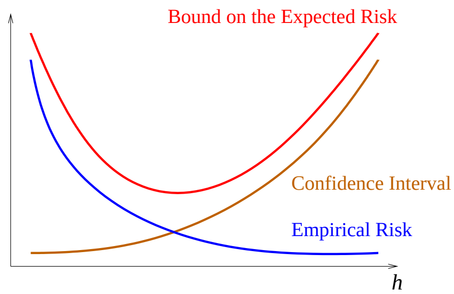
Statistical Machine Learning from Data
- Quick Reminder -

Samy Bengio
IDIAP Research Institute
CP 592, rue du Simplon 4, 1920 Martigny, Switzerland
bengio@idiap.ch
http://www.idiap.ch/~bengio

November 9, 2005

1 Statistical Learning Theory

- Let D_n be a training set of examples z_i drawn independently from unknown $p(z)$
- We need a set of functions $\mathcal{F}$. Example: linear functions $f(x) = a \cdot x + b$
- We need a loss function $L(z, f)$. Example: $L((x, y), f) = (f(x) - y)^2$
- Expected Risk: $R(f) = E_Z[L(z, f)] = \int_Z L(z, f)p(z)dz$ = generalization error
- Empirical Risk: $\hat{R}(f, D_n) = \frac{1}{n} \sum_{i=1}^{n} L(z_i, f)$
- Empirical Risk Minimization: $f^\star(D_n) = \arg\min_{f \in \mathcal{F}} \hat{R}(f, D_n)$
- Training error: $\hat{R}(f^\star(D_n), D_n)$
- Difference between Expected Risk and Empirical Risk bounded but depends on capacity
- Curves show that there is an optimal capacity:



- Methodology:
  - empirical risk minimization on a training set D^{tr}

    $$f^\star(D^{tr}) = \arg\min_{f \in \mathcal{F}} \hat{R}(f, D^{tr})$$

  - model selection on a validation set D^{va}

    $$\theta_m^\star = \arg\min_{\theta_m} R(f_{\theta_m}^\star(D^{tr}), D^{va})$$

  - estimation of the expected risk on a separate test set D^{te}

    $$R(f_{\theta_m}^\star(D^{tr} \cup D^{va}), D^{te})$$

  - if data is small, consider cross-validation