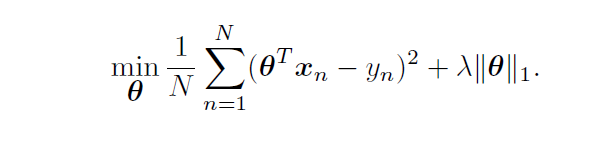
### Overall Embedded Feature selection

Risk:

### Feature selection with l1 norm (LASSO)



Loss:

Disadvantage:

1. Achieves through **shrinkage** of the coefficients

2. is restricted to **linear** models

Goal:

1. Avoid **shrinkage**

2. perform feature selection while learning a **nonlinear** function

### STG

* A **probabilistic** and computationally efficient neural network approach

***Why is that probabilistic?***

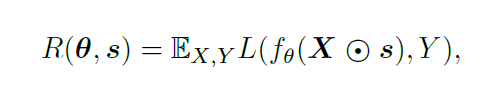
*1. l0 regularization is* ***computationally expensive*** *and* ***intractable*** *for* ***high dimensions***

*2. l0 norm cannot be incorporated into a gradient descent based optimization*

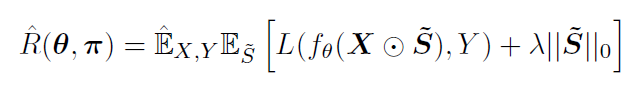
* **Bernoulli gates** applied to each of the d input nodes of a neural network

***What is Bernoulli gate?***

*A Bernoulli gate takes a logic signal (trigger or gate) as an input, and routes it to either of its two outputs according to a random coin toss.*

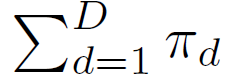


Overall Risk:



Bernoulli gate Risk:

For Bernoulli gates, whose entries are independent and satisfy

 boils down to the sum of Bernoulli parameters 

* **Problem Translation**

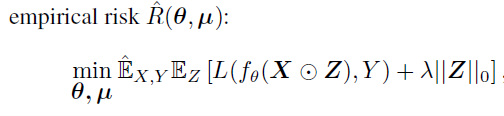
The problem of feature selection translates to finding Theta and Pi that minimize the empirical risk based on the above formulation

Develop an empirically superior continuous distribution that is fully differentiable and implemented only to activate or deactivate the gates linking each feature (node) to the rest of the network

### Bernoulli Continuous Relaxation for Feature Selection

**Problem**: Feature selection requires **stability** in the selected set of features, but Bernoulli variables due to the heavy-tailedness, which often leads to **inconsistency** in the set of selected features

**Solution**: A Gaussian-based continuous relaxation (**Approximation**) for the Bernoulli variables . We refer to each relaxed Bernoulli variable as a stochastic gate(STG) defined by 



Rewrite the risk:

