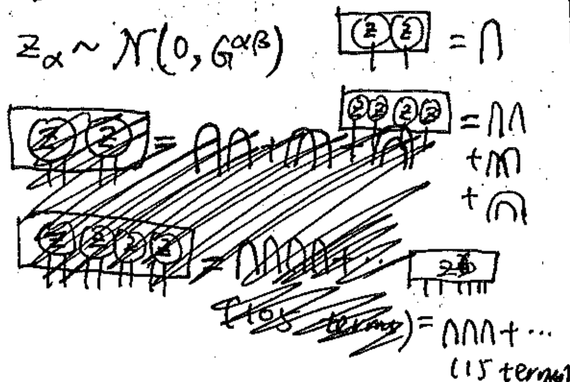
Ref sheet

Isserlis theorem

$z_\alpha \sim \mathcal{N}(0, G^{\alpha\beta})$



$z_{i\alpha} \sim \mathcal{N}(0, \delta^{ij} G^{\alpha\beta})$

Taylor expansions

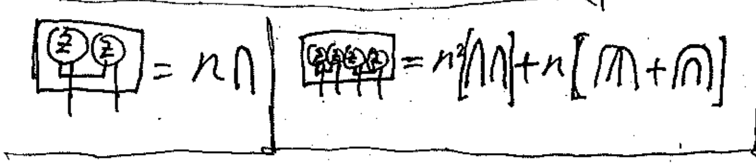
$$\frac{1}{1+x} = 1 - x + x^2 - x^3 \dots$$

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 \dots$$

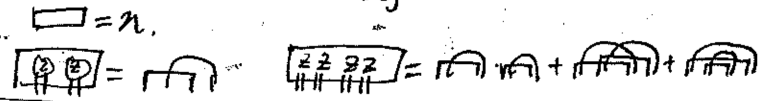
$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 \dots$$

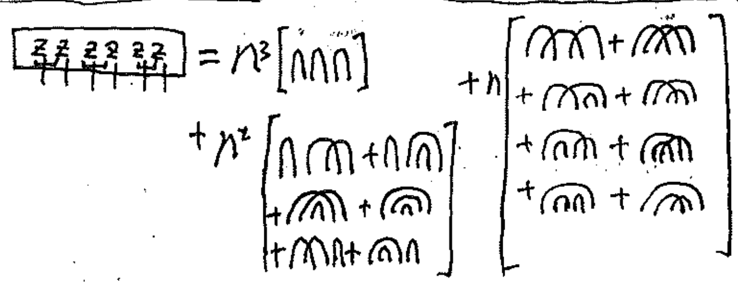
$$(1+x)^m = 1 + mx + \frac{m(m-1)}{2}x^2 + \frac{m(m-1)(m-2)}{6}x^3 \dots$$

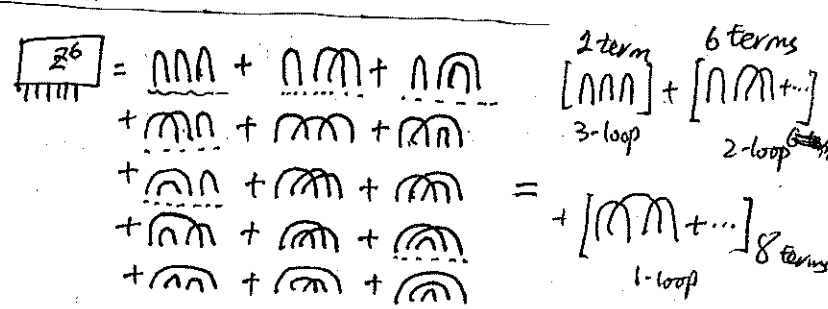
$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 \dots$$



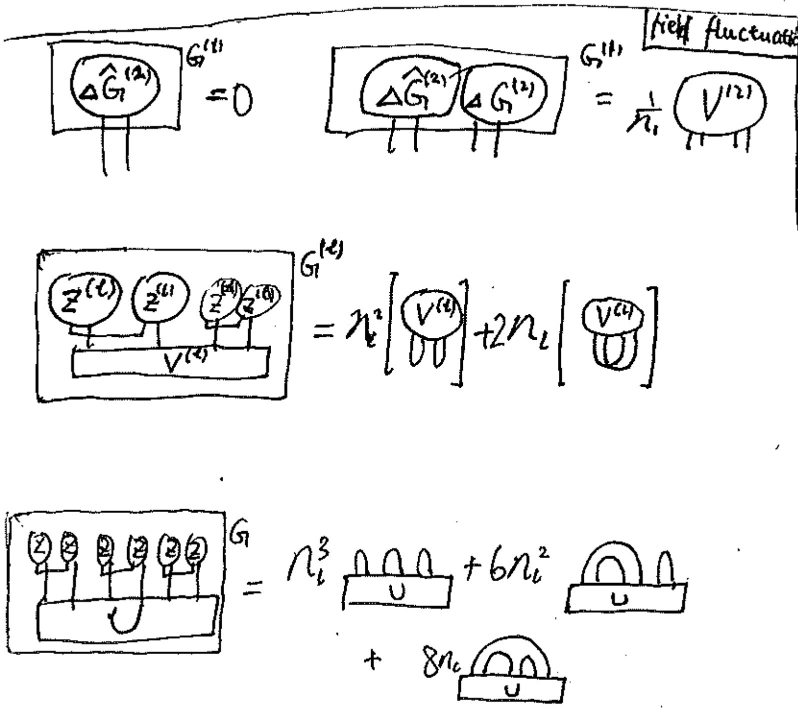
Kronecker delta: δ_{ij}







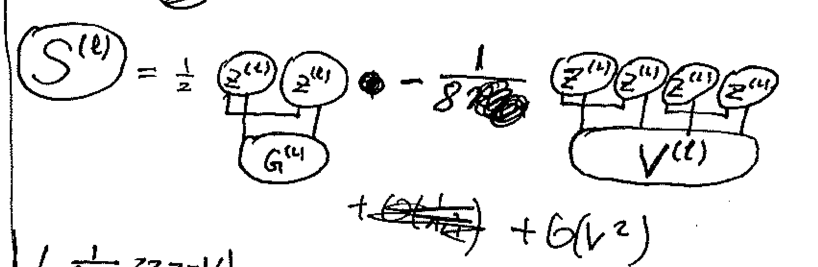
field fluctuation

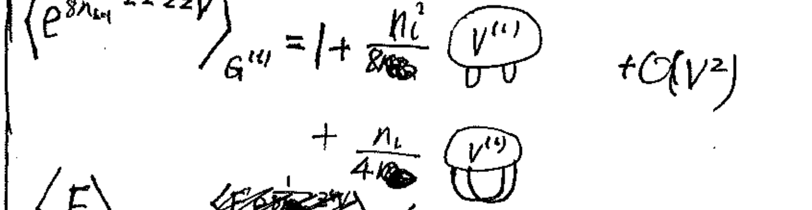


$$\hat{G}^{(\ell+1)}_{\alpha\beta} = C_b^{(\ell+1)} + C_W^{(\ell+1)} \frac{1}{n_\ell}\sum_i \sigma^\ell_{i\alpha}\sigma^\ell_{i\beta} = C_b^{(\ell+1)} + C_W^{(\ell+1)} \langle \hat{\sigma}^\ell_\alpha, \hat{\sigma}^\ell_\beta \rangle$$

$$\frac{1}{n_\ell} V^{(\ell+1)}_{\alpha_1\alpha_2,\beta_1\beta_2} = \left(C_W^{(\ell+1)}\right)^2 \mathrm{Cov}\left[\langle \hat{\sigma}^\ell_{\alpha_1}, \hat{\sigma}^\ell_{\alpha_2}\rangle, \langle \hat{\sigma}^\ell_{\beta_1}, \hat{\sigma}^\ell_{\beta_2}\rangle\right]$$

quartic action





$$\langle F \rangle_{S^{(\ell)}} = \frac{\langle F e^{\frac{1}{8} z^4 V}\rangle_{G}}{\langle e^{\frac{1}{8} z^4 V}\rangle_{G}}$$

$$= \langle F \rangle_{G^{(\ell)}} + \frac{1}{8}\mathrm{Cov}_{G^{(\ell)}}\left[F, zzzzV^{(\ell)}\right] + O(V^2)$$

$$\mathrm{Cov}_{S^{(\ell)}}[F, G] = \mathrm{Cov}_{G^{(\ell)}}[F, G] + \frac{1}{8}\left[\begin{array}{l}\langle FGz^4V\rangle - \langle F\rangle\langle Gz^4V\rangle - \langle G\rangle\langle Fz^4V\rangle \\ + 2\langle F\rangle\langle G\rangle\langle z^4V\rangle - \langle FG\rangle\langle z^4V\rangle\end{array}\right]$$

If $F(z^{(\ell)}_{i_1;A})$ and $G(z^{(\ell)}_{j_2;B})$

depend on two different neurons $i_1 \neq i_2$

then

$$\mathrm{Cov}_{z^{(\ell)}\sim S^{(\ell)}}[F, G] = \frac{1}{8}\left[\begin{array}{l}\langle F G\, z^4\rangle_{G^{(\ell)}\delta ij} - \langle F\rangle\langle G z^4\rangle \\ - \langle G\rangle\langle F z^4\rangle + \langle F\rangle\langle G\rangle\langle z^4\rangle\end{array}\right] V^{(\ell)} = \frac{1}{8}\sum_{\beta_1\cdots\beta_4}\sum_{j_1 j_2} V^{(\beta_1\beta_2)(\beta_3\beta_4)}\left(\begin{array}{l}\left\langle \begin{array}{l}F(z_{i_1})\, z_{j_1\beta_1} z_{j_1\beta_2} \\ G(z_{i_2})\, z_{j_2\beta_3} z_{j_2\beta_4}\end{array}\right\rangle_{G^{(\ell)}\delta ij} \\ - \cdots \\ - \cdots \\ + \cdots\end{array}\right)$$

By enumeration, the summation terms are zero, except when $\begin{array}{l} i_1 = j_1 \\ i_2 \neq j_2\end{array}$ or $\begin{array}{l} i_1 = j_2 \\ i_2 = j_1\end{array}$

giving us just 8 terms, which ~~equals~~ then gives

$$\mathrm{Cov}_{z^{(\ell)}\sim S^{(\ell)}}[F(z^{(\ell)}_{i_1;A_1}), G(z^{(\ell)}_{i_2;A_2})] = \frac{1}{4}\sum_{\beta_{1:4}} V^{(\beta_1\beta_2)(\beta_3\beta_4)} \left\langle F(z^{(\ell)}_{A_1})\left(z_{\beta_1} z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{G^{(\ell)}} \left\langle G(z^{(\ell)}_{A_2})\left(z_{\beta_3} z_{\beta_4} - G^{(\ell)}_{\beta_3\beta_4}\right)\right\rangle_{G^{(\ell)}}$$

