Effective Theory of Deep Learning
Beyond the Infinite-Width Limit

Dan Roberts\textsuperscript{a} and Sho Yaida\textsuperscript{b}

\textsuperscript{a}MIT, IAIFI, & Salesforce, \textsuperscript{b}Facebook AI Research

Deep Learning Theory Summer School at Princeton
July 27, 2021 – August 8, 2021
Lecture 1  Initialization, Linear Models
\[ \text{§0 + §7.1 + §10.4} \]

Lecture 2  Quadratic Models & Nearly-Kernel Methods
\[ \text{§11.4 (§7.2) + §∞.2.2} \]

Lecture 3  The Principle of Sparsity (Recurring)
\[ \text{§4, §8, §11.2, §∞.3} \]

Lecture 4  The Principle of Criticality
\[ \text{§5, §9, §11.3, §∞.1, §10.3.1} \]

Lecture 5  The End of Training, & More
\[ \text{§∞.2.3 + §10.3.2 + §A.3 + §ɛ} \]
The End of Training

\[ z_{i,j}(t = \infty) = z_{i,3} - \sum_{j,k,\tilde{\alpha}_1,\tilde{\alpha}_2} \hat{H}_{ij;\tilde{\alpha}_1,\tilde{\alpha}_2} \left( \hat{H}^{-1} \right)_{jk} \tilde{\alpha}_2 (z_{k,\tilde{\alpha}_2} - y_{k,\tilde{\alpha}_2}) \]

\[ + \sum_{j_1,j_2,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4} \left[ \frac{dH_{j_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2}}{d\tilde{\alpha}_1,\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6} H_{\tilde{\alpha}_5,\tilde{\alpha}_6} \hat{H}_{i_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2} \right] Z_{A}^{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4} (z_{j_1,\tilde{\alpha}_3} - y_{j_1,\tilde{\alpha}_3}) (z_{j_2,\tilde{\alpha}_4} - y_{j_2,\tilde{\alpha}_4}) \]

\[ + \sum_{j_1,j_2,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4} \left[ \frac{dH_{j_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2}}{d\tilde{\alpha}_1,\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5,\tilde{\alpha}_6} H_{\tilde{\alpha}_5,\tilde{\alpha}_6} \hat{H}_{i_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2} \right] Z_{B}^{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4} (z_{j_1,\tilde{\alpha}_3} - y_{j_1,\tilde{\alpha}_3}) (z_{j_2,\tilde{\alpha}_4} - y_{j_2,\tilde{\alpha}_4}) \]

\[ + \sum_{j_1,j_2,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6} \left[ \frac{ddH_{j_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2}}{d\tilde{\alpha}_1,\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_7,\tilde{\alpha}_8} H_{\tilde{\alpha}_7,\tilde{\alpha}_8} \hat{H}_{i_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2} \right] Z_{A}^{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6} (z_{j_1,\tilde{\alpha}_4} - y_{j_1,\tilde{\alpha}_4}) (z_{j_2,\tilde{\alpha}_5} - y_{j_2,\tilde{\alpha}_5}) (z_{j_3,\tilde{\alpha}_6} - y_{j_3,\tilde{\alpha}_6}) \]

\[ + \sum_{j_1,j_2,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6} \left[ \frac{ddH_{j_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2}}{d\tilde{\alpha}_1,\tilde{\alpha}_2} - \sum_{\tilde{\alpha}_7,\tilde{\alpha}_8} H_{\tilde{\alpha}_7,\tilde{\alpha}_8} \hat{H}_{i_1,j_2;\tilde{\alpha}_1,\tilde{\alpha}_2} \right] Z_{B}^{\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4,\tilde{\alpha}_5,\tilde{\alpha}_6} (z_{j_1,\tilde{\alpha}_4} - y_{j_1,\tilde{\alpha}_4}) (z_{j_2,\tilde{\alpha}_5} - y_{j_2,\tilde{\alpha}_5}) (z_{j_3,\tilde{\alpha}_6} - y_{j_3,\tilde{\alpha}_6}) \]

\[ + O \left( \frac{1}{n^2} \right) \]
The End of Training

\[ z_{i;\beta}(t = \infty) = z_{i;\beta} - \sum_{j,k,\alpha_1,\alpha_2} \hat{H}_{ij;\beta\alpha_1} (\hat{H}^{-1})_{jk} (z_{k;\alpha_2} - y_{k;\alpha_2}) \]

\[ + \sum_{j_1,j_2,\alpha_1,\alpha_2,\alpha_3,\alpha_4} dH_{j_1,j_2;\beta\alpha_1} \beta\alpha_2 - \sum_{\alpha_5,\alpha_6} H_{\beta\alpha_5} \hat{H}_{\alpha_6}\hat{dH}_{j_1,j_2;\alpha_1\alpha_6\alpha_2} \]

\[ + \sum_{j_1,j_2,\alpha_1,\alpha_2,\alpha_3,\alpha_4} \hat{dH}_{j_1,j_2;\beta\alpha_1} \beta\alpha_2 - \sum_{\alpha_5,\alpha_6} H_{\beta\alpha_5} \hat{H}_{\alpha_6}\hat{dH}_{j_1,j_2;\alpha_1\alpha_6\alpha_2} \]

\[ + \sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6} d\hat{dH}_{j_1,j_2,j_3;\beta\alpha_1\alpha_2\alpha_3 - \sum_{\alpha_7,\alpha_8} H_{\beta\alpha_7} \hat{H}_{\alpha_8}\hat{ddH}_{j_1,j_2,j_3;\alpha_1\alpha_8\alpha_2\alpha_3} \]

\[ + \sum_{j_1,j_2,j_3,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6} d\hat{ddH}_{j_1,j_2,j_3;\beta\alpha_1\alpha_2\alpha_3 - \sum_{\alpha_7,\alpha_8} H_{\beta\alpha_7} \hat{H}_{\alpha_8}\hat{ddH}_{j_1,j_2,j_3;\alpha_1\alpha_8\alpha_2\alpha_3} \]

\[ + O\left(\frac{1}{n^2}\right) \]
Complete Inverse of the Stochastic NTK

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

\[ \sum_{j,\tilde{\alpha}_2} (\hat{H}^{-1})_{ij} \hat{H}_{jk;\tilde{\alpha}_2\tilde{\alpha}_3} = \delta_{ik} \delta_{\tilde{\alpha}_1\tilde{\alpha}_3}. \]
Here, we needed the complete inverse of the stochastic NTK:

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

Defining a mean and fluctuation as

\[
\hat{H}_{i_1 i_2; \alpha_1 \alpha_2} \equiv \delta_{i_1 i_2} H_{\alpha_1 \alpha_2} + \Delta \hat{H}_{i_1 i_2; \alpha_1 \alpha_2} ,
\]

we can then expand around the fluctuation to get:

\[
\left( \hat{H}^{-1} \right)_{ij}^{\tilde{\alpha}_1 \tilde{\alpha}_2} = \delta_{ij} \hat{H}^{\tilde{\alpha}_1 \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_3, \tilde{\alpha}_4 \in A} \hat{H}^{\tilde{\alpha}_1 \tilde{\alpha}_3} \Delta \hat{H}_{ij; \tilde{\alpha}_3 \tilde{\alpha}_4} \hat{H}^{\tilde{\alpha}_4 \tilde{\alpha}_2} \\
+ \sum_{k=1}^{n_L} \sum_{\tilde{\alpha}_3, \ldots, \tilde{\alpha}_6 \in A} \hat{H}^{\tilde{\alpha}_1 \tilde{\alpha}_3} \Delta \hat{H}_{ik; \tilde{\alpha}_3 \tilde{\alpha}_4} \hat{H}^{\tilde{\alpha}_4 \tilde{\alpha}_5} \Delta \hat{H}_{kj; \tilde{\alpha}_5 \tilde{\alpha}_6} \hat{H}^{\tilde{\alpha}_6 \tilde{\alpha}_2} \\
+ O(\Delta^3) .
\]
The End of Training

\[ z_{i;\hat{t}}(t = \infty) = z_{i;\hat{t}} - \sum_{j,k,\hat{a}_1,\hat{a}_2} \hat{H}_{ij;\hat{t}\hat{a}_1} \left( \hat{H}^{-1} \right)_{jk} (z_k;\hat{a}_2 - y_k;\hat{a}_2) \]

\[ + \sum_{j,\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4} d\hat{H}_{ij;\hat{a}_1\hat{a}_2} - \sum_{\hat{a}_5,\hat{a}_6} \hat{H}_{ij;\hat{a}_5} \hat{H}_{ij;\hat{a}_6} d\hat{H}_{ij;\hat{a}_6\hat{a}_2} \]

\[ + \sum_{j,\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4} d\hat{H}_{ij;\hat{a}_1\hat{a}_2} - \sum_{\hat{a}_5,\hat{a}_6} \hat{H}_{ij;\hat{a}_5} \hat{H}_{ij;\hat{a}_6} d\hat{H}_{ij;\hat{a}_6\hat{a}_2} \]

\[ + \sum_{j,\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4,\hat{a}_5,\hat{a}_6} \]

\[ + O \left( \frac{1}{n^2} \right) \]
The End of Training

\[ z_{i;\beta}(t = \infty) = z_{i;\beta} - \sum_{\tilde{a}_1, \tilde{a}_2} H_{\beta \tilde{a}_1} \tilde{H}_{1 \tilde{a}_2}(z_{i;\tilde{a}_2} - y_{i;\tilde{a}_2}) \]

\[ + \sum_{j, \tilde{a}_1, \tilde{a}_2} \left[ \Delta H_{j;\beta \tilde{a}_1} - \sum_{\tilde{a}_3, \tilde{a}_4, \tilde{a}_5} H_{\beta \tilde{a}_3} \tilde{H}_{3 \tilde{a}_4} \Delta H_{j;\tilde{a}_4 \tilde{a}_1} \right] \tilde{H}_{1 \tilde{a}_2}(z_{j;\tilde{a}_2} - y_{j;\tilde{a}_2}) \]

\[ - \sum_{j, k, \tilde{a}_1, \ldots, \tilde{a}_4} \left[ \Delta H_{j;\beta \tilde{a}_1} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{\beta \tilde{a}_5} \tilde{H}_{5 \tilde{a}_6} \Delta H_{j;\tilde{a}_6 \tilde{a}_1} \right] \tilde{H}_{1 \tilde{a}_2} \Delta H_{k;2 \tilde{a}_3} \tilde{H}_{3 \tilde{a}_4}(z_{k;\tilde{a}_4} - y_{k;\tilde{a}_4}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} \tilde{H}_{j;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{\beta \tilde{a}_7} \tilde{H}_{7 \tilde{a}_8} \tilde{d} \tilde{H}_{j;\tilde{a}_7 \tilde{a}_8} \right] \tilde{Z}_{A}^{\tilde{a}_1 \tilde{a}_2 \tilde{a}_3 \tilde{a}_4}(z_{j;\tilde{a}_3} - y_{j;\tilde{a}_3})(z_{j;\tilde{a}_4} - y_{j;\tilde{a}_4}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} \tilde{H}_{j;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{\beta \tilde{a}_7} \tilde{H}_{7 \tilde{a}_8} \tilde{d} \tilde{H}_{j;\tilde{a}_7 \tilde{a}_8} \right] \tilde{Z}_{B}^{\tilde{a}_1 \tilde{a}_2 \tilde{a}_3 \tilde{a}_4}(z_{j;\tilde{a}_3} - y_{j;\tilde{a}_3})(z_{j;\tilde{a}_4} - y_{j;\tilde{a}_4}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{\beta \tilde{a}_7} \tilde{H}_{7 \tilde{a}_8} \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_7 \tilde{a}_8} \right] \tilde{Z}_{IA}^{\tilde{a}_1 \tilde{a}_2 \tilde{a}_3 \tilde{a}_4 \tilde{a}_5 \tilde{a}_6}(z_{j;\tilde{a}_4} - y_{j;\tilde{a}_4})(z_{j;\tilde{a}_5} - y_{j;\tilde{a}_5})(z_{j;\tilde{a}_6} - y_{j;\tilde{a}_6}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{\beta \tilde{a}_7} \tilde{H}_{7 \tilde{a}_8} \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_7 \tilde{a}_8} \right] \tilde{Z}_{IB}^{\tilde{a}_1 \tilde{a}_2 \tilde{a}_3 \tilde{a}_4 \tilde{a}_5 \tilde{a}_6}(z_{j;\tilde{a}_4} - y_{j;\tilde{a}_4})(z_{j;\tilde{a}_5} - y_{j;\tilde{a}_5})(z_{j;\tilde{a}_6} - y_{j;\tilde{a}_6}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{\beta \tilde{a}_7} \tilde{H}_{7 \tilde{a}_8} \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_7 \tilde{a}_8} \right] \tilde{Z}_{IA}^{\tilde{a}_1 \tilde{a}_2 \tilde{a}_3 \tilde{a}_4 \tilde{a}_5 \tilde{a}_6}(z_{j;\tilde{a}_4} - y_{j;\tilde{a}_4})(z_{j;\tilde{a}_5} - y_{j;\tilde{a}_5})(z_{j;\tilde{a}_6} - y_{j;\tilde{a}_6}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{\beta \tilde{a}_7} \tilde{H}_{7 \tilde{a}_8} \tilde{d} \tilde{d} \tilde{H}_{j;\tilde{a}_7 \tilde{a}_8} \right] \tilde{Z}_{IB}^{\tilde{a}_1 \tilde{a}_2 \tilde{a}_3 \tilde{a}_4 \tilde{a}_5 \tilde{a}_6}(z_{j;\tilde{a}_4} - y_{j;\tilde{a}_4})(z_{j;\tilde{a}_5} - y_{j;\tilde{a}_5})(z_{j;\tilde{a}_6} - y_{j;\tilde{a}_6}) \]

