Introduction

- Property of primary significance in nnet: **learn** from its environment, and **improve** its performance through learning.
- Iterative adjustment of synaptic weights.
- Learning: hard to define.
  - One definition by Mendel and McClaren: *Learning is a process by which the free parameters of a neural network are adapted through a process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place.*

Overview

Organization of this chapter:

1. Five basic learning rules
   - error correction, Hebbian, memory-based, competitive, and Boltzmann
2. Learning paradigms
   - credit assignment problem, supervised learning, unsupervised learning
3. Learning tasks, memory, and adaptation
4. Probabilistic and statistical aspects of learning
Error-Correction Learning

- Input $x(n)$, output $y_k(n)$, and desired response or target output $d_k(n)$.
- Error signal $e_k(n) = d_k(n) - y_k(n)$
- $e_k(n)$ actuates a control mechanism that gradually adjust the synaptic weights, to minimize the cost function (or index of performance):
  $$E(n) = \frac{1}{2} e_k^2(n)$$
- When synaptic weights reach a steady state, learning is stopped.

Memory-Based Learning

- All (or most) past experiences are explicitly stored, as input-target pairs $\{(x_i, d_i)\}_{i=1}^N$.
- Two classes $C_1, C_2$.
- Given a new input $x_{\text{test}}$, determine class based on local neighborhood of $x_{\text{test}}$.
  - Criterion used for determining the neighborhood
  - Learning rule applied to the neighborhood of the input, within the set of training examples.

<table>
<thead>
<tr>
<th>Error-Correction Learning</th>
<th>Memory-Based Learning: Nearest Neighbor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Widrow-Hoff rule, with learning rate $\eta$:</td>
<td>A set of instances observed so far:</td>
</tr>
<tr>
<td>$\Delta w_{kj}(n) = \eta e_k(n)x_j(n)$</td>
<td>$X = {x_1, x_2, ..., x_N}$</td>
</tr>
<tr>
<td>With that, we can update the weights:</td>
<td>Nearest neighbor $x'<em>N \in X$ of $x</em>{\text{test}}$:</td>
</tr>
<tr>
<td>$w_{kj}(n + 1) = w_{kj}(n) + \Delta w_{kj}(n)$</td>
<td>$\min_i d(x_i, x_{\text{test}}) = d(x'<em>N, x</em>{\text{test}})$</td>
</tr>
<tr>
<td>There is a sound theoretical reason for doing this, which we will discuss later.</td>
<td>where $d(\cdot, \cdot)$ is the Euclidean distance.</td>
</tr>
</tbody>
</table>

- $x_{\text{test}}$ is classified as the same class as $x'_N$.
- Cover and Hart (1967): error (for Nearest Neighbor) is at most twice that of the optimal (Bayes probability of error), given
  - The classified examples are independently and identically distributed.
  - The sample size $N$ is infinitely large.
Memory-Based Learning: $k$− Nearest Neighbor

- Identify $k$ classified patterns that lie nearest to the test vector $\mathbf{x}_{test}$, for some integer $k$.
- Assign $\mathbf{x}_{test}$ to the class that is most frequently represented by the $k$ neighbors (use majority vote).
- In effect, it is like averaging. It can deal with outliers. The input $\mathbf{x}$ above will be classified as 1.

Hebbian Learning


  When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic changes take place in one or both cells such that A's efficiency as one of the cells firing B, is increased.

- **Hebbian synapse**
  - If two neurons on either side of a synapse are activated simultaneously, the synapse is strengthened.
  - If they are activated asynchronously, the synapse is weakened or eliminated. (This part was not mentioned in Hebb.)

Hebbian Synapses

- Time-dependent mechanism
- Local mechanism
- Interactive mechanism
- Correlative/conjunctive mechanism

Strong evidence for Hebbian plasticity in the Hippocampus (brain region).

