Can an Autoencoder Learn a Projection in One Step?

Due: Wednesday, November 18 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

A linear autoencoder reconstructs its inputs through a k-dimensional bottleneck. We'll compare how close Muon and gradient descent get to the best possible reconstruction in one update.

Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ contain one training input per column, with $\text{rank}(\mathbf{X}) = k$ and $1 \le k < n$. We reconstruct the inputs using $\mathbf{W} \in \mathbb{R}^{n\times n}$, where

$$\text{rank}(\mathbf{W}) \le k, \qquad \mathcal{L}(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\mathbf{X} - \mathbf{X}\|_F^2.$$

Here $\mathbf{W}$ represents the combined encoder and decoder. We apply updates directly to $\mathbf{W}$, starting at $\mathbf{W} = \mathbf{0}$. Use idealized Muon without momentum or polynomial approximation.

Write the compact SVD of the data as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.

1. Both optimizers start from the same loss. Show that its gradient at any $\mathbf{W}$ is
$$\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}) = (\mathbf{W} - \mathbf{I}_n)\mathbf{X}\mathbf{X}^\top.$$
2. At $\mathbf{W} = \mathbf{0}$, Muon uses the polar factor of the gradient. Show that one update with learning rate 1 gives $\mathbf{W}_{\text{Muon}} = \mathbf{U}\mathbf{U}^\top$. Why does this update attain the smallest possible reconstruction loss under the rank constraint?
3. Now take one ordinary gradient-descent step from $\mathbf{W} = \mathbf{0}$. Under what condition on the nonzero singular values of $\mathbf{X}$ does some learning rate $\eta > 0$ attain the same minimum loss as Muon?

Takeaway: *Idealized Muon learns the projection onto the data's subspace in one step, yielding a perfect k-dimensional autoencoder. Gradient descent's first step scales each direction according to its strength in the training data; a single learning rate generally cannot correct all those scalings.*