\[ + O\left(\frac{1}{n^2}\right)\]
The End of Training

\[ z_i;\beta(t = \infty) = z_i;\beta - \sum_{\tilde{a}_1, \tilde{a}_2} H_{\beta;\tilde{a}_1} \tilde{H}^{\beta_1 \tilde{a}_2}(z_i;\tilde{a}_2 - y_i;\tilde{a}_2) \]

\[ + \sum_{j, \tilde{a}_1, \tilde{a}_2} \left[ \Delta H_{i;j;\beta;\tilde{a}_1} - \sum_{\tilde{a}_3, \tilde{a}_4 \in A} H_{\beta;\tilde{a}_3} \tilde{H}^{\beta_4 \tilde{a}_4} \Delta H_{i;j;\tilde{a}_4 \tilde{a}_1} \right] \tilde{H}^{\beta_1 \tilde{a}_2}(z_j;\tilde{a}_2 - y_j;\tilde{a}_2) \]

\[ - \sum_{j, k, \tilde{a}_1, \tilde{a}_2} \left[ \Delta H_{i;j;\beta;\tilde{a}_1} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{\beta;\tilde{a}_5} \tilde{H}^{\beta_6 \tilde{a}_6} \Delta H_{i;j;\tilde{a}_6 \tilde{a}_1} \right] \tilde{H}^{\beta_1 \tilde{a}_2} \Delta H_{j;k;\tilde{a}_2 \tilde{a}_3} \tilde{H}^{\beta_4 \tilde{a}_4}(z_k;\tilde{a}_4 - y_k;\tilde{a}_4) \]

\[ + \sum_{j_i, j_2, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} H_{i;j_1;j_2;\beta;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{\beta;\tilde{a}_5} \tilde{H}^{\beta_6 \tilde{a}_6} \tilde{d} H_{i;j_1;j_2;\tilde{a}_1 \tilde{a}_6 \tilde{a}_2} \right] \tilde{Z}_A^{\beta_1 \tilde{a}_2 \tilde{a}_4}(z_{j_1;\tilde{a}_3} - y_{j_1;\tilde{a}_3})(z_{j_2;\tilde{a}_4} - y_{j_2;\tilde{a}_4}) \]

\[ + \sum_{j_i, j_2, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d} H_{i;j_1;j_2;\beta;\tilde{a}_1 \tilde{a}_2 \tilde{a}_3} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{\beta;\tilde{a}_5} \tilde{H}^{\beta_6 \tilde{a}_6} \tilde{d} H_{i;j_1;j_2;\tilde{a}_1 \tilde{a}_6 \tilde{a}_2} \right] \tilde{Z}_B^{\beta_1 \tilde{a}_2 \tilde{a}_4}(z_{j_1;\tilde{a}_3} - y_{j_1;\tilde{a}_3})(z_{j_2;\tilde{a}_4} - y_{j_2;\tilde{a}_4}) \]

\[ + O\left(\frac{1}{n^2}\right) \]
Two Extra Differentials??

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

\[
\tilde{d}d II H^{(\ell)}_{ij i_1 i_2 i_3; \delta_1 \delta_2 \delta_3 \delta_4} \\
\equiv \sum_{\ell=1}^{\ell} \sum_{\mu_1, \nu_1,} \sum_{\mu_2, \nu_2,} \sum_{\mu_3, \nu_3} \lambda^{(\ell_1)}_{\mu_1 \nu_1} \lambda^{(\ell_2)}_{\mu_2 \nu_2} \lambda^{(\ell_3)}_{\mu_3 \nu_3} \frac{d^2 z^{(\ell)}_{i_1; \delta_1}}{d \theta^{(\ell_1)}_{\mu_1} d \theta^{(\ell_3)}_{\mu_3}} \frac{d^2 z^{(\ell)}_{i_2; \delta_2}}{d \theta^{(\ell_2)}_{\mu_2} d \theta^{(\ell_3)}_{\nu_3}} \frac{dz^{(\ell)}_{i_3; \delta_3}}{d \theta^{(\ell_1)}_{\nu_1}} \frac{dz^{(\ell)}_{i_4; \delta_4}}{d \theta^{(\ell_2)}_{\nu_2}}
\]
\[ z_{i;\beta}(t = \infty) = \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1} \tilde{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} [\tilde{\Delta} H_{j;\beta \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_3, \tilde{\alpha}_4 \in A} H_{\beta \tilde{\alpha}_3} \tilde{H} \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\Delta} H_{j;\tilde{\alpha}_4 \tilde{\alpha}_1}] \tilde{H} \tilde{\alpha}_1 \tilde{\alpha}_2 (z_{j;\tilde{\alpha}_2} - y_{j;\tilde{\alpha}_2}) - \sum_{j, k, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_4} [\tilde{\Delta} H_{j;\beta \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H} \tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\Delta} H_{j;\tilde{\alpha}_6 \tilde{\alpha}_1}] \tilde{H} \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\Delta} H_{k;\tilde{\alpha}_2 \tilde{\alpha}_3} \tilde{H} \tilde{\alpha}_3 \tilde{\alpha}_4 (z_{k;\tilde{\alpha}_4} - y_{k;\tilde{\alpha}_4}) + \sum_{j, j', \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} [\hat{d}H_{j, j';\tilde{\alpha}_1, \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H} \tilde{\alpha}_5 \tilde{\alpha}_6 \hat{d}H_{j, j';\tilde{\alpha}_6 \tilde{\alpha}_2}] \tilde{Z}_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, j', \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} [\hat{d}H_{j, j';\tilde{\alpha}_1, \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H} \tilde{\alpha}_5 \tilde{\alpha}_6 \hat{d}H_{j, j';\tilde{\alpha}_6 \tilde{\alpha}_2}] \tilde{Z}_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_{j, j', \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} [\hat{d}d_1 \hat{I} H_{j, j';\tilde{\alpha}_1, \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta \tilde{\alpha}_7} \tilde{H} \tilde{\alpha}_7 \tilde{\alpha}_8 \hat{d}d_1 \hat{I} H_{j, j';\tilde{\alpha}_7 \tilde{\alpha}_8 \tilde{\alpha}_2}] \tilde{Z}_A^{\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_{j, j', \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} [\hat{d}d_1 \hat{I} H_{j, j';\tilde{\alpha}_1, \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta \tilde{\alpha}_7} \tilde{H} \tilde{\alpha}_7 \tilde{\alpha}_8 \hat{d}d_1 \hat{I} H_{j, j';\tilde{\alpha}_7 \tilde{\alpha}_8 \tilde{\alpha}_2}] \tilde{Z}_B^{\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_{j, j', \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} [\hat{d}d_{11} \hat{I} H_{j, j';\tilde{\alpha}_1, \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta \tilde{\alpha}_7} \tilde{H} \tilde{\alpha}_7 \tilde{\alpha}_8 \hat{d}d_{11} \hat{I} H_{j, j';\tilde{\alpha}_7 \tilde{\alpha}_8 \tilde{\alpha}_2}] \tilde{Z}_A^{\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_{j, j', \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} [\hat{d}d_{11} \hat{I} H_{j, j';\tilde{\alpha}_1, \tilde{\alpha}_2} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta \tilde{\alpha}_7} \tilde{H} \tilde{\alpha}_7 \tilde{\alpha}_8 \hat{d}d_{11} \hat{I} H_{j, j';\tilde{\alpha}_7 \tilde{\alpha}_8 \tilde{\alpha}_2}] \tilde{Z}_B^{\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)\]
The End of Training

\[ z_{i;\bar{\alpha}}(t = \infty) = z_{i;\bar{\alpha}} - \sum_{\alpha_1,\alpha_2} H_{\beta_1\alpha_1} \bar{H}_{\alpha_1\alpha_2} (z_{i;\bar{\alpha}_2} - y_{i;\bar{\alpha}_2}) \]

\[ + \sum_{j,\alpha_1,\alpha_2} \left[ \Delta H_{ij;\bar{\alpha}_1} - \sum_{\alpha_3} H_{\beta_3\alpha_3} \bar{H}_{\alpha_3\alpha_4} \Delta H_{ij;\bar{\alpha}_4\alpha_1} \right] \bar{H}_{\alpha_1\alpha_2} (z_{j;\bar{\alpha}_2} - y_{j;\bar{\alpha}_2}) \]

\[ - \sum_{j,k \alpha_1,\ldots,\alpha_4} \left[ \Delta H_{ij;\bar{\alpha}_1} - \sum_{\alpha_5,\alpha_6} H_{\beta_5\alpha_5} \bar{H}_{\alpha_5\alpha_6} \Delta H_{ij;\bar{\alpha}_6\alpha_1} \right] \bar{H}_{\alpha_1\alpha_2} \Delta H_{jk;\bar{\alpha}_2\alpha_3} \bar{H}_{\alpha_3\alpha_4} (z_{k;\bar{\alpha}_4} - y_{k;\bar{\alpha}_4}) \]

\[ + \sum_{j_1,j_2,\alpha_1,\alpha_2,\alpha_3,\alpha_4} \left[ \tilde{d}_1 H_{j_1j_2;\alpha_2\alpha_3} - \sum_{\alpha_7,\alpha_8} H_{\beta_7\alpha_7} \bar{H}_{\alpha_7\alpha_8} \tilde{d}_1 H_{j_1j_2;\alpha_1\alpha_8\alpha_2\alpha_3} \right] Z_{A}^{\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6} (z_{j_1;\bar{\alpha}_4} - y_{j_1;\bar{\alpha}_4}) (z_{j_2;\bar{\alpha}_5} - y_{j_2;\bar{\alpha}_5}) (z_{j_3;\bar{\alpha}_6} - y_{j_3;\bar{\alpha}_6}) \]

\[ + \sum_{j_1,j_2,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\alpha_6} \left[ \tilde{d}_1 H_{j_1j_2;\alpha_2\alpha_3} - \sum_{\alpha_7,\alpha_8} H_{\beta_7\alpha_7} \bar{H}_{\alpha_7\alpha_8} \tilde{d}_1 H_{j_1j_2;\alpha_1\alpha_8\alpha_2\alpha_3} \right] Z_{B}^{\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6} (z_{j_1;\bar{\alpha}_4} - y_{j_1;\bar{\alpha}_4}) (z_{j_2;\bar{\alpha}_5} - y_{j_2;\bar{\alpha}_5}) (z_{j_3;\bar{\alpha}_6} - y_{j_3;\bar{\alpha}_6}) \]

\[ + O\left(\frac{1}{n^2}\right) \]
The End of Training

\[ z_{i:j}(t = \infty) = z_{i:j} - \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta_2^1, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_1, \tilde{\alpha}_2} (z_{i:\tilde{\alpha}_1} - y_{i:\tilde{\alpha}_1}) \]

