EECS 182 Deep Neural Networks
Fall 2025 Anant Sahai and Gireeja Ranade

Discussion 4

1. Maximal Update Learning Rates During Training

Assume we are using a minibatch size of 1. For simplicity, consider a neural network layer with input $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$ that is sampled from an i.i.d. unit Gaussian, and weights $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$.

(a) First, compute the (stochastic) gradient $\nabla_W \mathcal{L}$ for $\mathbf{y} = W\mathbf{x} + \mathbf{b}$ and downstream loss function $\mathcal{L}$ where the loss depends on $\mathbf{y}$. Your answer should be in terms of $\mathbf{x}_i$ and $\mathbf{g}_i = \nabla_{\mathbf{y}} \mathcal{L}$.

Solution: We can consider individual elements of the gradient. We have that the (i, j)-th element of the gradient is:

$$\frac{\partial}{\partial W_{ij}} \mathcal{L}(\mathbf{y}) = g_i x_j$$

This means that the whole gradient is given by:

$$\nabla_W \mathcal{L}(\mathbf{y}) = \begin{bmatrix} g_1 x_1 & g_1 x_2 & \cdots & g_1 x_{d_{\text{in}}} \\ \vdots & \vdots & \ddots & \vdots \\ g_{d_{\text{out}}} x_1 & g_{d_{\text{out}}} x_2 & \cdots & g_{d_{\text{out}}} x_{d_{\text{in}}} \end{bmatrix} = \mathbf{g}\mathbf{x}^T$$

(b) In SignGD, we know that W is updated as below:

$$W_{t+1} \leftarrow W_t + \eta \operatorname{sign}\left(\nabla_W \mathcal{L}(\mathbf{y})\right).$$

What is the expected RMS norm squared of the change in features $\Delta \mathbf{y} = \eta\, \mathbf{sign}(\nabla_W \mathcal{L}(\mathbf{y}))\mathbf{x}_i$? How does this scale with d_{out} or d_{in}? What constant should we multiply the update by to ensure that the expected RMS norm squared of $\Delta \mathbf{y}$ does not depend on either d_{out} or d_{in}?

Solution: Let's consider the update to a single output feature. Note that Δy_j is given by

$$\Delta y_j = \eta \operatorname{sign}(g_j x)^T x = \eta \operatorname{sign}(g_j) \sum_{i=1}^{d_{\text{in}}} |x_i|$$

which is a sum of d_{in} terms with positive expected value. Thus, we have the expected RMS norm squared equals:

$$\mathbb{E}\left[||\Delta y||_{RMS}^2\right] = \frac{1}{d_{\text{out}}} \sum_{j=1}^{d_{\text{out}}} \eta^2 \mathbb{E}\left[\left(\sum_{i=1}^{d_{\text{in}}} |x_i|\right)^2\right] = c d_{\text{in}}^2$$

To remove the dependence on d_{in} we multiply the update by $\frac{1}{d_{\text{in}}}$.

2. Understanding Newton-Schulz

Let us consider a parameter matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Define the degree-3 odd polynomial p as:

$$p(W) = \frac{1}{2}\left(3I_{d_{\text{out}}} - WW^T\right)W\,.$$

In this problem, we will study how the iteration $W_{k+1} = p(W_k)$ affects the singular values of W_k.

(a) **Show that the iteration acts only on the singular values of W.** i.e. if $W = U\Sigma V^T$ is the SVD, then show that

$$p(W) = Up(\Sigma)V^T\,.$$

Hint: First show that $WW^T = U(\Sigma\Sigma^T)U^T$.

Solution: First, we can show that

$$WW^\top = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T.$$

Substituting into $P(W)$ yields

$$\begin{aligned} p(W) &= \left(\frac{3}{2}I - \frac{1}{2}WW^T\right)W \\ &= \left(U\left(\frac{3}{2}I - \frac{1}{2}\Sigma^2\right)U^T\right)(U\Sigma V^T) \\ &= U\left(\frac{3}{2}I - \frac{1}{2}\Sigma^2\right)\Sigma V^T. \end{aligned}$$

Since Σ is diagonal, the product $\left(\frac{3}{2}I - \frac{1}{2}\Sigma^2\right)\Sigma$ is diagonal with entries

$$\left(\frac{3}{2} - \frac{1}{2}\sigma_i^2\right)\sigma_i = \frac{3}{2}\sigma_i - \frac{1}{2}\sigma_i^3 = p(\sigma_i).$$

Thus,

$$p(W) = U\,\mathrm{diag}\big(p(\sigma_1), \ldots, p(\sigma_r)\big)\,V^T.$$

(b) Write down the fixed point equation for $p(x) = \frac{3}{2}x - \frac{1}{2}x^3$. **Solve for all fixed points**, i.e. x^* such that $x^* = p(x^*)$.

Solution: Solving for

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

yields the fixed points $x = 0, 1, -1$.