Classification of Synaptic Plasticity

<table>
<thead>
<tr>
<th>Type</th>
<th>Positively correlated</th>
<th>Negatively correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hebbian</td>
<td>Strengthen</td>
<td>Weaken</td>
</tr>
<tr>
<td>Anti-Hebbian</td>
<td>Weaken</td>
<td>Strengthen</td>
</tr>
<tr>
<td>Non-Hebbian</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Mathematical Models of Synaptic Plasticity

- General form: \( \Delta w_{kj}(n) = F(y_k(n), x_j(n)) \)
- Hebbian learning (with learning rate \( \eta \)): \( \Delta w_{kj}(n) = \eta y_k(n) x_j(n) \)
- Covariance rule: \( \Delta w_{kj} = \eta (x_j - \bar{x})(y_k - \bar{y}) \)

Covariance Rule (Sejnowski 1977)

- Convergence to a nontrivial state
- Prediction of both potentiation and depression.
- Observations:
  - Weight enhanced when both pre- and post-synaptic activities are above average.
  - Weight depressed when
    * Presynaptic activity more than average, and postsynaptic activity less than average.
    * Presynaptic activity less than average, and postsynaptic activity more than average.

Competitive Learning

- Output neurons compete with each other for a chance to become active.
- Highly suited to discover statistically salient features (that may aid in classification).
- Three basic elements:
  - Same type of neurons with different weight sets, so that they respond differently to a given set of inputs.
  - A limit imposed on the strength of each neuron.
  - Competition mechanism, to choose one winner: winner-takes-all neuron.

Inputs and Weights Seen as Vectors in High-dimensional Space

- Inputs and weights can be seen as vectors: \( x \) and \( w_k \). Note that the weight vector belongs to a certain output neuron \( k \), and thus the index.
Competitive Learning: Example

- Single layer, feedforward excitatory, and lateral inhibitory connections

- Winner selection

\[ y_k = \begin{cases} 
1 & \text{if } v_k > v_j \text{ for all } j, j \neq k \\
0 & \text{otherwise}
\end{cases} \]

- Limit: \( \sum_j w_{kj} = 1 \) for all \( k \).

- Adaptation:

\[
\Delta w_{kj} = \begin{cases} 
\eta (x_j - w_{kj}) & \text{if } k \text{ is the winner} \\
0 & \text{otherwise}
\end{cases}
\]

* The synaptic weight vector \( w_k = (w_{k1}, w_{k2}, \ldots, w_{kn}) \) is moved toward the input vector.

Boltzmann Learning

- Stochastic learning algorithm rooted in statistical mechanics.

- Recurrent network, binary neurons (on: '+1', off: '-1').

- Energy function \( E \):

\[ E = -\frac{1}{2} \sum_j \sum_{k, k \neq j} w_{kj} x_k x_j \]

- Activation:
  - Choose a random neuron \( k \).
  - Flip state with a probability (given temperature \( T \))

\[ P(x_k \rightarrow -x_k) = \frac{1}{1 + \exp(-\Delta E_k / T)} \]

where \( \Delta E_k \) is the change in \( E \) due to the flip.

Boltzmann Machine

- Two types of neurons
  - Visible neurons: can be affected by the environment
  - Hidden neurons: isolated

- Two modes of operation
  - Clamped: visible neuron states are fixed by environmental input and held constant.
  - Free-running: all neurons are allowed to update their activity freely.
Boltzmann Machine: Learning and Operation

- **Learning:**
  - Correlation of activity during clamped condition $\rho_{kj}^+$
  - Correlation of activity during free-running condition $\rho_{kj}^-$
  - Weight update: $\Delta w_{kj} = \eta(\rho_{kj}^+ - \rho_{kj}^-), j \neq k$.

- **Train weights** $w_{kj}$ with various clamping input patterns.

- After training is completed, present new clamping input pattern that is a partial input of one of the known vectors.