\[ + \sum_{j, \tilde{\alpha}_1, \tilde{\alpha}_2} \left[ \Delta H_{j, j; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_3, \tilde{\alpha}_4} H_{\beta_3^3, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_3, \tilde{\alpha}_4} \Delta H_{j, j; \tilde{\alpha}_4} \right] \tilde{H}_{\tilde{\alpha}_1, \tilde{\alpha}_2} (z_{j: \tilde{\alpha}_1} - y_{j: \tilde{\alpha}_1}) \]

\[ - \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2} \left[ \Delta H_{j, k; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta_5^5, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_5, \tilde{\alpha}_6} \Delta H_{j, k; \tilde{\alpha}_6} \right] \tilde{H}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \Delta H_{k, k; \tilde{\alpha}_2} \tilde{H}_{\tilde{\alpha}_2, \tilde{\alpha}_4} (z_{k: \tilde{\alpha}_4} - y_{k: \tilde{\alpha}_4}) \]

\[ + \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ \hat{d} H_{j, j; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta_5^5, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_5, \tilde{\alpha}_6} \hat{d} H_{j, j; \tilde{\alpha}_6} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} (z_{j: \tilde{\alpha}_3} - y_{j: \tilde{\alpha}_3}) (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) \]

\[ + \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ \hat{d} H_{j, k; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta_5^5, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_5, \tilde{\alpha}_6} \hat{d} H_{j, k; \tilde{\alpha}_6} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} (z_{j: \tilde{\alpha}_3} - y_{j: \tilde{\alpha}_3}) (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) \]

\[ + \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ \hat{d} d H_{j, j; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta_7^7, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_7, \tilde{\alpha}_8} \hat{d} d H_{j, j; \tilde{\alpha}_8} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} \tilde{z}_{\tilde{\alpha}_5} (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) (z_{j: \tilde{\alpha}_5} - y_{j: \tilde{\alpha}_5}) (z_{j: \tilde{\alpha}_6} - y_{j: \tilde{\alpha}_6}) \]

\[ + \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ \hat{d} d H_{j, k; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta_7^7, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_7, \tilde{\alpha}_8} \hat{d} d H_{j, k; \tilde{\alpha}_8} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} \tilde{z}_{\tilde{\alpha}_5} (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) (z_{j: \tilde{\alpha}_5} - y_{j: \tilde{\alpha}_5}) (z_{j: \tilde{\alpha}_6} - y_{j: \tilde{\alpha}_6}) \]

\[ + \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ \hat{d} d d H_{j, j; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta_7^7, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_7, \tilde{\alpha}_8} \hat{d} d d H_{j, j; \tilde{\alpha}_8} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} \tilde{z}_{\tilde{\alpha}_5} \tilde{z}_{\tilde{\alpha}_6} (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) (z_{j: \tilde{\alpha}_5} - y_{j: \tilde{\alpha}_5}) (z_{j: \tilde{\alpha}_6} - y_{j: \tilde{\alpha}_6}) \]

\[ + \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ \hat{d} d d H_{j, k; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta_7^7, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_7, \tilde{\alpha}_8} \hat{d} d d H_{j, k; \tilde{\alpha}_8} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} \tilde{z}_{\tilde{\alpha}_5} \tilde{z}_{\tilde{\alpha}_6} (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) (z_{j: \tilde{\alpha}_5} - y_{j: \tilde{\alpha}_5}) (z_{j: \tilde{\alpha}_6} - y_{j: \tilde{\alpha}_6}) \]

\[ + \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4, \tilde{\alpha}_5, \tilde{\alpha}_6} \left[ \hat{d} d d d H_{j, j; \beta_1^1, \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_7, \tilde{\alpha}_8} H_{\beta_7^7, \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_7, \tilde{\alpha}_8} \hat{d} d d d H_{j, j; \tilde{\alpha}_8} \right] \tilde{z}_{\tilde{\alpha}_1, \tilde{\alpha}_2} \tilde{z}_{\tilde{\alpha}_3, \tilde{\alpha}_4} \tilde{z}_{\tilde{\alpha}_5} \tilde{z}_{\tilde{\alpha}_6} \tilde{z}_{\tilde{\alpha}_7} (z_{j: \tilde{\alpha}_4} - y_{j: \tilde{\alpha}_4}) (z_{j: \tilde{\alpha}_5} - y_{j: \tilde{\alpha}_5}) (z_{j: \tilde{\alpha}_6} - y_{j: \tilde{\alpha}_6}) \]

\[ + O\left(\frac{1}{n^2}\right) \]
The End of Training

\[ z_{i,j}(t = \infty) = z_{i,j} - \sum_{\tilde{a}_1, \tilde{a}_2} H_{j\tilde{a}_1} \tilde{H}_{i1\tilde{a}_2}(z_{i,\tilde{a}_2} - y_{i,\tilde{a}_2}) \]

\[ + \sum_{j, \tilde{a}_1, \tilde{a}_2} \left[ \tilde{\Delta}H_{j;\tilde{a}_1} - \sum_{\tilde{a}_3, \tilde{a}_4 \in A} H_{j\tilde{a}_3} \tilde{H}_{\tilde{a}_4\tilde{a}_1} \tilde{\Delta}H_{j,\tilde{a}_4\tilde{a}_1} \right] \tilde{H}_{i1\tilde{a}_2}(z_{j,\tilde{a}_2} - y_{j,\tilde{a}_2}) \]

\[ - \sum_{j, k, \tilde{a}_1, \ldots, \tilde{a}_4} \left[ \tilde{\Delta}H_{j;\tilde{a}_1} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{j\tilde{a}_5} \tilde{H}_{\tilde{a}_6\tilde{a}_1} \tilde{\Delta}H_{j,\tilde{a}_6\tilde{a}_1} \right] \tilde{H}_{i1\tilde{a}_2} \tilde{\Delta}H_{j,\tilde{a}_2\tilde{a}_3} \tilde{H}_{\tilde{a}_4}(z_{k,\tilde{a}_4} - y_{k,\tilde{a}_4}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d}H_{j;\tilde{a}_1\tilde{a}_2} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{j\tilde{a}_5} \tilde{H}_{\tilde{a}_6\tilde{a}_1} \tilde{d}H_{j,\tilde{a}_6\tilde{a}_1} \right] Z_{A}^{\tilde{a}_1\tilde{a}_2\tilde{a}_3\tilde{a}_4}(z_{j,\tilde{a}_3} - y_{j,\tilde{a}_3})(z_{j,\tilde{a}_4} - y_{j,\tilde{a}_4}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d}H_{j;\tilde{a}_1\tilde{a}_2} - \sum_{\tilde{a}_5, \tilde{a}_6} H_{j\tilde{a}_5} \tilde{H}_{\tilde{a}_6\tilde{a}_1} \tilde{d}H_{j,\tilde{a}_6\tilde{a}_1} \right] Z_{B}^{\tilde{a}_1\tilde{a}_2\tilde{a}_3\tilde{a}_4}(z_{j,\tilde{a}_3} - y_{j,\tilde{a}_3})(z_{j,\tilde{a}_4} - y_{j,\tilde{a}_4}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d}d_{1}H_{j;\tilde{a}_1\tilde{a}_2} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{j\tilde{a}_7} \tilde{H}_{\tilde{a}_8\tilde{a}_1} \tilde{d}d_{1}H_{j,\tilde{a}_8\tilde{a}_1} \right] Z_{A1}^{\tilde{a}_1\tilde{a}_2\tilde{a}_3\tilde{a}_4\tilde{a}_5\tilde{a}_6}(z_{j,\tilde{a}_4} - y_{j,\tilde{a}_4})(z_{j,\tilde{a}_5} - y_{j,\tilde{a}_5})(z_{j,\tilde{a}_6} - y_{j,\tilde{a}_6}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d}d_{1}H_{j;\tilde{a}_1\tilde{a}_2} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{j\tilde{a}_7} \tilde{H}_{\tilde{a}_8\tilde{a}_1} \tilde{d}d_{1}H_{j,\tilde{a}_8\tilde{a}_1} \right] Z_{A2}^{\tilde{a}_1\tilde{a}_2\tilde{a}_3\tilde{a}_4\tilde{a}_5\tilde{a}_6}(z_{j,\tilde{a}_4} - y_{j,\tilde{a}_4})(z_{j,\tilde{a}_5} - y_{j,\tilde{a}_5})(z_{j,\tilde{a}_6} - y_{j,\tilde{a}_6}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d}d_{1}H_{j;\tilde{a}_1\tilde{a}_2} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{j\tilde{a}_7} \tilde{H}_{\tilde{a}_8\tilde{a}_1} \tilde{d}d_{1}H_{j,\tilde{a}_8\tilde{a}_1} \right] Z_{A3}^{\tilde{a}_1\tilde{a}_2\tilde{a}_3\tilde{a}_4\tilde{a}_5\tilde{a}_6}(z_{j,\tilde{a}_4} - y_{j,\tilde{a}_4})(z_{j,\tilde{a}_5} - y_{j,\tilde{a}_5})(z_{j,\tilde{a}_6} - y_{j,\tilde{a}_6}) \]

\[ + \sum_{j, k, \tilde{a}_1, \tilde{a}_2, \tilde{a}_3, \tilde{a}_4} \left[ \tilde{d}d_{1}H_{j;\tilde{a}_1\tilde{a}_2} - \sum_{\tilde{a}_7, \tilde{a}_8} H_{j\tilde{a}_7} \tilde{H}_{\tilde{a}_8\tilde{a}_1} \tilde{d}d_{1}H_{j,\tilde{a}_8\tilde{a}_1} \right] Z_{A4}^{\tiltilde{a}_1\tilde{a}_2\tilde{a}_3\tilde{a}_4\tilde{a}_5\tilde{a}_6}(z_{j,\tilde{a}_4} - y_{j,\tilde{a}_4})(z_{j,\tilde{a}_5} - y_{j,\tilde{a}_5})(z_{j,\tilde{a}_6} - y_{j,\tilde{a}_6}) \]

\[ + O\left(\frac{1}{n^2}\right)\]
The End of Training

\[ z_i;\beta(t = \infty) \]
\[ = z_i;\beta - \sum_{\tilde{a}_1, \tilde{a}_2} H_{\tilde{\beta};\tilde{a}_1} \tilde{H}_{\tilde{a}_1 \tilde{a}_2} (z_i;\tilde{a}_2 - y_i;\tilde{a}_2) \]
\[ + \sum_{\tilde{a}_1, \tilde{a}_2} \left[ \tilde{H}_{\tilde{a}_1 \tilde{a}_2} \tilde{H}_{\tilde{a}_1 \tilde{a}_2} \left( \sum_{\tilde{a}_3, \tilde{a}_4 \in A} H_{\tilde{\beta};\tilde{a}_3} \tilde{H}_{\tilde{a}_3 \tilde{a}_4} \tilde{H}_{\tilde{a}_4 \tilde{a}_1} \right) \tilde{H}_{\tilde{a}_1 \tilde{a}_2} (z_j;\tilde{a}_2 - y_j;\tilde{a}_2) \right] \]
\[ - \sum_{\tilde{a}_1, \tilde{a}_2} \left[ \tilde{H}_{\tilde{a}_1 \tilde{a}_2} \tilde{H}_{\tilde{a}_1 \tilde{a}_2} \left( \sum_{\tilde{a}_3, \tilde{a}_4 \in A} H_{\tilde{\beta};\tilde{a}_3} \tilde{H}_{\tilde{a}_3 \tilde{a}_4} \tilde{H}_{\tilde{a}_4 \tilde{a}_1} \right) \tilde{H}_{\tilde{a}_1 \tilde{a}_2} (z_k;\tilde{a}_4 - y_k;\tilde{a}_4) \right] \]
\[ + \sum_{\tilde{a}_1, \tilde{a}_2} \left[ \tilde{H}_{\tilde{a}_1 \tilde{a}_2} \tilde{H}_{\tilde{a}_1 \tilde{a}_2} \left( \sum_{\tilde{a}_3, \tilde{a}_4 \in A} H_{\tilde{\beta};\tilde{a}_3} \tilde{H}_{\tilde{a}_3 \tilde{a}_4} \tilde{H}_{\tilde{a}_4 \tilde{a}_1} \right) \tilde{H}_{\tilde{a}_1 \tilde{a}_2} (z_j;\tilde{a}_4 - y_j;\tilde{a}_4) \right] \]
\[ + O\left( \frac{1}{n^2} \right) \]
The End of Training

