

applications, it is important to discriminate between elements that are exactly zero and elements that are small but nonzero. In these cases, we turn to a function that grows at the same rate in all locations, but retains mathematical simplicity: the L^1 norm. The L^1 norm may be simplified to

$$\|\mathbf{x}\|_1 = \sum_i |x_i|. \quad (2.31)$$

The L^1 norm is commonly used in machine learning when the difference between zero and nonzero elements is very important. Every time an element of \mathbf{x} moves away from 0 by ϵ , the L^1 norm increases by ϵ .

We sometimes measure the size of the vector by counting its number of nonzero elements. Some authors refer to this function as the “ L^0 norm,” but this is incorrect terminology. The number of non-zero entries in a vector is not a norm, because scaling the vector by α does not change the number of nonzero entries. The L^1 norm is often used as a substitute for the number of nonzero entries.

One other norm that commonly arises in machine learning is the L^∞ norm, also known as the **max norm**. This norm simplifies to the absolute value of the element with the largest magnitude in the vector,

$$\|\mathbf{x}\|_\infty = \max_i |x_i|. \quad (2.32)$$

Sometimes we may also wish to measure the size of a matrix. In the context of deep learning, the most common way to do this is with the otherwise obscure **Frobenius norm**:

$$\|A\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}, \quad (2.33)$$

which is analogous to the L^2 norm of a vector.

The dot product of two vectors can be rewritten in terms of norms. Specifically,

$$\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos \theta \quad (2.34)$$

where θ is the angle between \mathbf{x} and \mathbf{y} .

2.6 Special Kinds of Matrices and Vectors

Some special kinds of matrices and vectors are particularly useful.

Diagonal matrices consist mostly of zeros and have non-zero entries only along the main diagonal. Formally, a matrix \mathbf{D} is diagonal if and only if $D_{i,j} = 0$ for