- Let it run clamped on the new input (subset of visible neurons), and eventually it will complete the pattern (**pattern completion**).

\[
\text{Correl}(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}
\]

Credit-Assignment Problem

- How to assign **credit** or **blame** for overall outcome to **individual decisions** made by the learning machine.

- In many cases, the outcomes depend on a **sequence of actions**.
  - Assignment of credit for outcomes of actions (**temporal credit-assignment problem**): **When** does a particular action deserve credit.
  - Assignment of credit for actions to internal decisions (**structural credit-assignment problem**): assign credit to internal structures of actions generated by the system.

Credit-assignment problem routinely arises in neural network learning. Which neuron, which connection to credit or blame?

Learning Paradigms

- **Learning with a Teacher**
  - Also known as supervised learning
  - Teacher has knowledge, represented as **input–output examples**. The environment is **unknown** to the nnet.
  - Nnet tries to emulate the teacher gradually.
  - Error-correction learning is one way to achieve this.
  - Error surface, gradient, steepest descent, etc.

- **Learning with a Teacher**
  - Also known as supervised learning
  - Teacher has knowledge, represented as **input–output examples**. The environment is **unknown** to the nnet.
  - Nnet tries to emulate the teacher gradually.
  - Error-correction learning is one way to achieve this.
  - Error surface, gradient, steepest descent, etc.
Learning without a Teacher

Two classes

- Reinforcement learning (RL)/Neurodynamic programming
- Unsupervised learning/Self-organization

Learning without a Teacher: Reinforcement Learning

- Learning input-output mapping through continued interaction with the environment.
- **Actor-critic**: circuit converts primary reinforcement signal into higher-quality, heuristic reinforcement signal (Barto, Sutton, ...).
- Goal is to optimize the cumulative cost of actions.
- In many cases, learning is under delayed reinforcement. Delayed RL is difficult since (1) teacher does not provide desired action at each step, and (2) must solve temporal credit-assignment problem.
- Relation to dynamic programming, in the context of optimal control theory (Bellman).

Learning without a Teacher: Unsupervised Learning

- Learn based on task-independent measure of the quality of representation.
- Internal representations for encoding features of the input space.
- Competitive learning rule needed, such as winner-takes-all.

Learning Tasks, Memory, and Adaptation

Learning tasks

- Pattern association
- Pattern recognition
- Function approximation
- Control
- Filtering/Beamforming

Memory and adaptation
Pattern Association

- **Associative memory**: brain-like distributed memory that learns association. Storage and retrieval (recall).
- **Pattern association** ($x_k : \text{key pattern}, y_k : \text{memorized pattern}$):
  - autoassociation ($x_k = y_k$): given partial or corrupted version of stored pattern and retrieve the original.
  - heteroassociation ($x_k \neq y_k$): Learn arbitrary pattern pairs and retrieve them.

- Relevant issues: storage capacity vs. accuracy.

Function Approximation

- Nonlinear input-output mapping: $d = f(x)$ for an unknown $f$.
- Given a set of labeled examples $T = \{(x_i, d_i)\}_{i=1}^N$, estimate $F(\cdot)$ such that
  $$\|F(x) - f(x)\| < \epsilon, \text{ for all } x$$

Pattern Classification

- Mapping between input pattern and a prescribed number of classes (categories).
- Two general types:
  - Feature extraction (observation space to feature space: cf. dimensionality reduction), then classification (feature space to decision space).
  - Single step (observation space to decision space).

Function Approx: System Identification and Inverse System Modeling

- System identification: learn function of an unknown system.
  $$d = f(x)$$
- Inverse system modeling: learn inverse function:
  $$x = f^{-1}(d)$$
Control

- Control of a *plant*, a process or critical part of a system that is to be maintained in a controlled condition.
- Feedback controller: adjust plant input $u$ so that the output of the plant $y$ tracks the reference signal $d$. Learning is in the form of free-parameter adjustment in the controller.

Filtering, Smoothing, and Prediction

Extract information about a quantity of interest from a set of noisy data.