\[ z_i;\hat{\beta}(t = \infty) = z_i;\beta - \sum_{\hat{\alpha}_1, \hat{\alpha}_2} H_{\beta;\hat{\alpha}_1} \bar{H}_{\hat{\alpha}_1 \hat{\alpha}_2}(z_i;\hat{\alpha}_2 - y_i;\hat{\alpha}_2) \]

\[ + \sum_{j,\hat{\alpha}_1, \hat{\alpha}_2} \left[ \Delta H_{j_1;\beta;\hat{\alpha}_1} - \sum_{\hat{\alpha}_3, \hat{\alpha}_4 \in A} H_{\beta;\hat{\alpha}_3} \bar{H}_{\hat{\alpha}_3 \hat{\alpha}_4} \Delta H_{j_1;\hat{\alpha}_4;\hat{\alpha}_1} \right] \bar{H}_{\hat{\alpha}_1 \hat{\alpha}_2}(z_j;\hat{\alpha}_2 - y_j;\hat{\alpha}_2) \]

\[ - \sum_{j,k,\hat{\alpha}_1, \ldots, \hat{\alpha}_4} \left[ \Delta H_{j;\beta;\hat{\alpha}_1} - \sum_{\hat{\alpha}_5, \hat{\alpha}_6} H_{\beta;\hat{\alpha}_5} \bar{H}_{\hat{\alpha}_5 \hat{\alpha}_6} \Delta H_{j;\hat{\alpha}_6;\hat{\alpha}_1} \right] \bar{H}_{\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3 \hat{\alpha}_4}(z_k;\hat{\alpha}_4 - y_k;\hat{\alpha}_4) \]

\[ + \sum_{j_1, j_2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4} \left[ \tilde{d}H_{j_1 j_2;\beta;\hat{\alpha}_1 \hat{\alpha}_2;\hat{\alpha}_3} - \sum_{\hat{\alpha}_7, \hat{\alpha}_8} H_{\beta;\hat{\alpha}_7} \bar{H}_{\hat{\alpha}_7 \hat{\alpha}_8} \tilde{d}dH_{j_1 j_2 j_3;\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3} \right] Z_{\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3 \hat{\alpha}_4 \hat{\alpha}_5 \hat{\alpha}_6}(z_{j_1};\hat{\alpha}_4 - y_{j_1};\hat{\alpha}_4)(z_{j_2};\hat{\alpha}_5 - y_{j_2};\hat{\alpha}_5)(z_{j_3};\hat{\alpha}_6 - y_{j_3};\hat{\alpha}_6) \]

\[ + \sum_{j_1, j_2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\alpha}_5, \hat{\alpha}_6} \left[ \tilde{d}dH_{j_1 j_2 j_3;\beta;\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3} - \sum_{\hat{\alpha}_7, \hat{\alpha}_8} H_{\beta;\hat{\alpha}_7} \bar{H}_{\hat{\alpha}_7 \hat{\alpha}_8} \tilde{d}dH_{j_1 j_2 j_3;\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3} \right] Z_{\hat{\alpha}_1 \hat{\alpha}_2 \hat{\alpha}_3 \hat{\alpha}_4 \hat{\alpha}_5 \hat{\alpha}_6}(z_{j_1};\hat{\alpha}_4 - y_{j_1};\hat{\alpha}_4)(z_{j_2};\hat{\alpha}_5 - y_{j_2};\hat{\alpha}_5)(z_{j_3};\hat{\alpha}_6 - y_{j_3};\hat{\alpha}_6) \]

\[ + O\left(\frac{1}{n^2}\right) \]
The End of Training

$$z_{i;\beta}(t = \infty) = z_{i;\beta} - \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2}(z_{i;\tilde{\alpha}_2} - y_{i;\tilde{\alpha}_2})$$

$$+ \sum_{j, \tilde{\alpha}_1, \tilde{\alpha}_2} \left[ \Delta \tilde{H}_{j;\beta \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_3, \tilde{\alpha}_4 \in \mathcal{A}} H_{\beta \tilde{\alpha}_3} \tilde{H}^{3\tilde{\alpha}_4} \Delta \tilde{H}_{j;\tilde{\alpha}_4 \tilde{\alpha}_1} \right] \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2}(z_{j;\tilde{\alpha}_2} - y_{j;\tilde{\alpha}_2})$$

$$- \sum_{j, k, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_4} \left[ \Delta \tilde{H}_{j;\beta \tilde{\alpha}_1} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H}^{5\tilde{\alpha}_6} \Delta \tilde{H}_{j;\tilde{\alpha}_6 \tilde{\alpha}_1} \right] \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2} \Delta \tilde{H}_{j;\tilde{\alpha}_6 \tilde{\alpha}_1} \tilde{H}^{3\tilde{\alpha}_4}(z_{k;\tilde{\alpha}_4} - y_{k;\tilde{\alpha}_4})$$

$$+ \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ d\tilde{H}_{j;\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H}^{6\tilde{\alpha}_6} d\tilde{H}_{j;\tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_1 \tilde{\alpha}_2} \right] Z^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4}(z_{j;\tilde{\alpha}_3} - y_{j;\tilde{\alpha}_3})(z_{j;\tilde{\alpha}_4} - y_{j;\tilde{\alpha}_4})$$

$$+ \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ d\tilde{H}_{j;\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H}^{6\tilde{\alpha}_6} d\tilde{H}_{j;\tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_1 \tilde{\alpha}_2} \right] Z^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4}(z_{j;\tilde{\alpha}_3} - y_{j;\tilde{\alpha}_3})(z_{j;\tilde{\alpha}_4} - y_{j;\tilde{\alpha}_4})$$

$$+ \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ d\tilde{H}_{j;\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H}^{6\tilde{\alpha}_6} d\tilde{H}_{j;\tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_1 \tilde{\alpha}_2} \right] Z^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4}(z_{j;\tilde{\alpha}_3} - y_{j;\tilde{\alpha}_3})(z_{j;\tilde{\alpha}_4} - y_{j;\tilde{\alpha}_4})$$

$$+ \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ d\tilde{H}_{j;\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H}^{6\tilde{\alpha}_6} d\tilde{H}_{j;\tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_1 \tilde{\alpha}_2} \right] Z^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4}(z_{j;\tilde{\alpha}_3} - y_{j;\tilde{\alpha}_3})(z_{j;\tilde{\alpha}_4} - y_{j;\tilde{\alpha}_4})$$

$$+ \sum_{j, k, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4} \left[ d\tilde{H}_{j;\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} - \sum_{\tilde{\alpha}_5, \tilde{\alpha}_6} H_{\beta \tilde{\alpha}_5} \tilde{H}^{6\tilde{\alpha}_6} d\tilde{H}_{j;\tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_1 \tilde{\alpha}_2} \right] Z^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4}(z_{j;\tilde{\alpha}_3} - y_{j;\tilde{\alpha}_3})(z_{j;\tilde{\alpha}_4} - y_{j;\tilde{\alpha}_4})$$

$$+ O\left(\frac{1}{n^2}\right)$$
The Algorithm Projectors

**Gradient Descent:**

\[
\begin{align*}
\hat{Z}_A & = Y_2 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 , \\
\hat{Z}_B & = Y_2 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 + \eta_4 \nu_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4} , \\
\hat{Z}_{\text{IA}} & = - Y_3 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 , \\
\hat{Z}_{\text{IB}} & = - Y_3 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 - \frac{\eta_4}{2} X_{\text{III}} , \\
\hat{Z}_{\text{IA}} & = - Y_3 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 , \\
\hat{Z}_{\text{IB}} & = - Y_3 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 - Y_3 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6. 
\end{align*}
\]

Various Shorthands:

\[
\begin{align*}
Y_1 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 & = X_{\text{II}} \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6 - \sum_{\hat{a}_7} H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} , \\
Y_2 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 & = H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} - \sum_{\hat{a}_7} H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} , \\
Y_3 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 & = H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} - \sum_{\hat{a}_7} H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} + \sum_{\hat{a}_7} H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} , \\
Y_4 \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \alpha_5 \alpha_6 & = H_{\hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6} - \sum_{\hat{a}_7, \hat{a}_8} X_{\text{III}} \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_6. 
\end{align*}
\]

Inverting Tensors:

\[
\begin{align*}
\delta_{\hat{a}_5 \hat{a}_6} & = \sum_{\hat{a}_3, \hat{a}_4 \in A} X_{\text{II}} \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \left( H_{\hat{a}_5 \hat{a}_6} \delta_{\hat{a}_4 \hat{a}_6} + \delta_{\hat{a}_5 \hat{a}_6} \hat{H}_{\hat{a}_4 \hat{a}_6} - \eta \hat{H}_{\hat{a}_5 \hat{a}_6} \hat{H}_{\hat{a}_4 \hat{a}_6} \right) , \\
\delta_{\hat{a}_7 \hat{a}_8} & = \sum_{\hat{a}_4, \hat{a}_6, \hat{a}_9} X_{\text{III}} \hat{a}_1 \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5 \hat{a}_9 \left( H_{\hat{a}_4 \hat{a}_6} \delta_{\hat{a}_4 \hat{a}_6} + \delta_{\hat{a}_4 \hat{a}_6} \hat{H}_{\hat{a}_5 \hat{a}_6} \delta_{\hat{a}_4 \hat{a}_6} + \delta_{\hat{a}_4 \hat{a}_6} \hat{H}_{\hat{a}_5 \hat{a}_6} \right) \\
& - \eta \left( \hat{H}_{\hat{a}_5 \hat{a}_6} \hat{H}_{\hat{a}_5 \hat{a}_6} \delta_{\hat{a}_4 \hat{a}_6} + \hat{H}_{\hat{a}_4 \hat{a}_6} \delta_{\hat{a}_4 \hat{a}_6} \hat{H}_{\hat{a}_5 \hat{a}_6} + \delta_{\hat{a}_4 \hat{a}_6} \hat{H}_{\hat{a}_5 \hat{a}_6} \right) + \eta^2 \left( \hat{H}_{\hat{a}_4 \hat{a}_6} \hat{H}_{\hat{a}_5 \hat{a}_6} \hat{H}_{\hat{a}_6 \hat{a}_9} \right) .
\end{align*}
\]
The Algorithm Projectors

**Gradient Flow** (ODE limit):

\[
Z^A_1 \equiv Y_2 \langle \alpha_3 \rangle_{\alpha_4}, \\
Z^B_1 \equiv Y_2 \langle \overline{\alpha}_3 \rangle_{\overline{\alpha}_4}, \\
Z^A_{1A} \equiv -Y_3 \langle \overline{\alpha}_3 \rangle_{\alpha_5} \langle \overline{\alpha}_5 \rangle_{\alpha_4}, \\
Z^B_{1B} \equiv -Y_3 \langle \alpha_3 \rangle_{\overline{\alpha}_5} \langle \overline{\alpha}_5 \rangle_{\alpha_4}, \\
Z^A_{1B} \equiv -Y_3 \langle \bar{\alpha}_3 \rangle_{\alpha_5} \langle \overline{\alpha}_5 \rangle_{\alpha_4}, \\
Z^B_{1A} \equiv -Y_3 \langle \alpha_3 \rangle_{\overline{\alpha}_5} \langle \overline{\alpha}_5 \rangle_{\alpha_4}.
\]

Various Shorthands:

\[
Y_2 \langle \bar{\alpha}_3 \rangle_4 \equiv H^\bar{\alpha}_3 \overline{H}^\alpha_4 - \sum_{\bar{\alpha}_5} H^\alpha_5 X_{\bar{\alpha} \overline{\alpha}} \langle \alpha_3 \rangle_{\alpha_4}, \\
Y_3 \langle \bar{\alpha}_3 \rangle_6 \equiv \overline{H}^{\bar{\alpha}_3} \overline{H}^{\alpha_5} - \sum_{\bar{\alpha}_7} \overline{H}^{\bar{\alpha}_7} X_{\bar{\alpha} \overline{\alpha}} \langle \alpha_3 \rangle_{\alpha_5} + \sum_{\bar{\alpha}_7, \bar{\alpha}_8, \bar{\alpha}_9} \overline{H}^{\bar{\alpha}_9} X_{\bar{\alpha} \overline{\alpha}} \langle \alpha_3 \rangle_{\alpha_7} \langle \alpha_7 \rangle_{\alpha_6},
\]

