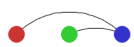
0. Diagrammatic Expansion of Losses

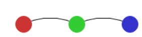
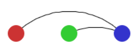
0.0. Use and Interpretation of Diagrams

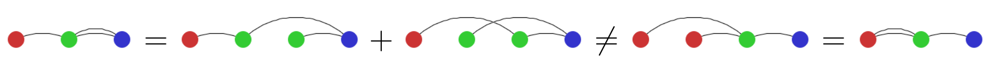
We use diagrams for book-keeping of the Taylor terms (of test loss at a fixed initial weight). Each color in a diagram represents a data value and thus selects a particular loss function from the data-parameterized distribution of loss functions. Each node in a diagram represents a tensor obtained from derivatives of the loss function corresponding to that node's color. It is a diagram's edges 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 assigments of data to the colors. Thus,

 $:= \mathbb{E}_{\text{red,green}} \nabla^a(l_{\text{red}}) \nabla^a(l_{\text{green}})$  $:= \mathbb{E}_{\text{red}} \nabla^a(l_{\text{red}}) \nabla^a(l_{\text{red}})$

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

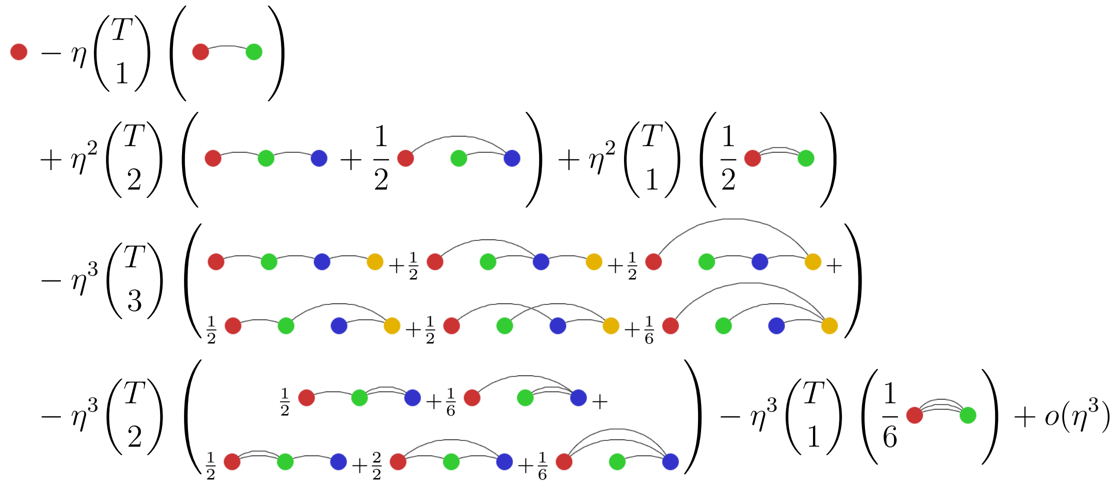
 $:= \mathbb{E}_{\text{red,green,blue}} \nabla^a(l_{\text{red}}) \nabla^b(l_{\text{green}}) \nabla^a \nabla^b(l_{\text{blue}})$

We see that  $-$  gives the trace of the covariance of gradients. Moreover,  $=$ , illustrating how diagram notation can streamline computation by helping to group terms. However, we caution that a diagram's value generally depends on that diagram's digraph structure, not just its undirected structure. For example:

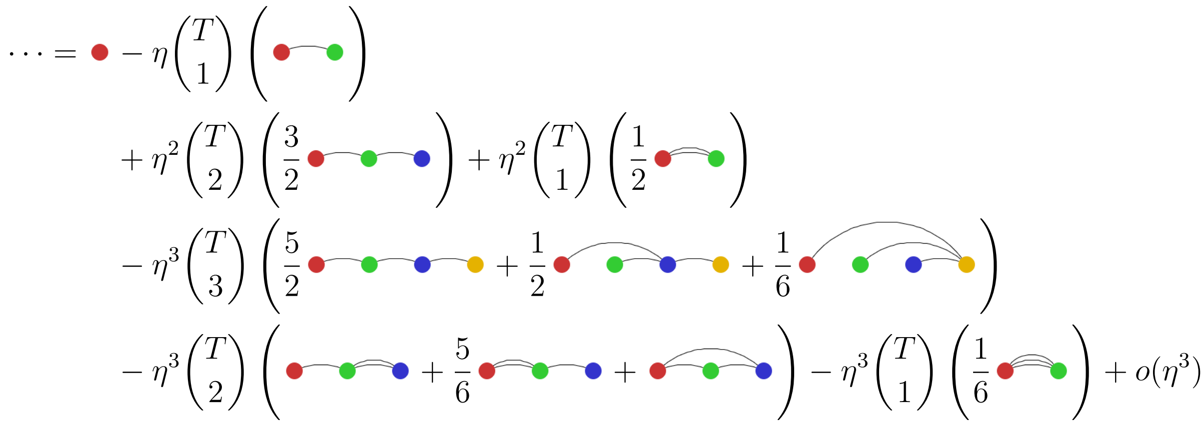


0.1. SGD Test Loss

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 with entirely distinct colors whose underlying poset has a unique rightmost element. Each node in the diagram contributes a symmetry factor $o!/i! \prod_k o_k!$ where o, i are the node's in- and out- degrees and the o_k count the out-edges to node k. 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:

$\mathbb{E}$ (SGD Test Loss) $=$ 

And a routine grouping of terms yields:



0.2. GD and Train Loss

To compute losses for non-stochastic gradient descent, we allow non-rightmost nodes to share colors with each other. To compute train losses, we allow the rightmost node to share a color with previous nodes with probability C/N, where C counts the number of non-rightmost colors in the diagram. We may thus compute (by subtraction) generalization gaps and the benefit of stochasticity.