(c) We define a fixed point x^* of $p(x)$ as *locally stable* if $|\frac{d}{dx}p(x^*)| < 1$. First, convince yourself that a stable fixed point means that the distance towards the fixed point decreases with more iterations. **Determine which fixed points of $p(x)$ are stable and which are unstable.**

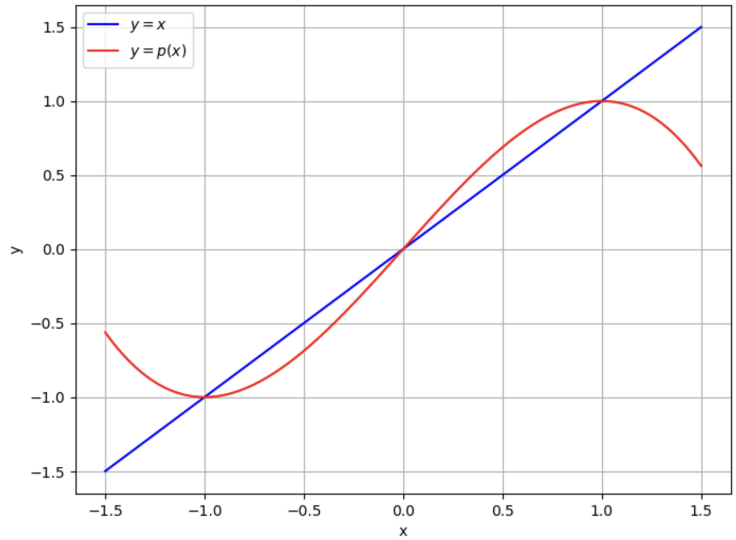
Solution: We have

$$\frac{d}{dx}p(x) = \frac{3}{2} - \frac{3}{2}x^2.$$

which we an use to evaluate

$$\frac{d}{dx}p(0) = 1.5 > 1 \quad \Rightarrow \quad \text{unstable}, \qquad \frac{d}{dx}p(\pm 1) = 0 \quad \Rightarrow \quad \text{stable}.$$

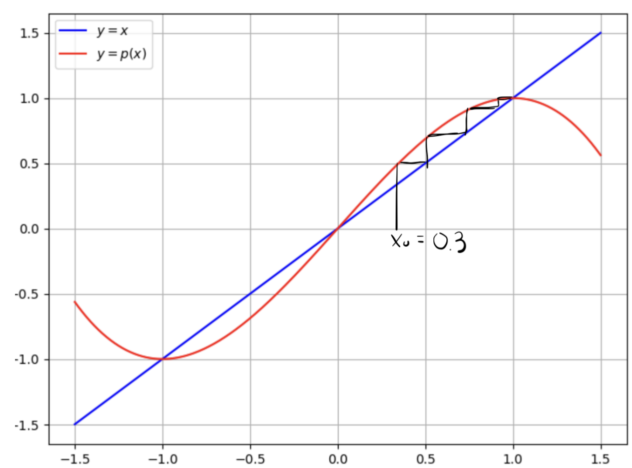
(d) Below are plots of $y = p(x)$ and $y = x$. Pick different starting points for x and **show graphically how iteration $x = p(x)$ eventually converges to a stable fixed point.**
Use "cobweb diagram" to show how x evolves over time.



Solution: A cobweb diagram is drawn as:

- Start at x_0 on the x-axis.
- Draw a vertical line up to the curve $y = p(x)$, ending at $(x_0, p(x_0))$.
- Draw a horizontal line from that point to the line, ending at $(p(x_0), p(x_0))$.
- Repeat: the above traces one iteration.

You should converge rapidly to ± 1.



(e) Suppose the singular value starts as $+\sigma$. **For which values of σ does it converge to $+1$? What does it do for other values?**

Solution: We notice that between $(0, \sqrt{3})$ the singular value σ will converge to 1. At $\sigma = \sqrt{3}$, we get $p(\sigma) = 0$ so σ goes to 0, then for $\sigma > \sqrt{3}$, we get $p(\sigma) < 0$, so σ will either converge to -1 or even diverge. Divergence occurs when $|p(\sigma)| > |\sigma|$, which we see occurs at $\sigma > \sqrt{5}$. At $\sigma = \sqrt{5}$, it will oscillate between $\pm\sqrt{5}$.

(f) **Explain why this iteration can be viewed as an approximate way to make W closer to an orthogonal matrix** (with singular values near ± 1) and what we must ensure before using the iterations.

Solution: Each iteration applies the transformation $p(\sigma_i) = \frac{3}{2}\sigma_i - \frac{1}{2}\sigma_i^3$ to each singular value σ_i for W. From the above analysis, we show that the singular values will eventually converge to ± 1, which is the goal of orthogonalization.

Contributors:

- Kevin Frans.
- Anant Sahai.
- Gireeja Ranade.
- Joey Hong.