Inverting Tensors:

\[
\delta^\bar{\alpha}_{\bar{\alpha}_5} \delta_{\overline{\alpha}_6} = \sum_{\bar{\alpha}_3, \overline{\alpha}_4 \in A} X_{\bar{\alpha} \overline{\alpha}} \langle \bar{\alpha}_3 \rangle_{\alpha_4} \langle \alpha_4 \rangle_{\overline{\alpha}_6} \left( H_{\bar{\alpha}_3 \alpha_5} \delta_{\alpha_5 \overline{\alpha}_6} + \delta_{\alpha_3 \overline{\alpha}_5} H_{\alpha_5 \overline{\alpha}_6} \right), \\
\delta^\bar{\alpha}_7 \delta_{\overline{\alpha}_8} \delta_{\overline{\alpha}_9} = \sum_{\bar{\alpha}_4, \overline{\alpha}_6, \overline{\alpha}_9} X_{\bar{\alpha} \overline{\alpha}} \langle \bar{\alpha}_7 \rangle_{\alpha_4} \langle \alpha_4 \rangle_{\overline{\alpha}_6} \langle \alpha_7 \rangle_{\overline{\alpha}_9} \left( H_{\bar{\alpha}_7 \alpha_5} \delta_{\alpha_5 \overline{\alpha}_6} \delta_{\alpha_6 \overline{\alpha}_9} + \delta_{\alpha_7 \overline{\alpha}_5} H_{\alpha_5 \overline{\alpha}_6} \delta_{\alpha_6 \overline{\alpha}_9} + \delta_{\alpha_7 \alpha_5} H_{\bar{\alpha}_5 \alpha_9} \right).
\]
The Algorithm Projectors

Direct Optimization:

\[ Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4}_A \equiv 0, \]
\[ Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4}_B \equiv \frac{1}{2} H^{\tilde{\alpha}_1\tilde{\alpha}_3} H^{\tilde{\alpha}_2\tilde{\alpha}_4}, \]
\[ Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6}_A \equiv 0, \]
\[ Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6}_B \equiv -\frac{1}{6} H^{\tilde{\alpha}_1\tilde{\alpha}_4} H^{\tilde{\alpha}_2\tilde{\alpha}_5} H^{\tilde{\alpha}_3\tilde{\alpha}_6}, \]
\[ Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6}_{\text{IA}} \equiv 0, \]
\[ Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6}_{\text{IB}} \equiv 0. \]

Various Shorthands:

Inverting Tensors:
Mean Prediction

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

\[
E \left[ z^{(L)}_{i,\beta}(\infty) \right] \equiv m_{i,\beta} \\
= m_{i,\beta}^{\text{NTK}} + \frac{1}{n_{L-1}} \left( m_{i,\beta}^{\Delta \text{NTK}} + m_{i,\beta}^{d \text{NTK}} + m_{i,\beta}^{d^2 \text{NTK-I}} + m_{i,\beta}^{d^2 \text{NTK-II}} \right) \\
- \frac{1}{n_{L-1}} \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1}^{(L)}(\tilde{\alpha}_2) \tilde{H}_{\beta \tilde{\alpha}}^{(L)}(\tilde{\alpha}_2) \left( m_{i,\tilde{\alpha}_2}^{\Delta \text{NTK}} + m_{i,\tilde{\alpha}_2}^{d \text{NTK}} + m_{i,\tilde{\alpha}_2}^{d^2 \text{NTK-I}} + m_{i,\tilde{\alpha}_2}^{d^2 \text{NTK-II}} \right)
\]
Mean Prediction

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

\[
\mathbb{E} \left[ z_{i;\beta}^{(L)}(\infty) \right] \equiv m_{i;\beta} = m_{i;\beta}^{\text{NTK}} + \frac{1}{n_{L-1}} \left( m_{i;\beta}^{\Delta \text{NTK}} + m_{i;\beta}^{d \text{NTK}} + m_{i;\beta}^{dd \text{NTK-I}} + m_{i;\beta}^{dd \text{NTK-II}} \right) \\
- \frac{1}{n_{L-1}} \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1}^{(L)}(L) \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2} \left( m_{i;\tilde{\alpha}_2}^{\Delta \text{NTK}} + m_{i;\tilde{\alpha}_2}^{d \text{NTK}} + m_{i;\tilde{\alpha}_2}^{dd \text{NTK-I}} + m_{i;\tilde{\alpha}_2}^{dd \text{NTK-II}} \right)
\]

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

\[
m_{i;\beta}^{\text{NTK}} \equiv \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1}^{(L)}(L) \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2} y_{i;\tilde{\alpha}_2}.
\]
Mean Prediction

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

$$
\mathbb{E} \left[ z^{(L)}_{i;\beta}(\infty) \right] \equiv m_{i;\beta} \\
= m_{i;\beta}^{\text{NTK}} + \frac{1}{n_{L-1}} \left( m_{i;\beta}^{\Delta \text{NTK}} + m_{i;\beta}^{d\text{NTK}} + m_{i;\beta}^{dd\text{NTK-I}} + m_{i;\beta}^{dd\text{NTK-II}} \right) \\
- \frac{1}{n_{L-1}} \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1}^{(L)} \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2}^{(L)} \left( m_{i;\tilde{\alpha}_2}^{\Delta \text{NTK}} + m_{i;\tilde{\alpha}_2}^{d\text{NTK}} + m_{i;\tilde{\alpha}_2}^{dd\text{NTK-I}} + m_{i;\tilde{\alpha}_2}^{dd\text{NTK-II}} \right)
$$

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

  $$
m_{i;\beta}^{\text{NTK}} \equiv \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H_{\beta \tilde{\alpha}_1}^{(L)} \tilde{H}_{\tilde{\alpha}_1 \tilde{\alpha}_2}^{(L)} y_{i;\tilde{\alpha}_2}.
$$

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

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

\[
\mathbb{E}\left[ z^{(L)}(\infty) \right] \equiv m_{i;\hat{\beta}}
\]

\[
= m_{i;\hat{\beta}}^{\text{NTK}} + \frac{1}{n_{L-1}} \left( m_{i;\hat{\beta}}^{\Delta\text{NTK}} + m_{i;\hat{\beta}}^{\text{dNTK}} + m_{i;\hat{\beta}}^{\text{ddNTK-I}} + m_{i;\hat{\beta}}^{\text{ddNTK-II}} \right)
\]

\[
- \frac{1}{n_{L-1}} \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2} H^{(L)}_{\beta\tilde{\alpha}_1} \tilde{H}_{\tilde{\alpha}_1\tilde{\alpha}_2}(L) \left( m_{i;\tilde{\alpha}_2}^{\Delta\text{NTK}} + m_{i;\tilde{\alpha}_2}^{\text{dNTK}} + m_{i;\tilde{\alpha}_2}^{\text{ddNTK-I}} + m_{i;\tilde{\alpha}_2}^{\text{ddNTK-II}} \right)
\]

- 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

$$\mathbb{E} \left[ z^{(L)}_{i;\beta}(\infty) \right] \equiv m_{i;\beta}$$

$$= m^{\text{NTK}}_{i;\beta} + \frac{1}{n_{L-1}} \left( m^{\Delta\text{NTK}}_{i;\beta} + m^{d\text{NTK}}_{i;\beta} + m^{dd\text{NTK-I}}_{i;\beta} + m^{dd\text{NTK-II}}_{i;\beta} \right)$$

$$- \frac{1}{n_{L-1}} \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2} H^{(L)}_{\beta\tilde{\alpha}_1} \tilde{H}^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2} \left( m^{\Delta\text{NTK}}_{i;\tilde{\alpha}_2} + m^{d\text{NTK}}_{i;\tilde{\alpha}_2} + m^{dd\text{NTK-I}}_{i;\tilde{\alpha}_2} + m^{dd\text{NTK-II}}_{i;\tilde{\alpha}_2} \right)$$

- 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

\[
m_{i;\delta}^{\Delta NTK} \equiv \sum_{\bar{\alpha}_1,\ldots,\bar{\alpha}_4} \left( A^{(L)}_{(\delta\bar{\alpha}_1)(\bar{\alpha}_2\bar{\alpha}_3)} + B^{(L)}_{\delta\bar{\alpha}_2\bar{\alpha}_1\bar{\alpha}_3} + n_L B^{(L)}_{\delta\bar{\alpha}_3\bar{\alpha}_1\bar{\alpha}_2} \right) \times \bar{H}_{(L)}(\bar{\alpha}_1\bar{\alpha}_2) \bar{H}_{(L)}(\bar{\alpha}_3\bar{\alpha}_4) y_{i;\bar{\alpha}_4},
\]

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

\[
\mathbb{E} \left[ \Delta H_{i_1i_2;\alpha_1\alpha_2}^{(L)} \Delta H_{i_3i_4;\alpha_3\alpha_4}^{(L)} \right] \\
\equiv \frac{1}{n_{L-1}} \left[ \delta_{i_1i_2} \delta_{i_3i_4} A^{(L)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1i_3} \delta_{i_2i_4} B^{(L)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1i_4} \delta_{i_2i_3} B^{(L)}_{\alpha_1\alpha_4\alpha_2\alpha_3} \right].
\]
Mean Prediction: \textit{NTK Variance}

The fluctuation of the NTK term gives

$$m_{i;\delta}^{\Delta\text{NTK}} \equiv \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_4} \left( A^{(L)}_{(\delta\tilde{\alpha}_1)(\tilde{\alpha}_2\tilde{\alpha}_3)} + B^{(L)}_{\delta\tilde{\alpha}_2\tilde{\alpha}_1\tilde{\alpha}_3} + n_L B^{(L)}_{\delta\tilde{\alpha}_3\tilde{\alpha}_1\tilde{\alpha}_2} \right)$$

$$\times \tilde{H}_{(L)} \tilde{H}_{(L)} y_{i;\tilde{\alpha}_4},$$

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

$$\mathbb{E} \left[ \Delta H_{i_1i_2;\alpha_1\alpha_2} \Delta H_{i_3i_4;\alpha_3\alpha_4} \right]$$

$$\equiv \frac{1}{n_{L-1}} \left[ \delta_{i_1i_2} \delta_{i_3i_4} A^{(L)}_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)} + \delta_{i_1i_3} \delta_{i_2i_4} B^{(L)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta_{i_1i_4} \delta_{i_2i_3} B^{(L)}_{\alpha_1\alpha_4\alpha_2\alpha_3} \right].$$
Mean Prediction: *dNTK-Preactivation Cross Correlation*

The dNTK term gives

\[
m^{d\text{NTK}}_{i;\delta} = \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_4} \left[ 2 \left( P^{(L)}_{\delta\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\delta\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} + n_L Q^{(L)}_{\delta\tilde{\alpha}_2\tilde{\alpha}_1\tilde{\alpha}_3} \right) Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4}_B \\
+ \left( n_L P^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\delta\tilde{\alpha}_3} \right) Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4}_A \\
+ \left( P^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} + n_L Q^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} + Q^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\delta\tilde{\alpha}_3} \right) Z^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_4\tilde{\alpha}_3}_A \right] y_{i;\tilde{\alpha}_4},
\]

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

\[
\mathbb{E} \left[ \hat{d}H^{(L)}_{i_0 i_1 i_2;\delta_0\delta_1\delta_2} z^{(L)}_{i_3;\delta_3} \right] \\
\equiv \frac{1}{n_{L-1}} \left[ \delta_{i_0 i_3} \delta_{i_1 i_2} P^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0 i_1} \delta_{i_2 i_3} Q^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0 i_2} \delta_{i_1 i_3} Q^{(L)}_{\delta_0\delta_2\delta_1\delta_3} \right].
\]
Mean Prediction: \textit{dNTK-Preactivation Cross Correlation}

