{\alpha}_2\tilde{\alpha}_5}\widetilde{H}^{\tilde{\alpha}_3\tilde{\alpha}_6} \qquad ,\\
&Z_{\mathrm{IIA}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} \equiv 0\,,\\
&Z_{\mathrm{IIB}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} \equiv 0\,.
\end{aligned}
$$

Various Shorthands:

Inverting Tensors:

Mean Prediction

The *mean* prediction of a deep MLP is given by

$$
\begin{aligned}
&\mathbb{E}\left[z^{(L)}_{i;\dot\beta}(\infty)\right] \equiv m_{i;\dot\beta} \\
=& m^{\mathrm{NTK}}_{i;\dot\beta} + \frac{1}{n_{L-1}}\left(m^{\Delta\mathrm{NTK}}_{i;\dot\beta} + m^{\mathrm{dNTK}}_{i;\dot\beta} + m^{\mathrm{ddNTK\text{-}I}}_{i;\dot\beta} + m^{\mathrm{ddNTK\text{-}II}}_{i;\dot\beta}\right) \\
&- \frac{1}{n_{L-1}}\sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(m^{\Delta\mathrm{NTK}}_{i;\tilde\alpha_2} + m^{\mathrm{dNTK}}_{i;\tilde\alpha_2} + m^{\mathrm{ddNTK\text{-}I}}_{i;\tilde\alpha_2} + m^{\mathrm{ddNTK\text{-}II}}_{i;\tilde\alpha_2}\right)
\end{aligned}
$$

Mean Prediction

The *mean* prediction of a deep MLP is given by

$$
\begin{aligned}
&\mathbb{E}\left[z^{(L)}_{i;\dot\beta}(\infty)\right] \equiv m_{i;\dot\beta} \\
=&m^{\text{NTK}}_{i;\dot\beta} + \frac{1}{n_{L-1}}\left(m^{\Delta\text{NTK}}_{i;\dot\beta} + m^{\text{dNTK}}_{i;\dot\beta} + m^{\text{ddNTK-I}}_{i;\dot\beta} + m^{\text{ddNTK-II}}_{i;\dot\beta}\right) \\
&- \frac{1}{n_{L-1}}\sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(m^{\Delta\text{NTK}}_{i;\tilde\alpha_2} + m^{\text{dNTK}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-I}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-II}}_{i;\tilde\alpha_2}\right)
\end{aligned}
$$

- The first term is the (neural tangent) *kernel prediction*:

$$
m^{\text{NTK}}_{i;\dot\beta} \equiv \sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)} y_{i;\tilde\alpha_2}\,.
$$

Mean Prediction

The *mean* prediction of a deep MLP is given by

$$
\begin{aligned}
&\mathbb{E}\left[z^{(L)}_{i;\dot\beta}(\infty)\right] \equiv m_{i;\dot\beta} \\
=& m^{\text{NTK}}_{i;\dot\beta} + \frac{1}{n_{L-1}}\left(m^{\Delta\text{NTK}}_{i;\dot\beta} + m^{\text{dNTK}}_{i;\dot\beta} + m^{\text{ddNTK-I}}_{i;\dot\beta} + m^{\text{ddNTK-II}}_{i;\dot\beta}\right) \\
&- \frac{1}{n_{L-1}} \sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1} \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)} \left(m^{\Delta\text{NTK}}_{i;\tilde\alpha_2} + m^{\text{dNTK}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-I}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-II}}_{i;\tilde\alpha_2}\right)
\end{aligned}
$$

- The first term is the (neural tangent) *kernel prediction*:

$$
m^{\text{NTK}}_{i;\dot\beta} \equiv \sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1} \widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)} y_{i;\tilde\alpha_2}\,.
$$

- The four other kinds of terms give the $O(1/n)$ corrections.

Mean Prediction

The *mean* prediction of a deep MLP is given by

$$\begin{aligned}
&\mathbb{E}\left[z^{(L)}_{i;\dot\beta}(\infty)\right] \equiv m_{i;\dot\beta} \\
=&m^{\text{NTK}}_{i;\dot\beta} + \frac{1}{n_{L-1}}\left(m^{\Delta\text{NTK}}_{i;\dot\beta} + m^{\text{dNTK}}_{i;\dot\beta} + m^{\text{ddNTK-I}}_{i;\dot\beta} + m^{\text{ddNTK-II}}_{i;\dot\beta}\right) \\
&-\frac{1}{n_{L-1}}\sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(m^{\Delta\text{NTK}}_{i;\tilde\alpha_2} + m^{\text{dNTK}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-I}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-II}}_{i;\tilde\alpha_2}\right)
\end{aligned}$$

- Each network fits the training data with its own *particular* NTK, and so the resulting fully-trained particular output depends on the **NTK fluctuation**.

Mean Prediction

The *mean* prediction of a deep MLP is given by

$$
\begin{aligned}
&\mathbb{E}\left[z^{(L)}_{i;\dot\beta}(\infty)\right] \equiv m_{i;\dot\beta} \\
=&m^{\text{NTK}}_{i;\dot\beta} + \frac{1}{n_{L-1}}\left(m^{\Delta\text{NTK}}_{i;\dot\beta} + m^{\text{dNTK}}_{i;\dot\beta} + m^{\text{ddNTK-I}}_{i;\dot\beta} + m^{\text{ddNTK-II}}_{i;\dot\beta}\right) \\
&- \frac{1}{n_{L-1}}\sum_{\tilde\alpha_1,\tilde\alpha_2} H^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{H}^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(m^{\Delta\text{NTK}}_{i;\tilde\alpha_2} + m^{\text{dNTK}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-I}}_{i;\tilde\alpha_2} + m^{\text{ddNTK-II}}_{i;\tilde\alpha_2}\right)
\end{aligned}
$$

- Each network fits the training data with its own *particular* NTK, and so the resulting fully-trained particular output depends on the **NTK fluctuation**.
- The dependence on the **NTK differentials** means there's nontrivial representation learning at finite width.

Mean Prediction: *NTK Variance*

The fluctuation of the NTK term gives

$$m^{\Delta\mathrm{NTK}}_{i;\delta} \equiv \sum_{\tilde{\alpha}_1,\dots,\tilde{\alpha}_4} \left(A^{(L)}_{(\delta\tilde{\alpha}_1)(\tilde{\alpha}_2\tilde{\alpha}_3)} + B^{(L)}_{\delta\tilde{\alpha}_2\tilde{\alpha}_1\tilde{\alpha}_3} + n_L B^{(L)}_{\delta\tilde{\alpha}_3\tilde{\alpha}_1\tilde{\alpha}_2} \right) \times \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_2}_{(L)} \widetilde{H}^{\tilde{\alpha}_3\tilde{\alpha}_4}_{(L)} y_{i;\tilde{\alpha}_4} \, ,$$

where we decomposed the *NTK variance* into $A^{(L)}$ and $B^{(L)}$:

$$\mathbb{E}\left[\widehat{\Delta H}^{(L)}_{i_1 i_2;\alpha_1\alpha_2} \widehat{\Delta H}^{(L)}_{i_3 i_4;\alpha_3\alpha_4} \right] \equiv \frac{1}{n_{L-1}} \left[\delta_{i_1 i_2}\delta_{i_3 i_4} A^{(L)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4} B^{(L)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1 i_4}\delta_{i_2 i_3} B^{(L)}_{\alpha_1\alpha_4\alpha_2\alpha_3} \right] .$$

Mean Prediction: *NTK Variance*

The fluctuation of the NTK term gives

$$m^{\Delta\mathrm{NTK}}_{i;\delta} \equiv \sum_{\tilde{\alpha}_1,\dots,\tilde{\alpha}_4} \left(A^{(L)}_{(\delta\tilde{\alpha}_1)(\tilde{\alpha}_2\tilde{\alpha}_3)} + B^{(L)}_{\delta\tilde{\alpha}_2\tilde{\alpha}_1\tilde{\alpha}_3} + n_L B^{(L)}_{\delta\tilde{\alpha}_3\tilde{\alpha}_1\tilde{\alpha}_2} \right) \times \widetilde{H}^{\tilde{\alpha}_1\tilde{\alpha}_2}_{(L)} \widetilde{H}^{\tilde{\alpha}_3\tilde{\alpha}_4}_{(L)} y_{i;\tilde{\alpha}_4} \,,$$

where we decomposed the *NTK variance* into $A^{(L)}$ and $B^{(L)}$:

$$\mathbb{E}\left[\widehat{\Delta H}^{(L)}_{i_1 i_2;\alpha_1\alpha_2}\widehat{\Delta H}^{(L)}_{i_3 i_4;\alpha_3\alpha_4}\right] \equiv \frac{1}{n_{L-1}} \left[\delta_{i_1 i_2}\delta_{i_3 i_4} A^{(L)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1 i_3}\delta_{i_2 i_4} B^{(L)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1 i_4}\delta_{i_2 i_3} B^{(L)}_{\alpha_1\alpha_4\alpha_2\alpha_3}\right] .$$