~~more succinctly,~~

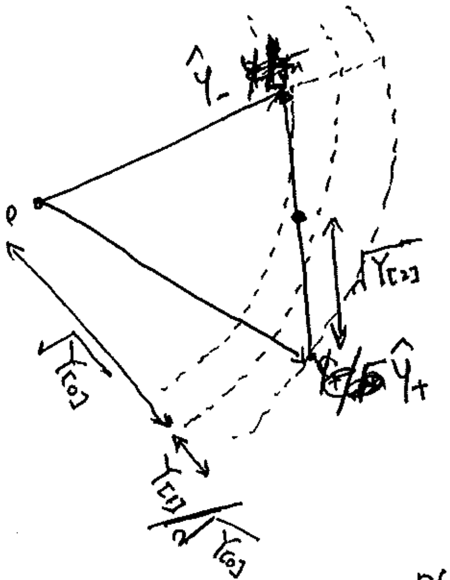
$$\cancel{\mathrm{Cov}_{S^{(\ell)}}[F(z^{(\ell)}_{i_1;A_1}), G(z^{(\ell)}_{i_2;A_2})] = \frac{1}{4} V \langle F_{A_1}(zz - g)\rangle\langle G_{A_2}(zz - g)\rangle}$$

In particular, we have 4-point covariance

$$\mathrm{Cov}[\sigma^{(\ell)}_{i_1\alpha_1}\sigma^{(\ell)}_{i_1\alpha_2}, \sigma^{(\ell)}_{i_2\alpha_3}\sigma^{(\ell)}_{i_2\alpha_4}] = \frac{1}{4}\sum_{\beta_{1:4}} V^{\beta_1\cdots\beta_4}\left\langle \sigma_{\alpha_1}\sigma_{\alpha_2}\left(z_{\beta_1}z_{\beta_2} - G^{(\ell)}_{\beta_1\beta_2}\right)\right\rangle_{G^{(\ell)}}\langle\cdots\rangle \qquad i_1 \neq i_2$$

We also have $\qquad i_1 = i_2$

$$\left\langle F(z^{(\ell)}_{i_1;A_1})\, G(z^{(\ell)}_{i_1;A_2})\right\rangle_{S^{(\ell)}} = \left\langle F^{(\ell)}_{A_1} G^{(\ell)}_{A_2}\right\rangle_{G^{(\ell)}} +$$



$$Y_{[0]} = \left\| \frac{1}{2}(y_+ + y_-)/\sqrt{n} \right\|^2 = \left\| \tfrac{1}{2}(\hat{y}_+ + \hat{y}_-) \right\|^2$$

$$Y_{[1]} = \left\| \frac{1}{2} y_+/\sqrt{n} \right\|^2 - \left\| \tfrac{1}{2} y_-/\sqrt{n} \right\|^2 = \left\| \tfrac{1}{2}\hat{y}_+ \right\|^2 - \left\| \tfrac{1}{2}\hat{y}_- \right\|^2$$

$$Y_{[2]} = \left\| \frac{1}{2}(y_+ - y_-)/\sqrt{n} \right\|^2 = \left\| \tfrac{1}{2}(\hat{y}_+ - \hat{y}_-) \right\|^2$$

$$K = \cancel{K_{[0]}Y_{[0]} + K_{[1]}Y_{[1]} + K_{[2]}Y_{[2]}}\ K_{[0]}\begin{bmatrix}1&1\\1&1\end{bmatrix} + K_{[1]}\begin{bmatrix}1&\\&-1\end{bmatrix} + K_{[2]}\begin{bmatrix}1&-1\\-1&1\end{bmatrix}$$

and $y \sim \mathcal{N}(0, \delta_{ij} K_{\alpha\beta})$

then ~~$Y_{[0]}, Y_{[1]}, Y_{[2]}$ has joint PDF~~

$$P(y_+, y_- | H) = \left(4\pi^2 (4K_0K_2 - K_1^2)\right)^{-\frac{n}{2}} \exp\left(-\frac{n}{4K_0K_2 - K_1^2}\left(2K_0Y_2 + 2K_2Y_0 - K_1Y_1\right)\right)$$

(equation 6.48)

Let $\hat{K}_{[i]} = \frac{1}{n}K_{[i]}$, $\hat{y}_\pm = \frac{1}{\sqrt{n}}y_\pm$, then $P(\hat{y}_+, \hat{y}_- | H) = \left(4\pi^2(4\hat{K}_0\hat{K}_2 - \hat{K}_1^2)\right)^{-\frac{n}{2}} \exp\left(-\frac{2\hat{K}_0Y_2 + 2\hat{K}_2Y_0 - \hat{K}_1Y_1}{4\hat{K}_0\hat{K}_2 - \hat{K}_1^2}\right)$

$\hat{\sigma}_\alpha^l = \sigma_\alpha^l / \sqrt{n_l}$

$\hat{G}_{\alpha\beta}^{(l+1)} = C_b^{(l+1)} + C_W^{(l+1)}(\hat{\sigma}_\alpha^l \cdot \hat{\sigma}_\beta^l)$

$K_{\alpha\beta}^{(l+1)} = \langle \hat{G}_{\alpha\beta}^{(l+1)} \rangle = C_b^{(l+1)} + C_W^{(l+1)}\langle \hat{\sigma}_\alpha^l \cdot \hat{\sigma}_\beta^l \rangle = \langle \hat{z}_\alpha^{(l+1)} \cdot \hat{z}_\beta^{(l+1)} \rangle$

$V_{\alpha_1\alpha_2\beta_1\beta_2}^{(l+1)\{0\}} = n_l (C_W^{(l+1)})^2 \mathrm{Cov}[\hat{\sigma}_{\alpha_1}^l \cdot \hat{\sigma}_{\alpha_2}^l, \hat{\sigma}_{\beta_1}^l \cdot \hat{\sigma}_{\beta_2}^l]$

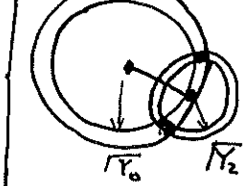
$= \frac{(C_W^{(l+1)})^2}{n_l} \mathrm{Cov}[\sigma_{\alpha_1}^l \sigma_{\alpha_2}^l, \sigma_{\beta_1}^l \sigma_{\beta_2}^l]$

If $K_{[1]} = 0$, and $Y_{[1]} = 0$,

then

$P(Y_{[0]}, Y_{[2]} | Y_{[1]} = 0)$ approximately

$$\propto Y_{[0]}^{\frac{n}{2}-1} Y_{[2]}^{\frac{n}{2}-1} \exp\left(-\frac{Y_{[0]}\hat{K}_{[2]} - Y_{[2]}\hat{K}_{[0]}}{2\hat{K}_0\hat{K}_2}\right)$$



which has maximum

$Y_{[0]} = (n-2)\hat{K}_{[0]} \approx K_{[0]}$

$Y_{[2]} = (n-2)\hat{K}_{[2]} \approx K_{[2]}$

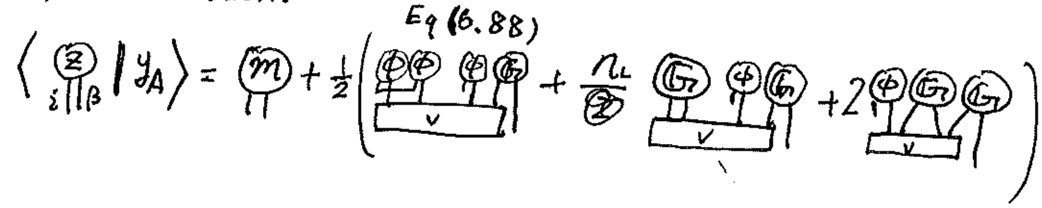
$$G^{(l)} = \underbrace{G^{\{0\}(l)}}_{:=K} + \frac{1}{n_{l-1}}G^{\{1\}(l)} + \frac{1}{n_{l-1}^2}G^{\{2\}(l)} + \cdots = \langle z_{i\alpha}^{(l)} z_{i\beta}^{(l)} \rangle = \langle \hat{z}_\alpha^{(l)} \cdot \hat{z}_\beta^{(l)} \rangle$$

$$V^{(l)} = V^{\{0\}(l)} + \frac{1}{n_{l-1}}V^{\{1\}(l)} + \frac{1}{n_{l-1}^2}V^{\{2\}(l)} + \cdots$$

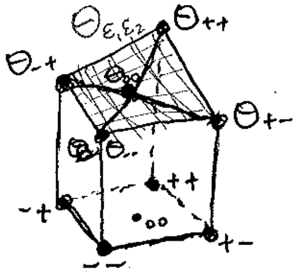
$$= \mathrm{Cov}[\hat{z}_{\alpha_1}^l \cdot \hat{z}_{\alpha_2}^l, \hat{z}_{\beta_1}^l \cdot \hat{z}_{\beta_2}^l] = \mathrm{Cov}[\hat{G}_{\alpha_1\alpha_2}^{(l)}, \hat{G}_{\beta_1\beta_2}^{(l)}]$$

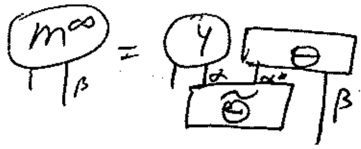
Bayesian correction.

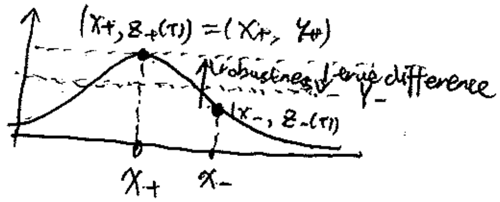
Eq (6.88)



Kernel learning

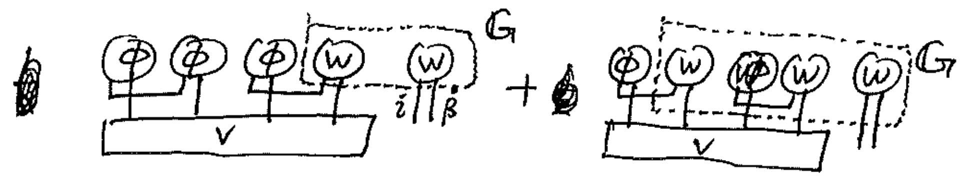




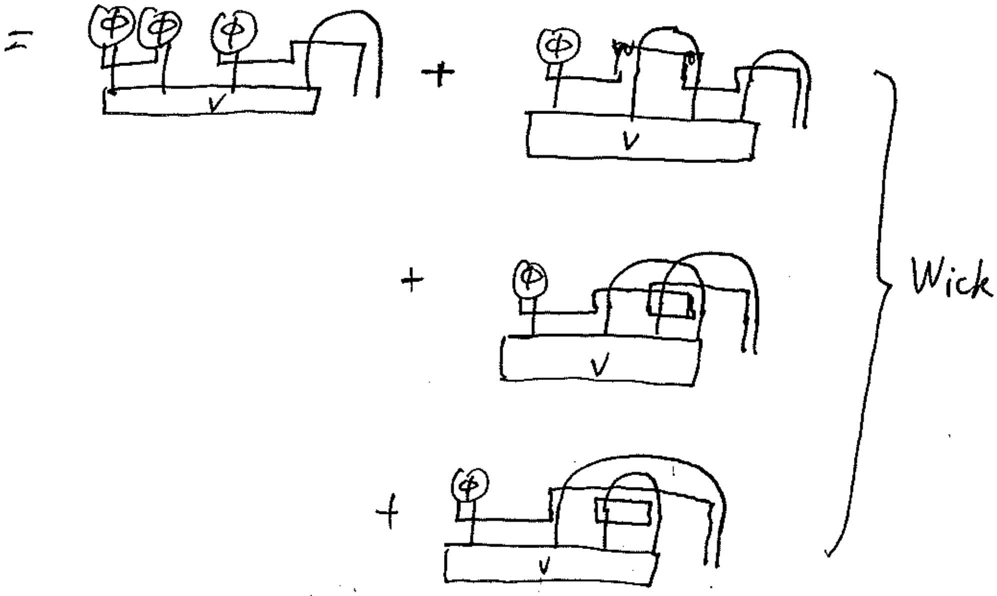


Example of using Feynmann diagram: evaluating (6.88).

