1. Mathematical Foundations of Machine Learning
1.1 Basic Concepts
Definition of Learning

**Definition [Mitchell (1997)]**

“A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$”

...recall that this is not pure mathematics...
The Task $T$

Classification

Compute $f : \mathbb{R}^n \rightarrow \{1, \ldots, k\}$ which maps data $x \in \mathbb{R}^n$ to a category in $\{1, \ldots, k\}$. Alternative: Compute $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$ which maps data $x \in \mathbb{R}^n$ to a histogram with respect to $k$ categories.

$x = \begin{array}{c}
3 & 4 & 2 & 1 & 9 & 5 & 6 & 2 & 1 & 8 \\
8 & 9 & 1 & 2 & 5 & 0 & 0 & 6 & 6 & 4 \\
6 & 7 & 0 & 1 & 6 & 3 & 6 & 3 & 7 & 0 \\
3 & 7 & 7 & 9 & 4 & 6 & 6 & 1 & 8 & 2 \\
2 & 9 & 3 & 4 & 3 & 8 & 7 & 2 & 5 \\
1 & 5 & 9 & 8 & 3 & 6 & 5 & 7 & 2 & 3 \\
9 & 3 & 1 & 9 & 1 & 5 & 8 & 0 & 8 & 9 \\
5 & 4 & 2 & 6 & 8 & 5 & 8 & 8 & 9 & 9 \\
3 & 7 & 7 & 0 & 9 & 8 & 5 & 4 & 3 \\
7 & 9 & 6 & 7 & 0 & 6 & 9 & 2 & 3
\end{array} 
\mapsto f(x) = 5.$
The Task $T$

Regression

Predict a numerical value $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

- Expected claim of insured person
- Algorithmic trading
The Task $T$

### Density Estimation

Estimate a probability density $p : \mathbb{R} \rightarrow \mathbb{R}_+$ which can be interpreted as a probability distribution on the space that the examples were drawn from.

- Useful for many tasks in data processing, for example if we observe corrupted data $\tilde{x}$ we may estimate the original $x$ as the argmax of $p(\tilde{x} | x)$. 
The experience typically consists of a dataset which consists of many examples (aka data points).

- If these data points are labeled (for example in the classification problem, if we know the classifier of our given data points) we speak of *supervised learning*.
- If these data points are not labeled (for example in the classification problem, the algorithm would have to find the clusters itself from the given dataset) we speak of *unsupervised learning*.
The Performance Measure $P$

In classification problems this is typically the accuracy, i.e., the proportion of examples for which the model produces the correct output.

- Often the given dataset is split into a training set on which the algorithm operates and a test set on which its performance is measured.
An Example: Linear Regression

The Task
Regression: Predict $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

The Experience
Training data $((x_{i}^{\text{train}}, y_{i}^{\text{train}}))_{i=1}^{m}$.

The Performance Measure
Given test data $((x_{i}^{\text{test}}, y_{i}^{\text{test}}))_{i=1}^{n}$ we evaluate the performance of an estimator $\hat{f} : \mathbb{R}^n \rightarrow \mathbb{R}$ as the mean squared error

$$\frac{1}{n} \sum_{i=1}^{n} |\hat{f}(x_{i}^{\text{test}}) - y_{i}^{\text{test}}|^2.$$
An Example: Linear Regression

The Computer Program

Define a **Hypothesis Space**

$$\mathcal{H} = \text{span}\{\varphi_1, \ldots, \varphi_l\} \subset C(\mathbb{R}^n)$$

and, given training data

$$z = (x_i, y_i)_{i=1}^m,$$

and the defining the **empirical risk**

$$\mathcal{E}_z(f) := \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2,$$

we let our algorithm find the minimizer (a.k.a. **empirical target function**)

$$f_{\mathcal{H},z} := \arg\min_{f \in \mathcal{H}} \mathcal{E}_z(f).$$
An Example: Linear Regression

Computing the Empirical Target Function

- Let
  $$A = (\varphi_j(x_i))_{i,j} \in \mathbb{R}^{m \times l}.$$  

- Every $$f \in \mathcal{H}$$ can be written as $$\sum_{i=1}^{m} w_i \varphi_i$$ and we denote $$\mathbf{w} := (w_i)_{i=1}^l.$$  