Mean Prediction: *dNTK-Preactivation Cross Correlation*

The dNTK term gives

$$
\begin{aligned}
& m^{\mathrm{dNTK}}_{i;\delta} \\
\equiv - \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_4} \Bigg[& 2\left(P^{(L)}_{\delta\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\delta\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} + n_L Q^{(L)}_{\delta\tilde{\alpha}_2\tilde{\alpha}_1\tilde{\alpha}_3} \right) Z_{\mathrm{B}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \\
& + \left(n_L P^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\delta\tilde{\alpha}_3} \right) Z_{\mathrm{A}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \\
& + \left(P^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} + n_L Q^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\delta\tilde{\alpha}_3} \right) Z_{\mathrm{A}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_4\tilde{\alpha}_3} \Bigg] y_{i;\tilde{\alpha}_4} \,,
\end{aligned}
$$

where we decomposed the *dNTK-preactivation cross correlators* into tensors $P^{(L)}$ and $Q^{(L)}$:

$$
\begin{aligned}
& \mathbb{E}\left[\widehat{\mathrm{d}H}^{(L)}_{i_0 i_1 i_2;\delta_0\delta_1\delta_2} z^{(L)}_{i_3;\delta_3} \right] \\
\equiv & \frac{1}{n_{L-1}} \left[\delta_{i_0 i_3}\delta_{i_1 i_2} P^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0 i_1}\delta_{i_2 i_3} Q^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0 i_2}\delta_{i_1 i_3} Q^{(L)}_{\delta_0\delta_2\delta_1\delta_3} \right] .
\end{aligned}
$$

Mean Prediction: *dNTK-Preactivation Cross Correlation*

The dNTK term gives

$$
\begin{aligned}
& m^{\text{dNTK}}_{i;\delta} \\
\equiv - \sum_{\tilde\alpha_1,\ldots,\tilde\alpha_4} \Bigg[& 2\left(P^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} + Q^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} + n_L Q^{(L)}_{\delta\tilde\alpha_2\tilde\alpha_1\tilde\alpha_3}\right) Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \\
& + \left(n_L P^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3} + Q^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3} + Q^{(L)}_{\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3}\right) Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \\
& + \left(P^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3} + n_L Q^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3} + Q^{(L)}_{\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3}\right) Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_4\tilde\alpha_3} \Bigg] y_{i;\tilde\alpha_4} \,,
\end{aligned}
$$

where we decomposed the *dNTK-preactivation cross correlators* into tensors $P^{(L)}$ and $Q^{(L)}$:

$$
\begin{aligned}
& \mathbb{E}\left[\widehat{\mathrm{d}H}^{(L)}_{i_0 i_1 i_2;\delta_0\delta_1\delta_2} z^{(L)}_{i_3;\delta_3}\right] \\
\equiv & \frac{1}{n_{L-1}} \left[\delta_{i_0 i_3}\delta_{i_1 i_2} P^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0 i_1}\delta_{i_2 i_3} Q^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0 i_2}\delta_{i_1 i_3} Q^{(L)}_{\delta_0\delta_2\delta_1\delta_3}\right] .
\end{aligned}
$$

Mean Prediction: *ddNTK$_I$ Mean*

The first ddNTK term gives

$$
m_{i;\delta}^{\text{ddNTK-I}} \equiv
$$

$$
-\sum_{\tilde\alpha_1,\ldots,\tilde\alpha_6}\Big[R^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\Big(Z_{\text{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}+Z_{\text{IB}}^{\tilde\alpha_2\tilde\alpha_3\tilde\alpha_1\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4}+Z_{\text{IB}}^{\tilde\alpha_3\tilde\alpha_1\tilde\alpha_2\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5}\Big)+
$$

$$
R^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3}Z_{\text{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}+R^{(L)}_{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\delta}Z_{\text{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4}+R^{(L)}_{\tilde\alpha_1\tilde\alpha_3\delta\tilde\alpha_2}Z_{\text{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5}
$$

$$
\Big]\times\Big[y_{i;\tilde\alpha_4}\Big(\sum_j y_{j;\tilde\alpha_5}y_{j;\tilde\alpha_6}+n_L K^{(L)}_{\tilde\alpha_5\tilde\alpha_6}\Big)+y_{i;\tilde\alpha_5}K^{(L)}_{\tilde\alpha_6\tilde\alpha_4}+y_{i;\tilde\alpha_6}K^{(L)}_{\tilde\alpha_4\tilde\alpha_5}\Big]
$$

where we decomposed the *dNTK$_I$ mean* into the tensor $R^{(L)}$:

$$
\mathbb{E}\left[\widehat{\mathrm{dd_I}H}^{(L)}_{i_0 i_1 i_2 i_3;\delta_0\delta_1\delta_2\delta_3}\right]
\equiv\frac{1}{n_{L-1}}\left[\delta_{i_0 i_1}\delta_{i_2 i_3}R^{(L)}_{\delta_0\delta_1\delta_2\delta_3}+\delta_{i_0 i_2}\delta_{i_3 i_1}R^{(L)}_{\delta_0\delta_2\delta_3\delta_1}+\delta_{i_0 i_3}\delta_{i_1 i_2}R^{(L)}_{\delta_0\delta_3\delta_1\delta_2}\right].
$$

Mean Prediction: *ddNTK$_I$ Mean*

The first ddNTK term gives

$$
\begin{aligned}
&m_{i;\delta}^{\text{ddNTK-I}} \equiv \\
&-\sum_{\tilde\alpha_1,\dots,\tilde\alpha_6}\Big[R^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\Big(Z_{\text{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}+Z_{\text{IB}}^{\tilde\alpha_2\tilde\alpha_3\tilde\alpha_1\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4}+Z_{\text{IB}}^{\tilde\alpha_3\tilde\alpha_1\tilde\alpha_2\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5}\Big)+ \\
&R^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3}Z_{\text{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}+R^{(L)}_{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\delta}Z_{\text{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4}+R^{(L)}_{\tilde\alpha_1\tilde\alpha_3\delta\tilde\alpha_2}Z_{\text{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5} \\
&\Big]\times\Big[y_{i;\tilde\alpha_4}\Big(\sum_j y_{j;\tilde\alpha_5}y_{j;\tilde\alpha_6}+n_L K^{(L)}_{\tilde\alpha_5\tilde\alpha_6}\Big)+y_{i;\tilde\alpha_5}K^{(L)}_{\tilde\alpha_6\tilde\alpha_4}+y_{i;\tilde\alpha_6}K^{(L)}_{\tilde\alpha_4\tilde\alpha_5}\Big]
\end{aligned}
$$

where we decomposed the *dNTK$_I$ mean* into the tensor $R^{(L)}$:

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\text{dd}_{\text{I}}H}^{(L)}_{i_0i_1i_2i_3;\delta_0\delta_1\delta_2\delta_3}\right] \\
&\equiv\frac{1}{n_{L-1}}\left[\delta_{i_0i_1}\delta_{i_2i_3}R^{(L)}_{\delta_0\delta_1\delta_2\delta_3}+\delta_{i_0i_2}\delta_{i_3i_1}R^{(L)}_{\delta_0\delta_2\delta_3\delta_1}+\delta_{i_0i_3}\delta_{i_1i_2}R^{(L)}_{\delta_0\delta_3\delta_1\delta_2}\right].
\end{aligned}
$$