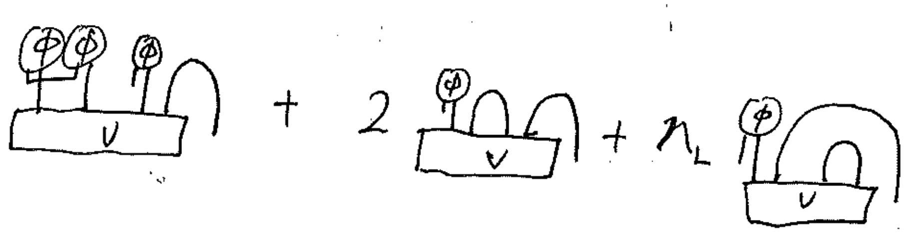
$2\langle\langle W_{i;\beta}\, Q(\omega)\rangle\rangle_G =$



$=$



$=$



$\square = n_L = \sum_{i,j} \delta_{ij}\delta^{ij}$

$\gamma^{[0]}$ $\gamma^{[1]}$ $\gamma^{[2]}$

$\begin{matrix}+&+\\+&+\end{matrix}$ $\quad$ $\begin{matrix}+&0\\0&-\end{matrix}$ $\quad$ $\begin{matrix}+&-\\-&+\end{matrix}$

$K_{00}^{(\ell)} = \langle \frac{1}{n_\ell} \| z_0^{(\ell)} \|^2 \rangle$

$R^{(\ell)} = \langle \frac{1}{n_\ell} \| z_+^{(\ell)} \|^2 - \frac{1}{n_\ell} \| z_-^{(\ell)} \|^2 \rangle$

$D^{(\ell)} = \langle \frac{1}{n_\ell} \| z_+^{(\ell)} - z_-^{(\ell)} \|^2 \rangle$

$$\begin{bmatrix} K_{[0]} \\ K_{[1]} \\ K_{[2]} \\ R \\ D \end{bmatrix} = \begin{bmatrix} 1/4 & 1/4 & 2/4 \\ 2/4 & -2/4 & 0 \\ 1/4 & 1/4 & -2/4 \\ 1 & -1 & 0 \\ 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} K_{++} \\ K_{--} \\ K_{+-} \end{bmatrix}$$

$$K = K_{[0]} \gamma^{[0]} + K_{[1]} \gamma^{[1]} + K_{[2]} \gamma^{[2]}$$

$$g(K) := \frac{1}{\sqrt{2\pi K}} \int dz\, e^{-\frac{z^2}{2K}} \sigma(z) \sigma(z) = \mathbb{E}_{z \sim \mathcal{N}(0,K)} [\sigma(z)^2]$$

~~trivial~~ fixed-point of 1 input.

$$\begin{cases} K_{00}^{(\ell+1)} \leftarrow C_b + C_W\, g(K_{00}^{(\ell)}) \\ K_{00}^{*(\ell+1)} = K_{00}^{*(\ell)} = K_{00}^{*} \end{cases}$$

parallel sus. $K_{00}^{(\ell+1)} = K_{00}^{*} + \Delta K_{00}^{(\ell+1)}$

$$\Delta K_{00}^{(\ell+1)} = \chi_{\parallel}(K_{00}^{*}) \cdot \Delta K_{00}^{(\ell)} + O(\delta^2)$$

$$\chi_{\parallel}(K) = \frac{C_W}{2K^2} \langle \sigma(z) \sigma(z) (z^2 - K) \rangle_{z \sim \mathcal{N}(0,K)}$$

coincidence-limit of 2 inputs / coincident fixed-point

$K_{00}^{*} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $R = D = 0$

is a fixedpoint of $\boxed{K_{\alpha\beta}^{(\ell+1)} \leftarrow C_b + C_W \langle \sigma(z_\alpha) \sigma(z_\beta) \rangle_{K^{(\ell)}}}$

$$K_{\alpha\beta}^{(\ell)} := \langle \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} z_{i,\alpha}^{(\ell)} z_{i,\beta}^{(\ell)} \rangle$$

(kernel recursion formula)

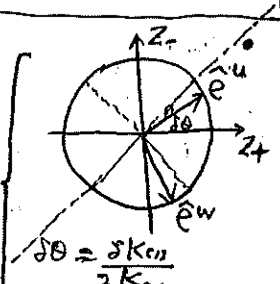
3 perturbations

$$\begin{cases} \Delta K_{00} \\ \Delta R = \Delta(K_{++} - K_{--}) \end{cases} \text{single input}$$

$$\Delta D = \Delta(K_{++} + K_{--} - 2K_{+-}) \quad \text{something new}$$

~~$\langle F(z_+, z_-) \rangle_{K^{(\ell)}} =$~~

$$\begin{cases} K = (K_{00} + \delta^2 K_{[0]}) \gamma^{[0]} \\ \quad + (\delta K_{[1]}) \gamma^{[1]} \\ \quad + (\delta^2 K_{[2]}) \gamma^{[2]}, \\ K \hat{e}^u = \lambda_u \hat{e}^u, \; K \hat{e}^w = \lambda_w \hat{e}^w \end{cases} \Rightarrow$$



$$\delta\theta = \frac{\delta K_{[1]}}{2K_{00}}$$

$$\lambda_u = 2K_{00} + 2\delta\delta K_{[0]} + \frac{(\delta K_{[1]})^2}{2K_{00}}$$

$$\lambda_w = 2\delta\delta K_{[2]} - \frac{(\delta K_{[1]})^2}{2K_{00}}$$

~~input~~ midpoint perturbation.

let x_0 be a whatever input, ("midpoint input")

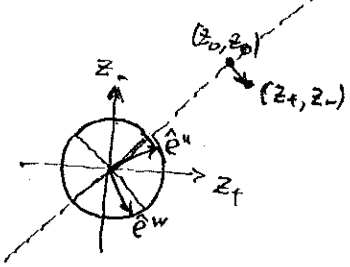
let δx be a whatever input. ("perturbation")

Define $x_\pm := x_0 \pm \delta x$ (infinitesimally-separated input points)

Define $K_{00}^{(\ell)}$ to be the $K_{00}^{(\ell)}$ if we use x_0 as input.

$$K_{00}^{(\ell)} := \langle \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} [z_{i,\alpha}^{(\ell)}(x_0)]^2 \rangle$$

"midpoint kernel"



Strategy:

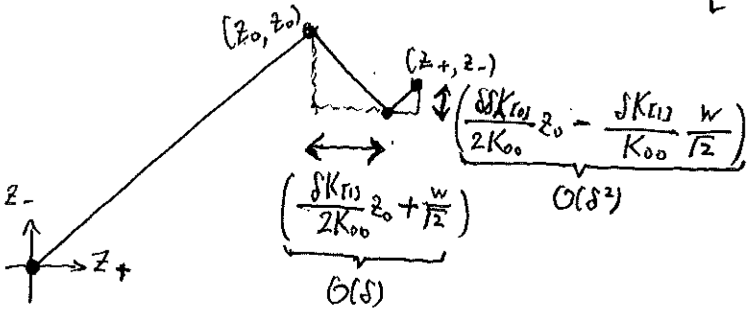
① Let $x_0 = \frac{1}{2}(x_+ + x_-)$, $x_\pm = x_0 \pm \delta x$

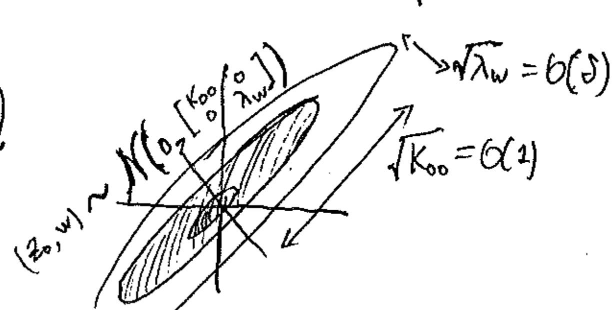
1-input: ② analyze midpoint kernel $K_{00}^{(\ell)}$, $\Delta K_{00}^{(\ell)}$ to order δ^2.

③ analyze $\Delta R^{(\ell)} = \Delta K_{++}^{(\ell)} - \Delta K_{--}^{(\ell)}$ to order δ^2

2-input: ④ analyze ΔD. to order δ^2.

$$\langle F(z_+, z_-) \rangle_{z \sim K} = \langle F(z_0, z_0) \rangle_{K_{00}} + \left(\frac{\delta K_{[1]}}{2K_{00}} \right) \langle z_0 (\partial_+ - \partial_-) F \rangle_{K_{00}} + \left[\frac{\delta\delta K_{[0]}}{2K_{00}} \langle z_0 (\partial_+ + \partial_-) F \rangle_{K_{00}} + \frac{1}{2} \left\langle \left[\delta\delta K_{[2]} + \left(\frac{\delta K_{[1]}}{2K_{00}} \right)^2 (z_0^2 - K_{00}) \right] (\partial_+ - \partial_-)^2 F \right\rangle_{K_{00}} \right]$$