The dNTK term gives

\[
m_{i;\delta}^{d\text{NTK}} \equiv - \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_4} \left[ 2 \left( P_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} + Q_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} + n_L Q_{\delta \tilde{\alpha}_2 \tilde{\alpha}_1 \tilde{\alpha}_3}^{(L)} \right) Z_{B}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} \\
+ \left( n_L P_{\tilde{\alpha}_1 \delta \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} + Q_{\tilde{\alpha}_1 \delta \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} + Q_{\tilde{\alpha}_1 \tilde{\alpha}_2 \delta \tilde{\alpha}_3}^{(L)} \right) Z_{A}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} \\
+ \left( P_{\tilde{\alpha}_1 \delta \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} + n_L Q_{\tilde{\alpha}_1 \delta \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} + Q_{\tilde{\alpha}_1 \tilde{\alpha}_2 \delta \tilde{\alpha}_3}^{(L)} \right) Z_{A}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_4 \tilde{\alpha}_3} \right] y_{i;\tilde{\alpha}_4},
\]

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

\[
\mathbb{E} \left[ \hat{d}H_{i_0 i_1 i_2;\delta_0 \delta_1 \delta_2}^{(L)} z_{i_3;\delta_3}^{(L)} \right] \\
\equiv \frac{1}{n_{L-1}} \left[ \delta_{i_0 i_3} \delta_{i_1 i_2} P_{\delta_0 \delta_1 \delta_2 \delta_3}^{(L)} + \delta_{i_0 i_1} \delta_{i_2 i_3} Q_{\delta_0 \delta_1 \delta_2 \delta_3}^{(L)} + \delta_{i_0 i_2} \delta_{i_1 i_3} Q_{\delta_0 \delta_2 \delta_1 \delta_3}^{(L)} \right].
\]
Mean Prediction: $ddNTK_i$ Mean

The first $ddNTK$ term gives

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

$$- \sum_{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_6} \left[ R^{(L)}_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3} \left( Z_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6}^{IB} + Z_{\tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_4}^{IB} + Z_{\tilde{\alpha}_3 \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_6 \tilde{\alpha}_4 \tilde{\alpha}_5}^{IB} \right) \right] +$$

$$R^{(L)}_{\tilde{\alpha}_1 \delta \tilde{\alpha}_2 \tilde{\alpha}_3} Z_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6}^{IA} + R^{(L)}_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \delta} Z_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_4}^{IA} + R^{(L)}_{\tilde{\alpha}_1 \tilde{\alpha}_3 \delta \tilde{\alpha}_2} Z_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_6 \tilde{\alpha}_4 \tilde{\alpha}_5}^{IA}$$

$$\times \left[ y_i;\tilde{\alpha}_4 \left( \sum_j y_j;\tilde{\alpha}_5 y_j;\tilde{\alpha}_6 + n_L K_{\tilde{\alpha}_5 \tilde{\alpha}_6}^{(L)} \right) + y_i;\tilde{\alpha}_5 K_{\tilde{\alpha}_6 \tilde{\alpha}_4}^{(L)} + y_i;\tilde{\alpha}_6 K_{\tilde{\alpha}_4 \tilde{\alpha}_5}^{(L)} \right]$$

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

$$\mathbb{E} \left[ \widehat{ddI H}_{i_0 i_1 i_2 i_3; \delta_0 \delta_1 \delta_2 \delta_3}^{(L)} \right]$$

$$\equiv \frac{1}{n_{L-1}} \left[ \delta_{i_0 i_1} \delta_{i_2 i_3} R_{\delta_0 \delta_1 \delta_2 \delta_3}^{(L)} + \delta_{i_0 i_2} \delta_{i_3 i_1} R_{\delta_0 \delta_2 \delta_3 \delta_1}^{(L)} + \delta_{i_0 i_3} \delta_{i_1 i_2} R_{\delta_0 \delta_3 \delta_1 \delta_2}^{(L)} \right] .$$
Mean Prediction: \(ddNTK_I\) Mean

The first \(ddNTK\) term gives

\[
m_{i;\delta}^{ddNTK-I} \equiv \sum_{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_6} \left[ R^{(L)}_{\delta\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3} \left( Z_{IB}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} + Z_{IB}^{\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_1\tilde{\alpha}_5\tilde{\alpha}_6\tilde{\alpha}_4} + Z_{IB}^{\tilde{\alpha}_3\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_6\tilde{\alpha}_4\tilde{\alpha}_5} \right) + 
R^{(L)}_{\tilde{\alpha}_1\delta\tilde{\alpha}_2\tilde{\alpha}_3} Z_{IA}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4\tilde{\alpha}_5\tilde{\alpha}_6} + R^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\delta} Z_{IA}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_5\tilde{\alpha}_6\tilde{\alpha}_4} + R^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_3\delta\tilde{\alpha}_2} Z_{IA}^{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_6\tilde{\alpha}_4\tilde{\alpha}_5} \right] \times \left[ y_{i;\tilde{\alpha}_4} \left( \sum_{j} y_{j;\tilde{\alpha}_5} y_{j;\tilde{\alpha}_6} + n_L K^{(L)}_{\tilde{\alpha}_5\tilde{\alpha}_6} \right) + y_{i;\tilde{\alpha}_5} K^{(L)}_{\tilde{\alpha}_6\tilde{\alpha}_4} + y_{i;\tilde{\alpha}_6} K^{(L)}_{\tilde{\alpha}_4\tilde{\alpha}_5} \right]
\]

where we decomposed the \(dNTK_I\) mean into the tensor \(R^{(L)}\):

\[
\mathbb{E} \left[ dd_1 H^{(L)}_{i_0i_1i_2i_3;\delta_0\delta_1\delta_2\delta_3} \right] \equiv \frac{1}{n_{L-1}} \left[ \delta_{i_0i_1} \delta_{i_2i_3} R^{(L)}_{\delta_0\delta_1\delta_2\delta_3} + \delta_{i_0i_2} \delta_{i_3i_1} R^{(L)}_{\delta_0\delta_2\delta_3\delta_1} + \delta_{i_0i_3} \delta_{i_1i_2} R^{(L)}_{\delta_0\delta_3\delta_1\delta_2} \right].
\]
Mean Prediction: $ddNTK_{||}$ Mean

The second $ddNTK$ term gives

$$m_{i;\alpha}^{ddNTK-||}$$

\[= - \sum_{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_6} \left[ S_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} Z_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6} \right. \]

\[\quad + U_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3}^{(L)} Z_{\tilde{\alpha}_3 \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_6 \tilde{\alpha}_4 \tilde{\alpha}_5} + S_{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \alpha_5 \tilde{\alpha}_6 \tilde{\alpha}_4}^{(L)} \]

\[\quad + U_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6}^{(L)} + U_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5}^{(L)} \]

\[\left. \times \left[ y_{i;\tilde{\alpha}_4} \left( \sum_j y_{j;\tilde{\alpha}_5} y_{j;\tilde{\alpha}_6} + n_L K^{(L)}_{\tilde{\alpha}_5 \tilde{\alpha}_6} \right) + y_{i;\tilde{\alpha}_5} K^{(L)}_{\tilde{\alpha}_6 \tilde{\alpha}_4} + y_{i;\tilde{\alpha}_6} K^{(L)}_{\tilde{\alpha}_4 \tilde{\alpha}_5} \right] \right] \]

where we decomposed the $dNTK_{||}$ mean into $S^{(L)}$, $T^{(L)}$, and $U^{(L)}$:

$$\mathbb{E}\left[ \dd_{||} H_{i_1 i_2 i_3 i_4;\delta_1 \delta_2 \delta_3 \delta_4}^{(L)} \right]$$

\[= \frac{1}{n_L - 1} \left[ \delta_{i_1 i_2} \delta_{i_3 i_4} S_{\delta_1 \delta_2 \delta_3 \delta_4}^{(L)} + \delta_{i_1 i_3} \delta_{i_4 i_2} T_{\delta_1 \delta_3 \delta_4 \delta_2}^{(L)} + \delta_{i_1 i_4} \delta_{i_2 i_3} U_{\delta_1 \delta_4 \delta_2 \delta_3}^{(L)} \right]. \]
Mean Prediction: $ddNTK_{II}$ Mean

The second $ddNTK$ term gives

$$m_{i;\delta}^{ddNTK-II} \equiv - \sum_{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_6} \left[ S^{(L)}_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3} Z_{\text{IIB}}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6} + T^{(L)}_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3} Z_{\text{IIB}}^{\tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_1 \tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_4} \\
+ U^{(L)}_{\delta \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3} Z_{\text{IIB}}^{\tilde{\alpha}_3 \tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_6 \tilde{\alpha}_4 \tilde{\alpha}_5} + S^{(L)}_{\tilde{\alpha}_1 \tilde{\alpha}_2 \delta \tilde{\alpha}_3} Z_{\text{IIA}}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_5 \tilde{\alpha}_6 \tilde{\alpha}_4} \\
+ T^{(L)}_{\tilde{\alpha}_1 \delta \tilde{\alpha}_3 \tilde{\alpha}_2} Z_{\text{IIA}}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4 \tilde{\alpha}_5 \tilde{\alpha}_6} + U^{(L)}_{\tilde{\alpha}_1 \tilde{\alpha}_3 \tilde{\alpha}_2 \delta} Z_{\text{IIA}}^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_6 \tilde{\alpha}_4 \tilde{\alpha}_5} \right] \times \left[ y_{i;\tilde{\alpha}_4} \left( \sum_j y_{j;\tilde{\alpha}_5} y_{j;\tilde{\alpha}_6} + n_L K^{(L)}_{\tilde{\alpha}_5 \tilde{\alpha}_6} \right) + y_{i;\tilde{\alpha}_5} K^{(L)}_{\tilde{\alpha}_6 \tilde{\alpha}_4} + y_{i;\tilde{\alpha}_6} K^{(L)}_{\tilde{\alpha}_4 \tilde{\alpha}_5} \right]$$

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

$$\mathbb{E} \left[ \hat{dd}_{II} H^{(L)}_{i_1 i_2 i_3 i_4; \delta_1 \delta_2 \delta_3 \delta_4} \right] \equiv \frac{1}{n_L - 1} \left[ \delta_{i_1 i_2} \delta_{i_3 i_4} S^{(L)}_{\delta_1 \delta_2 \delta_3 \delta_4} + \delta_{i_1 i_3} \delta_{i_4 i_2} T^{(L)}_{\delta_1 \delta_3 \delta_4 \delta_2} + \delta_{i_1 i_4} \delta_{i_2 i_3} U^{(L)}_{\delta_1 \delta_4 \delta_2 \delta_3} \right].$$
In addition to the ensemble mean, we can consider other statistics:

\[
\text{Cov}\left[ z_{i_1; \hat{\beta}_1}(\infty), z_{i_2; \hat{\beta}_2}(\infty) \right] \\
\equiv \mathbb{E} \left[ z_{i_1; \hat{\beta}_1}(\infty) z_{i_2; \hat{\beta}_2}(\infty) \right] - \mathbb{E} \left[ z_{i_1; \hat{\beta}_1}(\infty) \right] \mathbb{E} \left[ z_{i_2; \hat{\beta}_2}(\infty) \right].
\]
In addition to the ensemble mean, we can consider other statistics:

\[
\text{Cov} \left[ z^{(L)}_{i_1;\beta_1} (\infty), z^{(L)}_{i_2;\beta_2} (\infty) \right] \\
\equiv \mathbb{E} \left[ z^{(L)}_{i_1;\beta_1} (\infty) z^{(L)}_{i_2;\beta_2} (\infty) \right] - \mathbb{E} \left[ z^{(L)}_{i_1;\beta_1} (\infty) \right] \mathbb{E} \left[ z^{(L)}_{i_2;\beta_2} (\infty) \right] .
\]

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

\[
\sum_{\tilde{\alpha}_1, \ldots, \tilde{\alpha}_4} \mathbb{E} \left[ z_{i_2; \beta_2}^{(L)} \hat{d} H_{i_1 j_1 j_2; \beta_1 \tilde{\alpha}_1 \tilde{\alpha}_2}^{(L)} \right] Z_B^{\tilde{\alpha}_1 \tilde{\alpha}_2 \tilde{\alpha}_3 \tilde{\alpha}_4} \times \mathbb{E} \left[ \left( z_{j_1; \tilde{\alpha}_3}^{(L)} - y_{j_1; \tilde{\alpha}_3} \right) \left( z_{j_2; \tilde{\alpha}_4}^{(L)} - y_{j_2; \tilde{\alpha}_4} \right) \right]
\]
Prediction Variance

To see another manifestation of output “wiring,” consider

\[ \sum_{j_1, j_2} \mathbb{E} \left[ z_{i_2; \hat{\beta}_2}^{(L)} \, \hat{d}H_{i_1j_1j_2; \hat{\beta}_1\hat{\alpha}_1\hat{\alpha}_2}^{(L)} \right] Z_{\hat{\alpha}_1\hat{\alpha}_2\hat{\alpha}_3\hat{\alpha}_4} \times \mathbb{E} \left[ \left( z_{j_1; \hat{\alpha}_3}^{(L)} - y_{j_1; \hat{\alpha}_3} \right) \left( z_{j_2; \hat{\alpha}_4}^{(L)} - y_{j_2; \hat{\alpha}_4} \right) \right] \]

\[ = \frac{1}{n_{L-1}} \delta_{i_1 i_2} \sum_{\hat{\alpha}_1, \ldots, \hat{\alpha}_4} \left\{ P_{\beta_1\hat{\alpha}_1\hat{\alpha}_2\hat{\beta}_2}^{(L)} \left( \sum_j y_j; \hat{\alpha}_3 y_j; \hat{\alpha}_4 \right) \right. 