Mean Prediction: *ddNTK$_{II}$ Mean*

The second ddNTK term gives

$$
\begin{aligned}
& m_{i;\delta}^{\text{ddNTK-II}} \\
\equiv & -\sum_{\tilde\alpha_1,\dots,\tilde\alpha_6} \Big[S^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} Z_{\text{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} + T^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} Z_{\text{IIB}}^{\tilde\alpha_2\tilde\alpha_3\tilde\alpha_1\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4} \\
& \qquad + U^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} Z_{\text{IIB}}^{\tilde\alpha_3\tilde\alpha_1\tilde\alpha_2\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5} + S^{(L)}_{\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3} Z_{\text{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4} \\
& \qquad + T^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_3\tilde\alpha_2} Z_{\text{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6} + U^{(L)}_{\tilde\alpha_1\tilde\alpha_3\tilde\alpha_2\delta} Z_{\text{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5} \\
& \Big] \times \Big[y_{i;\tilde\alpha_4} \Big(\sum_j y_{j;\tilde\alpha_5} y_{j;\tilde\alpha_6} + n_L K^{(L)}_{\tilde\alpha_5\tilde\alpha_6} \Big) + y_{i;\tilde\alpha_5} K^{(L)}_{\tilde\alpha_6\tilde\alpha_4} + y_{i;\tilde\alpha_6} K^{(L)}_{\tilde\alpha_4\tilde\alpha_5} \Big]
\end{aligned}
$$

where we decomposed the *dNTK$_{II}$ mean* into $S^{(L)}$, $T^{(L)}$, $U^{(L)}$:

$$
\begin{aligned}
& \mathbb{E}\left[\widehat{\mathrm{dd}_{\mathrm{II}} H}^{(L)}_{i_1 i_2 i_3 i_4;\delta_1\delta_2\delta_3\delta_4} \right] \\
\equiv & \frac{1}{n_{L-1}} \left[\delta_{i_1 i_2}\delta_{i_3 i_4} S^{(L)}_{\delta_1\delta_2\delta_3\delta_4} + \delta_{i_1 i_3}\delta_{i_4 i_2} T^{(L)}_{\delta_1\delta_3\delta_4\delta_2} + \delta_{i_1 i_4}\delta_{i_2 i_3} U^{(L)}_{\delta_1\delta_4\delta_2\delta_3} \right] .
\end{aligned}
$$

Mean Prediction: *ddNTK$_{II}$ Mean*

The second ddNTK term gives

$$
\begin{aligned}
&m^{\text{ddNTK-II}}_{i;\delta} \\
&\equiv -\sum_{\tilde\alpha_1,\dots,\tilde\alpha_6}\Big[S^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}Z^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}_{\text{IIB}} + T^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}Z^{\tilde\alpha_2\tilde\alpha_3\tilde\alpha_1\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4}_{\text{IIB}} \\
&\qquad + U^{(L)}_{\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}Z^{\tilde\alpha_3\tilde\alpha_1\tilde\alpha_2\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5}_{\text{IIB}} + S^{(L)}_{\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3}Z^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_5\tilde\alpha_6\tilde\alpha_4}_{\text{IIA}} \\
&\qquad + T^{(L)}_{\tilde\alpha_1\delta\tilde\alpha_3\tilde\alpha_2}Z^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}_{\text{IIA}} + U^{(L)}_{\tilde\alpha_1\tilde\alpha_3\tilde\alpha_2\delta}Z^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_6\tilde\alpha_4\tilde\alpha_5}_{\text{IIA}} \\
&\Big] \times \Big[y_{i;\tilde\alpha_4}\Big(\sum_j y_{j;\tilde\alpha_5}y_{j;\tilde\alpha_6} + n_L K^{(L)}_{\tilde\alpha_5\tilde\alpha_6}\Big) + y_{i;\tilde\alpha_5}K^{(L)}_{\tilde\alpha_6\tilde\alpha_4} + y_{i;\tilde\alpha_6}K^{(L)}_{\tilde\alpha_4\tilde\alpha_5}\Big]
\end{aligned}
$$

where we decomposed the *dNTK$_{II}$ mean* into $S^{(L)}$, $T^{(L)}$, $U^{(L)}$:

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{\text{dd}_{\text{II}}H}^{(L)}_{i_1 i_2 i_3 i_4;\delta_1\delta_2\delta_3\delta_4}\right] \\
&\equiv \frac{1}{n_{L-1}}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}S^{(L)}_{\delta_1\delta_2\delta_3\delta_4} + \delta_{i_1 i_3}\delta_{i_4 i_2}T^{(L)}_{\delta_1\delta_3\delta_4\delta_2} + \delta_{i_1 i_4}\delta_{i_2 i_3}U^{(L)}_{\delta_1\delta_4\delta_2\delta_3}\right].
\end{aligned}
$$

Prediction Variance

In addition to the ensemble mean, we can consider other statistics:

$$
\begin{aligned}
&\mathrm{Cov}\Big[z^{(L)}_{i_1;\dot\beta_1}(\infty),\, z^{(L)}_{i_2;\dot\beta_2}(\infty)\Big] \\
\equiv&\mathbb{E}\left[z^{(L)}_{i_1;\dot\beta_1}(\infty)\; z^{(L)}_{i_2;\dot\beta_2}(\infty)\right] - \mathbb{E}\left[z^{(L)}_{i_1;\dot\beta_1}(\infty)\right]\mathbb{E}\left[z^{(L)}_{i_2;\dot\beta_2}(\infty)\right] .
\end{aligned}
$$

Prediction Variance

In addition to the ensemble mean, we can consider other statistics:

$$
\begin{aligned}
&\mathrm{Cov}\Big[z^{(L)}_{i_1;\dot{\beta}_1}(\infty),\, z^{(L)}_{i_2;\dot{\beta}_2}(\infty)\Big] \\
\equiv&\mathbb{E}\left[z^{(L)}_{i_1;\dot{\beta}_1}(\infty)\; z^{(L)}_{i_2;\dot{\beta}_2}(\infty)\right] - \mathbb{E}\left[z^{(L)}_{i_1;\dot{\beta}_1}(\infty)\right]\mathbb{E}\left[z^{(L)}_{i_2;\dot{\beta}_2}(\infty)\right]\,.
\end{aligned}
$$

While we won't print this quantity in full – the full expression doesn't really play nicely with the constraints of the slides – you can easily extract insight by considering specific contributions.

Prediction Variance

To see another manifestation of output "wiring," consider

$$\sum_{\substack{j_1,j_2\\ \tilde{\alpha}_1,\ldots,\tilde{\alpha}_4}} \mathbb{E}\left[z^{(L)}_{i_2;\dot{\beta}_2}\widehat{\mathrm{d}H}^{(L)}_{i_1j_1j_2;\dot{\beta}_1\tilde{\alpha}_1\tilde{\alpha}_2}\right] Z_{\mathrm{B}}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4}$$

$$\times\, \mathbb{E}\left[\left(z^{(L)}_{j_1;\tilde{\alpha}_3} - y_{j_1;\tilde{\alpha}_3}\right)\left(z^{(L)}_{j_2;\tilde{\alpha}_4} - y_{j_2;\tilde{\alpha}_4}\right)\right]$$

Prediction Variance

To see another manifestation of output "wiring," consider

$$
\begin{aligned}
&\sum_{\substack{j_1,j_2\\ \tilde\alpha_1,\dots,\tilde\alpha_4}} \mathbb{E}\left[z^{(L)}_{i_2;\dot\beta_2}\widehat{\mathrm{d}H}^{(L)}_{i_1j_1j_2;\dot\beta_1\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \\
&\qquad\qquad \times \mathbb{E}\left[\left(z^{(L)}_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z^{(L)}_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\right] \\
=&\frac{1}{n_{L-1}}\delta_{i_1i_2}\sum_{\tilde\alpha_1,\dots,\tilde\alpha_4}\Bigg[P^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}\Big(\sum_j y_{j;\tilde\alpha_3}y_{j;\tilde\alpha_4}\Big) \\
&\qquad + \left(n_L P^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}+Q^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}+Q^{(L)}_{\dot\beta_1\tilde\alpha_2\tilde\alpha_1\dot\beta_2}\right)G^{(L)}_{\tilde\alpha_3\tilde\alpha_4}\Bigg] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4} \\
&+\frac{1}{n_{L-1}}\sum_{\tilde\alpha_1,\dots,\tilde\alpha_4} Q^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2} Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(y_{i_1;\tilde\alpha_3}y_{i_2;\tilde\alpha_4}+y_{i_1;\tilde\alpha_4}y_{i_2;\tilde\alpha_3}\right),
\end{aligned}
$$

