The section talks about two cases of optimization of state-of-the-art models that have big datasets and a lot of model parameters, and the problems encountered while trying to optimize as well as determine the use of appropriate prediction functions for large scale models like these.

2.1 Text Classification via Convex Optimization.

We are introduced to a basic problem encountered which is the inability of the model to correctly classify the topic of interest accurately upon introduction of new documents that didn't follow the previously established rules of classification. The contradiction of rules led to limiting the applications of such models to simple tasks. A simple statistical approach with a prediction function that minimized the frequency of misclassifications was used however, this could lead to the function memorizing the examples and so the need for a prediction function arose that generalizes the concepts rather than rote learning the examples. Choosing prediction functions can be carried out by using cross-validation procedures such as splitting the sets into 3 (ie. addition of a validation set) and seeing which function's performance is best on the validation set. The bag of words approach seems to work well for this. In this approach a document is represented by a feature vector linked to a specific set of words. This encoding method leads to sparse vectors and scaling ensures the difference in document lengths is overcome. Prediction functions of the form $h(x; w, \tau) = w^T x - \tau$ work well in this case. Accuracy is calculated by counting the no. of times $\text{sign}(h(x; w, \tau))$ matches the correct label but large scale optimization gets difficult to do due to the discontinuous nature of the sign function. To overcome this a log-loss function of the form $(h, y) = \log(1 + \exp(-hy))$ is used with a regularization term parameterized by a scalar $\lambda > 0$. One can also use an 1-norm regularizer. The choice of loss function is done based on experimentation. This approach works theoretically as well as experimentally well for text classification problems however doesn't translate to working for deep neural networks which gives rise to large scale non-linear nonconvex optimization problems discussed ahead.

2.2 Perceptual Tasks via Deep Neural Networks

Perceptual tasks aren't well performed in an automated manner using computer programs based on sets of prescribed rules. In DNNs the value of the prediction function is computed by applying successive transformations to a given input vector in different layers along with the use of an activation function which gives the ultimate prediction value as the output vector with all parameters of successive layers. Optimization problem here too has the collection of a training set and the choice of a loss function. However, this optimization problem is highly nonlinear and nonconvex, making it difficult to solve to global optimality. An elegant solution is to use gradient-based methods along with back propagation. The number of layers and the size of each layer is chosen based on performance on validation set. Convolutional neural networks are effective for computer vision and signal processing. Such a network is composed of convolutional layers, wherein the parameter matrix is a circulant matrix and the input is construed as a multichannel image. The product then computes the convolution of the image by a trainable filter while the activation function (which are piecewise linear functions) can perform more complex operations like image rectification, contrast normalization, or subsampling. Although DNNs are extremely successful their training process requires skill and care. One can do anything useful with such large, highly nonlinear, and nonconvex model.

References:

*Bottou et al. Optimization Methods for Large-Scale Machine Learning. Siam 2018. (**Sections 2.1 and 2.2**)*