- Filtering: estimate quantity at time $n$, based on measurements up to time $n$.
- Smoothing: estimate quantity at time $n$, based on measurements up to time $n + \alpha$ ($\alpha > 0$).
- Prediction: estimate quantity at time $n + \alpha$, based on measurements up to time $n$ ($\alpha > 0$).

Blind Source Separation and Nonlinear Prediction

- Blind source separation: recover $u(n)$ from distorted signal $x(n)$ when mixing matrix $A$ is unknown.
  \[ x(n) = Au(n) \]
- Nonlinear prediction: given $x(n - T), x(n - 2T), \ldots$, estimate $x(n)$ ($\hat{x}(n)$ is the estimated value).

Linear Algebra Tip: Partitioned (or Block) Matrices

When multiplying matrices or matrix and a vector, partitioning them and multiplying the corresponding partitions can be very convenient.

- When multiplying matrices or matrix and a vector, partitioning them and multiplying the corresponding partitions can be very convenient.

Consider the $4 \times 5$ matrix above (let's call it $X$). If you have another $5 \times 4$ matrix partitioned similarly into $\begin{bmatrix} E & F \\ G & H \end{bmatrix}$ (let's call it $Y$), then you can calculate the product as another block matrix:

\[
XY = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} AE + BG, AF + BH \\ CE + DG, CF + DH \end{bmatrix}
\]

Memory

- **Memory**: relatively enduring neural alterations induced by an organism’s interaction with the environment.
- Memory needs to be accessible by the nervous system to influence behavior.
- Activity patterns need to be stored through a learning process.
- Types of memory: short-term and long-term memory.

Associative Memory (cont’d)

- Weight matrix:

  \[ y_k = W(k)x_k, \text{for } k = 1, 2, \ldots, q \]
  \[ y_{ki} = \sum_{j=1}^{m} w_{ij}(k)x_{kj}, \text{for } i = 1, 2, \ldots, m \]

  \[ y_{ki} = [w_{i1}(k), w_{i2}(k), \ldots, w_{im}(k)] \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{km} \end{bmatrix}, \text{for } i = 1, 2, \ldots, m \]

  \[ \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{km} \end{bmatrix} = \begin{bmatrix} w_{11}(k), w_{12}(k), \ldots, w_{1m}(k) \\ w_{21}(k), w_{22}(k), \ldots, w_{2m}(k) \\ \vdots \\ w_{m1}(k), w_{m2}(k), \ldots, w_{mm}(k) \end{bmatrix} \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{km} \end{bmatrix} \]

- With a single \( W(k) \), we can only represent one mapping (\( x_k \) to \( y_k \)). For all pairs (\( x_k, y_k \) \( k = 1, 2, \ldots, q \)), we need \( q \) such weight matrices.

  \[ y_k = W(k)x_k, \text{for } k = 1, 2, \ldots, q \]

- One strategy is to combine all \( W(k) \) into a single memory matrix \( M \) by simple summation:

  \[ M = \sum_{k=1}^{q} W(k) \]

- Will such a simple strategy work? That is, can the following be possible with \( M \)?

  \[ y_k \approx Mx_k, \text{for } k = 1, 2, \ldots, q \]
**Associative Memory: Example – Storing Multiple Mappings**

With fixed set of key vectors $x_k$, an $m \times m$ matrix can store $m$ arbitrary output vectors $y_k$.

- Let $x_k = [0, 0, \ldots, 1, \ldots, 0]^T$ where only the $k$-th element is 1 and all the rest is 0.
- Construct a memory matrix $M$ with each column representing the arbitrary output vectors $y_k$ (use the block matrix technique):

  $$M = \begin{bmatrix} y_1, y_2, \ldots, y_m \end{bmatrix}$$

- Then, $y_k = Mx_k$, for all $k = 1, 2, \ldots, m$.
- But, we want $x_k$ to be arbitrary too!