Prediction Variance

To see another manifestation of output "wiring," consider

$$
\begin{aligned}
&\sum_{\substack{j_1,j_2\\ \tilde\alpha_1,\dots,\tilde\alpha_4}} \mathbb{E}\left[z^{(L)}_{i_2;\dot\beta_2}\widehat{\mathrm{d}H}^{(L)}_{i_1j_1j_2;\dot\beta_1\tilde\alpha_1\tilde\alpha_2}\right] Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\\
&\qquad\qquad\times \mathbb{E}\left[\left(z^{(L)}_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z^{(L)}_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\right]\\
=&\frac{1}{n_{L-1}}\delta_{i_1i_2}\sum_{\tilde\alpha_1,\dots,\tilde\alpha_4}\Bigg[P^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}\Big(\sum_j y_{j;\tilde\alpha_3}y_{j;\tilde\alpha_4}\Big)\\
&\qquad+\left(n_L P^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}+Q^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}+Q^{(L)}_{\dot\beta_1\tilde\alpha_2\tilde\alpha_1\dot\beta_2}\right)G^{(L)}_{\tilde\alpha_3\tilde\alpha_4}\Bigg]Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\\
&+\frac{1}{n_{L-1}}\sum_{\tilde\alpha_1,\dots,\tilde\alpha_4}Q^{(L)}_{\dot\beta_1\tilde\alpha_1\tilde\alpha_2\dot\beta_2}Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(y_{i_1;\tilde\alpha_3}y_{i_2;\tilde\alpha_4}+y_{i_1;\tilde\alpha_4}y_{i_2;\tilde\alpha_3}\right),
\end{aligned}
$$

- *Wiring* is exhibited when $y_{i_1}\neq 0$ and $y_{i_2}\neq 0$.

Prediction is Nearly-Gaussian

At finite width, $p\left(z^{(L)}(\infty)\,\middle|\,\mathcal{D}\right)$ now has non-Gaussian statistics:

Prediction is Nearly-Gaussian

At finite width, $p\Big(z^{(L)}(\infty)\,\Big|\mathcal{D}\Big)$ now has non-Gaussian statistics:

- $z_{i;\dot\beta}(\infty) \equiv z_{i;\dot\beta}(\infty)\left[z^{(L)},\ \widehat{H}^{(L)},\ \widehat{\mathrm{d}H}^{(L)},\ \widehat{\mathrm{dd_I}H}^{(L)},\ \widehat{\mathrm{dd_{II}}H}^{(L)}\right].$

Prediction is Nearly-Gaussian

At finite width, $p\Big(z^{(L)}(\infty)\,\Big|\mathcal{D}\Big)$ now has non-Gaussian statistics:

- $z_{i;\dot\beta}(\infty) \equiv z_{i;\dot\beta}(\infty)\left[z^{(L)},\, \widehat{H}^{(L)},\, \widehat{\mathrm{d}H}^{(L)},\, \widehat{\mathrm{dd_I}H}^{(L)},\, \widehat{\mathrm{dd_{II}}H}^{(L)}\right]$.
- $p\Big(z^{(L)},\, \widehat{H}^{(L)},\, \widehat{\mathrm{d}H}^{(L)},\, \widehat{\mathrm{dd_I}H}^{(L)},\, \widehat{\mathrm{dd_{II}}H}^{(L)}\Big|\mathcal{D}\Big)$ is *nearly-Gaussian*.

Prediction is Nearly-Gaussian

At finite width, $p\Big(z^{(L)}(\infty)\,\Big|\mathcal{D}\Big)$ now has non-Gaussian statistics:

- $z_{i;\dot\beta}(\infty) \equiv z_{i;\dot\beta}(\infty)\left[z^{(L)},\, \widehat{H}^{(L)},\, \widehat{\mathrm{d}H}^{(L)},\, \widehat{\mathrm{dd_I}H}^{(L)},\, \widehat{\mathrm{dd_{II}}H}^{(L)}\right]$.
- $p\Big(z^{(L)},\, \widehat{H}^{(L)},\, \widehat{\mathrm{d}H}^{(L)},\, \widehat{\mathrm{dd_I}H}^{(L)},\, \widehat{\mathrm{dd_{II}}H}^{(L)}\Big|\mathcal{D}\Big)$ is *nearly-Gaussian*.
- Explicit expressions of higher-point correlators are challenging to display in any media format. . .

Generalization

The **generalization error** is a quantitative measure of how well a network is really approximating the desired function:

$$\mathcal{E} \equiv \mathcal{L}_{\mathcal{B}} - \mathcal{L}_{\mathcal{A}} \,.$$

Generalization

The **generalization error** is a quantitative measure of how well a network is really approximating the desired function:

$$\mathcal{E} \equiv \mathcal{L}_{\mathcal{B}} \, .$$

Generalization

The **generalization error** is a quantitative measure of how well a network is really approximating the desired function:

$$\mathcal{E} \equiv \mathcal{L}_{\mathcal{B}} \,.$$

Let's evaluate the **MSE test loss**, averaged over an ensemble of fully-trained networks:

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{B}}(T)\right] = \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{n_L}\sum_{\dot{\beta}\in\mathcal{B}}\left(z^{(L)}_{i;\dot{\beta}}(\infty) - y_{i;\dot{\beta}}\right)^2\right] \,.$$

Generalization

The **generalization error** is a quantitative measure of how well a network is really approximating the desired function:

$$\mathcal{E} \equiv \mathcal{L}_{\mathcal{B}} \,.$$

Let's evaluate the **MSE test loss**, averaged over an ensemble of fully-trained networks:

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{B}}(T)\right] = \frac{1}{2} \sum_{\dot{\beta}\in\mathcal{B}} \left\{ \sum_{i=1}^{n_L} \left(m_{i;\dot{\beta}} - y_{i;\dot{\beta}} \right)^2 + \sum_{i=1}^{n_L} \mathrm{Cov}\left[z^{(L)}_{i;\dot{\beta}}(\infty),\, z^{(L)}_{i;\dot{\beta}}(\infty) \right] \right\} .$$

Generalization

The **generalization error** is a quantitative measure of how well a network is really approximating the desired function:

$$\mathcal{E} \equiv \mathcal{L}_{\mathcal{B}}\,.$$

Let's evaluate the **MSE test loss**, averaged over an ensemble of fully-trained networks:

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{B}}(T)\right] = \frac{1}{2}\sum_{\dot{\beta}\in\mathcal{B}}\left\{\sum_{i=1}^{n_L}\left(m_{i;\dot{\beta}} - y_{i;\dot{\beta}}\right)^2 + \sum_{i=1}^{n_L}\mathrm{Cov}\left[z^{(L)}_{i;\dot{\beta}}(\infty),\, z^{(L)}_{i;\dot{\beta}}(\infty)\right]\right\}\,.$$

- This illustrates a ***generalized*** **bias-variance tradeoff**: different settings of the hyperparameters will decrease one term at the cost of increasing the other.

Generalization

The **generalization error** is a quantitative measure of how well a network is really approximating the desired function:

$$\mathcal{E} \equiv \mathcal{L}_{\mathcal{B}} \,.$$

Let's evaluate the **MSE test loss**, averaged over an ensemble of fully-trained networks:

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{B}}(T)\right] = \frac{1}{2} \sum_{\dot{\beta} \in \mathcal{B}} \left\{ \sum_{i=1}^{n_L} \left(m_{i;\dot{\beta}} - y_{i;\dot{\beta}} \right)^2 + \sum_{i=1}^{n_L} \mathrm{Cov}\left[z^{(L)}_{i;\dot{\beta}}(\infty),\, z^{(L)}_{i;\dot{\beta}}(\infty) \right] \right\} .$$

- This illustrates a ***generalized*** **bias-variance tradeoff**: different settings of the hyperparameters will decrease one term at the cost of increasing the other.
- Importantly, we're choosing between *ensembles* not *models*.

Generalization: Interpolation and Extrapolation

Given the true outputs $y_{i;\pm}$ for *two* inputs $x_{i;\pm} = x_{i;0} \pm \frac{\delta x_i}{2}$, what is the prediction for a one-parameter family of test inputs,

$$s x_{i;+} + (1-s) x_{i;-} = x_{i;0} + \frac{(2s-1)}{2} \delta x_i \equiv x_{i;(2s-1)} \,,$$

that sit on a line passing through $x_{i;+}$ and $x_{i;-}$?

