Lecture 7

- **Muon** .

Recall:

Optimizer recepie.

$$\underbrace{\operatorname*{argmin}_{\|\Delta W\| \leq \eta}}_{\text{Choose appropriate norm.}} \left\langle \underbrace{\nabla_W \mathcal{L}(W)}_{\text{Gradient of Loss}}, \underbrace{\Delta W}_{\text{change in } W.} \right\rangle.$$

Choose spectral norm.

$$\operatorname*{argmin}_{\|\Delta W\|_2 \leq \eta} \left\langle \nabla_W \mathcal{L}(W), \Delta W \right\rangle = \eta \cdot U_r V_r^T$$

$$\text{if } \nabla_W \mathcal{L}(W) = U \Sigma V^T = U_r \Sigma_r V_r^T$$

$\Rightarrow$ Semi-orthogonal update

$$W_{t+1} = W_t - \eta \cdot U_r V_r^T$$

[Key update of Shampoo (without accumulation)]

Shampoo:

$$L_t = L_{t-1} + G_t G_t^T \qquad \nabla_{W_t} \mathcal{L}(W_t) = G_t.$$

$$R_t = R_{t-1} + G_t^T G_t$$

$$W_{t+1} = W_t - \eta \, L_t^{-1/4} G_t R_t^{-1/4}.$$

Shampoo without accumulation:

$$\boxed{W_{t+1} = W_t - \eta \cdot (G_t G_t^T)^{-1/4} \, G_t \, (G_t^T G_t)^{-1/4}.}$$

$$G_t = U \Sigma V^T$$

$$G_t G_t^T = U \Sigma V^T V \Sigma^T U^T = U \Sigma \Sigma^T U^T$$

$$G_t^T G_t = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T$$

$$G_t = d_{out} \times \overset{d_{in}}{\boxed{\quad\quad\quad\quad}} \Longrightarrow \Sigma = \begin{bmatrix} \Sigma_r & 0 \end{bmatrix}$$

Assume full row rank.

$$\Longrightarrow \quad \Sigma\Sigma^T = \underset{d_{out}}{\begin{bmatrix} \Sigma_r & 0 \end{bmatrix}} \overset{d_{out}}{\begin{bmatrix} \Sigma_r^T \\ \\ \end{bmatrix}} = \Sigma_r \Sigma_r^T = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \end{bmatrix}$$

$$(G_t G_t^T)^{-\frac{1}{4}} = U \begin{bmatrix} \sigma_1^{-\frac{1}{2}} & \\ & \ddots \end{bmatrix} U^T$$

$$(G_t^T G_t)^{-\frac{1}{4}} = V \begin{bmatrix} \sigma_1^{-\frac{1}{2}} & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & \ddots \end{bmatrix} V^T$$

$$(G_t G_t^T)^{-\frac{1}{4}} G_t \, (G_t^T G_t)^{-\frac{1}{4}} = U \begin{bmatrix} \sigma_1^{-\frac{1}{2}} & \\ & \\ \end{bmatrix} U^T U \Sigma V^T V \begin{bmatrix} \sigma_1^{-\frac{1}{2}} & \\ & \ddots \\ & \end{bmatrix} V^T$$

$$= U \begin{bmatrix} \Sigma_r^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \end{bmatrix} \begin{bmatrix} \Sigma_r^{-\frac{1}{2}} \\ 0 \end{bmatrix} V^T = U_r V_r^T$$

Semi-orthogonal matrices:

$$\{ A \in \mathbb{R}^{m \times n} : AA^T = I_m \text{ or } A^TA = I_n \}$$

What if we had the RMS-RMS norm instead? $W \in \mathbb{R}^{d_{out} \times d_{in}}$

$$\underset{\|\Delta W\|_{RMS \to RMS} \leq \eta}{\operatorname{argmin}} \langle \nabla_W \mathcal{L}(W), \Delta W \rangle$$

$$= \underset{\|\Delta W\|_2 \leq \sqrt{\frac{d_{out}}{d_{in}}} \cdot \eta}{\operatorname{argmin}} \langle \nabla_W \mathcal{L}(W), \Delta W \rangle$$

Muon key idea 1.

$$\Rightarrow \quad \Delta W^* = -\eta \cdot \sqrt{\frac{d_{out}}{d_{in}}} \; U_r V_r^T$$

Why might semi-orthogonal matrices be a good idea?

→ Condition number $U_r V_r^T = 1$. Uniform step in all directions.

→ Not dominated by largest singular values.

But computing UV^T is expensive.

$\rightarrow$ Needs SVD.

Muon solves this issue through two observations

(1) Getting the direction approximately correct is good enough.

(2) Newton-Schulz iterations. $\leftarrow$ Muon key idea 2.

$$\underline{U\Sigma V^T} \longrightarrow \underline{UV^T}.$$

i.e. replace all the singular values by 1.