- We let $$\mathbf{y} := (y_i)_{i=1}^m.$$  

- We get that
  $$\mathcal{E}_z(f) = \|A\mathbf{w} - \mathbf{y}\|^2.$$  

- A minimizer is given by $$\mathbf{w}_* := A^\dagger\mathbf{y},$$ and we get our estimate
  $$f_* := \sum_{i=1}^{l} (w_*)_i \varphi_i.$$
Degree too low: underfitting. Degree too high: overfitting!
Bias-Variance Problem

“Capacity” of the hypothesis space has to be adapted to the complexity of the target function and the sample size!
Please brush up on your probability knowledge. Required notions:
Probability measure, expectation, conditional measure, marginal
measure, ...

If you need help, start with Chapter I.3 of
www.deeplearningbook.org.

You may want to read Chapter I.1 of [Cucker-Smale, On the
Mathematical Foundations of Learning]
1.2 Mathematical Foundations of General Regression Problems
1.2.1 Basic Definitions
The Mathematical Learning Problem

Given (compact) $X (\subset \mathbb{R}^d)$ and $Y = \mathbb{R}^k$ and a probability measure $\rho$ on $Z := X \times Y$.

Define the least squares error

$$\mathcal{E}(f) := \int_Z (f(x) - y)^2 d\rho(x, y).$$

The learning problem asks for the function $f$ which minimizes $\mathcal{E}(f)$. 
Example 1: Regression

- Suppose our data is generated from noisy observations
  \[ y = f(x) + \xi, \]
  where \( \xi \) is a r.v. with \( \eta \) as probability density on \( Y = \mathbb{R}^k \) with
  \[ \int_Y y \eta(y) dy = 0. \]
- Then the density of \( \rho \) is given by
  \[ p_\rho(x, y) = \frac{1}{|X|} \eta(y - f(x)). \]
- We have that
  \[ \mathcal{E}(g) = \frac{1}{|X|} \int_{X \times Y} (g(x) - y)^2 \eta(y - f(x)) dy dx = \]
  \[ \frac{1}{|X|} \int_{X \times Y} (g(x) - f(x) - y)^2 \eta(y) dy dx = \frac{1}{|X|} \int_{X \times Y} (g(x) - f(x))^2 \eta(y) dy dx \]
  \[ + \frac{1}{|X|} \int_{X \times Y} y^2 \eta(y) dy dx. \]
- The learning problem finds \( f! \)
Example 2: Classifications

- Suppose that there is a function $f$ which maps a matrix $x \in [0, 1]^{256 \times 256}$ to a histogram $f(x) \in \mathbb{R}^{10}_+$. We consider the vector $f(x) \cdot \sum_{i=1}^{10} f(x)_i$ as a histogram describing which digit the image $x$ represents.

- Let $\rho$ be any measure on $Z = X \times Y$ which generates the measurement data we get to see ($\rho$ will not be known to us!!!)

- Now, a function $f$ as above will in general not exist for our problem. But we can look for the function $f_\rho$ which minimizes the least squares error $\mathcal{E}$ – this will be the optimal explanation of the measurements in terms of a functional relation between $X$ and $Y$!
Regression Function

Conditional Probability

Let \( p_\rho \) be the density of \( \rho \). Then the measure \( \rho(\cdot|x) \) on \( Y \) has density

\[
\frac{1}{\int_Y p_\rho(x, y) \, dy} p_\rho(x, \cdot).
\]

Regression Function

The \textit{regression function} is defined as

\[
f_\rho(x) := \int_Y y \, d\rho(y|x)
\]

Variance

The \textit{variance} is defined as

\[
\sigma^2_\rho := \mathcal{E}(f_\rho).
\]
The Regression Function solves the Learning Problem

Marginal
The marginal of $\rho$ on $X$ is defined as the measure

$$\rho_X(A) := \rho(A \times Y).$$

Proposition
For every $f : X \rightarrow Y$ it holds that

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 d\rho_X(x) + \sigma^2_\rho.$$

Corollary
The regression function solves the learning problem!

Are we done??

😢 We don’t know $\rho$!!!
The Next Goal

We do have access to a finite number of random samples of $\rho$.