+ \left. \left( n_L P_{\beta_1\hat{\alpha}_1\hat{\alpha}_2\hat{\beta}_2}^{(L)} + Q_{\beta_1\hat{\alpha}_1\hat{\alpha}_2\hat{\beta}_2}^{(L)} + Q_{\beta_1\hat{\alpha}_2\hat{\alpha}_1\hat{\beta}_2}^{(L)} \right) G_{\hat{\alpha}_3\hat{\alpha}_4}^{(L)} \right\} Z_{B}^{\hat{\alpha}_1\hat{\alpha}_2\hat{\alpha}_3\hat{\alpha}_4} 

+ \frac{1}{n_{L-1}} \sum_{\hat{\alpha}_1, \ldots, \hat{\alpha}_4} Q_{\beta_1\hat{\alpha}_1\hat{\alpha}_2\hat{\beta}_2}^{(L)} Z_{B}^{\hat{\alpha}_1\hat{\alpha}_2\hat{\alpha}_3\hat{\alpha}_4} \left( y_{i_1; \hat{\alpha}_3 y_{i_2; \hat{\alpha}_4} + y_{i_1; \hat{\alpha}_4 y_{i_2; \hat{\alpha}_3}} \right) \right\} \]
Prediction Variance

To see another manifestation of output “wiring,” consider

\[
\sum_{j_1,j_2} \mathbb{E} \left[ z_{i_2;\beta_2}^{(L)} dH_{i_1j_1j_2;\beta_1\tilde{\alpha}_1\tilde{\alpha}_2}^{(L)} \right] Z_{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} B
\]

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

\[
= \frac{1}{nL-1} \delta_{i_1i_2} \sum_{\tilde{\alpha}_1,...,\tilde{\alpha}_4} \left[ P_{\beta_1\tilde{\alpha}_1\tilde{\alpha}_2\beta_2}^{(L)} \left( \sum_j y_j;\tilde{\alpha}_3 y_j;\tilde{\alpha}_4 \right) \right]
\]

\[
+ \left( nL P_{\beta_1\tilde{\alpha}_1\tilde{\alpha}_2\beta_2}^{(L)} + Q_{\beta_1\tilde{\alpha}_1\tilde{\alpha}_2\beta_2}^{(L)} + Q_{\beta_1\tilde{\alpha}_2\tilde{\alpha}_1\beta_2}^{(L)} \right) G_{\tilde{\alpha}_3\tilde{\alpha}_4}^{(L)} Z_{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} B
\]

\[
+ \frac{1}{nL-1} \sum_{\tilde{\alpha}_1,...,\tilde{\alpha}_4} Q_{\beta_1\tilde{\alpha}_1\tilde{\alpha}_2\beta_2}^{(L)} Z_{\tilde{\alpha}_1\tilde{\alpha}_2\tilde{\alpha}_3\tilde{\alpha}_4} \left( y_{i_1;\tilde{\alpha}_3} y_{i_2;\tilde{\alpha}_4} + y_{i_1;\tilde{\alpha}_4} y_{i_2;\tilde{\alpha}_3} \right)
\]

\[\rightarrow \text{Wiring is exhibited when } y_{i_1} \neq 0 \text{ and } y_{i_2} \neq 0.\]
At finite width, $p\left(z^{(L)}(\infty) \mid D\right)$ now has non-Gaussian statistics:
Prediction is Nearly-Gaussian

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

\[
\begin{align*}
\uparrow z_{i;\hat{\beta}}(\infty) & \equiv z_{i;\hat{\beta}}(\infty) \left[ z^{(L)}, \hat{H}^{(L)}, \hat{d}H^{(L)}, \hat{dd}_{I}H^{(L)}, \hat{dd}_{II}H^{(L)} \right].
\end{align*}
\]
At finite width, \( p\left(z^{(L)}(\infty) \mid \mathcal{D}\right) \) now has non-Gaussian statistics:

\[
\begin{align*}
\triangleright z_{i;\hat{\beta}}(\infty) &\equiv z_{i;\hat{\beta}}(\infty) \left[z^{(L)}, \hat{H}^{(L)}, \hat{dH}^{(L)}, \hat{ddI}H^{(L)}, \hat{ddII}H^{(L)}\right]. \\
\triangleright p\left(z^{(L)}, \hat{H}^{(L)}, \hat{dH}^{(L)}, \hat{ddI}H^{(L)}, \hat{ddII}H^{(L)} \mid \mathcal{D}\right) \text{ is nearly-Gaussian.}
\end{align*}
\]
Prediction is Nearly-Gaussian

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

\[
\begin{align*}
\uparrow & \quad z_{i;\hat{\beta}}(\infty) \equiv z_{i;\hat{\beta}}(\infty) \left[ z^{(L)}, \hat{H}^{(L)}, \hat{dH}^{(L)}, \hat{dd\!\!I\!\!H}^{(L)}, \hat{dd\!\!II\!\!H}^{(L)} \right]. \\
\uparrow & \quad p\left( z^{(L)}, \hat{H}^{(L)}, \hat{dH}^{(L)}, \hat{dd\!\!I\!\!H}^{(L)}, \hat{dd\!\!II\!\!H}^{(L)} \mid D \right) \text{ is nearly-Gaussian.} \\
\uparrow & \quad \text{Explicit expressions of higher-point correlators are challenging to display in any media format...}
\end{align*}
\]
Generalization

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

\[ \mathcal{E} \equiv \mathcal{L}_B - \mathcal{L}_A. \]
The generalization error is a quantitative measure of how well a network is really approximating the desired function:

\[ \mathcal{E} \equiv \mathcal{L}_B. \]
Generalization

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

$$
\mathcal{E} \equiv \mathcal{L}_B.
$$

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

$$
\mathbb{E} \left[ \mathcal{L}_B \left( T \right) \right] = \mathbb{E} \left[ \frac{1}{2} \sum_{i=1}^{n_L} \sum_{\hat{\beta} \in \mathcal{B}} \left( z_{i;\hat{\beta}}^{(L)}(\infty) - y_{i;\hat{\beta}} \right)^2 \right].
$$
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:

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

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

\[ \mathcal{E} \equiv \mathcal{L}_B. \]

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

\[
\mathbb{E} \left[ \mathcal{L}_B(T) \right] = \frac{1}{2} \sum_{\beta \in B} \left\{ \sum_{i=1}^{n_L} \left( m_{i;\hat{\beta}} - y_{i;\hat{\beta}} \right)^2 + \sum_{i=1}^{n_L} \text{Cov} \left[ z_{i;\hat{\beta}}^{(L)}(\infty), z_{i;\hat{\beta}}^{(L)}(\infty) \right] \right\}.
\]

▶ 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:

\[ \mathcal{E} \equiv \mathcal{L}_B. \]

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

\[
\mathbb{E} [\mathcal{L}_B(T)] = \frac{1}{2} \sum_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n_L} \left( m_i;\beta - y_i;\beta \right)^2 + \sum_{i=1}^{n_L} \text{Cov} \left[ z^{(L)}_{i;\beta}(\infty), z^{(L)}_{i;\beta}(\infty) \right] \right\}.
\]

- 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,

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

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,

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

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 **ut-polation**.
*-polation at Infinite Width

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

\[ z_{i;\beta}^{(L)}(\infty) = z_{i;\beta}^{(L)} - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2 \in A} \Theta_{\beta\tilde{\alpha}_1}^{(L)} \tilde{\Theta}_{\tilde{\alpha}_1\tilde{\alpha}_2}^{(L)} \left( z_{i;\tilde{\alpha}_2}^{(L)} - y_{i;\tilde{\alpha}_2} \right). \]
To understand *-polation at infinite by smooth activations, we need to evaluate:

\[ z^{(L)}_{i;\beta}(\infty) = z^{(L)}_{i;\beta} - \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2 \in A} \Theta^{(L)}_{\beta \tilde{\alpha}_1} \tilde{\Theta}^{(L)}_{\tilde{\alpha}_1 \tilde{\alpha}_2} \left( z^{(L)}_{i;\tilde{\alpha}_2} - y_{i;\tilde{\alpha}_2} \right). \]

▶ We need to invert the two-by-two submatrix of the NTK on the training set only, \( \tilde{\Theta}^{(L)}_{\tilde{\alpha}_1 \tilde{\alpha}_2} \).

▶ 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:

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

The inverse of the submatrix can be written as:

$$\widetilde{\Theta}_{\tilde{\alpha}_1 \tilde{\alpha}_2} = \frac{1}{\Theta_{++} \Theta_{--} - \Theta_{+-}^2} \left( \begin{array}{cc} \Theta_{--} & -\Theta_{+-} \\ -\Theta_{+-} & \Theta_{++} \end{array} \right).$$
*-polation at Infinite Width

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

\[ z^{(L)}_{i;\hat{\beta}}(\infty) = z^{(L)}_{i;\hat{\beta}} - \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2 \in A} \Theta^{(L)}_{\beta\tilde{\alpha}_1} \tilde{\Theta}^{(L)}_{\tilde{\alpha}_1\tilde{\alpha}_2} \left( z^{(L)}_{i;\tilde{\alpha}_2} - y_{i;\tilde{\alpha}_2} \right) . \]

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

\[ \delta\delta\Theta[0] \equiv \frac{1}{4} \left[ \Theta_{++} + \Theta_{--} + 2\Theta_{+-} \right] - \Theta_{00} + O\left(\delta^4\right) . \]
As an illustration, consider two training inputs with the same norm:

$$z^{(L)}_{i;(2s-1)}(\infty) = \left[ z^{(L)}_{i;(2s-1)} - sz^{(L)}_{i;+} - (1 - s)z^{(L)}_{i;-} \right] + \left[ sy_{i;+} + (1 - s)y_{i;-} \right]$$

$$- 4s(1 - s) \left( \frac{\delta \delta \Theta[0]}{\Theta_{00}} \right) \left( z^{(L)}_{i;+} + z^{(L)}_{i;-} + y_{i;+} + y_{i;-} \right)$$

$$+ O(\delta^3).$$
As an illustration, consider two training inputs with the same norm:

\[
\begin{align*}
    z_i^{(L)}_{(2s-1)}(\infty) &= \left[ z_i^{(L)}_{(2s-1)} - sz_i^{(L)}_+ - (1 - s)z_i^{(L)}_- \right] + \left[ sy_i^+ + (1 - s)y_i^- \right] \\
    &- 4s(1 - s) \left( \frac{\delta\delta\Theta_{[0]}}{\Theta_{00}} \right) \left( z_i^{(L)}_+ + z_i^{(L)}_- + y_i^+ + y_i^- \right) \\
    &+ O\left(\delta^3\right).
\end{align*}
\]

- Nonlinear networks can nonlinearily \(*-polute!\)
The prediction of the ensemble has a bias:

\[
m_{i; (2s-1)}^\infty - y_{i; (2s-1)} = \left[ y_{i; +} + (1 - s)y_{i; -} - y_{i; (2s-1)} \right] \\
- 4s(1 - s) \left( \frac{\delta \delta \Theta[0]}{\Theta_{00}} \right) (y_{i; +} + y_{i; -}) + O(\delta^3) .
\]
Interpolation at Infinite Width

