**Summary:** We guess a neural network’s gradients without computing a loss or knowing the label.

Backpropagation requires a lot of memory and is not biologically plausible. **Directional descent** is a previously proposed alternative:

1. Pick a random direction \( \epsilon \)
2. Find directional derivative along \( \epsilon \) using forward-mode automatic differentiation (cheap!).
3. Scale \( \epsilon \) by directional derivative:
   \[
   w_{t+1} = w_t - \alpha (\epsilon \cdot \nabla L)\epsilon
   \]

**Pros:**
- Unbiased estimator of \( \nabla L \)
- Guaranteed to be within 90° of true gradient.
- Doesn’t require storing activations like backprop

**Cons:**
- Cosine similarity decreases with guess dimensionality as \( O\left(\frac{1}{\sqrt{N}}\right) \)

**Main Question:** How can we narrow guess space?

**Answer:** Use local feature/architecture knowledge!

### Results – MLPs

**Directional Descent:** Random guess for weights

<table>
<thead>
<tr>
<th>Method</th>
<th>Cosine Similarity</th>
<th>1-step effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backprop (Oracle)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Directional Descent</td>
<td>0.0003</td>
<td>1 x 10⁶</td>
</tr>
<tr>
<td>Activation Perturbation</td>
<td>0.016</td>
<td>6.9 x 10⁴</td>
</tr>
<tr>
<td>Activation Mixing</td>
<td>0.025</td>
<td>3.4 x 10⁻³</td>
</tr>
<tr>
<td>( W^T )</td>
<td>0.030</td>
<td>1.7 x 10⁻³</td>
</tr>
</tbody>
</table>

### Results – LocalMixer

<table>
<thead>
<tr>
<th>Method</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backprop</td>
<td>66.4</td>
<td>69.3</td>
</tr>
<tr>
<td>Ren et al. (2022)</td>
<td>71.2</td>
<td>71.2</td>
</tr>
<tr>
<td>Mixing</td>
<td>68.8</td>
<td>72.5 (+1.3)</td>
</tr>
</tbody>
</table>

**Augmentation (500 epochs):**
- Backprop: 76.4
- Ren et al.: 72.2
- Mixing: 68.2
- Ours: 74.4 (+1.2)

**Augmentation (5000 epochs):**
- Backprop: 77.6
- Ren et al.: 76
- Mixing: 69.4
- Ours: 77.4 (+1.4)

### Gradient guess bias vs. optimization

**Matched Cosine Similarity**

**Interaction between momentum and bias**

Bias impedes convergence when momentum is used.

**How bias leads to better guesses over time**

Low-rank guess \( \rightarrow \) low-rank weight \( \rightarrow \) smaller guess space \( \rightarrow \) higher cosine similarity.