**Goal**

Approximate (or “learn”) the regression function $f_\rho$ from a finite number of random samples w.r.t. the measure $\rho$ on $Z$. 
1.2.2 Empirical Minimization, Hypothesis Space, and Bias-Variance Tradeoff
Empirical Error

Given $z := ((x_1, y_1), \ldots, (x_m, y_m)) \in Z^m$ be i.i.d. w.r.t. $\rho$ samples. Define the empirical error

$$\mathcal{E}_z(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$  

The empirical error can actually be computed!

Defect

The defect of $f$ is defined as

$$L_z(f) := \mathcal{E}(f) - \mathcal{E}_z(f).$$  

Can we control the defect? If yes, we actually have some hope of approximating the regression function.
Concentration Inequalities

Bernstein Inequality

Suppose that $\xi$ is a r.v. on a probability space $(Z, \rho)$ with mean $\mathbb{E}(\xi) = \mu$ and variance $\mathbb{V}(\xi) = \sigma^2$. Suppose that $|\xi(z) - \mu| \leq M$ with probability 1. Then

$$\mathbb{P}_{z \in Z^m} \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \varepsilon \right\} \leq 2e^{-\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3} M\varepsilon)}}.$$
Bounding the Defect

Theorem

For \( f : X \rightarrow Y \) define \( f_Y : Z \rightarrow Y \) via \( f_Y(x, y) = f(x) - y \) and let \( \sigma^2 = \text{Var}(f_Y^2) \). Suppose that \( |f(x) - y| \leq M \) almost everywhere. Then, for all \( \varepsilon > 0 \),

\[
\mathbb{P}_{z \in Z^m} \{ |L_z(f)| \leq \varepsilon \} \geq 1 - 2e^{-\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3} M \varepsilon)}}.
\]

Are we done??

😃 Any \( f \) vanishing on the sample points makes the empirical error vanish!!!
Hypothesis Space

**Definition**
Let $\mathcal{H}$ be a compact subset of the Banach space $\{f : X \to Y, \text{ continuous}\}$ with norm $\|f\| := \max_{x \in X} |f(x)|$. We call $\mathcal{H}$ hypothesis space or model space.

**Target Function in $\mathcal{H}$**
Define the *target function in $\mathcal{H}$* via

$$f_{\mathcal{H}} := \arg\min_{f \in \mathcal{H}} \mathcal{E}(f).$$

**Empirical Target Function**
Given $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m)) \in Z^m$ random samples, define the *empirical target function* as

$$f_{\mathcal{H}, \mathbf{z}} := \arg\min_{\mathcal{H}} \mathcal{E}_\mathbf{z}(f).$$

The empirical target function can be computed!
Sample- and Approximation Error

For a hypothesis class $\mathcal{H}$ let

$$f_\mathcal{H} := \arg\min_{f \in \mathcal{H}} \mathcal{E}(f) = \arg\min_{f \in \mathcal{H}} \int_X (f(x) - f_\rho(x))d\rho_X(x).$$

For $f \in \mathcal{H}$ let

$$\mathcal{E}_\mathcal{H}(f) := \mathcal{E}(f) - \mathcal{E}(f_\mathcal{H}) \geq 0.$$

The empirical error $\mathcal{E}(f_{\mathcal{H},z})$ decomposes into

$$\mathcal{E}(f_{\mathcal{H},z}) = \mathcal{E}_\mathcal{H}(f_{\mathcal{H},z}) + \mathcal{E}(f_\mathcal{H}).$$

The first term is called sample error and the second term is called approximation error.

- Our goal is to make the empirical error as small as possible!
- Think about what happens if we fix $m$, the number of samples and enlarge the hypothesis space $\mathcal{H}$. 
Figure: Blue: $f_{\mathcal{H}}$, Red: $f_{\mathcal{H},z}$, $m = 10$, $\mathcal{H} =$ polynomials of degree 5, 15, 20 (from top left to bottom).

If $\mathcal{H}$ is too complex, the sampling error increases.
The Bias-Variance Trade-Off

If we keep the sample size $m$ fixed and enlarge the hypothesis space $\mathcal{H}$, the approximation error will certainly decrease, **BUT** the sample error will increase – this is exactly what we observed experimentally!

Bishop [Neural Networks for Pattern Recognition (1995)]

“A model which is too simple, or too inflexible, will have a large bias, while one which has too much flexibility in relation to the particular data set will have a large variance. Bias and variance are complementary quantities, and the best generalization is obtained when we have the best compromise between the conflicting requirements of small bias and small variance.”