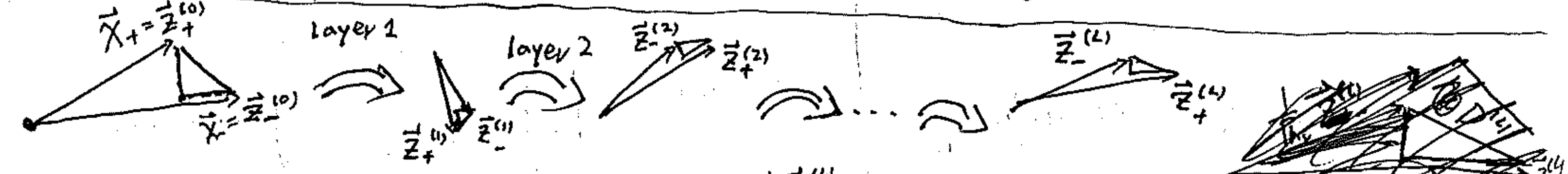


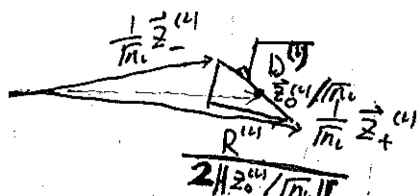


$$\begin{cases} K_{00}^{(\ell+1)} = C_b + C_W \langle\sigma\sigma\rangle_{K_{00}^{(\ell)}} = K_{00}^{*} + \chi_\parallel(K_{00}^{*})\Delta K_{00}^{(\ell)} + O(\Delta^2) = C_b + C_W\, g(K_{00}^{(\ell)}) \\ \delta K_{[1]}^{(\ell+1)} = \left(\dfrac{\delta K_{[1]}^{(\ell)}}{K_{00}^{(\ell)}}\langle z\sigma'\sigma\rangle_{K_{00}^{(\ell)}}\right) C_W = \frac12 R^{(\ell+1)} = \chi_\parallel(K_{00}^{(\ell)})\,\delta K_{[1]}^{(\ell)} \\ \delta\delta K_{[2]}^{(\ell+1)} = \left(\delta\delta K_{[2]}^{(\ell)}\langle\sigma'\sigma'\rangle_{K_{00}^{(\ell)}} + \left(\dfrac{\delta K_{[1]}^{(\ell)}}{2K_{00}^{(\ell)}}\right)^2 \langle (z^2 - K_{00}^{(\ell)})\sigma'\sigma'\rangle_{K_{00}^{(\ell)}}\right) C_W = \frac14 D^{(\ell+1)} = \chi_\perp(K_{00}^{(\ell)})\,\delta\delta K_{[2]}^{(\ell)} + h(K_{00}^{(\ell)})\left(\delta K_{[1]}^{(\ell)}\right)^2 \end{cases}$$

$$\begin{cases} g(K) = \langle\sigma\sigma\rangle_{z\sim\mathcal{N}(0,K)} \\ \chi_\parallel(K) = C_W g'(K) = \dfrac{C_W}{2K^2}\langle\sigma\sigma(z^2-K)\rangle_K = \dfrac{C_W}{K}\langle z\sigma'\sigma\rangle_K = C_W\langle\sigma''\sigma+\sigma'\sigma'\rangle_K \\ \chi_\perp(K) = C_W\langle\sigma'\sigma'\rangle_K \\ h(K) = \dfrac{C_W}{4K^2}\langle\sigma'\sigma'(z^2-K)\rangle_K = \frac12\chi_\perp'(K) = \dfrac{C_W}{2K}\langle z\sigma''\sigma'\rangle_K \end{cases}$$

$$\begin{cases} K_{00}^{(\ell)} \doteq \text{midpoint norm}^2 \\ \delta K_{[1]}^{(\ell)} = \frac12 R^{(\ell)} : \Delta(\text{norm}^2) \\ \delta\delta K_{[2]}^{(\ell)} = \frac14 D^{(\ell)} : \text{norm}^2(\Delta) \end{cases}$$

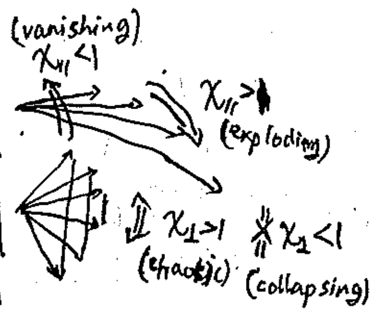




$$\begin{cases} R^{(\ell+1)} = \chi_\parallel R^{(\ell)} \\ D^{(\ell+1)} = \chi_\perp D^{(\ell)} + h(R^{(\ell)})^2 \\ K_{00}^{(\ell)} = \langle \|\vec{z}_0^{(\ell)}/\sqrt{n_\ell}\|^2\rangle + O(\varepsilon^2) \end{cases}$$

criticality equations

$$K_{00}^{*} = C_b + C_W\, g(K_{00}^{*})$$
$$\chi_\parallel(K_{00}^{*}) = 1$$
$$\chi_\perp(K_{00}^{*}) = 1$$



$\sigma = \sum \frac{\sigma_k}{k!} z^k$ with $\sigma(0)=0$, $\sigma'(1)\neq 0$

$$g = \langle\sigma\sigma\rangle_K = \sigma_1^2\left[K + a_1K^2 + a_2K^3 + O(K^4)\right]$$
$$\chi_\parallel = C_W\sigma_1^2\left[1 + 2a_1K + 3a_2K^2 + O(K^3)\right]$$
$$\chi_\perp = C_W\sigma_1^2\left[1 + b_1K + O(K^2)\right]$$

$$\begin{cases} a_1 = \bar\sigma_3 + \frac34\bar\sigma_2^2 \\ a_2 = \frac{1}{24}(6\bar\sigma_5 + 15\bar\sigma_4\bar\sigma_2 + 10\bar\sigma_3^2) \\ b_1 = \bar\sigma_3 + \bar\sigma_2^2 \end{cases}$$

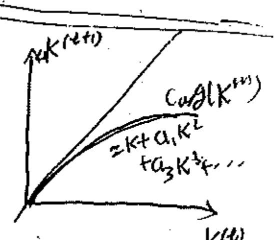
$$h = \frac12 C_W\sigma_1^2(b_1 + O(K))$$

$$K_{00}^{(\ell)} \sim -\frac{1}{a_1\ell}, \quad \delta K_{[1]}^{(\ell)} \sim \frac{\delta}{\ell^2}, \quad \delta\delta K_{[2]}^{(\ell)} \sim \frac{\delta^2}{\ell^{b_1/a_1}}$$

If $a_1 = b_1$, then we have angle preservation on a ball.

Criticality at $C_b = 0$, $C_W = 1/\sigma_1^2$

$$h = \frac12 b_1 + O(K)$$
$$\Delta K^{(\ell)} = \frac{1}{-a_1}\frac{1}{\ell} + O\left(\frac{\log \ell}{\ell^2}\right)$$
$$V^{(\ell)} = \frac{2}{3a_1^2}\frac{1}{\ell} + O\left(\frac{\log\ell}{\ell^2}\right)$$
$$\frac{V^{\ell}}{n(K^{\ell})^2} = \frac23\frac{\ell}{n} + O\left(\frac{\log\ell}{n}\right)$$
$$g^{(\ell)} = \frac{\sigma_1^2}{-a_1}\frac{1}{\ell} + O\left(\frac{\log\ell}{\ell^2}\right)$$



NTK

$$H_{ij\alpha\beta} = \lambda_{\mu\nu}\, \partial_{\theta_\mu} z_{i\alpha}\, \partial_{\theta_\nu} z_{j\beta}$$

$\mathcal{H} := (z)(z)$ with λ — for output $z = z^{(L)}$

$$\varepsilon^{i,\tilde\alpha} := \partial_{z_{i,\tilde\alpha}} \mathcal{L}_A$$

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \lambda_{\mu\nu}\, \partial_{\theta_\nu} \mathcal{L}_A + \mathcal{O}(\eta^2)$$

$$\mathcal{O}(t+1) = \mathcal{O}(t) - \eta\, \partial_{z_{i,\tilde\alpha}} \mathcal{L}_A\, \partial_{z_{j,\delta}} \mathcal{O}\, H_{ij\tilde\alpha\delta} + \mathcal{O}(\eta^2)$$

$\mathcal{O}(t+1) - \mathcal{O}(t) = -\eta\, \mathcal{H}\,(\mathcal{L}_A)_{z_A}\,(\mathcal{O})_{z_D} = -\eta\, \mathcal{H}\,(\varepsilon_A)\,(\mathcal{O})_{z_D}$

---

more generally,

$\mathcal{H}^{(l)} = (z^{(l)})(z^{(l)})$ with λ, θ

$$\lambda = \begin{cases} \delta\, \lambda_b^{(l)} \\ \delta\delta\, \dfrac{\lambda_W^{(l)}}{n_{l-1}} \end{cases}$$

$$\mathcal{H}^{(l)}_{i_1 i_2 \alpha_1 \alpha_2} = \lambda_{\mu\nu}\, \partial_{\theta_\mu} z^{l}_{i_1\alpha_1}\, \partial_{\theta_\nu} z^{l}_{i_2\alpha_2}$$

$\hat{H}$ means "at initialization"

$$H^{(l+1)}_{i_1 i_2 \alpha_1 \alpha_2} = \delta_{i_1 i_2}\left[\lambda_b^{(l+1)} + \frac{\lambda_W^{(l+1)}}{n_l}\left(\sigma^{(l)}_{\alpha_1} \cdot \sigma^{(l)}_{\alpha_2}\right)\right] \text{ new weights}$$
$$\underbrace{+ \sum_{j_1 j_2} H^{(l)}_{j_1 j_2 \alpha_1 \alpha_2}\, W^{(l+1)}_{i_1 j_1} (\sigma')^{(l)}_{j_1\alpha_1}\, W^{(l+1)}_{i_2 j_2} (\sigma')^{(l)}_{j_2\alpha_2}}_{\text{chain rule forward prop.}}$$

RN flow

layer 1

$$\hat{G}^{(1)}_{\alpha_1\alpha_2} = C_b^{1} + C_W^{1} (\vec{x}_{\alpha_1} \cdot \vec{x}_{\alpha_2})/n_0$$

$$P(z^{(1)}, \hat{H}^{(1)} | \mathcal{D}) = P_{\hat{G}^{(1)}}(z^{(1)}) \prod \delta\left(\hat{H}^{(1)}_{i_1 i_2 \alpha_1 \alpha_2} - \delta_{i_1 i_2} H^{(1)}_{\alpha_1\alpha_2}\right)$$

$$H^{(1)}_{\alpha_1\alpha_2} = \lambda_b^{(1)} + \lambda_W^{(1)} (\vec{x}_{\alpha_1} \cdot \vec{x}_{\alpha_2})/n_0$$

$$\mathcal{H}^{(1)} = \delta \left(\lambda_b^{(1)} + \frac{\lambda_W^{(1)}}{n_0} (x)(x)\right)$$

$$\langle \hat{H}^{(2)} \rangle = \delta\, H^{(2)} \qquad H^{(2)} = \lambda_b^{(2)} + \lambda_W^{(2)} \langle \sigma\sigma \rangle_{G^{(1)}} + C_W^{(2)} \langle H^{(1)} \sigma'\sigma' \rangle_{G^{(1)}}$$

$$H^{(2)}_{\alpha_1\alpha_2} = \lambda_b^{(2)} + \lambda_W^{(2)} \langle \sigma^{(1)}_{\alpha_1} \sigma^{(1)}_{\alpha_2} \rangle_{G^{(1)}} + C_W^{(2)} H^{(1)}_{\alpha_1\alpha_2} \langle \sigma'^{(1)}_{\alpha_1} \sigma'^{(1)}_{\alpha_2} \rangle_{G^{(1)}}$$

$$\mathrm{Cov}[\hat{H}^{(2)}, \hat{H}^{(2)}] = \frac{1}{n_1}\left(\delta\delta\, A^{(2)} + \delta\delta\, B^{(2)}\ B^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta\delta\, B^{(2)}\ B^{(2)}_{\alpha_1\alpha_4\alpha_2\alpha_3}\right)$$

$$B^{(2)} = (C_W^{(2)})^2 \langle H^{(1)} \sigma'\sigma'\ H^{(1)} \sigma'\sigma' \rangle_{G^{(1)}}$$

$$B^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = (C_W^{(2)})^2 H^{(1)}_{\alpha_1\alpha_2} H^{(1)}_{\alpha_3\alpha_4} \langle \sigma'_{\alpha_1} \sigma'_{\alpha_2} \sigma'_{\alpha_3} \sigma'_{\alpha_4} \rangle_{G^{(1)}}$$

