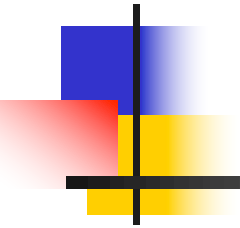
CMSC 510
Regularization Methods for Machine Learning

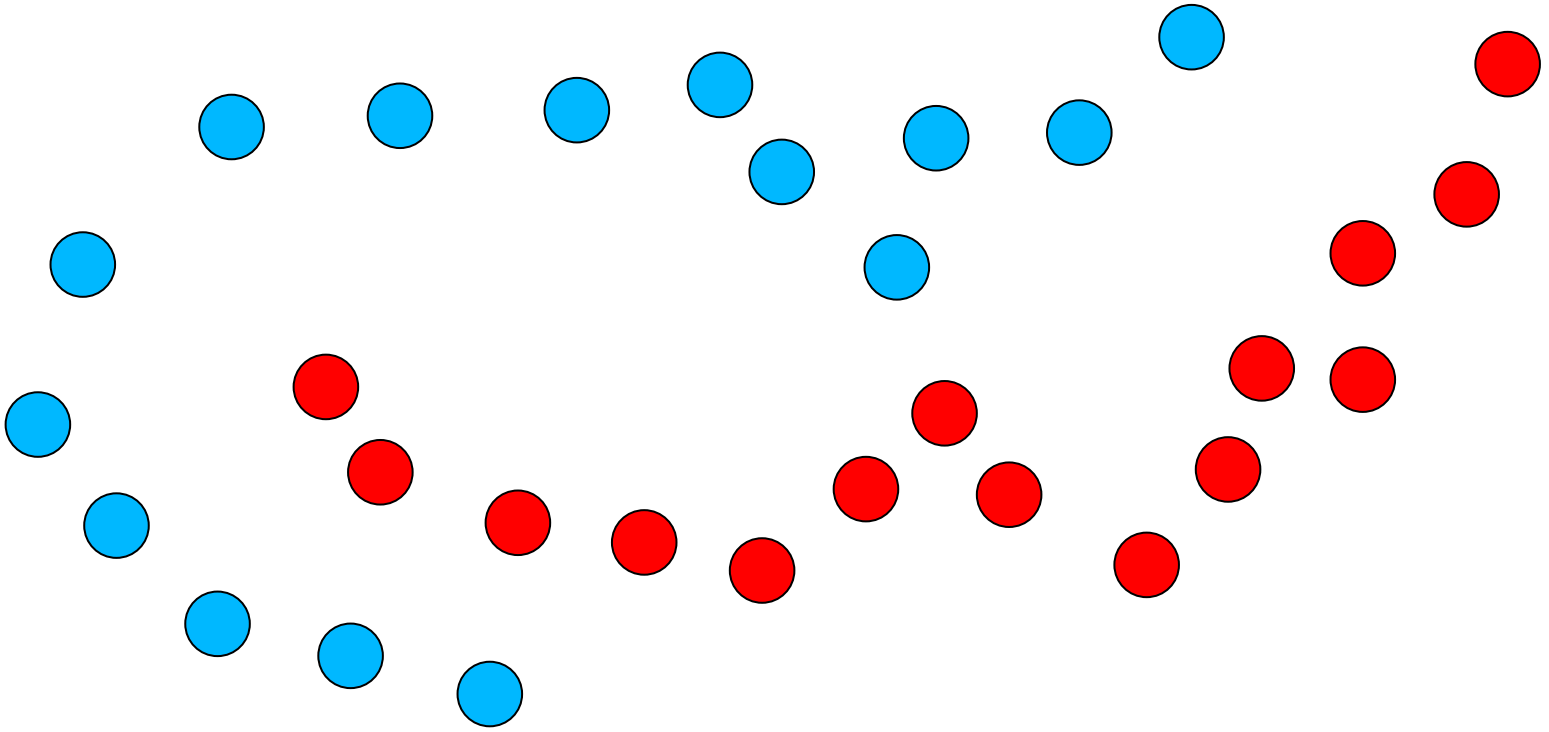


Representer Theorem

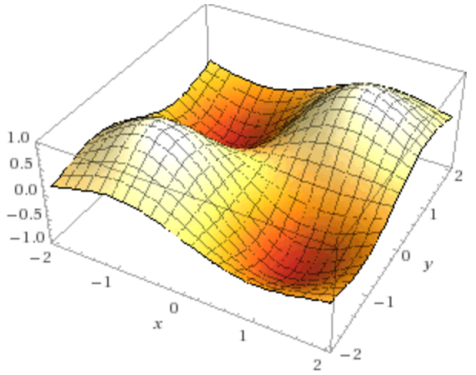
Instructor:
Dr. Tom Arodz

Naïve approach

- We have training samples



Naïve approach



- We place a Gaussian (or other function, depending on kernel) at each sample