Bias-Variance Problem

What are the precise relations between the number of samples $m$ and the “capacity” of our hypothesis space $\mathcal{H}$?
Covering Numbers

**Definition**

Let $S$ be a metric space and $s > 0$. Define the *covering number* $\mathcal{N}(S, s)$ to be the minimal $l \in \mathbb{N}$ such that there exist $l$ disks in $S$ with radius $s$ covering $S$.

Scaling of $\mathcal{N}(S, s)$ with $s$ is a measure of complexity of $S$ termed *metric entropy*. 
Abstract Analysis of Sample Error

**Theorem**
Let $\mathcal{H} \subset C(X)$ be a hypothesis class. Assume that for all $f \in \mathcal{H}$ it holds that $|f(x) - y| < M$ a.e. Then, for all $\varepsilon > 0$,

$$\mathbb{P}_{z \in Z^m} \left( \sup_{f \in \mathcal{H}} |L_z(f)| \leq \varepsilon \right) \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{8M}) 2e^{-\frac{m\varepsilon^2}{4(2\sigma^2 + \frac{1}{3}M^2 \varepsilon)}},$$

where $\sigma^2 := \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$.

Proof is left as an exercise!
Abstract Analysis of Sample Error

**Lemma**

Let $\varepsilon > 0$ and $0 < \delta < 1$ such that

$$P_{z \in Z^m} \left( \sup_{f \in \mathcal{H}} |L_z(f)| \leq \varepsilon \right) \geq 1 - \delta.$$

Then

$$P_{z \in Z^m} (\mathcal{E}_\mathcal{H}(f_{\mathcal{H}, z}) \leq 2\varepsilon) \geq 1 - \delta.$$

**Theorem**

Let $\mathcal{H}$ be a hypothesis class. Assume that for all $f \in \mathcal{H}$ it holds that $|f(x) - y| < M$ a.e. Then, for all $\varepsilon > 0$,

$$P_{z \in Z^m} (\mathcal{E}_\mathcal{H}(f_{\mathcal{H}, z}) \leq \varepsilon) \geq 1 - N(\mathcal{H}, \frac{\varepsilon}{16M}) 2e^{-\frac{m\varepsilon^2}{8(2\sigma^2 + \frac{1}{3}M^2 \varepsilon)}},$$

where $\sigma^2 := \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$. 
Abstract Analysis of Sample Error

**Question**

Given $\varepsilon, \delta > 0$, how many samples $m$ do we need such that the probability that the sample error is $\leq \varepsilon$ is at least $1 - \delta$?

**Answer**

By the previous theorem it suffices to choose

$$m \geq \frac{8(4\sigma^2 + \frac{1}{3}M^2\varepsilon)}{\varepsilon^2} \left( \ln(2\mathcal{N}(\mathcal{H}, \frac{\varepsilon}{16M})) + \ln\left(\frac{1}{\delta}\right) \right).$$

**Question**

How to bound the covering number?
Recall

$$\mathcal{H}_R = \text{span}\{\varphi_1, \ldots, \varphi_l\} \cap \{f \in C(X): \|f\| \leq R\} \subset C(X).$$

**Theorem**

Let $T := \|\sum_{j=1}^{l} |\varphi_j|\|$. Then

$$\ln(N(\mathcal{H}_R, \eta)) \leq l \cdot \ln \left(\frac{4RT}{\eta}\right).$$

In the motivational section on linear regression we have seen that $f_{\mathcal{H},z}$ can be found by solving an $l$-dimensional linear system.
1.2.3 Reproducing Kernel Hilbert Spaces (RKHS)
Reproducing Kernel Hilbert Spaces (RKHS)

Motivation
Suppose we have a ‘similarity measure’ $K : X \times X \to \mathbb{R}$ on $X$ and we would like to do things like nearest neighbour search, PCA etc. with respect to this measure of similarity. One idea is to associate each $x \in X$ with a feature map $\Phi_x$ which is an element of a high-dimensional inner-product-space, and which ‘linearizes’ the similarity measure in the sense that

$$K(x, x') = \langle \Phi_x, \Phi_{x'} \rangle.$$

Question
Which conditions on $K$ guarantee the existence of a feature map?
Mercer Theorem

**Definition**

\( K : X \times X \to \mathbb{R} \) is symmetric if \( K(x, x') = K(x', x) \) for all \( x, x' \in X \). Let \( x = \{x_1, \ldots, x_k\} \subset X \) and \( K[x] \in \mathbb{R}^{k \times k} \) with entries \( K(x_i, x_j) \) the Gramian of \( K \) at \( x \). \( K \) is called positive semidefinite if its Gramian is always positive semidefinite. \( K \) is called a **Mercer kernel** if it is symmetric, positive semidefinite and continuous.

**Theorem**

There exists a unique Hilbert space \( \mathcal{H}_K \) of functions on \( X \) satisfying

1. The functions \( K_x : x' \mapsto K(x, x') \) are in \( \mathcal{H}_K \),
2. the span of the \( K_x \)'s is dense, and
3. for all \( f \in \mathcal{H}_K \) we have \( f(x) = \langle f, K_k \rangle \).

In particular, \( K(x, x') = \langle K_x, K_{x'} \rangle \) and in this sense, the RKHS \( \mathcal{H}_K \) can be regarded as a feature space.
Examples I

**Dot-Product Kernels**

Let $X$ be the ball of radius $T$ in $\mathbb{R}^n$ and $K(x, x') = \sum_{d=1}^{\infty} a_d (x \cdot x')$, where $a_d \geq 0$ for all $d$ and $\sum_d a_d T^{2d} < \infty$. Then $K$ is a mercer kernel on $X$ called a *dot product kernel*.

**Example**

Suppose that $X$ is as above and $K(x, x') = 1 + x \cdot x'$. Then $\{1, x_1, \ldots, x_n\}$ constitutes an ONB of $\mathcal{H}_K$. 
Translation-Invariant Kernels

Suppose that \( k : \mathbb{R}^n \to \mathbb{R} \) is such that its Fourier transform is real-valued and non-negative. Then \( K(x, x') := k(x - x') \) is a Mercer kernel, called a translation-invariant kernel.

Example

Let \( k = \chi_{[-1,1]} \ast \chi_{[-1,1]} \) the cardinal B-spline of degree one. Then

\[
K(x, x') = \begin{cases} 
1 - \frac{|x - x'|}{2} & |x - x'| \leq 2 \\
0 & \text{else}
\end{cases}
\]
Suppose that $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ is completely monotonic (i.e. $(-1)^k f^{(k)}(x) \geq 0$). Then $K(x, x') := f(|x - x'|^2)$ is a mercer kernel, called a \textit{RBF kernel}.

\textbf{Example}

A Gaussian $f(t) := e^{-t/c^2}$ and an inverse multiquadric $(c^2 + |t|)^{-\alpha}, \alpha > 0$ are completely monotonic and define corresponding RBF kernels.
Covering Numbers

**Theorem**

For $R > 0$ denote $B_R$ the ball of radius $R$ in a RKHS $\mathcal{H}_K$. Then $B_R$ is a compact subset of $C(X)$ and thus a valid hypothesis space.

**Theorem**

If $K \in C^s$ then

$$\ln(\mathcal{N}(B_R, \eta)) \leq C \cdot \text{diam}(X)^n \|K\|_{C^s}^{n/s} \left(\frac{R}{\eta}\right)^{2n/s}.$$  

For specific kernels (such as Gaussian RBF kernels), much better results exist.
Computational Issues

Question
How can we determine $f_{\mathcal{H}, z}$?

Let $\mathcal{H}_{K,z} := \text{span}\{K_{x_1}, \ldots, K_{x_m}\}$ and $P : \mathcal{H}_K \to \mathcal{H}_{K,z}$ the orthogonal projection. Then, since $f(x_i) = \langle f, K_{x_i} \rangle = \langle P(f), K_{x_i} \rangle = P(f)(x_i)$ we have $\mathcal{E}_z(f) = \mathcal{E}_z(P(f))$!!!

Theorem
We have $f_{\mathcal{H}, z} = \sum_{i=1}^{m} c_i^* K_{x_i}$, where $(c_i^*)$ is a minimizer of

$$\frac{1}{m} \sum_{j=1}^{m} \left( \sum_{i=1}^{m} c_i K(x_i, x_j) - y_j \right)^2 \quad \text{s.t.} \quad c^T K[x] c \leq R^2.$$ 

This is a convex quadratic program that can be efficiently solved by interior point methods! Check out http://cvxr.com/cvx/!
1.2.4 Regularization
Theorem

Let $U$ be a Hilbert space and $F, H : U \to \mathbb{R}$ smooth. Let $c \in U$ be a solution of the problem

$$\min F(f) \quad \text{s.t.} \quad H(f) \leq 0.$$ 

Then there exist $\mu, \lambda \in \mathbb{R}$ with

$$\mu DF(c) + \lambda DH(c) = 0.$$ 

If $H(c) < 0$ then $\lambda = 0$ and $\mu \neq 0$ (we would have an unconstrained stationary point). If $DH(c) \neq 0$ then we can take $\mu = 1$ and $\lambda \geq 0$. 
Let $z$ be such that $K[x]$ is invertible and denote $f_{BR,z}$ as above as minimizer of $F = \mathcal{E}_z$ s.t. $H(f) \leq 0$ with

$$H(f) := \|f\|^2_{\mathcal{H}_K} - R^2.$$

Let $f^*$ be the unconstrained minimizer of $\mathcal{E}_z$ in $\mathcal{H}_K$ such that $R_0 := \|f^*\|_{\mathcal{H}_K}$ is minimal.

Since $K[x]$ is invertible, one can show that for all $R < R_0$ we have $DF(f_{BR,z}) \neq 0$ and $DH(f_{BR,z}) \neq 0$, and therefore, by the result on Lagrange multipliers, there exists $\lambda(R) > 0$ with

$$DF(f_{BR,z}) + \lambda(R)DH(f_{BR,z}) = 0.$$

**Theorem**

The function $f_{BR,z}$ is also a solution to the unconstrained regularized problem to minimize

$$\mathcal{E}_z(f) + \lambda(R)\|f\|^2_{\mathcal{H}_K}.$$
Let $z$ be such that $K[x]$ is invertible and denote $f_{\lambda,z}$ as above as the minimizer of the convex unconstrained problem to minimize

$$\mathcal{E}_z(f) + \lambda \cdot \|f\|_{\mathcal{H}_K}^2.$$ 

Let $R(\lambda) := \|f_{\lambda,z}\|_{\mathcal{H}_K}$.

**Theorem**

The function $f_{\lambda,z}$ is also a solution to the constrained regularized problem to minimize

$$\mathcal{E}_z(f) \quad \text{s.t.} \quad \|f\|_{\mathcal{H}_K}^2 \leq R(\lambda)^2.$$ 

Regularized and Constrained Least Squares are equivalent!
Numerical Solution of Regularized Least Squares

Theorem

We have \( f_{\lambda,z} = \sum_{i=1}^{m} c^*_i Kx_i \), where \( c^* = (c_i^*) \) is a solution of the regular linear equation

\[
(\lambda mI + K[x]) \mathbf{c} = \mathbf{y}.
\]

Extremely easy to implement!
**Figure:** Example for Kernel regression with Gaussian kernel. In clockwise order: Correct choice of $\lambda$, overfitting ($\lambda$ too small), underfitting ($\lambda$ too large).
It’s really simple

```matlab
function G = GramAssemble(x, kernel)

% Assemble the Gram matrix associated with data points x.
% The parameters are a data vector x and an inline function kernel.
% Output is the Gram matrix

n = size(x, 2);
G = zeros(n);
for i = 1:n
    for j = 1:n
        G(i, j) = kernel(x(:, i), x(:, j));
    end
end
```
It’s really simple

```matlab
function v=KernelInterpol(x,y,t,kernel,gamma)

%the regularized least squares problem is solved.
%inputs are sample data x,y, an inline function kernel which describes
%the kernel and a regularization parameter gamma.
%output is a vector v which is fiven by the estimated regression function,
evaluated
%on a vector t of test points.

n = length(x);

%solve regularized least squares
G = GramAssemble(x,kernel);
A = G + n*gamma*eye(n);
c = pinv(A)*y';

%create function values of empirical target function at grid t
v = zeros(size(t,2),1);
for i = 1:n
    v = v+c(i)*kernel(repmat(squeeze(x(:,i)),1,size(t,2)),t);
end
```
1.2.5 A Bayesian Interpretation
Bayes’ Theorem

- Suppose that the probability of measuring $f \in \mathcal{H}_K$ is equal to
  $$\mathbb{P}(f) = c \cdot \exp(-\|f\|_H^2).$$

- Suppose that we observe a function $f$, corrupted with Gaussian noise. Then the probability of making the observations $z = ((x_1, y_1), \ldots, (x_m, y_m))$, from the signal $f$ is equal to
  $$\mathbb{P}(z|f) = d \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (f(x_i) - y_i)^2\right).$$

- Bayes’ Theorem yields that
  $$\mathbb{P}(f|z) = \frac{\mathbb{P}(z|f) \cdot \mathbb{P}(f)}{\mathbb{P}(z)} \sim \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (f(x_i) - y_i)^2 - \|f\|_H^2\right).$$
The MAP Estimate maximizes the a posteriori probability $\mathbb{P}(f|z)$, given an a priori distribution on $f$ and on the noise.

For the a priori distribution $\mathbb{P}(f) = c \cdot \exp(-\|f\|_H^2)$ and Gaussian noise, the solution of the regularized least squares problem is also the MAP estimate!
1.3 Classification
1.3.1 Regularized Classification
We now aim at classifying data into two classes and thus look for \( f : X \to \{-1, 1\} \). Therefore, let’s put \( Y = \{-1, 1\} \) and \( Z := X \times Y \).

**Misclassification Error**

Given a probability measure \( \rho \) on \( Z \) and \( f : X \to Y \), define the *misclassification error* as

\[
\mathcal{R}(f) := \mathbb{P}_{z \in Z}(f(x) \neq y) = \int_X \mathbb{P}(f(x) \neq y | x) d\rho_X
\]

The classification problem asks to minimize the misclassification error.
Bayes Rule

Define the *Bayes rule* as

\[ f_c := \text{sgn}(f_\rho). \]

Theorem

The Bayes rule minimizes the misclassification error

We can re–use everything!
Figure: Bayes Rule for Gaussian Kernel regression. Left: sample data. Right: Estimate using Bayes Rule.
Case Study: Breast Cancer Detection
Size of dataset $m = 683$ and dimensionality of feature space $d = 10$.

We obtain 95 percent classification accuracy from only 68 training samples and linear kernel (see MATLAB codes on the webpage)!
Visualizing Data

Data is very well-clustered...

But how did we obtain this visualization of our 10-dimensional dataset?
1.3.2 (Kernel) Principal Component Analysis (PCA)
Dimensionality Reduction

Dimensionality Reduction Problem

- Given dataset \( \mathbf{x} = (x_i)_{i=1}^{m} \subset \mathbb{R}^d \) with \( d \) large.
- Goal: Construct map \( \Phi : \mathbb{R}^d \to \mathbb{R}^s \) with \( s \ll d \) such that the features of the dataset \( \mathbf{x} \) are preserved under the mapping \( \Phi \).
- Useful for reduction in computational complexity, visualization (if \( s \leq 3 \)) or de-noising/compression.

Simplest case: \( \Phi \) is orthogonal projection onto affine subspace \( \sim \) PCA.
What is a good projection?

Pick subspace which maximizes variance of the projected dataset.
PCA

PCA Problem

Look for $s$-dimensional affine subspace with associated orthogonal projection $\Phi$ such that the variance $\sum_{i=1}^{m} |\Phi(x_i - \frac{1}{m} \sum_{j=1}^{m} x_j)|^2$ is maximized.

PCA Solution

- Suppose w.l.o.g. (why?) that the data is centered, i.e., $\frac{1}{m} \sum_{j=1}^{m} x_j = 0$.
- Define the empirical covariance matrix
  $$G = \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T \in \mathbb{R}^{d \times d}.$$  
- Then the solution is given by the subspace spanned by the first $s$ normalized Eigenvectors $u_1, \ldots, u_s$ of $G$ and
  $$\Phi(x) = \sum_{l=1}^{s} (x \cdot u_l) u_l.$$
It’s really simple

function z=pca(X)
%project data X on its %principal components.
C=cov(X);
[U,D,pc] = svd(C);
z = center(X)*pc;
scatter(z(:,1),z(:,2));%plot 2d projection
When PCA fails...
When PCA fails...

Projection on the primary eigenvector

- Original data
- Projected data
- Projection error
Construct nonlinear $\Phi$ by applying linear PCA on RKHS $\mathcal{H}_K$ by mapping data points $x_i$ to their feature vectors $Kx_i$!