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

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

▶ This decomposes into a part measuring the **nonlinearity in the labels** and the **network curvature** around \((y_{i;+} + y_{i;-})/2\).
The prediction of the ensemble has a **bias**:

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

- 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:

$$m_i;(2s-1) - y_i;(2s-1) = \left[ y_i;+ + (1-s)y_i;- - y_i;(2s-1) \right] + O(y^3)$$

$$- 4s(1-s) \left( \frac{\delta \delta \Theta[0]}{\Theta_{00}} \right) (y_i;+ + y_i;-) + O(\delta^3) .$$

- 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:

\[
\begin{align*}
A^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^2}, \\
B^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^2}, \\
P^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^2}, \\
Q^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^2}, \\
R^{(L)} K^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^3}, \\
S^{(L)} K^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^3}, \\
T^{(L)} K^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^3}, \\
U^{(L)} K^{(L)} & \sim \frac{1}{n_{L-1} \left( \tilde{H}^{(L)} \right)^3}.
\end{align*}
\]
All terms are proportional to one of these dimensionless ratios:

\[
\begin{align*}
\frac{A^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^2}, \quad & \frac{B^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^2}, \quad \frac{P^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^2}, \quad \frac{Q^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^2}, \\
\frac{R^{(L)} K^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^3}, \quad & \frac{S^{(L)} K^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^3}, \quad \frac{T^{(L)} K^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^3}, \quad \frac{U^{(L)} K^{(L)}}{n_{L-1} \left( \tilde{H}(L) \right)^3}.
\end{align*}
\]

Overall we should find for the finite-width corrections:

\[
m_{i;\hat{\beta}} - m_{i;\hat{\beta}}^\infty = O \left( \frac{L}{n} \right).
\]
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 $O(L^2/n^2)$ 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

\[ S[p(x)] \equiv - \sum_x p(x) \log p(x) . \]
Aside: Information Theory

The entropy of a probability distribution is given by

\[ S[p(x)] \equiv - \sum_x p(x) \log p(x). \]

- 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

\[
S[p(x)] \equiv - \sum_x p(x) \log p(x).
\]

- 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 $p(x, y) = p(x) p(y)$:

$$S[p(x, y)] = S[p(x)] + S[p(y)].$$
Aside: Information Theory

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

\[
S[p(x, y)] = - \sum_{x,y} p(x, y) \log p(x, y)
\]

\[
= - \sum_{x,y} p(x)p(y) \left[ \log p(x) + \log p(y) \right]
\]

\[
= - \sum_{x} p(x) \log p(x) - \sum_{y \in Y} p(y) \log p(y)
\]

\[
= S[p(x)] + S[p(y)].
\]
Aside: Information Theory

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

$$S[p(x, y)] = S[p(x)] + S[p(y)].$$
Aside: Information Theory

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

$$S[p(x, y)] = S[p(x)] + S[p(y)].$$

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

$$S[p(x, y)] < S[p(x)] + S[p(y)].$$
Aside: Information Theory

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

$$S[p(x, y)] = S[p(x)] + S[p(y)].$$

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

$$S[p(x, y)] < S[p(x)] + S[p(y)].$$

This is physically intuitive: for dependent variables observing $y$ doesn’t inform as much as it would have if we didn’t know $x$. 
The **mutual information** (MI) between two random variables is

\[
\mathcal{I}[p(x, y)] \equiv S[p(x)] + S[p(y)] - S[p(x, y)]
\]

\[
= \sum_{x,y} p(x, y) \log \left[ \frac{p(x, y)}{p(x) p(y)} \right].
\]
The **mutual information** (MI) between two random variables is

\[
\mathcal{I}[p(x, y)] \equiv S[p(x)] + S[p(y)] - S[p(x, y)]
\]

\[
= \sum_{x,y} p(x, y) \log \left[ \frac{p(x, y)}{p(x) p(y)} \right].
\]

▶ A functional of a joint probability distribution.
The **mutual information** (MI) between two random variables is

\[ I[p(x, y)] \equiv S[p(x)] + S[p(y)] - S[p(x, y)] \]

\[ = \sum_{x, y} p(x, y) \log \left[ \frac{p(x, y)}{p(x)p(y)} \right]. \]

▶ 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:

$$I[p(x, y)] \geq 0.$$
Aside: Information Theory

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

$$I[p(x, y)] \geq 0.$$ 

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*.
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 \textit{representation learning} against the negative effect of \textit{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 \textbf{inductive bias} in the \textit{prior distribution}.
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 \( L \):

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

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

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

$$r^* \equiv \arg \max_{\{r, \mathcal{M}_1, \mathcal{M}_2\}} I[p(\mathcal{M}_1, \mathcal{M}_2|x)].$$
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.
Using a variational principle, we can compute $S$ to $O(1/n^3)$:

$$S[p(z_1, \ldots, z_m|x)] = \frac{m}{2} \log(2\pi eG) - \frac{(m^2 + 2m)}{16} \left( \frac{V}{nG^2} \right)^2$$
$$+ \frac{(m^3 + 10m^2 + 16m)}{48} \left( \frac{V}{nG^2} \right)^3 + O\left(\frac{1}{n^4}\right).$$
Using a **variational principle**, we can compute $S$ to $O(1/n^3)$:

$$S[p(z_1, \ldots, z_m|x)] = \frac{m}{2} \log(2\pi eG) - \frac{(m^2 + 2m)}{16} \left( \frac{V}{nG^2} \right)^2 \right.$$

$$\left. + \frac{(m^3 + 10m^2 + 16m)}{48} \left( \frac{V}{nG^2} \right)^3 + O\left( \frac{1}{n^4} \right) \right.$$

▶ Note that the correction is definitely negative.
Using a variational principle, we can compute $S$ to $O(1/n^3)$:

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

- 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,

\[
\frac{V^{(L)}}{n_{L-1} (G^{(L)})^2} \equiv \nu r,
\]

with \( \nu \) an activation function–dependent constant, we get:

\[
\mathcal{I} [p(M_1, M_2|x)] = \frac{m_1 m_2}{8} \nu^2 r^2 - \frac{m_1 m_2 (20 + 3m_1 + 3m_2)}{48} \nu^3 r^3 + O(r^4)
\]
Optimal Aspect Ratio

Defining for the four-point vertex ratio,

\[ \frac{V^{(L)}}{n_{L-1} (G^{(L)})^2} \equiv \nu r , \]

with \( \nu \) an activation function–dependent constant, we get:

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

Noting the differing signs, we find:

\[ r^* = \left( \frac{4}{20 + 3n_L} \right) \frac{1}{\nu} . \]
Optimal Aspect Ratio

Defining for the four-point vertex ratio,

\[
\frac{V^{(L)}}{n_{L-1}(G^{(L)})^2} \equiv \nu r,
\]

with \( \nu \) an activation function–dependent constant, we get:

\[
\mathcal{I}[p(M_1, M_2|x)] = \frac{m_1 m_2}{8} \nu^2 r^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48} \nu^3 r^3 + O(r^4)
\]

Noting the differing signs, we find:

\[
r^* = \left( \frac{4}{20 + 3n_L} \right) \frac{1}{\nu}.
\]

\[\text{For } n_L = 10, \ r^* = .12 \ (\text{tanh}) \text{ and } r^* = .016 \ (\text{ReLU}).\]
Optimal Aspect Ratio

Defining for the four-point vertex ratio,
\[
V^{(L)}(n_L) \equiv \nu r,
\]
with \(\nu\) an activation function–dependent constant, we get:
\[
\mathcal{I}[p(M_1, M_2|x)] = \frac{m_1 m_2}{8} \nu^2 r^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48} \nu^3 r^3 + O(r^4)
\]

Noting the differing signs, we find:
\[
r^* = \left( \frac{4}{20 + 3n_L} \right) \frac{1}{\nu}
\]

- For \(n_L = 10\), \(r^* = .12\) (tanh) and \(r^* = .016\) (ReLU).
- **Residual connections** let us push \(r^*\) 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**: 
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**,

\[ r \equiv \frac{L}{n}, \]

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

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

This is sparse, depending on a few objects in a simple way.
At infinite width, we found a **Gaussian** trained distribution:

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

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

▶ This is sparse, depending on a few objects in a simple way.
Sparsity at Infinite Width

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

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

▶ 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 $\Theta_{\delta_1 \delta_2}^{(L)}$, while the variance is only a function of $K_{\delta_1 \delta_2}^{(L)}$ and $\Theta_{\delta_1 \delta_2}^{(L)}$.

▶ 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:

\[ p(z(\infty)) \equiv p(z(\infty) \mid y, G, H, V, A, B, D, F, P, Q, R, S, T, U) + O\left(\frac{1}{n^2}\right) . \]
Near-Sparsity at Finite Width

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

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

In addition to \( G \) and \( H \), here we’re accounting for the finite-width data-dependent couplings arising from:

\[ \mathbb{E} [zzzz]_{\text{connected}} , \quad \mathbb{E} [\Delta H_{zz}] , \quad \mathbb{E} [\Delta H^2] , \]
\[ \mathbb{E} [\hat{d}Hz] , \quad \mathbb{E} [\hat{d}\hat{d}_1H] , \quad \mathbb{E} [\hat{d}\hat{d}_1\parallel H] . \]
Near-Sparsity at Finite Width

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

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

- In addition to \(G\) and \(H\), here we’re accounting for the finite-width data-dependent couplings arising from:

  \[ \mathbb{E}\left[zzzz\right]_{\text{connected}}, \quad \mathbb{E}\left[\hat{\Delta}Hzz\right], \quad \mathbb{E}\left[\hat{\Delta}H^2\right], \]
  \[ \mathbb{E}\left[\hat{d}Hz\right], \quad \mathbb{E}\left[\hat{dd}_I H\right], \quad \mathbb{E}\left[\hat{dd}_{II} H\right]. \]

- 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 $N_D$:

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

  $$n_{\text{out}} N_A + \left[ \frac{N_D(N_D + 1)}{2} \right] + \left[ \frac{N_D(N_D + 1)}{2} \right] = O\left( N_D^2 \right)$$

  numbers in order to completely specify the distribution.
Consider a fixed combined training and test dataset of size $N_D$:

- For the finite-width nearly-Gaussian distribution with $0 < r \ll 1$, we will instead need $O(N_D^4)$ numbers, with the counting dominated by the finite-width tensors.
Consider a fixed combined training and test dataset of size $N_D$:

For an accuracy $O\left(\frac{L^k}{n^k}\right)$, a *macroscopic description*

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

will need $O\left(N_D^{2k}\right)$ numbers in general.
Consider a fixed combined training and test dataset of size $N_D$:

- For an accuracy $O\left(\frac{L^k}{n^k}\right)$, a *macroscopic description*

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

will need $O\left(N_D^{2k}\right)$ numbers in general.

- The 1/$n$ *expansion* gives a sequence of effective theories with increasing accuracy at the cost of increasing complexity.
What we have found here is the manifestation of the \textbf{microscopic-macroscopic duality}:
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*.
What we have found here is the manifestation of the \textit{microscopic-macroscopic duality}:

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

- In the \textit{overparameterized regime}, this indicates that we should identify the \textit{model complexity} with the \textit{data-dependent couplings} rather than the \textit{model parameters}.
As $r$ increases, we’ll need to include more of these higher-order terms, making our macroscopic description more complex:
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 $O(N_D^2)$ Gaussian description of the infinite-width limit will be accurate.
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 \textit{sparse} $O(N_D^2)$ \textbf{Gaussian} description of the infinite-width limit will be accurate.
- In the regime $0 < r \sim r^* \ll 1$, the \textit{nearly-sparse} $O(N_D^4)$ \textbf{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* $O(N_D^2)$ **Gaussian** description of the infinite-width limit will be accurate.
- In the regime $0 < r \sim r^* \ll 1$, the *nearly-sparse* $O(N_D^4)$ **nearly-Gaussian** description of the finite-width effective theory truncated at order $1/n$ will be accurate.
- For larger $r$, a more generic $O(N_D^{2k})$ **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!