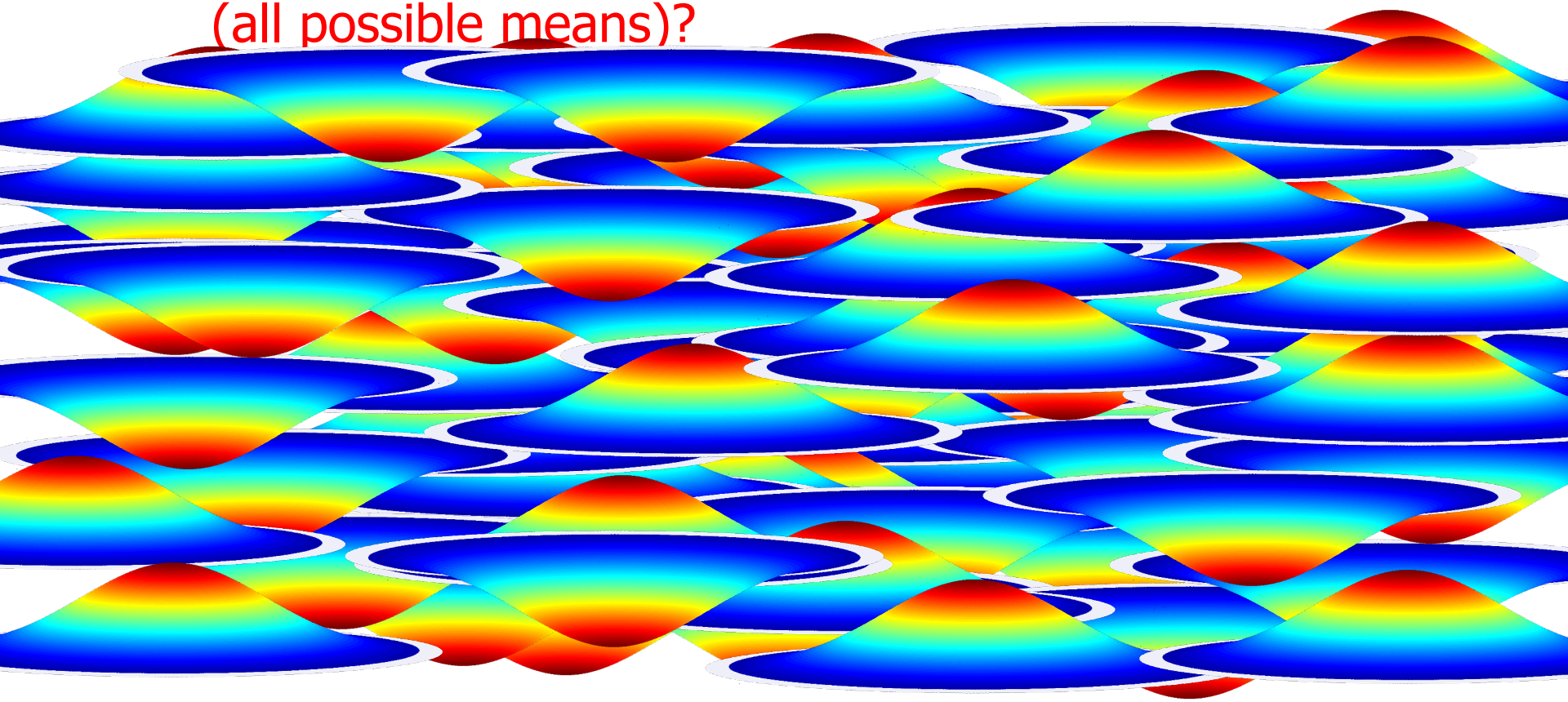
$h(x)=\sum_j a_j\, y_j \exp(-||x-x_j||^2)$

Problems:

- Gaussians – why only at samples?
- We can increase a_j towards infinity (taller Gaussians) and get higher values of h(x)
  - And reduce our loss

Naïve approach

- Do we need to search through the whole space H?
  - functions **h'**: e.g. combinations of all possible Gaussians (all possible means)?