This is a powerful idea, called *kernelization*!

Define the matrix $G = K[x] - 1_mK[x] - K[x]1_m + 1_mK[x]1_m \in \mathbb{R}^{m \times m}$ and $(1_m)_{i,j} = \frac{1}{m}$ for $i, j \in \{1, \ldots, m\}$ and denote $u_1, \ldots, u_s$ the first $s$ normalized (w.r.t. the inner product $u^T K[x] u$) Eigenvectors of $G$. Then the projection $\Phi$ is defined as

$$\Phi(x) = \left( \sum_{i=1}^{m} (u_1)_i K(x_i, x), \ldots, \sum_{i=1}^{m} (u_s)_i K(x_i, x) \right)^T.$$
Example

Concentric circles

First principal component after Linear PCA

gamma = 15

First principal component after RBF Kernel PCA

gamma = 15
Kernel PCA Denoising

<table>
<thead>
<tr>
<th>Gaussian noise</th>
<th>‘speckle’ noise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>orig.</strong></td>
<td>0123456789</td>
</tr>
<tr>
<td><strong>noisy</strong></td>
<td>0124567890</td>
</tr>
<tr>
<td><strong>n = 1</strong></td>
<td>0123456789</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>0123456789</td>
</tr>
<tr>
<td><strong>16</strong></td>
<td>0123456789</td>
</tr>
<tr>
<td><strong>64</strong></td>
<td>0123456789</td>
</tr>
<tr>
<td><strong>256</strong></td>
<td>0123456789</td>
</tr>
</tbody>
</table>

**Figure:** Top: Linear PCA reconstruction from $n$ principal components. Bottom: Gaussian Kernel Reconstruction from $n$ principal components (find $z$ with $\|Kz - \Phi(x)\|_{\mathcal{H}_K}$ minimal).

Other methods: Multidimensional Scaling, Isomap, Diffusion Maps, ...

Try to appreciate the power of kernelization!

Go to [https://archive.ics.uci.edu/ml/datasets.html](https://archive.ics.uci.edu/ml/datasets.html) for further datasets and play around with them!
1.3.3 (Kernel) Support Vector Machine (SVM)
Basic Idea

- Suppose that data points \((x_i)_{i=1}^m \subset \mathbb{R}^n\) to be classified are *linearly separable*, i.e. there exists a separating hyperplane defined by \(w \in \mathbb{R}^n, |w| = 1\) and \(b \in \mathbb{R}\) such that

  \[ y_i = 1 \iff w \cdot x_i > b. \]

- Define the *margin* of a separating hyperplane defined by \(w, b\) as above by

  \[ \Delta(w, b) := \min_{i=1}^m |w \cdot x_i - b|. \]

*Try to find separating hyperplane with maximal margin!*
The SVM problem can be formalized by the following minimization problem

\[
\arg\min_{w,b} |w| \quad \text{s.t.} \quad y_i (w \cdot x_i - b) \geq 1 \quad \text{for all} \quad i \in \{1, \ldots, m\}.
\]

Data is in general not linearly separable...

**Soft Margin SVM**

Relax to

\[
\arg\min_{w,b} \frac{1}{m} \sum_{i=1}^{m} \Phi_{hl}(y_i (w \cdot x_i - b)) + \lambda |w|^2,
\]

where \( \Phi_{hl}(t) := \max(0, 1 - t) \), the **hinge loss**.

This is a convex quadratic program that can be efficiently solved!
Kernelization yields the problem

$$\text{argmin}_{f \in \mathcal{H}_K, b} \frac{1}{m} \sum_{i=1}^{m} \Phi_{hl}(y_i(f(x_i) - b)) + \lambda \| f \|_2^2$$

Separable Measures

This provably works for separable measures $\rho$ in the sense that there is $f_s \in \mathcal{H}_K$ with $y f(x) > 0$ almost surely. It means that data is separated by the zero level set of $f_s$. Clearly, the Bayes classifier is that equal to $\text{sgn}(f_s)$.

For most data points the hinge loss will be zero which implies that $f$ will be sparse in $\{K_{x_i} : i = 1, \ldots, m\}$, resulting in potentially big computational savings!

Main advantage as compared to classification method in Section 1.2 is that solution will be sparse.

Experiment and Compare!