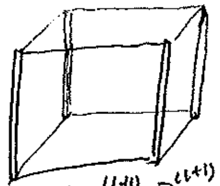
$$A^{(2)} = \mathrm{Cov}[\hat{\Omega}^{(2)}, \hat{\Omega}^{(2)}]_{G^{(1)}}$$

$$\hat{\Omega}^{(2)} = \lambda_W^{(2)} (\sigma^{(1)})(\sigma^{(1)}) + C_W^{(2)} H^{(1)} \sigma'^{(1)} \sigma'^{(1)}$$

$$\mathrm{Cov}[z^{(2)} z^{(2)}, \hat{H}^{(2)}] = \frac{1}{n_1} \delta\delta \left(\lambda_W^{(2)} C_W^{(2)}\, \mathrm{Cov}[\sigma\sigma, \sigma\sigma] + (C_W^{(2)})^2\, \mathrm{Cov}[\sigma\sigma, H^{(1)}\sigma'\sigma']\right)$$
$$+ \frac{1}{n_1} (\delta\delta + \delta\delta) (C_W^{(2)})^2 \langle \sigma\sigma\ \sigma'\sigma' \rangle H^{(1)}$$

$$= \frac{1}{n_1}\left(\delta\delta\, D^{(2)} + \delta\delta\, F_{\alpha_1\alpha_3\alpha_2\alpha_4} + \delta\delta\, F^{(2)}_{\alpha_1\alpha_4\alpha_2\alpha_3}\right)$$

$$D^{(2)} = C_W^{(2)}\, \mathrm{Cov}[\sigma\sigma, \hat{\Omega}] \qquad F^{(2)}_{\alpha_1\alpha_3\alpha_2\alpha_4} = (C_W^{(2)})^2 H^{(1)} \langle \sigma\sigma\ \sigma'\sigma' \rangle$$



$z^{(l)}$ $\hat{H}^{(l)}$ $W^{(l+1)}$ $z^{(l+1)}$

$\mathcal{Q}$ $\mathcal{P}$ $\mathcal{O}$

$$\left\langle \mathcal{Q}(z^l, \hat{H}^l)\cdot P(W^{l+1})\cdot \mathcal{O}(z^{l+1})\right\rangle = e^{\frac{C_W^{(l+1)}}{2n_l}\|J\|_2^2}\left\langle \mathcal{Q}(z^l,\hat{H}^l)\cdot \left\langle\!\left\langle \mathcal{O}\left(z^{l+1}_{i;\alpha} + C_W^{l+1}\frac{1}{n_l}\sum_j J_{ij}\,\sigma^{(l)}_{j;\alpha}\right)\right\rangle\!\right\rangle_{\hat{G}^{(l+1)}}\right\rangle$$

$$P(W^l) = e^{J_{ij} W_{ij}^{(l+1)}}$$

$$\langle \mathcal{Q}\,\mathcal{O}\rangle = \left\langle \mathcal{Q}(z^{(l)},\hat{H}^{(l)})\cdot \langle\!\langle \mathcal{O}(z^{(l+1)})\rangle\!\rangle_{\hat{G}^{(l+1)}}\right\rangle$$

$$\langle \mathcal{Q}\, W_{ij}^{(l+1)}\mathcal{O}\rangle = \sum_\alpha \left\langle \mathcal{Q}\,\langle\!\langle \partial_{z^{(l+1)}_{i\alpha}}\mathcal{O}\rangle\!\rangle_{\hat{G}^{(l+1)}}\,\sigma^{(l)}_{j\alpha}\right\rangle\cdot\frac{C_W^{(l+1)}}{n_l}$$

$$\mathrm{Cov}[\mathcal{Q},\mathcal{O}] = \left\langle \mathcal{Q}\left(\langle\!\langle \mathcal{O}\rangle\!\rangle_{\hat{G}} - \langle \mathcal{O}\rangle\right)\right\rangle$$

$$\langle \mathcal{Q}\, W_{i_3j_3} W_{i_4j_4}\mathcal{O}\rangle = \langle \mathcal{Q}\,\boxed{W}\,\boxed{W}\,\mathcal{O}\rangle$$

$$= \frac{C_W^{(l+1)}}{n_l}\,\left\langle \mathcal{Q}\,\langle\!\langle \mathcal{O}\rangle\!\rangle_{\hat{G}^{(l+1)}}\right\rangle + \left(\frac{C_W^{(l+1)}}{n_l}\right)^2\left\langle \mathcal{Q}\,\left\langle\!\left\langle \left(z^{(l+1)}\sigma^{(l)}\, z^{(l+1)}\sigma^{(l)} - \sigma\sigma\right)\mathcal{O}\right\rangle\!\right\rangle_{\hat{G}^{(l+1)}}\right\rangle$$

$$\mathrm{Cov}[\mathcal{Q}, \boxed{W}\boxed{W}\mathcal{O}] = \frac{C_W^{(l+1)}}{n_l}\,\mathrm{Cov}[\mathcal{Q},\langle\!\langle \mathcal{O}\rangle\!\rangle] + \left(\frac{C_W}{n_l}\right)^2\mathrm{Cov}\left[\mathcal{Q}, \left\langle\!\left\langle \left(z^{(l+1)}\sigma'\, z^{(l+1)}\sigma' - \sigma'\sigma'\right)\mathcal{O}\right\rangle\!\right\rangle_{\hat{G}^{(l+1)}}\right]$$

$$\frac{1}{n_l}\boxed{D^{(l+1)}} = \frac{\lambda_W^{(l+1)}}{C_W^{(l+1)}}\langle \widehat{\Delta G}^{(l+1)}\,\widehat{\Delta G}^{(l+1)}\rangle + C_W^{(l+1)}\left\langle \boxed{\widehat{\Delta G}^{(l+1)}}\ \sigma'\,\sigma'\ \boxed{\hat{H}^{(l)}}\right\rangle$$

$$\boxed{\hat{G}^{(l+1)}} = C_b^{(l+1)} + \frac{C_W^{(l+1)}}{n_l}\sigma^{(l)}\sigma^{(l)}$$

$$\frac{1}{n_l}\boxed{F^{(l+1)}} = \cancel{C_W^2\langle \sigma'\sigma'\sigma'\sigma'\,\hat{H}\rangle}\ \left(C_W^{(l+1)}\right)^2\left\langle \sigma'^{(l)}\,\sigma'^{(l)}\,\sigma'^{(l)}\,\sigma'^{(l)}\ \boxed{\hat{H}^{(l)}}\right\rangle$$

$$\mathrm{Cov}[\mathcal{O}(z^{(l)}), \hat{H}^{(l)}] = \frac{1}{2n_{l-1}}\ \boxed{D^{(l)}}\ \langle\!\langle (z\,z - \,)\,\mathcal{O}(z^{(l)})\rangle\!\rangle_{G^{(l)}} + \frac{1}{n_{l-1}}\ \boxed{F^{(l)}}\ \langle\!\langle (z\,z - \,)\,\mathcal{O}(z^{(l)})\rangle\!\rangle$$

$$\boxed{\hat{\Omega}^{(l+1)}} = \lambda_W^{(l+1)}\sigma^{(l)}\sigma^{(l)} + C_W^{(l+1)}\boxed{H^{(l)}}\,\sigma'^{(l)}\,\sigma'^{(l)}$$

$$\mathrm{Cov}[\mathcal{F}(z^{(l)}_{i_1A_1}), \mathcal{G}(z^{(l)}_{i_2A_2})] \quad (i_1\neq i_2)$$

$$= \frac{1}{4n_{l-1}}\left\langle (z\,z - \,)\mathcal{F}(z_{A_1})\right\rangle_{G^{(l)}}\left\langle (z\,z - \,)\mathcal{G}(z_{A_2})\right\rangle_{G^{(l)}}\ \boxed{V^{(l)}} + O\!\left(\frac{1}{n^2}\right)$$

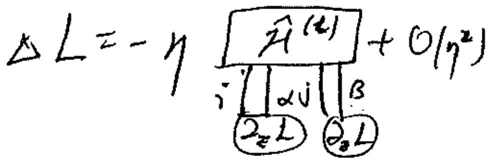
Equivalence principle: All ~~parameters~~ parameter ~~types~~ types should contribute equally to learning.

$$H^{(\ell)}_{ij\alpha\beta} = \sum_{\ell' \le \ell} \sum_{\mu\nu} \frac{\partial z^{(\ell)}_{i\alpha}}{\partial \theta^{(\ell')}_{\mu}} \frac{\partial z^{(\ell)}_{j\beta}}{\partial \theta^{(\ell')}_{\nu}} \lambda^{(\ell')}_{\mu\nu}$$

~~Each of $\ell' \le \ell$,~~ $\sum_{kk'} \frac{\partial z_{i\alpha}}{\partial b^{(\ell')}_{k}} \frac{\partial z_{j\beta}}{\partial b^{(\ell')}_{k'}} \lambda_{b,k,k'}$

$$\sum \frac{\partial z^{(\ell)}_{i\alpha}}{\partial W^{(\ell')}_{ks}} \frac{\partial z^{(\ell)}_{j\beta}}{\partial W^{(\ell')}_{k's'}} \lambda^{(\ell')}_{W,ksk's'}$$

should have the same order of magnitude. for each of $\ell' = 1, 2, \dots, \ell$

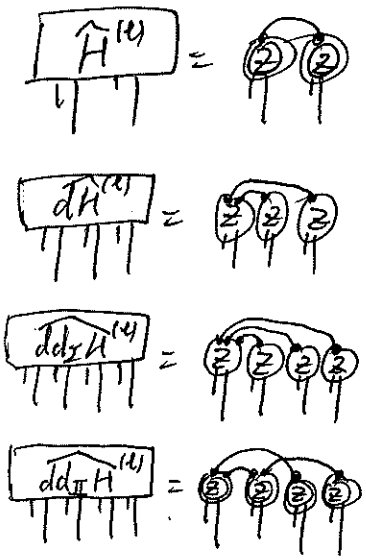


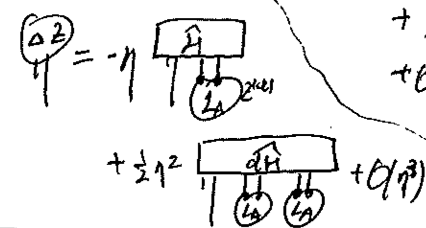
so trainability means $\Delta L \sim O(1)$, or, $\hat{H}^{(L)} \sim O(1)$.

NTK learning.