Generalization: Interpolation and Extrapolation

Given the true outputs $y_{i;\pm}$ for *two* inputs $x_{i;\pm} = x_{i;0} \pm \frac{\delta x_i}{2}$, what is the prediction for a one-parameter family of test inputs,

$$s x_{i;+} + (1-s) x_{i;-} = x_{i;0} + \frac{(2s-1)}{2}\delta x_i \equiv x_{i;(2s-1)}\,,$$

that sit on a line passing through $x_{i;+}$ and $x_{i;-}$?

- When our parameter s is inside the unit interval $s \in [0,1]$, this is a question about neural-network **interpolation**.
- For s outside the unit interval, it's **extrapolation**.
- For general s, let's refer to this collectively as $*$-**polation**.

$*$-polation at Infinite Width

To understand $*$-**polation** at infinite by *smooth activations*, we need to evaluate:

$$z_{i;\dot{\beta}}^{(L)}(\infty) = z_{i;\dot{\beta}}^{(L)} - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2\in\mathcal{A}} \Theta_{\dot{\beta}\tilde{\alpha}_1}^{(L)} \widetilde{\Theta}_{(L)}^{\tilde{\alpha}_1\tilde{\alpha}_2} \left(z_{i;\tilde{\alpha}_2}^{(L)} - y_{i;\tilde{\alpha}_2} \right) .$$

$*$-polation at Infinite Width

To understand $*$-**polation** at infinite by *smooth activations*, we need to evaluate:

$$z^{(L)}_{i;\dot\beta}(\infty) = z^{(L)}_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2\in\mathcal{A}} \Theta^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde\Theta^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(z^{(L)}_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}\right) .$$

- We need to invert the two-by-two submatrix of the NTK on the training set only, $\widetilde\Theta^{\tilde\alpha_1\tilde\alpha_2}_{(L)}$.
- We will need to evaluate elements of the NTK between our test and training set, $\Theta^{(L)}_{(2s-1)\pm}$.

$*$-polation at Infinite Width

To understand $*$-**polation** at infinite by *smooth activations*, we need to evaluate:

$$z^{(L)}_{i;\dot\beta}(\infty) = z^{(L)}_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2\in\mathcal{A}} \Theta^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{\Theta}^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(z^{(L)}_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}\right) .$$

The inverse of the submatrix can be written as:

$$\widetilde{\Theta}^{\tilde\alpha_1\tilde\alpha_2} = \frac{1}{\Theta_{++}\Theta_{--} - \Theta_{+-}^2}\begin{pmatrix} \Theta_{--} & -\Theta_{+-} \\ -\Theta_{+-} & \Theta_{++} \end{pmatrix} .$$

$*$-polation at Infinite Width

To understand $*$-**polation** at infinite by *smooth activations*, we need to evaluate:

$$z^{(L)}_{i;\dot\beta}(\infty) = z^{(L)}_{i;\dot\beta} - \sum_{\tilde\alpha_1,\tilde\alpha_2\in\mathcal{A}} \Theta^{(L)}_{\dot\beta\tilde\alpha_1}\widetilde{\Theta}^{\tilde\alpha_1\tilde\alpha_2}_{(L)}\left(z^{(L)}_{i;\tilde\alpha_2} - y_{i;\tilde\alpha_2}\right) .$$

The test-train NTK can be evaluated as

$$\Theta_{(2s-1)\pm} = s\Theta_{\pm+} + (1-s)\Theta_{\pm-} - 2s(1-s)\delta\delta\Theta_{[0]} + O\left(\delta^3\right) ,$$

where $\delta\delta\Theta_{[0]}$ is the difference

$$\delta\delta\Theta_{[0]} \equiv \frac{1}{4}\Big[\Theta_{++} + \Theta_{--} + 2\Theta_{+-}\Big] - \Theta_{00} + O\left(\delta^4\right) .$$

$*$-polation at Infinite Width

As an illustration, consider two training inputs with the same norm:

$$
\begin{aligned}
z^{(L)}_{i;(2s-1)}(\infty) = & \left[z^{(L)}_{i;(2s-1)} - s z^{(L)}_{i;+} - (1-s) z^{(L)}_{i;-}\right] + \left[s y_{i;+} + (1-s) y_{i;-}\right] \\
& - 4s(1-s)\left(\frac{\delta\delta\Theta_{[0]}}{\Theta_{00}}\right)\left(z^{(L)}_{i;+} + z^{(L)}_{i;-} + y_{i;+} + y_{i;-}\right) \\
& + O\!\left(\delta^3\right) .
\end{aligned}
$$

$*$-polation at Infinite Width

As an illustration, consider two training inputs with the same norm:

$$
\begin{aligned}
z^{(L)}_{i;(2s-1)}(\infty) = &\left[z^{(L)}_{i;(2s-1)} - s z^{(L)}_{i;+} - (1-s) z^{(L)}_{i;-}\right] + \left[s y_{i;+} + (1-s) y_{i;-}\right] \\
&- 4s(1-s)\left(\frac{\delta\delta\Theta_{[0]}}{\Theta_{00}}\right)\left(z^{(L)}_{i;+} + z^{(L)}_{i;-} + y_{i;+} + y_{i;-}\right) \\
&+ O\!\left(\delta^3\right) .
\end{aligned}
$$

- *Nonlinear* networks can *nonlinearly* $*$-polate!

$*$-polation at Infinite Width

The prediction of the ensemble has a **bias**:

$$m^{\infty}_{i;(2s-1)} - y_{i;(2s-1)} = \Big[y_{i;+} + (1-s) y_{i;-} - y_{i;(2s-1)} \Big] \\ - 4s(1-s)\left(\frac{\delta\delta\Theta_{[0]}}{\Theta_{00}}\right)(y_{i;+} + y_{i;-}) + O\Big(\delta^3\Big) \; .$$

$*$-polation at Infinite Width

The prediction of the ensemble has a **bias**:

$$
\begin{aligned}
m^{\infty}_{i;(2s-1)} - y_{i;(2s-1)} = & \left[y_{i;+} + (1-s) y_{i;-} - y_{i;(2s-1)} \right] \\
& - 4s(1-s) \left(\frac{\delta\delta\Theta_{[0]}}{\Theta_{00}} \right) \left(y_{i;+} + y_{i;-} \right) + O\left(\delta^3\right) .
\end{aligned}
$$

- This decomposes into a part measuring the nonlinearity in the labels and the **network curvature** around $\left(y_{i;+} + y_{i;-} \right)/2$.

$*$-polation at Infinite Width

The prediction of the ensemble has a **bias**:

$$m^{\infty}_{i;(2s-1)} - y_{i;(2s-1)} = \Big[y_{i;+} + (1-s)y_{i;-} - y_{i;(2s-1)}\Big] \\ - 4s(1-s)\left(\frac{\delta\delta\Theta_{[0]}}{\Theta_{00}}\right)(y_{i;+} + y_{i;-}) + O\Big(\delta^3\Big)\,.$$

- This decomposes into a part measuring the nonlinearity in the labels and the **network curvature** around $(y_{i;+} + y_{i;-})/2$.
- This curvature encodes the **inductive bias** of the function computed by the network.

$*$-polation at Finite Width

The prediction of the ensemble has a **bias**:

$$m_{i;(2s-1)} - y_{i;(2s-1)} = \Big[y_{i;+} + (1-s)y_{i;-} - y_{i;(2s-1)}\Big] + O\Big(y^3\Big) \\ - 4s(1-s)\left(\frac{\delta\delta\Theta_{[0]}}{\Theta_{00}}\right)(y_{i;+} + y_{i;-}) + O\Big(\delta^3\Big) \,.$$

- This decomposes into a part measuring the nonlinearity in the labels and the **network curvature** around $(y_{i;+} + y_{i;-})/2$.
- This curvature encodes the **inductive bias** of the function computed by the network.
- At *finite width*, we'd have a **cubic** $*$-polation.

