

## 0. Diagrams for SGD Test Loss

We use diagrams for book-keeping of the Taylor terms (of test loss at a fixed initial weight). Each color in a diagram represents a data value and thus selects a particular loss function from the data-parameterized distribution of loss functions. Each node in a diagram represents a tensor obtained from derivatives of the loss function corresponding to that node's color. It is a diagram's edges in a diagram that specify those derivative tensors. We understand the edges as directed from left to right, with the source acting on the target by differentiation. Finally, a diagram's value is the expectation over all iid assignments of data to the colors. Thus,

$$\begin{aligned}
 \text{red} \text{---} \text{green} &:= \mathbb{E}_{\text{red,green}} \nabla^a(l_{\text{red}}) \nabla^a(l_{\text{green}}) & \text{red} \text{---} \text{red} &:= \mathbb{E}_{\text{red}} \nabla^a(l_{\text{red}}) \nabla^a(l_{\text{red}}) \\
 \text{red} \text{---} \text{green} \text{---} \text{blue} &:= \mathbb{E}_{\text{red,green,blue}} \nabla^a(l_{\text{red}}) \nabla^a \nabla^b(l_{\text{green}}) \nabla^b(l_{\text{blue}}) \\
 \text{red} \text{---} \text{green} \text{---} \text{blue} &:= \mathbb{E}_{\text{red,green,blue}} \nabla^a(l_{\text{red}}) \nabla^b(l_{\text{green}}) \nabla^a \nabla^b(l_{\text{blue}})
 \end{aligned}$$

We see that  $\text{red} \text{---} \text{red} - \text{red} \text{---} \text{green}$  gives the trace of the covariance of gradients. Moreover,

$\text{red} \text{---} \text{green} \text{---} \text{blue} = \text{red} \text{---} \text{green} \text{---} \text{blue}$ , illustrating how diagram notation can streamline computation by helping to group terms. However, we caution that a diagram's value generally depends on a diagram's digraph structure, not just its undirected structure. For example:

$$\text{red} \text{---} \text{green} \text{---} \text{blue} = \text{red} \text{---} \text{green} \text{---} \text{blue} + \text{red} \text{---} \text{green} \text{---} \text{blue} \neq \text{red} \text{---} \text{red} \text{---} \text{green} \text{---} \text{blue} = \text{red} \text{---} \text{green} \text{---} \text{blue}$$

Thus prepared, we may expand the test loss after  $T$  updates, each with batch-size 1 sampled without replacement. The recipe is to draw all the diagrams whose underlying poset has a unique rightmost element. Each node in the diagram contributes a symmetry factor  $1/i!$  where  $i$  is the node's in-degree. On top of that, a diagram with  $a$  edges and  $v$  vertices has an overall combinatorial weight of  $(-\eta)^a \binom{T}{v-1}$ . We obtain:

$$\begin{aligned}
 \mathbb{E}(\text{SGD Test Loss}) &= \text{red} - \eta \binom{T}{1} \left( \text{red} \text{---} \text{green} \right) \\
 &+ \eta^2 \binom{T}{2} \left( \text{red} \text{---} \text{green} \text{---} \text{blue} + \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} \right) + \eta^2 \binom{T}{1} \left( \frac{1}{2} \text{red} \text{---} \text{green} \right) \\
 &- \eta^3 \binom{T}{3} \left( \text{red} \text{---} \text{green} \text{---} \text{blue} \text{---} \text{yellow} + \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} \text{---} \text{yellow} + \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} \text{---} \text{yellow} + \right. \\
 &\quad \left. \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} \text{---} \text{yellow} + \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} \text{---} \text{yellow} + \frac{1}{6} \text{red} \text{---} \text{green} \text{---} \text{blue} \text{---} \text{yellow} \right) \\
 &- \eta^3 \binom{T}{2} \left( \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} + \frac{1}{2} \text{red} \text{---} \text{green} \text{---} \text{blue} + \frac{1}{6} \text{red} \text{---} \text{green} \text{---} \text{blue} + \frac{1}{6} \text{red} \text{---} \text{green} \text{---} \text{blue} \right) \\
 &- \eta^3 \binom{T}{1} \left( \frac{1}{6} \text{red} \text{---} \text{green} \right) + o(\eta^3)
 \end{aligned}$$

And a routine grouping of terms yields:

$$\begin{aligned}
\cdots = & \bullet - \eta \binom{T}{1} \left( \bullet \text{---} \bullet \right) \\
& + \eta^2 \binom{T}{2} \left( \frac{3}{2} \bullet \text{---} \bullet \text{---} \bullet \right) + \eta^2 \binom{T}{1} \left( \frac{1}{2} \bullet \text{---} \bullet \right) \\
& - \eta^3 \binom{T}{3} \left( \frac{5}{2} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet + \frac{1}{2} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet + \frac{1}{6} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \right) \\
& - \eta^3 \binom{T}{2} \left( \bullet \text{---} \bullet \text{---} \bullet + \frac{5}{6} \bullet \text{---} \bullet \text{---} \bullet \right) \\
& - \eta^3 \binom{T}{1} \left( \frac{1}{6} \bullet \text{---} \bullet \right) + o(\eta^3)
\end{aligned}$$