**Correlation Matrix Memory**

- With $q$ pairs $(x_k, y_k)$, we can construct a candidate memory matrix that stores all $q$ mappings as:

  $$\hat{M} = \sum_{k=1}^{q} y_k x_k^T = y_1 x_1^T + y_2 x_2^T + \ldots + y_q x_q^T,$$

  where $y_k x_k^T$ represents the outer product of vectors that results in a matrix, i.e.,

  $$(y_k x_k^T)_{ij} = y_k x_k^T.$$

- A more convenient notation is:

  $$\hat{M} = \begin{bmatrix} y_1, y_2, \ldots, y_q \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix} = YX^T.$$  

  This can be verified easily using partitioned matrices.

**Correlation Matrix Memory: Recall**

- Will $\hat{M}x_k$ give $y_k$?

- For convenience, let’s say

  $$\hat{M} = \sum_{k=1}^{q} y_k x_k^T = \sum_{k=1}^{q} W(k).$$

- First, consider $W(k) = y_k x_k^T$ only.

  Check if $W(k)x_k = y_k$:

  $$W(k)x_k = y_k x_k^T x_k = y_k (x_k x_k^T) = cy_k$$

  where $c = x_k x_k^T$, a scalar value (the length of vector $x_k$ squared). If all $x_k$s were normalized to have length 1, $W(k)x_k = y_k$ will hold!

**Correlation Matrix Memory: Recall (cont’d)**

- Now, back to $\hat{M}$: under what condition will $\hat{M}x_j$ give $y_j$ for all $j$? Let’s begin by assuming $x_k^T x_k = 1$ (key vectors are normalized).

- We can decompose $\hat{M}x_j$ as follows:

  $$\hat{M}x_j = \sum_{k=1}^{q} y_k x_k^T x_j = y_j x_j^T x_j + \sum_{k=1, k \neq j}^{q} y_k x_k^T x_j.$$

- We know $y_j x_j^T x_j = y_j$, so it now becomes:

  $$\hat{M}x_j = y_j + \sum_{k=1, k \neq j}^{q} y_k x_k^T x_j.$$  

  If all keys are orthogonal (perpendicular to each other), then for an arbitrary $k \neq j$, $x_k^T x_j = ||x_k|| ||x_j|| \cos(\theta_{k,j}) = 1 \times 1 \times 0 = 0$, so the noise term becomes 0, and hence $\hat{M}x_j = y_j + 0 = y_j$. The example in page 41 is one such case, an extreme one!
Correlation Matrix Memory: Recall (cont’d)

- We can also ask how many items can be stored in $\hat{M}$, i.e., its capacity.
- The capacity is closely related with the rank of the matrix $\hat{M}$. The rank means the number of linearly independent column vectors (or row vectors) in the matrix.
- Linear independence means a linear combination of the vectors can be zero only when the coefficients are all zero:
  \[
  c_1 x_1 + c_2 x_2 + \ldots + c_n x_n = 0
  \]
  only when $c_i = 0$ for all $i = 1, 2, \ldots, n$.
- The above and the examples in the previous pages are best understood by running simple calculations in Octave or Matlab. See the src/ directory for example scripts.

Adaptation

- When the environment is stationary (the statistic characteristics do not change over time), supervised learning can be used to obtain a relatively stable set of parameters.
- If the environment is nonstationary, the parameters need to be adapted over time, on an on-going basis (continuous learning or learning-on-the-fly).
- If the signal is locally stationary (pseudostationary), then the parameters can repeatedly be retrained based on a small window of samples, assuming these are stationary: continual training with time-ordered samples.