Generalization at Finite Width

All terms are proportional to one of these dimensionless ratios:

$$\frac{A^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}, \quad \frac{B^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}, \quad \frac{P^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}, \quad \frac{Q^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2},$$

$$\frac{R^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}, \quad \frac{S^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}, \quad \frac{T^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}, \quad \frac{U^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}.$$

Generalization at Finite Width

All terms are proportional to one of these dimensionless ratios:

$$\frac{A^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}\,,\quad \frac{B^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}\,,\quad \frac{P^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}\,,\quad \frac{Q^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^2}\,,$$

$$\frac{R^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}\,,\quad \frac{S^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}\,,\quad \frac{T^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}\,,\quad \frac{U^{(L)}K^{(L)}}{n_{L-1}\left(\widetilde{H}^{(L)}\right)^3}\,.$$

Overall we should find for the finite-width corrections:

$$m_{i;\dot\beta} - m^{\infty}_{i;\dot\beta} = O\left(\frac{L}{n}\right)\,.$$

Optimal Aspect Ratio

In deep networks, we want to balance the positive effect of *representation learning* against the negative effect of *fluctuations*.

Optimal Aspect Ratio

In deep networks, we want to balance the positive effect of *representation learning* against the negative effect of *fluctuations*.

- We can look at the finite-width generalization error – and someone should(!) – but requires studying $O(L^2/n^2)$ effects.

Optimal Aspect Ratio

In deep networks, we want to balance the positive effect of *representation learning* against the negative effect of *fluctuations*.

- We can look at the finite-width generalization error – and someone should(!) – but requires studying $O\left(L^2/n^2\right)$ effects.
- Can try to pick a simpler quantity – perhaps one defined at initialization – where we can compute higher-order effects.

Aside: Information Theory

The **entropy** of a probability distribution is given by

$$\mathcal{S}[p(x)] \equiv -\sum_{x} p(x) \log p(x) \,.$$

Aside: Information Theory

The **entropy** of a probability distribution is given by

$$\mathcal{S}[p(x)] \equiv -\sum_x p(x) \log p(x) \,.$$

- Entropy is *functional* of the distribution, taking a distribution as an argument and outputting a number.

Aside: Information Theory

The **entropy** of a probability distribution is given by

$$\mathcal{S}[p(x)] \equiv -\sum_{x} p(x) \log p(x)\,.$$

- Entropy is *functional* of the distribution, taking a distribution as an argument and outputting a number.
- Quantitative measure of how much expected **information** is gained after making an *observation*.

Aside: Information Theory

The entropy is **additive** for two independent random variables x and y, with $p(x, y) = p(x)\, p(y)$:

$$\mathcal{S}[p(x, y)] = \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\,.$$

Aside: Information Theory

The entropy is **additive** for two independent random variables x and y, with $p(x,y) = p(x)\,p(y)$:

$$
\begin{aligned}
\mathcal{S}[p(x,y)] &= -\sum_{x,y} p(x,y) \log p(x,y) \\
&= -\sum_{x,y} p(x)p(y) \left[\log p(x) + \log p(y)\right] \\
&= -\sum_{x} p(x) \log p(x) - \sum_{y \in \mathcal{Y}} p(y) \log p(y) \\
&= \mathcal{S}[p(x)] + \mathcal{S}[p(y)] \ .
\end{aligned}
$$

Aside: Information Theory

The entropy is **additive** for two independent random variables x and y, with $p(x, y) = p(x)\, p(y)$:

$$\mathcal{S}[p(x, y)] = \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\,.$$

Aside: Information Theory

The entropy is **additive** for two independent random variables x and y, with $p(x,y) = p(x)\,p(y)$:

$$\mathcal{S}[p(x,y)] = \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\,.$$

For two statistically *dependent* observables, constrained by a nonzero interaction, the entropy is **subadditive**:

$$\mathcal{S}[p(x,y)] < \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\ .$$

Aside: Information Theory

The entropy is **additive** for two independent random variables x and y, with $p(x, y) = p(x)\, p(y)$:

$$\mathcal{S}[p(x, y)] = \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\,.$$

For two statistically *dependent* observables, constrained by a nonzero interaction, the entropy is **subadditive**:

$$\mathcal{S}[p(x, y)] < \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\ .$$

This is physically intuitive: for *dependent* variables observing y doesn't inform as much as it would have if we didn't know x.

Aside: Information Theory

The **mutual information** (MI) between two random variables is

$$\begin{aligned}\mathcal{I}\left[p(x,y)\right] &\equiv \mathcal{S}[p(x)] + \mathcal{S}[p(y)] - \mathcal{S}[p(x,y)] \\ &= \sum_{x,y} p(x,y) \log\left[\frac{p(x,y)}{p(x)\,p(y)}\right] \,.\end{aligned}$$

Aside: Information Theory

The **mutual information** (MI) between two random variables is

$$\begin{aligned}\mathcal{I}\,[p(x,y)] &\equiv \mathcal{S}[p(x)] + \mathcal{S}[p(y)] - \mathcal{S}[p(x,y)] \\ &= \sum_{x,y} p(x,y) \log\left[\frac{p(x,y)}{p(x)\,p(y)}\right] \;.\end{aligned}$$

- A functional of a joint probability distribution.

Aside: Information Theory

The **mutual information** (MI) between two random variables is

$$\mathcal{I}\left[p(x,y)\right] \equiv \mathcal{S}[p(x)] + \mathcal{S}[p(y)] - \mathcal{S}[p(x,y)]$$
$$= \sum_{x,y} p(x,y) \log\left[\frac{p(x,y)}{p(x)\,p(y)}\right] .$$

- A functional of a joint probability distribution.
- An average measure of how much information an observation of x conveys about an observation of y, and vice versa.

Aside: Information Theory

Rearranging, we see that the *subadditivity* of the entropy implies the nonnegativity of the mutual information:

$$\mathcal{I}\left[p(x,y)\right] \geq 0\,.$$

Aside: Information Theory

Rearranging, we see that the *subadditivity* of the entropy implies the nonnegativity of the mutual information:

$$\mathcal{I}\left[p(x, y)\right] \geq 0\,.$$

The MI of a joint distribution is telling us about the *interactions* that create the nontrivial non-Gaussian correlations, making it a *diagnostic of statistical dependence*.

Optimal Aspect Ratio

In deep networks, we want to balance the positive effect of *representation learning* against the negative effect of *fluctuations*.

Optimal Aspect Ratio

In deep networks, we want to balance the positive effect of *representation learning* against the negative effect of *fluctuations*.

- Intuitively, correlation between deep-layer neurons – or "wiring" – is a consequence of NTK differentials.

Optimal Aspect Ratio

In deep networks, we want to balance the positive effect of *representation learning* against the negative effect of *fluctuations*.

- Intuitively, correlation between deep-layer neurons – or "wiring" – is a consequence of NTK differentials.
- The MI could provide an accessible measure of the potential for such **inductive bias** in the *prior distribution*.

Optimal Aspect Ratio

Consider the *single-input* MI between two sets of neurons, $\mathcal{M}_1 = \{1, \ldots, m_1\}$ and $\mathcal{M}_2 = \{m_1 + 1, \ldots, m_1 + m_2\}$, in layer L:

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] \equiv \mathcal{S}[p(\mathcal{M}_1|x)] + \mathcal{S}[p(\mathcal{M}_2|x)] - \mathcal{S}[p(\mathcal{M}_1, \mathcal{M}_2|x)] \; .$$

Optimal Aspect Ratio

Consider the *single-input* MI between two sets of neurons, $\mathcal{M}_1 = \{1, \dots, m_1\}$ and $\mathcal{M}_2 = \{m_1 + 1, \dots, m_1 + m_2\}$, in layer L:

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] \equiv \mathcal{S}[p(\mathcal{M}_1|x)] + \mathcal{S}[p(\mathcal{M}_2|x)] - \mathcal{S}[p(\mathcal{M}_1, \mathcal{M}_2|x)] \ .$$

With $r \equiv L/n$, we can *estimate* the optimal aspect ratio as

$$r^\star \equiv \operatorname*{arg\,max}_{\{r, \mathcal{M}_1, \mathcal{M}_2\}} \mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] \ .$$

Aside: Unsupervised Learning

Maximizing this MI is related to **unsupervised learning** objectives.

Aside: Unsupervised Learning

Maximizing this MI is related to **unsupervised learning** objectives.

- The **InfoMax principle** recommends maximizing the mutual information between the input x and a *representation* $z(x)$.
- A related notion involves maximizing the mutual information between different representations, $z_1(x)$ and $z_2(x)$, for the same input x. This latter notion can be shown to *lower bound* the InfoMax objective and thus motivates our analysis here.

