Perturbative Analysis of SGD

0. Datatype of the Loss Landscape

1. Coefficients of Generalization

Let

$$INT = \langle (a) \rangle \langle (a) \rangle$$

$$UNC = \langle (a)(a) \rangle - \langle (a) \rangle \langle (a) \rangle$$

$$PAS = 2 \langle (a) \rangle \langle (ab)(b) \rangle$$

$$TEM = 2 \langle (a) \rangle \left(\langle (ab)(b) \rangle - \langle (ab) \rangle \langle (b) \rangle \right)$$

$$PER = \langle (ab) \rangle \left(\langle (a)(b) \rangle - \langle (a) \rangle \langle (b) \rangle \right)$$

One finds:

$$\mathbb{E}L_{\text{SGD}}(\eta) = () + \eta \binom{T}{1} INT + \eta^2 \left(\binom{T}{2} \left(\frac{PAS}{2} + \frac{PAS}{4} \right) + \binom{T}{1} \left(\frac{PAS}{4} + \frac{PER}{2} \right) \right) + \cdots$$

while:

$$\mathbb{E}L_{\text{GD}}(\eta) = () + \eta \binom{T}{1} INT + \eta^2 \left(\binom{T}{2} \left(\frac{PAS}{2} + \frac{TEM}{2N} + \frac{PAS}{4} + \frac{PER}{2N} \right) + \binom{T}{1} \left(\frac{PAS}{4} + \frac{PER}{2N} \right) \right) + \cdots$$

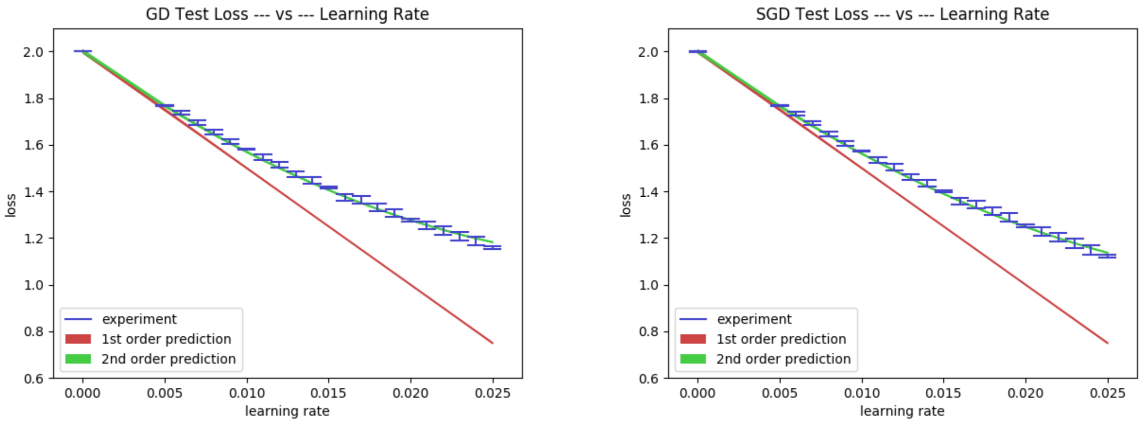
2. Toy Examples

2.0. Shifting Valleys

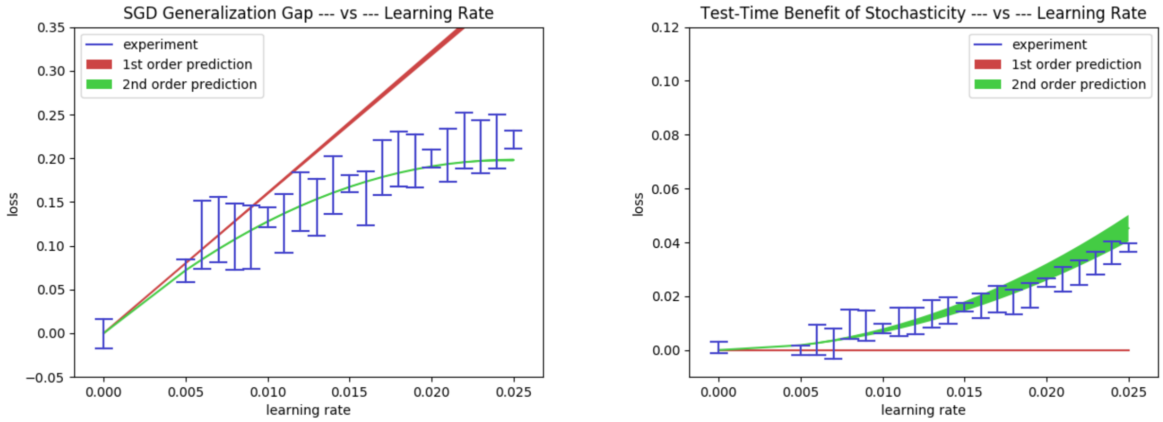
Let $L(x, (A, B)) = A + (B - Ax)^2 - A^2$, where the data samples $x \in \mathbb{R}^1$ obey a standard normal law. The weights $(A, B) \in \mathbb{R}^2$ we initialize to $(1, 1)$. The expected loss is $\mathbb{E}L(A, B) = A + B^2$.

2.1. Valley of Death

Let $L(x, (A, B)) = A + (B - Ax)^4 - 3A^4$, where the data samples $x \in \mathbb{R}^1$ obey a standard normal law. The weights $(A, B) \in \mathbb{R}^2$ we initialize to $(1, 1)$. The expected loss is $\mathbb{E}L(A, B) = A + 2A^2B^2 + B^4$.



Above: for the Valley Task, our 2nd order corrections for test-time loss match experiment (for $T = 10$ and $\eta \leq 0.025$; η of this magnitude suffice to halve the test loss).



Above: for the Valley Task, experiments verify (**left**) the predicted dependence of generalization gap on `uncertainty` $\langle(a)(a)\rangle - \langle(a)\rangle\langle(a)\rangle$ and (**right**) the resulting dependence of SGD's test-time outperformance of GD on `temerity` $2\langle(a)\rangle\left(\langle(ab)(b)\rangle - \langle(ab)\rangle\langle(b)\rangle\right)$ and `peril` $\langle(ab)\rangle\left(\langle(a)(b)\rangle - \langle(a)\rangle\langle(b)\rangle\right)$.