Statistical Nature of Learning

- Deviation between the target function $f(x)$ and the neural network realization of the function $F(x, w)$ can be expressed in statistical terms (note $F(\cdot, \cdot)$ is parametrized by the weight $w$).
- Random input vectors $X \in \{x_i\}_{i=1}^N$ and random output scalar values $D \in \{d_i\}_{i=1}^N$.
- Suppose we have a training set $T = \{(x_i, d_i)\}_{i=1}^N$. The problem is that the target values $D$ in the training set may only be approximate ($D \approx f(X)$, i.e., $D \neq f(X)$).
- So, we end up with a regressive model:
  \[
  D = f(X) + \epsilon,
  \]
  where $f(\cdot)$ is deterministic and $\epsilon$ is a random expectational error representing our ignorance.

Statistical Nature of Learning (cont’d)

- The error term $\epsilon$ is typically assumed to have a zero mean:
  \[
  E[\epsilon|x] = 0.\ (E[\cdot|\cdot] \text{ is the expected value of a random variable.})
  \]
- In this light, $f(x)$ can be expressed in statistical terms: $f(x) = E[D|x]$, since from
  \[
  D = f(X) + \epsilon, \text{ we can get}
  \]
  \[
  E[D|x] = E[f(X) + \epsilon] = f(x) + E[\epsilon|x] = f(x).
  \]
- A property that can be derived from the above is that the expectational error term is independent of the regressive function: $E[\epsilon f(X)] = 0$. This will become useful in the following.
Statistical Nature of Learning (cont’d)

- Neural network realization of the regressive model:
  
  \[ Y = F(\mathbf{X}, \mathbf{w}). \]

  We want to map the knowledge in the training data \( \mathcal{T} \) into the weights \( \mathbf{w} \).

- We can now define the cost function:

  \[ \mathcal{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{N} (d_i - F(\mathbf{x}_i, \mathbf{w}))^2 \]

  which can be written equivalently as an average over the training set \( \mathcal{E}_T[\cdot] \):

  \[ \mathcal{E}(\mathbf{w}) = \frac{1}{2} \mathcal{E}_T[(d_i - F(\mathbf{x}, \mathcal{T}))^2] \]

  This reduces to

\[
\begin{align*}
\text{Bias:} & \quad \mathcal{E}_T[(\mathbb{E}[D|x] - F(\mathbf{x}, \mathcal{T}))^2] \\
\text{Variance:} & \quad \mathcal{E}_T[(\mathbb{E}[F(\mathbf{x}, \mathcal{T})] - \mathbb{E}[D|x])^2 + \mathcal{E}_T[(\mathbb{E}[F(\mathbf{x}, \mathcal{T})] - \mathbb{E}[D|x])^2]. \\
\end{align*}
\]

The last step above is obtained using \( \mathcal{E}_T[\mathbb{E}[D|x]^2] = \mathbb{E}[D|x]^2 \), \( \mathcal{E}_T[\mathbb{E}[F(\mathbf{x}, \mathcal{T})]^2] = \mathbb{E}[F(\mathbf{x}, \mathcal{T})]^2 \), and \( \mathcal{E}_T[\mathbb{E}[D|x]F(\mathbf{x}, \mathcal{T})] = \mathbb{E}[D|x] \mathbb{E}[F(\mathbf{x}, \mathcal{T})] \).

* Note: \( \mathbb{E}[c] = c \) and \( \mathbb{E}[cX] = c \mathbb{E}[X] \) for constant \( c \) and random variable \( X \).

Statistical Nature of Learning: Bias/Variance Dilemma

The cost function we derived

\[ \mathcal{E}_T[(f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T}))^2] \]

can be rewritten, knowing \( f(\mathbf{x}) = \mathbb{E}[D|x] \):

\[
\begin{align*}
& \mathcal{E}_T[(\mathbb{E}[D|x] - F(\mathbf{x}, \mathcal{T}))^2] \\
= & \mathcal{E}_T[(\mathbb{E}[D|x] - \mathbb{E}[F(\mathbf{x}, \mathcal{T})] + \mathbb{E}[F(\mathbf{x}, \mathcal{T})] - F(\mathbf{x}, \mathcal{T}))^2] \\
= & (\mathcal{E}_T[F(\mathbf{x}, \mathcal{T})] - \mathbb{E}[D|x])^2 + \mathcal{E}_T[(F(\mathbf{x}, \mathcal{T}) - \mathbb{E}[D|x])^2]. \\
\end{align*}
\]