Slide unavailable.

Third-Order Result: Entropy

Using a **variational principle**, we can compute $\mathcal{S}$ to $O(1/n^3)$:

$$\begin{aligned}\mathcal{S}[p(z_1,\ldots,z_m|x)] =& \frac{m}{2}\log(2\pi e G) - \frac{(m^2+2m)}{16}\left(\frac{V}{nG^2}\right)^2 \\ &+ \frac{(m^3+10m^2+16m)}{48}\left(\frac{V}{nG^2}\right)^3 + O\left(\frac{1}{n^4}\right) .\end{aligned}$$

Third-Order Result: Entropy

Using a **variational principle**, we can compute $\mathcal{S}$ to $O(1/n^3)$:

$$
\begin{aligned}
\mathcal{S}[p(z_1, \ldots, z_m|x)] = & \frac{m}{2}\log(2\pi e G) - \frac{(m^2+2m)}{16}\left(\frac{V}{nG^2}\right)^2 \\
& + \frac{(m^3+10m^2+16m)}{48}\left(\frac{V}{nG^2}\right)^3 + O\left(\frac{1}{n^4}\right) .
\end{aligned}
$$

- Note that the correction is definitely negative.

Third-Order Result: Entropy

Using a **variational principle**, we can compute $\mathcal{S}$ to $O(1/n^3)$:

$$\mathcal{S}[p(z_1,\ldots,z_m|x)] = \frac{m}{2}\log(2\pi e G) - \frac{(m^2+2m)}{16}\left(\frac{V}{nG^2}\right)^2 + \frac{(m^3+10m^2+16m)}{48}\left(\frac{V}{nG^2}\right)^3 + O\left(\frac{1}{n^4}\right).$$

- Note that the correction is definitely negative.
- Note that unlike all our previous results, the leading correction here is *second order* in the inverse layer width $\sim V^2/n^2$.

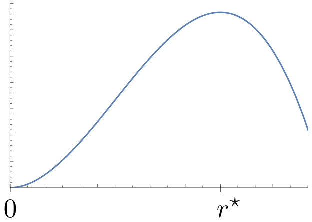
Optimal Aspect Ratio

Defining for the four-point vertex ratio,

$$\frac{V^{(L)}}{n_{L-1}\left(G^{(L)}\right)^2} \equiv \nu r\,,$$

with ν an *activation function–dependent constant*, we get:

$$\mathcal{I}\left[p(\mathcal{M}_1,\mathcal{M}_2|x)\right] = \frac{m_1 m_2}{8}\nu^2 r^2 - \frac{m_1 m_2(20+3m_1+3m_2)}{48}\nu^3 r^3 + O\left(r^4\right)$$



Optimal Aspect Ratio

Defining for the four-point vertex ratio,

$$\frac{V^{(L)}}{n_{L-1}\left(G^{(L)}\right)^2} \equiv \nu r\,,$$

with ν an *activation function–dependent constant*, we get:

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] = \frac{m_1 m_2}{8}\nu^2 r^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48}\nu^3 r^3 + O\Big(r^4\Big)$$

Noting the differing signs, we find:

$$r^\star = \left(\frac{4}{20 + 3n_L}\right)\frac{1}{\nu}\,.$$

Optimal Aspect Ratio

Defining for the four-point vertex ratio,

$$\frac{V^{(L)}}{n_{L-1}\left(G^{(L)}\right)^2} \equiv \nu r\,,$$

with ν an *activation function–dependent constant*, we get:

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] = \frac{m_1 m_2}{8}\nu^2 r^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48}\nu^3 r^3 + O\Big(r^4\Big)$$

Noting the differing signs, we find:

$$r^\star = \left(\frac{4}{20 + 3n_L}\right)\frac{1}{\nu}\,.$$

- For $n_L = 10$, $r^\star = .12$ ($\texttt{tanh}$) and $r^\star = .016$ ($\texttt{ReLU}$).

Optimal Aspect Ratio

Defining for the four-point vertex ratio,

$$\frac{V^{(L)}}{n_{L-1}\left(G^{(L)}\right)^2} \equiv \nu r\,,$$

with ν an *activation function–dependent constant*, we get:

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] = \frac{m_1 m_2}{8}\nu^2 r^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48}\nu^3 r^3 + O\Big(r^4\Big)$$

Noting the differing signs, we find:

$$r^\star = \left(\frac{4}{20 + 3n_L}\right)\frac{1}{\nu}\,.$$

- For $n_L = 10$, $r^\star = .12$ ($\texttt{tanh}$) and $r^\star = .016$ ($\texttt{ReLU}$).
- **Residual connections** let us push $r^\star$ to arbitrary depths.

Epilogue: Model Complexity

> *According to the hype of 1987, neural networks were meant to be intelligent models that discovered features and patterns in data. Gaussian processes in contrast are simply smoothing devices. How can Gaussian processes possibly replace neural networks? Were neural networks over-hyped, or have we underestimated the power of smoothing methods?*
>
> David MacKay

- The success of **overparameterized** models with far more parameters than training data has led many to conjecture that "more is better" when it comes to deep learning.
- There's mounting *empirical* evidence that a **scaling hypothesis** captures the behavior of deep neural networks, signaling the optimality of the overparameterized regime.

Epilogue: Model Complexity

The **Occam's razor** *principle of sparsity* posits that we should favor the simplest hypothesis that explains our observations:

- In the context of machine learning, this is usually interpreted to mean that we should prefer models with fewer parameters when comparing models performing the same tasks.
- We expect that models with fewer parameters will have smaller *generalization errors*, and will *overfit* less.

Epilogue: Model Complexity

The **Occam's razor** *principle of sparsity* posits that we should favor the simplest hypothesis that explains our observations:

- In the context of machine learning, this is usually interpreted to mean that we should prefer models with fewer parameters when comparing models performing the same tasks.
- We expect that models with fewer parameters will have smaller *generalization errors*, and will *overfit* less.

To conclude our lectures, we're going to see how to resolve this puzzling within our framework.

Epilogue: Model Complexity

This hinges on the notion of **model complexity**:

Epilogue: Model Complexity

This hinges on the notion of **model complexity**:

- On the one hand, the orthodox discussion of generalization takes a **microscopic perspective** – focusing on how a network works in terms of its low-level components – and wants to identify model complexity with *model parameters*.

Epilogue: Model Complexity

This hinges on the notion of **model complexity**:

- On the one hand, the orthodox discussion of generalization takes a **microscopic perspective** – focusing on how a network works in terms of its low-level components – and wants to identify model complexity with *model parameters*.
- On the other hand, in these lectures we integrated out the model parameters and developed a **macroscopic perspective** – providing an effective theory description of the predictions of realistic fully-trained networks – for which this notion of model complexity is completely *reversed*.

Epilogue: Model Complexity

We now know it's the **depth-to-width aspect ratio**,

$$r \equiv L/n \,,$$

controlling the complexity of overparameterized neural networks.

- It's the number of **data-dependent couplings** specifying the truncated nearly-Gaussian distribution – and *not* the number of *model parameters* – that ultimately define the model complexity in deep learning.

Sparsity at Infinite Width

At *infinite width*, we found a **Gaussian** trained distribution:

$$\lim_{n\to\infty} p\Big(z(\infty)\Big) \equiv p\Big(z(\infty)\Big|y_{\tilde{\alpha}}, K_{\delta_1\delta_2}, \Theta_{\delta_1\delta_2}\Big)\ .$$

Sparsity at Infinite Width

At *infinite width*, we found a **Gaussian** trained distribution:

$$\lim_{n\to\infty} p\Big(z(\infty)\Big) \equiv p\Big(z(\infty)\Big| y_{\tilde{\alpha}}, K_{\delta_1\delta_2}, \Theta_{\delta_1\delta_2}\Big) \ .$$

- The reason for writing it as a *conditional distribution* in this way is that the mean is only a function of $y_{\tilde{\alpha}}$ and $\Theta^{(L)}_{\delta_1\delta_2}$, while the variance is only a function of $K^{(L)}_{\delta_1\delta_2}$ and $\Theta^{(L)}_{\delta_1\delta_2}$.

