

# The Datatype of Loss Landscapes

Dan Roberts (roberts@ias.edu) and Samuel Tenka (coli@mit.edu)

2019-01-22

## 0. Loss Landscapes

What is the natural setting for a stochastic gradient-based optimizer such as SGD? A typical implementation might have this pseudocode:

```
Weight learn(DataN X, Weight w0, float(*l)(Data, Weight), int E) {
    Weight w = w0;
    for (int e = 0; e ≠ E; e = e + 1) {
        shuffle(X);
        for (Data x: X) {
            Covector g = ∇wl(x, w);
            w = expw(-transpose(g));
        }
    }
    return w;
}
```

We thus see that the key ingredients are

- a probability space  $X$  of data;
- a manifold  $W$  of weights that is equipped with
- a Riemannian (inverse) metric  $\text{transpose} : T^*W \rightarrow TW$  to turn covectors into vectors (the learning rate is part of this data) that in turn induces
- the flow  $\exp : TW \rightarrow W$  to update along a vector; and
- a loss function  $l : X \times W \rightarrow \mathbb{R}$ .

One wishes for  $W$  to be metrically complete, for  $w \mapsto l(x, w)$  to be smooth for all  $x$ , and for each random variable  $\nabla_w^a \nabla_w^b \cdots \nabla_w^z l(x, w)$  to be subgaussian for all  $w$ . In this case, let us call the listed data  $(X, W, l)$  a **loss landscape**. Traditionally,  $W$  has been either curved according to a Fisher metric or else flat. The purpose of this note is to unify and clarify these and all other reasonable possibilities.

Let us write  $\langle f(x) \rangle$  for the expectation of  $f(x)$ , and let us notate derivatives such as  $\nabla_w^a \nabla_w^b \nabla_w^c \nabla_w^d l(x, w)$  by parenthesized sequences of indices such as  $(abcd)$ . By subgaussianity, every grammatical expression of  $\langle \rangle$ s and  $()$ s has a finite value. For example, we may consider the **hessian**  $H = \langle (ab) \rangle$  or the **covariance**  $C = \langle (a)(b) \rangle - \langle (a) \rangle \langle (b) \rangle$ . Though both  $H$  and  $C$  are symmetric 2-tensors on  $TW$ , their meanings greatly differ. For instance,  $H$  scales linearly with  $l$  while  $C$  scales quadratically. Table 0 makes such scalings vivid by imagining that the loss has units of dollars.

## 1. Germs

## 2. Taylor Expansion in the Metric

DIMENSIONS	length <sup>-2</sup>	length <sup>-1</sup>	length <sup>0</sup>	length <sup>1</sup>
dollars <sup>-1</sup>			learning rate	gas mileage
dollars <sup>0</sup>	metric		unitless	weight update
dollars <sup>1</sup>	hessian	gradient	loss	
dollars <sup>2</sup>	covariance		trace covariance	

Table 0: This table indicates the scaling properties of selected objects so that the reader may orient her intuition for dimensional analysis. Many interesting objects are left unlisted: for instance, the variance  $\langle()()\rangle - \langle()\rangle\langle()\rangle$  of the loss would inhabit the trace-covariance’s cell, and the intensity  $\langle(a)\rangle\langle(a)\rangle$  of the mean gradient would inhabit the covariance’s cell.

### 3. Example: The Valley of Death

### 4. Geometry as Prior