Bias/Variance Dilemma (cont’d)

- The bias indicates how much \( F(\mathbf{x}, \mathcal{T}) \) differs from the true function \( f(\mathbf{x}) \): approximation error

- The variance indicates the variance in \( F(\mathbf{x}, \mathcal{T}) \) over the entire training set \( \mathcal{T} \): estimation error

- Typically, achieving smaller bias leads to higher variance, and smaller variance leads to higher bias.

\[ d - F(\mathbf{x}, \mathcal{T}) = d - f(\mathbf{x}) + f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T}) = \epsilon + (f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T})). \]

With that,

\[ \mathcal{E}(\mathbf{w}) = \frac{1}{2} \mathcal{E}_T[(d_i - F(\mathbf{x}, \mathcal{T}))^2] \]

becomes

\[ \frac{1}{2} \mathcal{E}_T[(\epsilon + (f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T}))^2] \]

\[ \frac{1}{2} \mathcal{E}_T[\epsilon^2 + 2\epsilon(f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T})) + (f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T}))^2] \]

\[ \frac{1}{2} \mathcal{E}_T[\epsilon^2] + \mathcal{E}_T[\epsilon(f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T})]] + \frac{1}{2} \mathcal{E}_T[(f(\mathbf{x}) - F(\mathbf{x}, \mathcal{T}))^2] \]

Intrinsic error This reduces to 0 We’re interested in this!
Statistical Learning Theory

- Statistical learning theory addresses the fundamental issue of how to control the generalization ability of a neural network in mathematical terms.

- Certain quantities such as sample size and the Vapnik-Chervonenkis dimension (VC dimension) is closely related to the bounds on generalization error.

- The probably approximately correct (PAC) learning model is another framework to study such bounds. In this case, the confidence $\delta$ (probably) and tolerable error level $\epsilon$ (approximately correct) are important quantities. Given these, and other measures such as the VC dimension, we can calculate the sample complexity (how many samples are needed to achieve that level of correctness $\epsilon$ with that much confidence $\delta$).

Three Instances Shattered

Definition: a dichotomy of a set $S$ is a partition of $S$ into two disjoint subsets.

Definition: a set of instances $S$ is shattered by a function class $F$ if and only if for every dichotomy of $S$ there exists some function in $F$ consistent with this dichotomy.

Appendix on VC Dimension

- The concept of Shattering
- VC dimension

Each closed contour indicates one dichotomy. What kind of classifier function can shatter the instances?
The Vapnik-Chervonenkis Dimension

Definition: The Vapnik-Chervonenkis dimension, \( VC(\mathcal{F}) \), of function class \( \mathcal{F} \) defined over sample space \( X \) is the size of the largest finite subset of \( X \) shattered by \( \mathcal{F} \). If arbitrarily large finite sets of \( X \) can be shattered by \( \mathcal{F} \), then \( VC(\mathcal{F}) \equiv \infty \).

Note that \( |\mathcal{F}| \) can be infinite, while \( VC(H) \) finite!

Uses of VC Dimension

- Training error decreases monotonically as the VC dimension is increased.
- Confidence interval increases monotonically as the VC dimension is increased.
- Sample complexity (in PAC framework) increases as VC dimension increases.

VC Dim. of Linear Decision Surfaces

- When \( \mathcal{F} \) is a set of lines, and \( S \) a set of points, \( VC(\mathcal{F}) = 3 \).
- \((a)\) can be shattered, but \((b)\) cannot be. However, if at least one subset of size 3 can be shattered, that's fine.
- Set of size 4 cannot be shattered, for any combination of points (think about an XOR-like situation).