Sparsity at Infinite Width

At *infinite width*, we found a **Gaussian** trained distribution:

$$\lim_{n\to\infty} p\Big(z(\infty)\Big) \equiv p\Big(z(\infty)\Big| y_{\tilde{\alpha}}, K_{\delta_1\delta_2}, \Theta_{\delta_1\delta_2}\Big) \ .$$

- The reason for writing it as a *conditional distribution* in this way is that the mean is only a function of $y_{\tilde{\alpha}}$ and $\Theta^{(L)}_{\delta_1\delta_2}$, while the variance is only a function of $K^{(L)}_{\delta_1\delta_2}$ and $\Theta^{(L)}_{\delta_1\delta_2}$.
- This is **sparse**, depending on a few objects in a simple way.

Near-Sparsity at Finite Width

At *finite width*, we found a **nearly-Gaussian** trained distribution:

$$p\Big(z(\infty)\Big) \equiv p\Big(z(\infty)\Big| y, G, H, V, A, B, D, F, P, Q, R, S, T, U\Big) + O\left(\frac{1}{n^2}\right) .$$

Near-Sparsity at Finite Width

At *finite width*, we found a **nearly-Gaussian** trained distribution:

$$p\Big(z(\infty)\Big) \equiv p\Big(z(\infty)\Big| y, G, H, V, A, B, D, F, P, Q, R, S, T, U\Big) + O\left(\frac{1}{n^2}\right) .$$

- In addition to G and H, here we're accounting for the finite-width *data-dependent couplings* arising from:

$$\mathbb{E}\left[zzzz\right]_{\text{connected}} , \qquad \mathbb{E}\left[\widehat{\Delta H} zz\right] , \qquad \mathbb{E}\left[\widehat{\Delta H}^2\right] ,$$
$$\mathbb{E}\left[\widehat{\mathrm{d}H} z\right] , \qquad \mathbb{E}\left[\widehat{\mathrm{dd_I} H}\right] , \qquad \mathbb{E}\left[\widehat{\mathrm{dd_{II}} H}\right] .$$

Near-Sparsity at Finite Width

At *finite width*, we found a **nearly-Gaussian** trained distribution:

$$p\Big(z(\infty)\Big) \equiv p\Big(z(\infty)\Big| y, G, H, V, A, B, D, F, P, Q, R, S, T, U\Big) + O\left(\frac{1}{n^2}\right) .$$

- In addition to G and H, here we're accounting for the finite-width *data-dependent couplings* arising from:

$$\mathbb{E}\left[zzzz\right]_{\text{connected}} , \qquad \mathbb{E}\left[\widehat{\Delta H} zz\right] , \qquad \mathbb{E}\left[\widehat{\Delta H}^2\right] ,$$
$$\mathbb{E}\left[\widehat{\mathrm{d}H} z\right] , \qquad \mathbb{E}\left[\widehat{\mathrm{dd}_{\mathrm{I}} H}\right] , \qquad \mathbb{E}\left[\widehat{\mathrm{dd}_{\mathrm{II}} H}\right] .$$

- This is **nearly-sparse**, depending only on two-hands-full of objects in a *nearly-simple* way.

Model Complexity of Fully-Trained Neural Networks

Consider a fixed combined training and test dataset of size $N_{\mathcal{D}}$:

- For the *infinite-width* **Gaussian distribution**, we only need

$$n_{\text{out}} N_{\mathcal{A}} + \left[\frac{N_{\mathcal{D}}(N_{\mathcal{D}}+1)}{2}\right] + \left[\frac{N_{\mathcal{D}}(N_{\mathcal{D}}+1)}{2}\right] = O\Big(N_{\mathcal{D}}^2\Big)$$

  numbers in order to completely specify the distribution.

Model Complexity of Fully-Trained Neural Networks

Consider a fixed combined training and test dataset of size $N_{\mathcal{D}}$:

- For the *finite-width* **nearly-Gaussian distribution** with $0 < r \ll 1$, we will instead need $O\left(N_{\mathcal{D}}^4\right)$ numbers, with the counting dominated by the finite-width tensors.

Model Complexity of Fully-Trained Neural Networks

Consider a fixed combined training and test dataset of size $N_{\mathcal{D}}$:

- For an accuracy $O\Big(L^k/n^k\Big)$, a *macroscopic description*

$$p\Big(z(\infty)\Big) = \sum_{m=0}^{k} \frac{p^{\{m\}}\Big(z(\infty)\Big)}{n^m} + O\left(\frac{L^{k+1}}{n^{k+1}}\right),$$

will need $O\Big(N_{\mathcal{D}}^{2k}\Big)$ numbers in general.

Model Complexity of Fully-Trained Neural Networks

Consider a fixed combined training and test dataset of size $N_{\mathcal{D}}$:

- For an accuracy $O\Big(L^k/n^k\Big)$, a *macroscopic description*

$$p\Big(z(\infty)\Big) = \sum_{m=0}^{k} \frac{p^{\{m\}}\Big(z(\infty)\Big)}{n^m} + O\left(\frac{L^{k+1}}{n^{k+1}}\right),$$

  will need $O\Big(N_{\mathcal{D}}^{2k}\Big)$ numbers in general.
- The $\boldsymbol{1/n}$ **expansion** gives a sequence of effective theories with increasing accuracy at the cost of increasing complexity.

Model Complexity of Fully-Trained Neural Networks

What we have found here is the manifestation of the **microscopic-macroscopic duality**:

Model Complexity of Fully-Trained Neural Networks

What we have found here is the manifestation of the **microscopic-macroscopic duality**:

- Complexity in *parameter space* is traded into simplicity in *sample space*, and density in *model parameters* is exchanged for sparsity in *data-dependent couplings*.

Model Complexity of Fully-Trained Neural Networks

What we have found here is the manifestation of the **microscopic-macroscopic duality**:

- Complexity in *parameter space* is traded into simplicity in *sample space*, and density in *model parameters* is exchanged for sparsity in *data-dependent couplings*.
- In the *overparameterized regime*, this indicates that we should identify the *model complexity* with the **data-dependent couplings** rather than the **model parameters**.

Model Complexity of Fully-Trained Neural Networks

As r increases, we'll need to include more of these higher-order terms, making our macroscopic description more complex:

Model Complexity of Fully-Trained Neural Networks

As r increases, we'll need to include more of these higher-order terms, making our macroscopic description more complex:

- In the strict limit $r \to 0$, the *sparse* $O(N_{\mathcal{D}}^2)$ **Gaussian** description of the infinite-width limit will be accurate.

Model Complexity of Fully-Trained Neural Networks

As r increases, we'll need to include more of these higher-order terms, making our macroscopic description more complex:

- In the strict limit $r \to 0$, the *sparse* $O(N_{\mathcal{D}}^2)$ **Gaussian** description of the infinite-width limit will be accurate.
- In the regime $0 < r \sim r^\star \ll 1$, the *nearly-sparse* $O(N_{\mathcal{D}}^4)$ **nearly-Gaussian** description of the finite-width effective theory truncated at order $1/n$ will be accurate.

Model Complexity of Fully-Trained Neural Networks

As r increases, we'll need to include more of these higher-order terms, making our macroscopic description more complex:

- In the strict limit $r \to 0$, the *sparse* $O\left(N_{\mathcal{D}}^2\right)$ **Gaussian** description of the infinite-width limit will be accurate.
- In the regime $0 < r \sim r^\star \ll 1$, the *nearly-sparse* $O\left(N_{\mathcal{D}}^4\right)$ **nearly-Gaussian** description of the finite-width effective theory truncated at order $1/n$ will be accurate.
- For larger r, a more generic $O\left(N_{\mathcal{D}}^{2k}\right)$ **non-Gaussian** description would in principle be necessary.

Conclusion

The practical success of deep learning in the *overparameterized* regime and the empirical accuracy of a simple *scaling hypothesis* is really telling us that useful neural networks should be **sparse** – hence the preference for larger and larger models – but not too sparse – so that they are also **deep**. Thus, from the macroscopic perspective, a **nearly-sparse** model complexity is perhaps the most important inductive bias of deep learning.

Conclusion

The practical success of deep learning in the *overparameterized* regime and the empirical accuracy of a simple *scaling hypothesis* is really telling us that useful neural networks should be **sparse** – hence the preference for larger and larger models – but not too sparse – so that they are also **deep**. Thus, from the macroscopic perspective, a **nearly-sparse** model complexity is perhaps the most important inductive bias of deep learning.

Thank You!