How?

$$f(U\Sigma V^T) \approx UV^T$$

Newton-Schulz.

① Odd polynomials commute with the SVD.

$$p(X) = a_0 X + a_1 X X^T X + a_2 (X X^T)^2 X + \cdots + a_n (X X^T)^n \cdot X.$$

e.g. $p(X) = \frac{3}{2} \cdot X - \frac{1}{2} X X^T X.$

$$p(U \Sigma V^T) = \frac{3}{2} U \Sigma V^T - \frac{1}{2} (U \Sigma V^T)(V \Sigma^T U^T) \cdot U \Sigma V^T$$

$$= \frac{3}{2} U \Sigma V^T - \frac{1}{2} U \Sigma \Sigma^T \Sigma V^T$$

$$= U \left[\frac{3}{2} \Sigma - \frac{1}{2} \Sigma \Sigma^T \Sigma \right] V^T$$

$$= U \, p(\Sigma) \cdot V^T$$

So $p(X) = U p(\Sigma) V^T.$

So can apply to a matrix without changing the singular vectors.

(2) Can we find $p(x)$ such that

$p(x) \to 1$ for $x > 0$

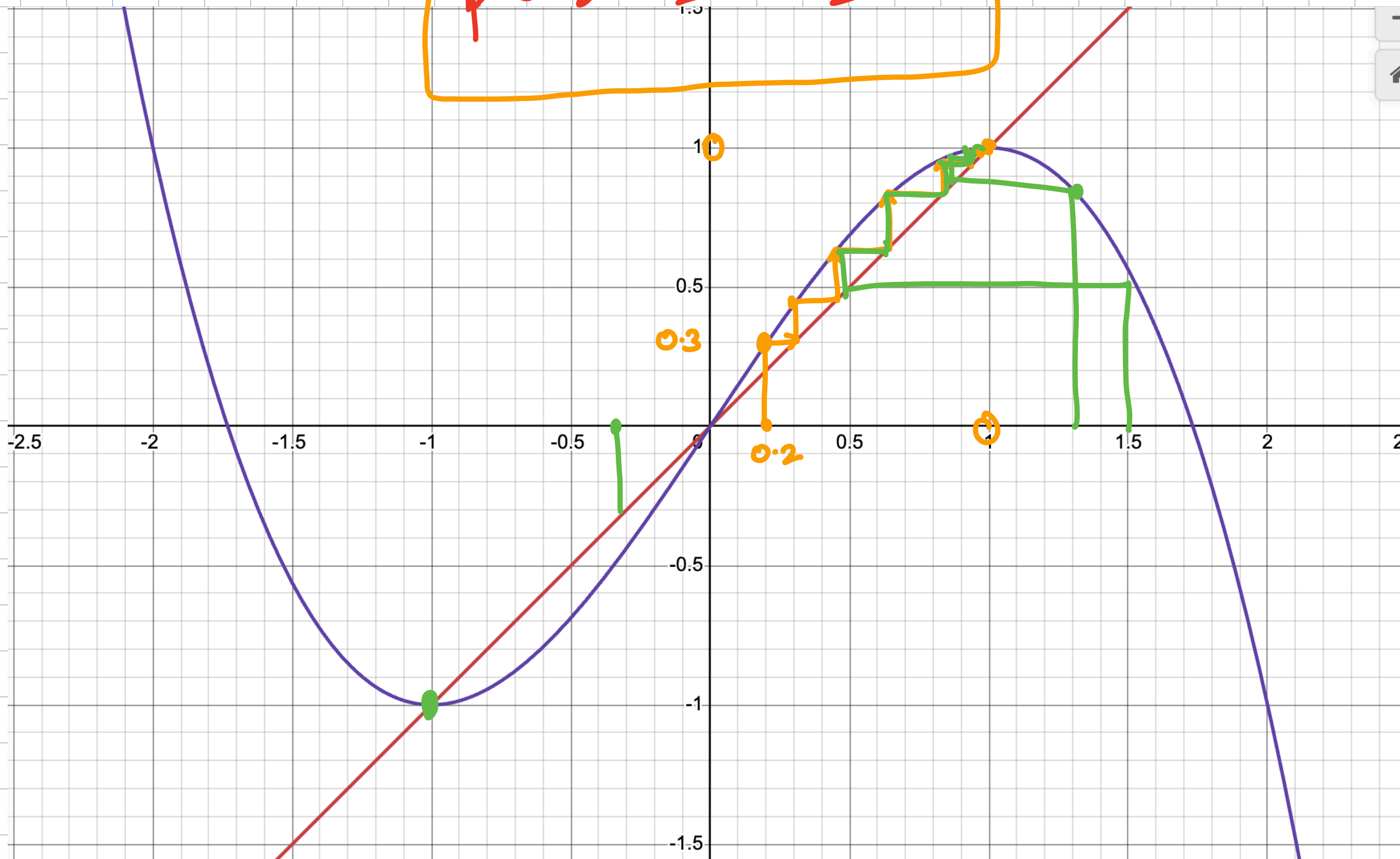
$p(x) \to \bigcup p(\Sigma) \forall T$

$p(p(x)) \to \bigcup p(p(\Sigma)) \forall T$

$\vdots$

Then, iteratively apply p.