~~$\Delta\theta(z^{(\ell)})$~~ $\Delta z^{(\ell)}_{i\delta} = -\eta \sum_{j\alpha} \hat{H}^{(\ell)}_{ij\delta\alpha} \partial_{z^{(\ell)}_{j\alpha}} L_A$

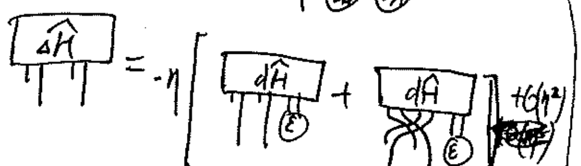
$+ \frac{1}{2}\eta^2 \sum_{j_1 j_2 \alpha_1 \alpha_2} dH_{i j_1 j_2 \delta \alpha_1 \alpha_2} \partial_{z^{(\ell)}_{j_1\alpha_1}} L_A \, \partial_{z^{(\ell)}_{j_2\alpha_2}} L_A$
$+ O(\eta^3)$

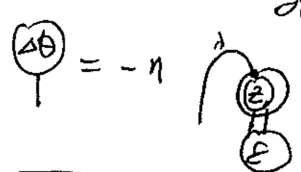


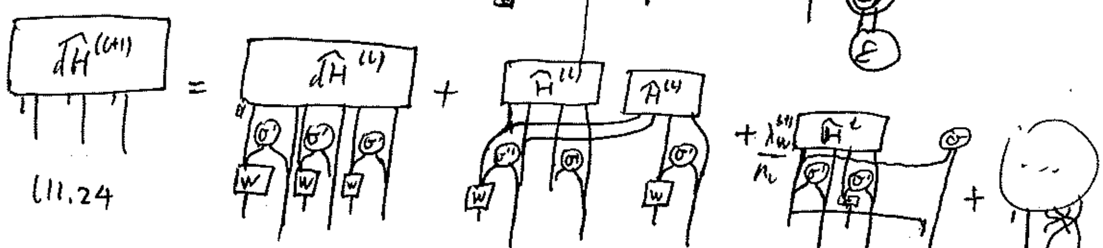


$$\Delta\theta_\mu = -\eta \sum_{\nu j\alpha} \lambda_{\mu\nu} \frac{\partial L_A}{\partial z^{(\ell)}_{j\alpha}} \frac{\partial z^{(\ell)}_{j\alpha}}{\partial \theta_\nu}$$

$$= -\eta \sum_{j\alpha\nu} \lambda_{\mu\nu} \frac{\partial z^{(\ell)}_{j\alpha}}{\partial \theta_\nu} \varepsilon^{(\ell)}_{j\alpha}$$

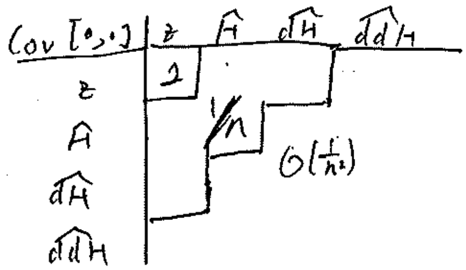


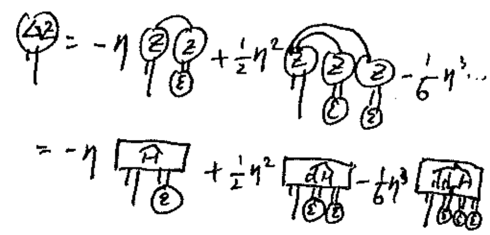




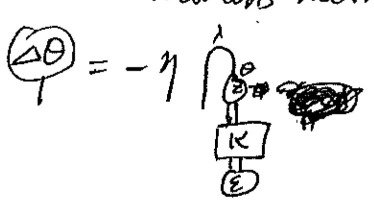
(11.24

Cov[·,·]	z	$\hat{H}$	$\widehat{dH}$	$\widehat{dd_I H}$	$\widehat{dd_{II} H}$
z	1	0	1/n		1/n²
$\hat{H}$	1/n	1/n		1/n²	
$\widehat{dH}$	1/n		1/n²		
$\widehat{dd_I H}$	1/n²	1/n²		1/n²	
$\widehat{dd_{II} H}$	1/n²				1/n²





generalized Newton's method



$$\Delta\theta_\mu = -\eta \sum_{\nu\alpha_1\alpha_2 ij} \lambda_{\mu\nu} \partial_{\theta_\nu} z_{i\alpha_1} \varepsilon^{j\alpha_2} K^{i}{}_{j}{}^{\alpha_1}{}_{\alpha_2}$$

ReLU (universality NTK) $A_2 = \frac{a_+^2 + a_-^2}{2}$, $A_4 = \frac{a_+^4 + a_-^4}{2}$

$$\begin{cases} \langle \sigma^4 \rangle_{K^{(\ell)}} = 3A_4 K^{(\ell)2} \\ \langle \sigma\sigma\sigma'\sigma' \rangle = A_4 K^{(\ell)} \\ \langle \sigma'^4 \rangle = A_4 \end{cases}$$

$$\begin{cases} \langle \sigma^2 \rangle = A_2 K^{(\ell)} \\ \langle \sigma'^2 \rangle = A_2 \\ g = A_2 K^{(\ell)} \\ \cancel{h = 0} \\ \chi_{\parallel} = C_W A_2 \\ \chi_{\perp} = C_W A_2 \\ h = 0 \end{cases}$$

point-correlations at criticality

$$\begin{cases} \chi_{\parallel} = 1 \\ \chi_{\perp} = 1 \\ K^{(\ell)} \text{ not explode} \end{cases} \Rightarrow \begin{cases} C_b = 0 \\ C_W A_2 = 1 \\ K^{(\ell)} = K^{(0)} = \frac{1}{A_2}\left(\frac{\sum x_i^2}{n_0}\right) = \frac{1}{A_2}\|\overline{x}\|_2^2 \\ V^{(\ell)} = (\ell-1)\left(\frac{3A_4}{A_2^2} - 1\right)(K^*)^2 \\ g = K^* A_2 \\ h = 0 \end{cases}$$

NTK at criticality

$$\begin{cases} \Theta^{(\ell)} = \sum_{s=1}^{\ell} (\lambda_b^{(s)} + \lambda_W^{(s)} K^* A_2) = \ell(\lambda_b + \lambda_W K^* A_2) \\ D^{(\ell)} = \frac{\ell(\ell-1)}{2} \times \left[\lambda_b\left(\frac{A_4}{A_2^2} - 1\right) + \lambda_W A_2 \left(\frac{4A_4}{A_2^2} - \cancel{2}\right) K^*\right] K^* \\ F^{(\ell)} = \frac{\ell(\ell-1)}{2} \times \frac{A_4}{A_2^2} (\lambda_b + \lambda_W A_2 K^*) K^* \\ B^{(\ell)} = \frac{\ell(\ell+1)(2\ell+1)}{6} \times \frac{A_4}{A_2^2} (\lambda_b + \lambda_W A_2 K^*)^2 \\ A^{(\ell)} = \frac{1}{3}\ell^3 \times \left(\left(\frac{A_4}{A_2^2} - 1\right)\lambda_b^2 + 3\left(A_4/A_2^2 - 1\right)\lambda_b \lambda_W A_2 K^* + (5A_4 - 3A_2^2)\lambda_W^2 (K^{*2}) \right) + O(\ell) \end{cases}$$

scaling laws.

$$\frac{\frac{1}{n} A^{(\ell)}}{\Theta^{(\ell)2}} \sim \frac{\ell}{n}, \quad \frac{\frac{1}{n} B^{(\ell)}}{\Theta^{(\ell)2}} \sim \frac{\ell}{n},$$

$$\frac{\frac{1}{n} D^{(\ell)}}{K^{(\ell)}\Theta^{(\ell)}} \sim \frac{\ell}{n}, \quad \frac{\frac{1}{n} F^{(\ell)}}{K^{(\ell)}\Theta^{(\ell)}} \sim \frac{\ell}{n}$$

$K^* = 0$ universality (tanh)

$C_b = 0$, $C_W = 1/\sigma_1^2$

$\sigma(z) = \sum \frac{\sigma_n}{n!} z^n$, $\sigma_0 = 0$.

$$\begin{cases} \langle \sigma^4 \rangle = 3\sigma_1^4 K^2 + O(K^3) \\ \langle \sigma\sigma\sigma'\sigma' \rangle = \sigma_1^4 K + O(K^2) \\ \langle \sigma'^4 \rangle = \sigma_1^4 + O(K) \end{cases}$$

$$K^{(\ell)} = \frac{1}{-a_1}\frac{1}{\ell} + O\left(\frac{\ln \ell}{\ell^2}\right)$$

$$V^{(\ell)} = \frac{2}{3a_1^2}\frac{1}{\ell} + O\left(\frac{\ln \ell}{\ell^2}\right)$$

$$\frac{V^{(\ell)}}{n(K^{(\ell)})^2} = \frac{2\ell}{3n} + O\left(\frac{\ln \ell}{n}\right)$$

$$\prod_{s=1}^{\ell} \chi_{\perp}^{(s)} \propto \ell^{-\frac{b_1}{a_1}}$$

$b_1/a_1 := p_\perp$

$$\Theta^{(\ell)} = \sum_{s=1}^{\ell} \left(\lambda_b^{(s)} + \lambda_W^{(s)} \frac{\sigma_1^2}{-a_1}\frac{1}{s}\right)\left(\frac{s}{\ell}\right)^{p_\perp} + \cdots$$

$$= \frac{\sigma_3 + \sigma_2^2}{\sigma_3 + \frac{3}{4}\sigma_2^2}$$

$\cancel{\chi_{\parallel} = 1 + 2a_1 K}$

$\chi_{\parallel} = 1 - 2/\ell + \cdots$

$\chi_{\perp} = 1 - p_\perp/\ell + \cdots$

$C_W g = \frac{1}{-a_1}\frac{1}{\ell} + \cdots$

$h = \frac{1}{2} b_1 + \cdots$

$C_W^2 \langle \sigma\sigma\sigma'\sigma' \rangle = \frac{1}{-a_1}\frac{1}{\ell} + \cdots$

$C_W^2 \langle \sigma'^4 \rangle = 1 + \cdots$

Depth-scaling of learning rates

$$\lambda_b^{(\ell)} = \ell^{-p_\perp}\tilde{\lambda}_b, \quad \lambda_W^{(\ell)} = \ell^{-p_\perp + 1}\tilde{\lambda}_W \Rightarrow \boxed{\Theta^{(\ell)} = \left(\tilde{\lambda}_b + \frac{\sigma_1^2}{-a_1}\tilde{\lambda}_W\right)\ell^{-p_\perp + 1}}$$

$$F^{(\ell)} = \frac{1}{5 - p_\perp}(\cdots)\left(\frac{1}{\ell}\right)^{p_\perp - 1}$$

$$D^{(\ell)} = \frac{-2}{9(-a_1)}(\cdots)\left(\frac{1}{\ell}\right)^{p_\perp + 1}$$

