Approach to Model Fitting

- Models differ by their ability to control model properties **Capacity, Variance, Bias, Smoothness**

- First, capacity should be sufficient

- Second, Variance and Bias should be addressed separately for optimal performance.

Is it easy to control these properties for networks?

Few good words about NN

- Simple model - composition of ridge function
- Natural for imposing bias via Projection Pursuit constraints
- Ideal for high dimensional space
- Ideal when linear projections are useful, e.g., for image recognition
- Simple interpretability as an extension of logistic regression

Specific problems to NN estimation

- Nonidentifiable model: Variability due to local minima
- Requires special care for high dimensional optimization
  - Adaptation of acceleration methods for gradient search
  - Methods for finding (nearly) global minimum
- Since works well in high dim, one tends to apply directly to the large data representation (other data representations)

Variance/Bias Decomposition for Ensembles

$$\bar{f}(x) = \frac{1}{Q}\sum_{i=1}^{Q} f_i(x).$$

$$\begin{aligned} E[(\bar{f} - E[\bar{f}])^2] &= E[(\frac{1}{Q}\sum f_i - E[\frac{1}{Q}\sum f_i])^2] \\ &= E[(\frac{1}{Q}\sum f_i)^2] - \left(E[\frac{1}{Q}\sum f_i]\right)^2. \end{aligned} \qquad (1)$$

The first RHS term can we rewritten as

$$E[(\frac{1}{Q}\sum f_i)^2] = \frac{1}{Q^2}\sum E[f_i^2] + \frac{2}{Q^2}\sum_{i<j} E[f_i f_j],$$

and the second term gives,

$$\left(E[\frac{1}{Q}\sum f_i]\right)^2 = \frac{1}{Q^2}\sum \left(E[f_i^2]\right)^2 + \frac{2}{Q^2}\sum_{i<j} E[f_i]E[f_j].$$

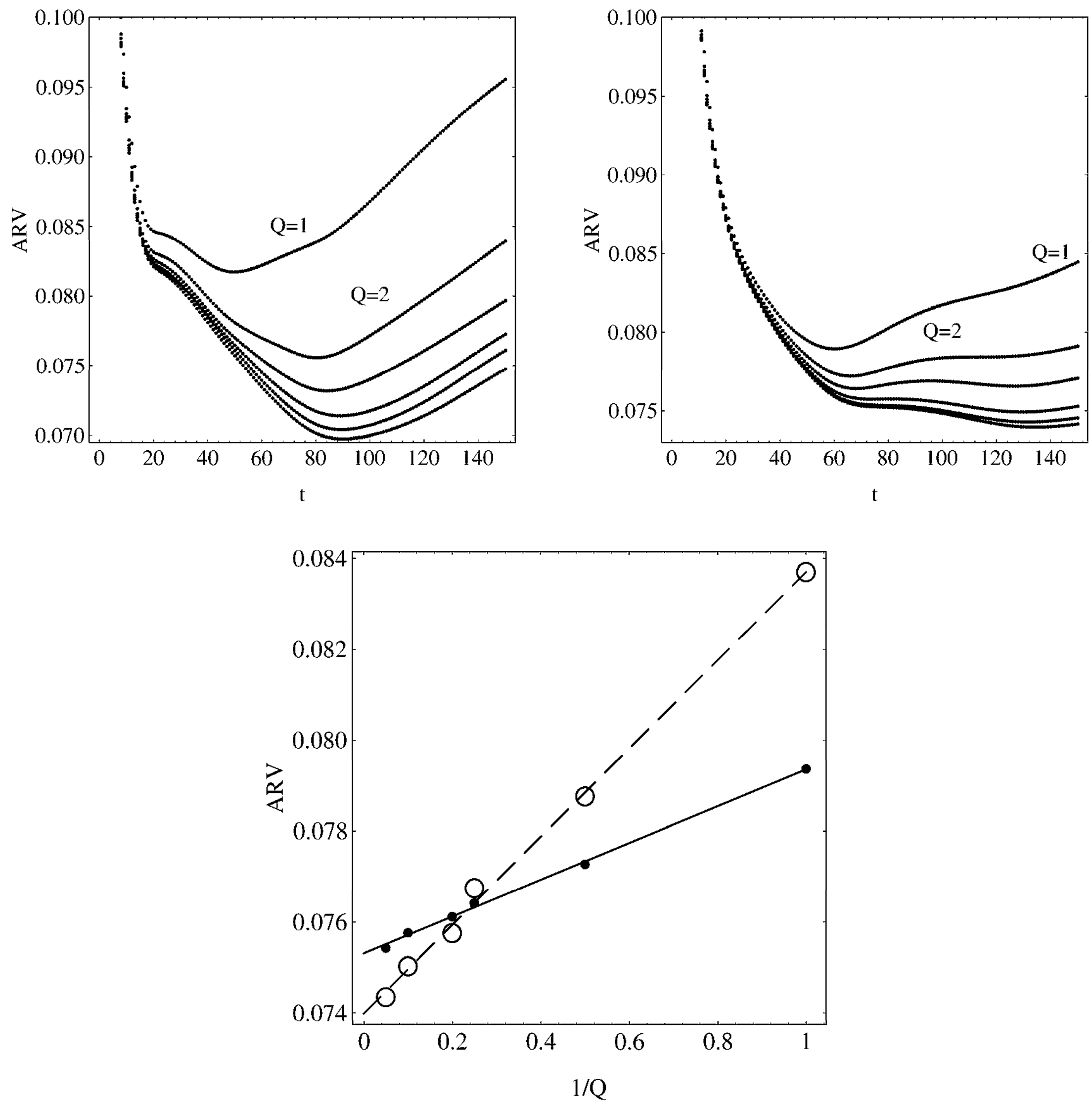
Plugging these equalities into (1) gives

$$E[(\bar{f} - E[\bar{f}])^2] = \frac{1}{Q^2}\sum \{E[f_i^2] - \left(E[f_i]\right)^2\} + \frac{2}{Q^2}\sum_{i<j}\{E[f_i f_j] - E[f_i]E[f_j]\}.$$

Set $\gamma = \text{Var}(f_i) + (Q-1)\text{max}_{i,j}(E[f_i f_j] - E[f_i]E[f_j])$.
It follows $[ab \le \frac{a^2+b^2}{2} \Rightarrow E[f_i f_j] - E[f_i]E[f_j] \le \text{max}_i \text{Var}(f_i)]$ that

$$\text{Var}(\bar{f}) \le \frac{1}{Q}\gamma \le \max_i \text{Var}(f_i). \qquad (2)$$

ARV as a function of ensemble size and training time



Different ensembles of two predictors as a function of training time. The variance goes down as 1/Q.

Bias Control

- The idea: Introduce Bias based on prior knowledge

- Since we want to use in NN which perform projection of the input space onto weight space, we want to find interesting directions in the data

- Statistical framework for prior knowledge - Exploratory Projection Pursuit EPP