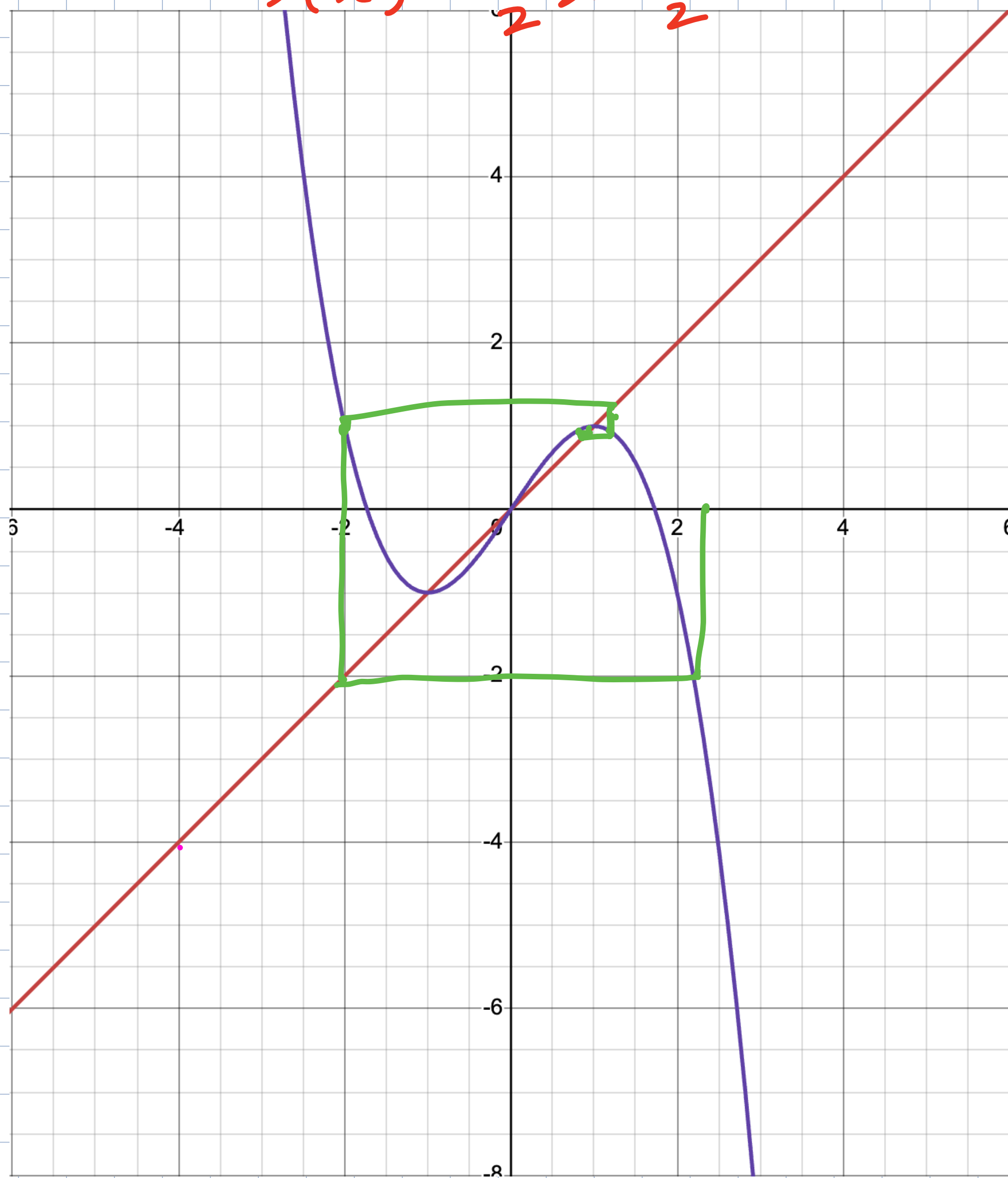
$$\boxed{p(x) = \frac{3}{2}x - \frac{1}{2}x^3}$$



$p(p(p(p(\sigma)))) \to p^n(\sigma) \to 1$

$n \to \infty$

$P(x) = \frac{3}{2}x - \frac{1}{2}x^2$ Zoomed out



→ Need to have singular values $\in (0, 1]$ before applying!

→ How? Normalize by the Frobenius norm : $\frac{G}{\|G\|_F}$.

Muon (Momentum Orthogonalized by Newton Schulz)

$$B_t = \mu \cdot B_{t-1} + \nabla_w \mathcal{L}(w)$$

$$O_t = \text{NewtonSchulz}(B_t)$$

$$W_t = W_{t-1} - \eta \cdot O_t \, .$$

How to choose $p(x)$?

$$p(x) = ax + bx^3 + cx^5 + \dots$$

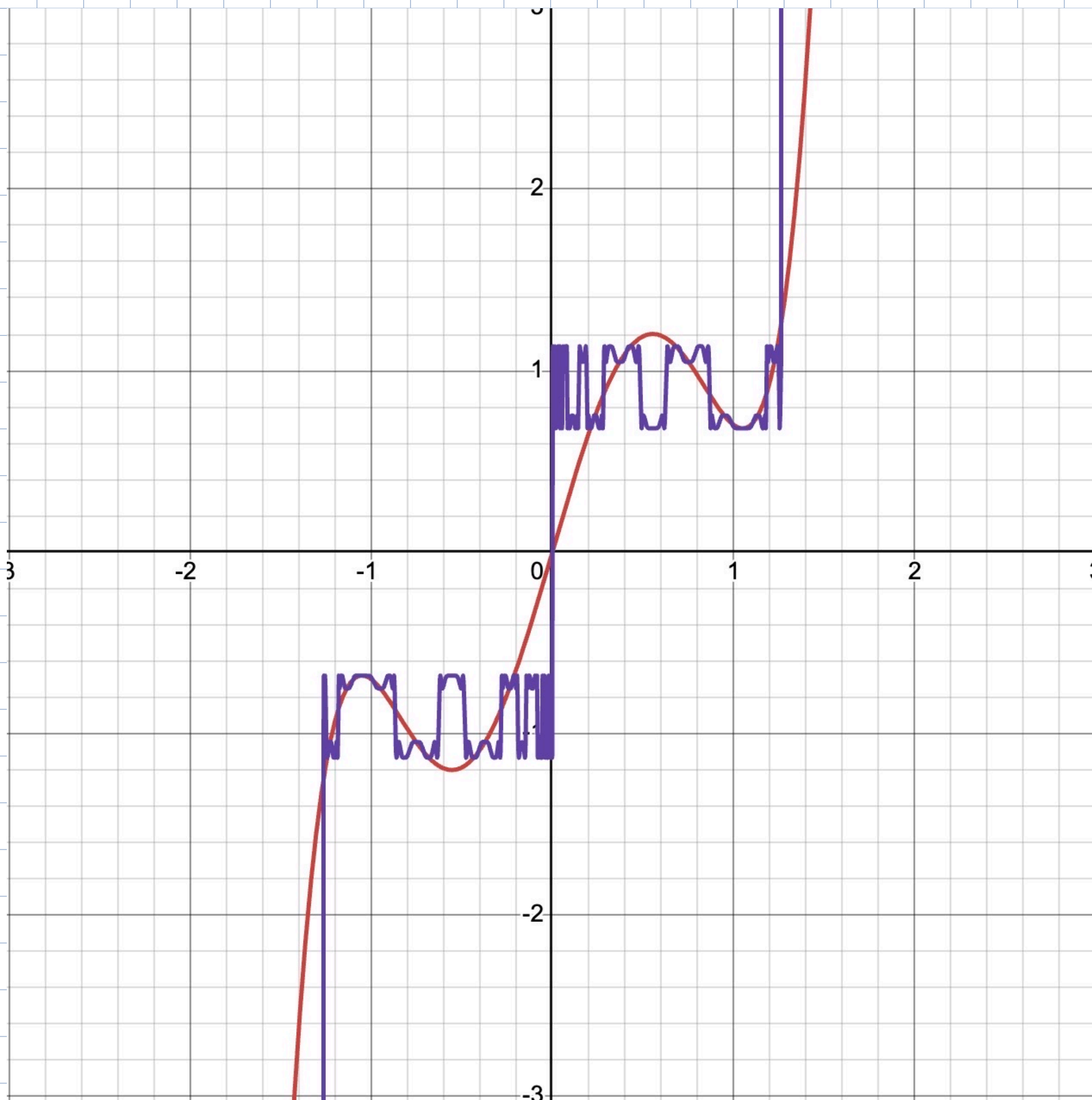
Can "tune" co-efficients for specific characteristics.

- Higher-order may converge faster, but each step is more expensive.

Do you have to converge?

NanoGPT speedrun : $f(x) = 3.4445x - 4.7750x^3 + 2.0315x^5$

$$f(1) \neq 1$$



2
1
-2
-1
0
1
2
-2
-3