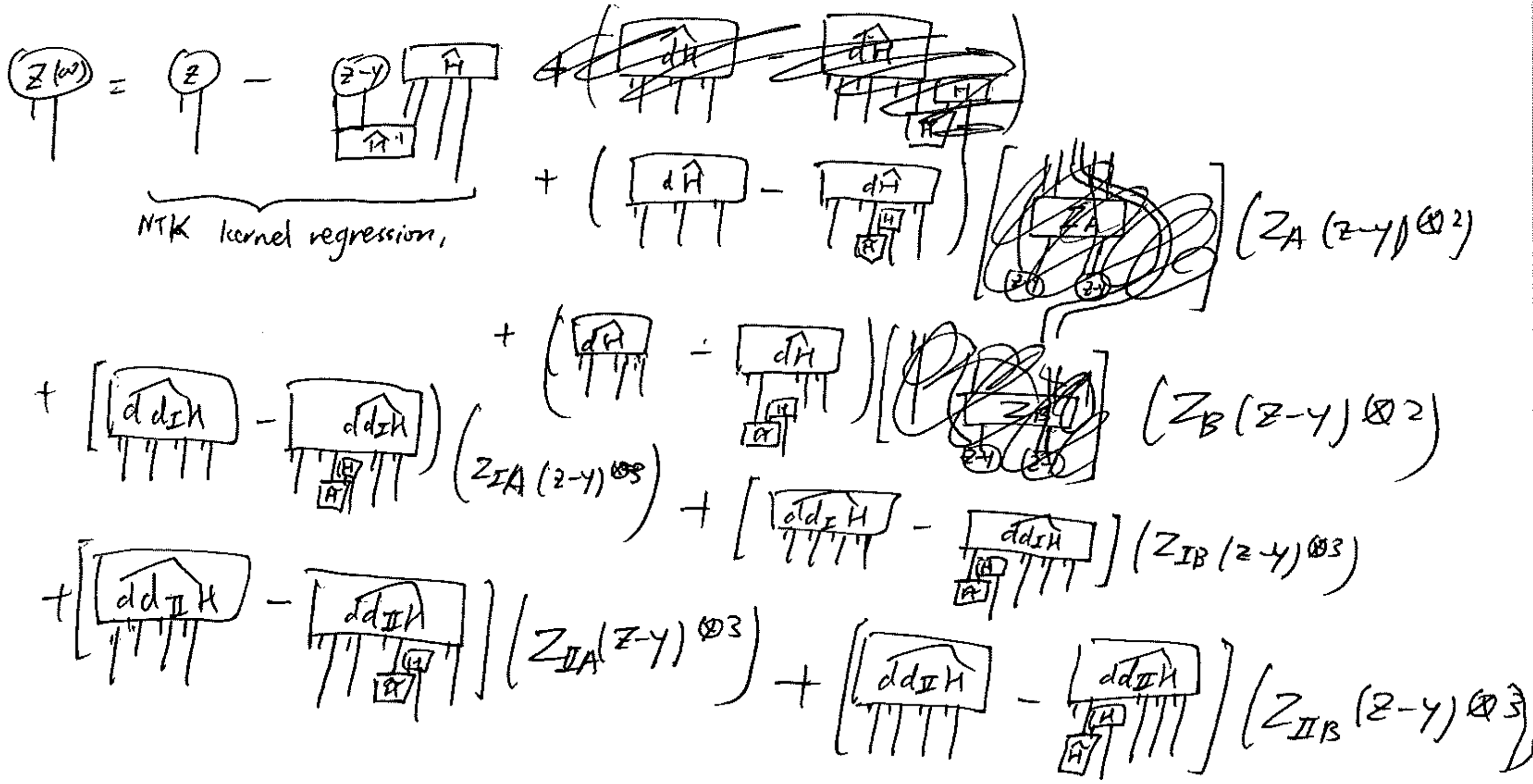
$$B^{(\ell)} = \frac{1}{3}(\cdots)^2\left(\frac{1}{\ell}\right)^{2p_\perp - 3}$$

$$A^{(\ell)} = (\cdots)\left(\frac{1}{\ell}\right)^{2p_\perp - 3}$$

~~(assume $p_\perp < 5$)~~ $(a_1 < 0)$ universality

(9.81) ~ (9.86)

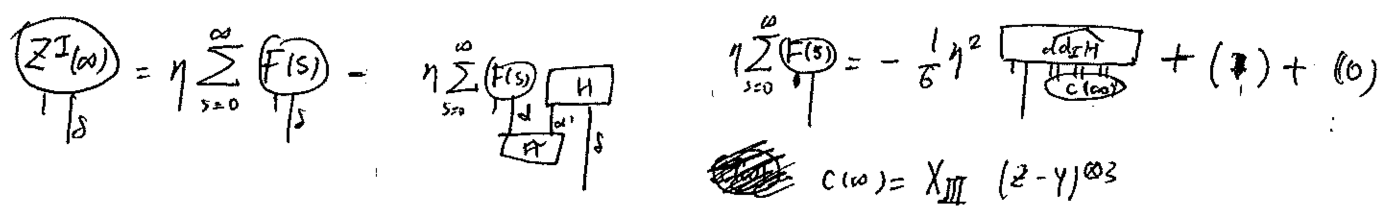
Scaling is still $\frac{\frac{1}{n}A^{(\ell)}}{(\Theta^{(\ell)})^2} \sim \frac{\ell}{n}$, ...



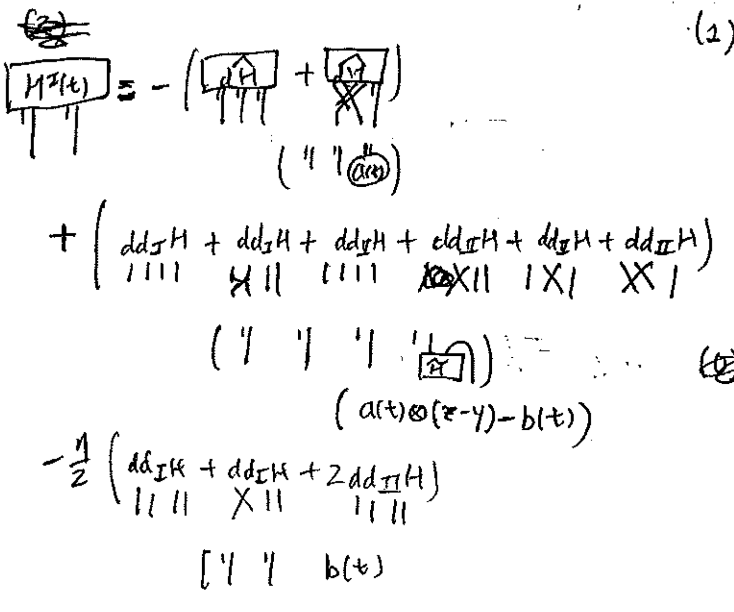
$= \left(\infty\text{ width NTK }\overset{\text{kernel}}{\text{regression}}\right) + \left(O\left(\tfrac{1}{n}\right)\ \overset{\text{correction to}}{\text{NTK}}\ \text{fluctuation}\ \overset{\text{kernel}}{\text{regression}}\right)$

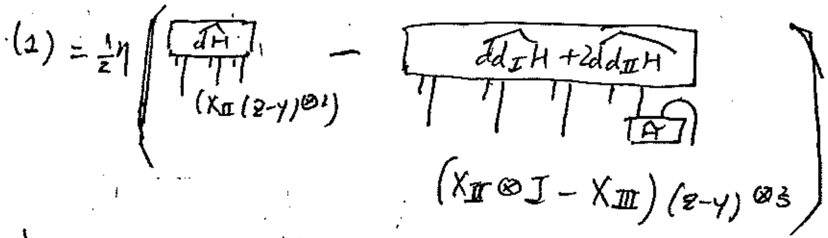
$+ [d\hat{H}]\ (Z_A, Z_B$ projections of ~~original~~ initial error of prediction)

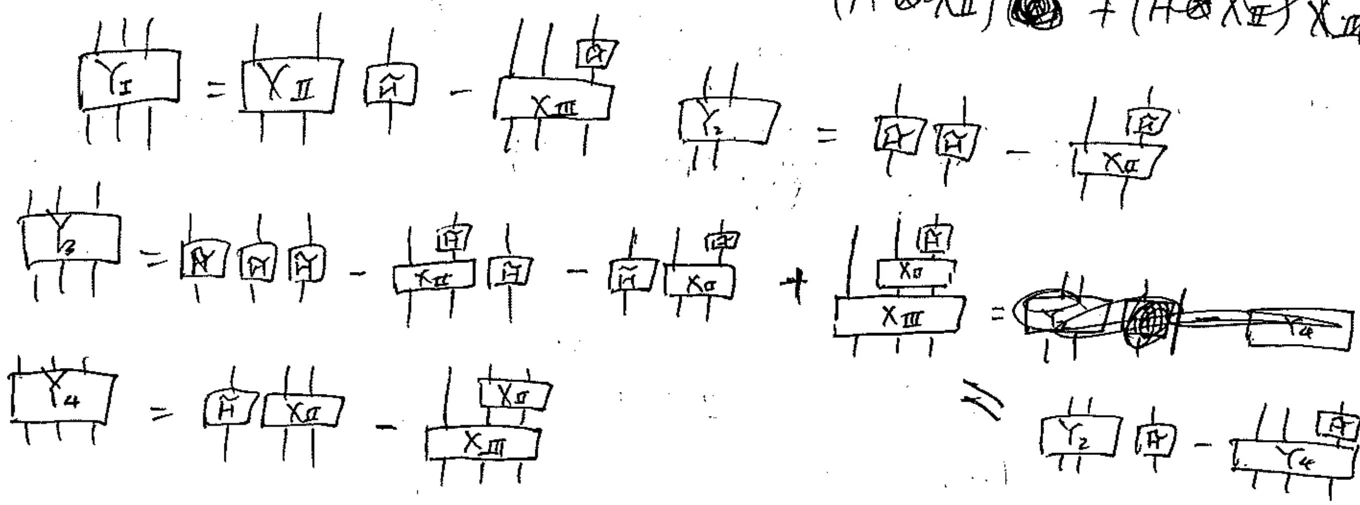
$+ [dd_I\hat{H}, dd_{II}\hat{H}]\ (Z_{IA}, Z_{IB}, Z_{IIA}, Z_{IIB}$ projections of initial errors of prediction)



$c(\infty) = X_{III} \; (z-y)^{\otimes 3}$







$Z_A = Y_2$

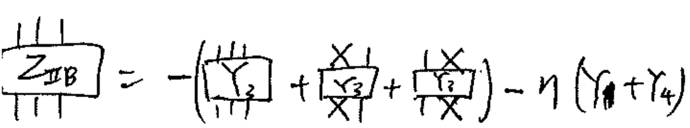
$Z_B = Y_2 + \frac{\eta}{2} X_{II}$

$Z_{IA} = -Y_3 - \frac{\eta}{2} Y_4$

$Z_{IB} = -Y_3 - \frac{\eta}{2}(Y_1 + Y_4) - \frac{\eta^2}{6} X_{III}$

$Z_{IIA} = -Y_3$

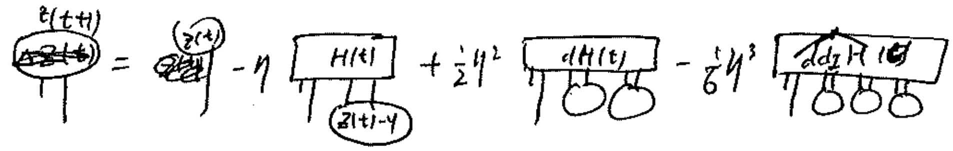
$\cancel{Z_{IIB} = -Y_3}$

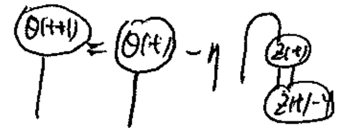


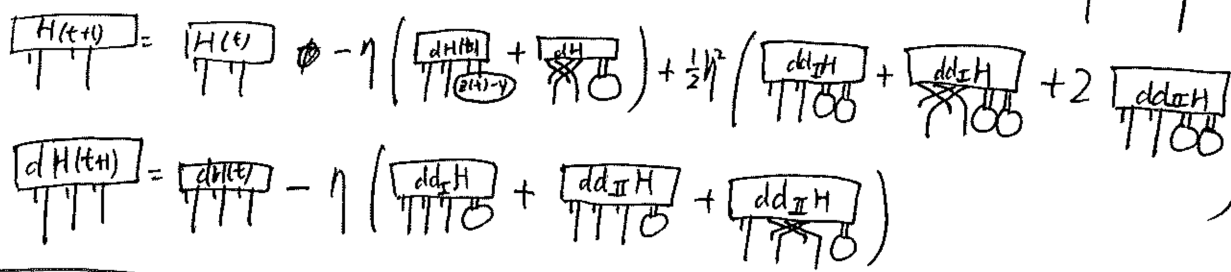
grad. desc.

const. η

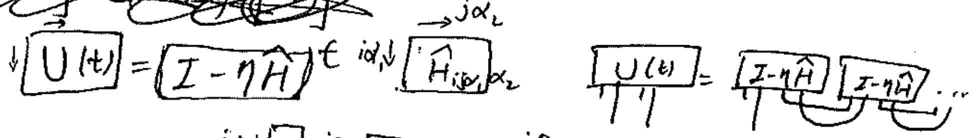
MSE.



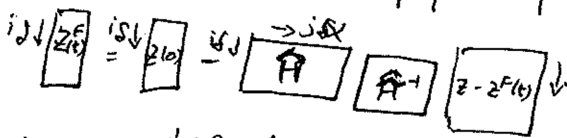




$$(z^F(t)-y)_{i\alpha} = \sum_{j\alpha_2} U^{(t)}_{ij\,\alpha,\alpha_2}(z(0)-y)_{j\alpha_2}$$



$$z^F(t)_{i\delta} = z(0)_{i\delta} - \sum_{j\alpha,k\alpha_2} \hat{H}_{i\delta j\alpha}(\hat{H}^{-1})_{j\alpha,k\alpha_2}(z-z^F(t))_{k\alpha_2}$$



$$\begin{cases} z = z^F(t) + z^I(t) \\ H(t) = \hat{H} + H^I(t) \end{cases}$$

free dynamic + interacting dynamic

~~dynamic free + dynamic interaction~~

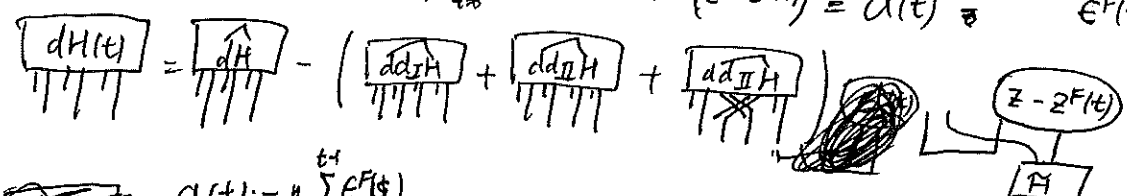
free dynamic is due to ~~exponential~~ $\hat{H}$

interacting dynamic is due to $\hat{dH}$, $\hat{dd_I H}$, $\hat{dd_{II} H}$

$$\eta \sum_{t=0}^{t-1}(z^F(t)-y) = \hat{H}^{-1}(z-z^F(t)) = a(t)$$

$$\epsilon^F(t) := z^F(t) - y = (I-\eta\hat{H})^t(z-y)$$

$$\epsilon^F(\infty) = 0$$



$$a(t) := \eta\sum_{s=0}^{t-1}\epsilon^F(s) = X_I\left((z-y)-\epsilon^F(t)\right)$$

$$X_I = \eta\left(I-(I-\eta\hat{H})\right)^{-1} = \hat{H}^{-1}$$

$$b(t) := \eta\sum_{s=0}^{t-1}\epsilon^F(s)\otimes\epsilon^F(s) = X_{II}\left((z-y)\otimes(z-y) - \epsilon^F(t)\otimes\epsilon^F(t)\right)$$

$$X_{II} = \eta\left(I\otimes I - (I-\eta\hat{H})\otimes(I-\eta\hat{H})\right)^{-1}$$

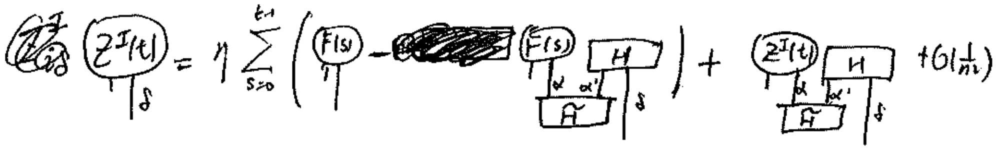
$$c(t) := \eta\sum_{s=0}^{t-1}\epsilon^F(s)^{\otimes 3} = X_{III}\left((z-y)^{\otimes 3} - \epsilon^F(t)^{\otimes 3}\right)$$

$$X_{III} = \eta\left(I^{\otimes 3} - (I-\eta\hat{H})^{\otimes 3}\right)^{-1}$$

$$\begin{cases} z^I(t+1) = (I-\eta\hat{H})z^I(t) + \eta F(t) \\ z^I(0) = 0 \end{cases}$$

$$z^I(t) = \eta\sum_{s=0}^{t-1}\left(F(s) - \hat{H}z^I(s)\right) = \eta\left(F(t-1) + UF(t-2) + \cdots + U^{t-1}F(0)\right)$$

$$\eta\left(z^I(t-1)+\cdots+z^I(0)\right) = \hat{H}^{-1}\left(\eta\left(F(t-1)+\cdots+F(0)\right) - z^I(t)\right)$$



$z^I(t) \sim e^{-(t-1)\eta\hat{H}}$

for training set.

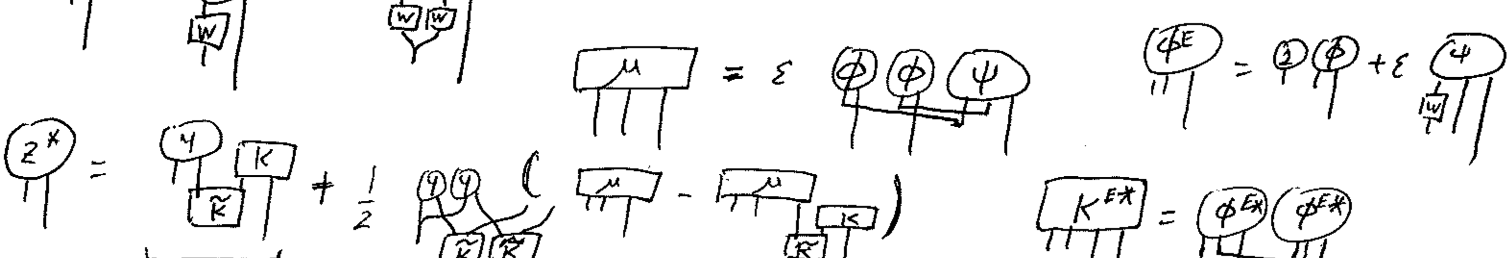
$$a(\infty) = X_I(z-y)$$

$$b(\infty) = X_{II}(z-y)^{\otimes 2}$$

$$c(\infty) = X_{III}(z-y)^{\otimes 3}$$

meta kernel, feature learning..

$\phi^E_{ij}(x_\delta) = \phi_j(x_\delta) + \varepsilon \sum_k W_{ik} \psi_{kj}(x_\delta)$



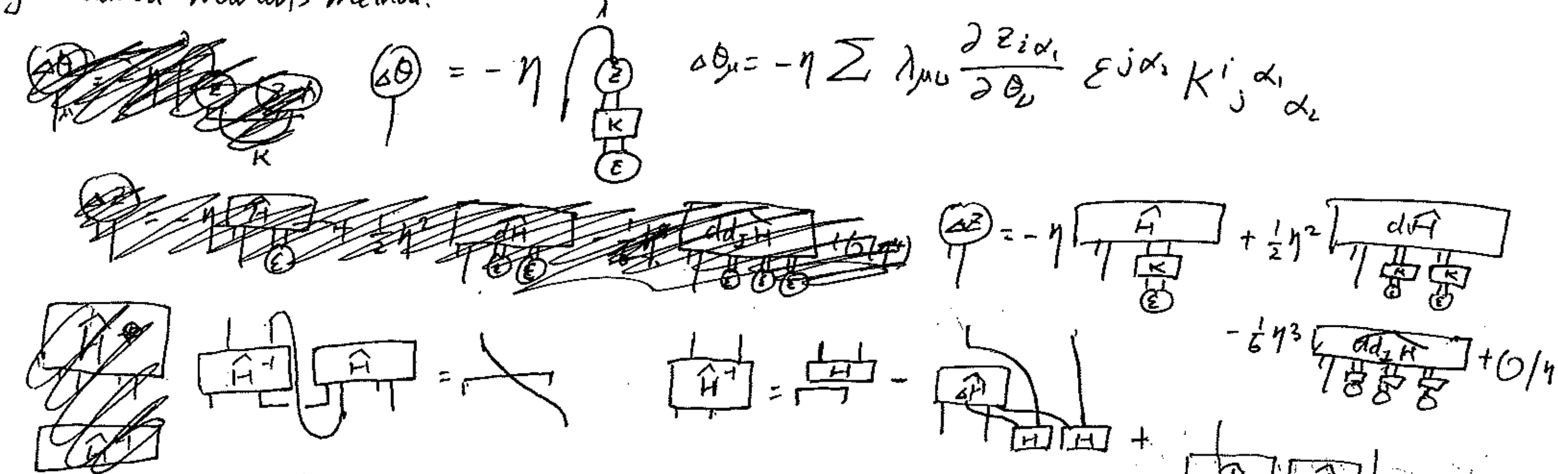
kernel learning

metakernel feature learning

$W^* = W^F + W^I$

generalized Newton's method.

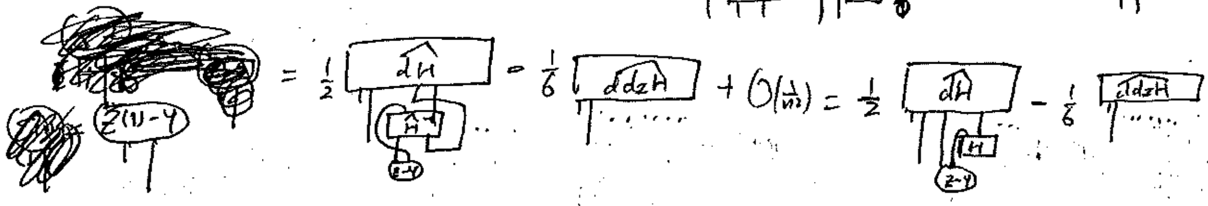
$\Delta\theta_\mu = -\eta \sum \lambda_{\mu\nu} \frac{\partial z_{i\alpha_1}}{\partial \theta_\nu} \varepsilon^{j\alpha_2} K^i{}_j{}^{\alpha_1}{}_{\alpha_2}$



step 1: $K = \frac{1}{\eta}\hat{H}^{-1}$. This removes the $-\eta \hat{H} K \varepsilon$ term.

step 2: order $O(\frac{1}{n})$ update to z, $K = \frac{1}{\eta}\delta H$

To order $O(\frac{1}{n})$, we have



I quit. see (∞. 79)