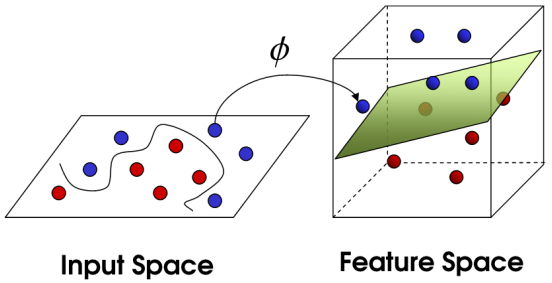


Representer Theorem

- What is a linear classifier?
  - $h(x)=0.3*\langle x_1,x\rangle + 0.2\langle x_2,x\rangle + \ldots$ is a linear classifier with $w=0.3x_1+0.2x_2+\ldots$
- If we only consider $w= \sum_j c_j x_j$ can we represent any w?
- No: if we have three samples: $x_1=[1,1,0]$, $x_2=[1, 2, 0]$, $x_3=[.5, 1.7, 0]$ we can't produce $w=[1,1,1]$.
- To allow arbitrary w we need to add orthogonal complement to x's:

  $$w= \sum_j c_j x_j + w^{\perp}$$

  - Here: $w^{\perp} = [0,0,\text{something}]$
    - Any vector v that has $\langle v,x_i\rangle=0$ for all x_i

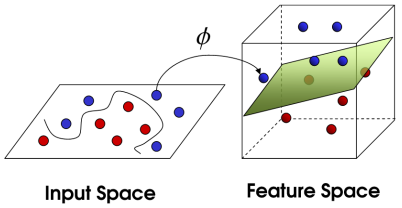


Say w is 3D, but x_i are on the plane. *Orthogonal complement* is all vectors that go straight out of the screen at

Representer Theorem in RKHS

- We're solving:

$$\min_{h \in \mathcal{H}} \quad C \sum_{i=1}^{m} \ell(y_i, h(x_i), b) + \frac{1}{2}||h||_{\mathcal{H}}^2$$



- To search through whole space H, we need to add $h^{\perp}$.

$$h'(x) = \sum_{j=1}^{m} c_j K_{x_j}(x) + h^{\perp}(x) \qquad h' = h + h^{\perp}$$

- We took Gaussians centered at samples: $K_{xj} = K(x_j,x) = \exp(-||x-x_j||^2)$
- And added orthogonal complement to all of them

$$\langle K_{x_j}, h^{\perp} \rangle_{\mathcal{H}} = 0.$$

  - Orthogonal complement of subspace spanned by K_{xi}: any function that is orthogonal to all K_{xi}
  - Together, they span the whole space H
- Now any function in H is included in the search
- Let's compare solution quality for h (from subspace) and for h' (from whole H)

$$h' = h + h^{\perp}$$

Say H is 3D, but K_t are on the plane. Orthogonal complement is all vectors that go straight out of the screen at

Representer Theorem

- We're solving:

$$\min_{h \in \mathcal{H}} \quad C \sum_{i=1}^{m} \ell(y_i, h(x_i), b) + \frac{1}{2}||h||_{\mathcal{H}}^2$$

- Using functions:

$$h'(x) = \sum_{j=1}^{m} c_j K_{x_j}(x) + h^{\perp}(x)$$

- Value of the loss:

$$h'(x_i) = \langle K_{x_i}, \sum_{j=1}^{m} a_j K_{x_j}(x_i) + h^{\perp}(x_i) \rangle_{\mathcal{H}} = \sum_{j=1}^{m} a_j \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}} + \langle K_{x_i}, h^{\perp} \rangle_{\mathcal{H}}$$

$$= \sum_{j=1}^{m} a_j \langle K_{x_i}, K_{x_j} \rangle_{\mathcal{H}} = \langle K_{x_i}, \sum_{j=1}^{m} a_j K_{x_j} \rangle_{\mathcal{H}} = h(x_i),$$

- $h'(x_i)=h(x_i)$, same loss on training set!

Representer Theorem

- We're solving:

$$\min_{h \in \mathcal{H}} \quad C \sum_{i=1}^{m} \ell(y_i, h(x_i), b) + \frac{1}{2}||h||_{\mathcal{H}}^2$$

- Value of squared norm ||h||^2

$$||h||_{\mathcal{H}} = \sqrt{\langle h, h \rangle_{\mathcal{H}}}$$

$$h' = h + h^{\perp}$$

<,> can be negative in general, but we have orthogonality!

$$||h'||_{\mathcal{H}}^2 = ||h + h^{\perp}||_{\mathcal{H}}^2 = ||h||_{\mathcal{H}}^2 + 2\langle h, h^{\perp} \rangle_{\mathcal{H}} + ||h^{\perp}||_{\mathcal{H}}^2 = ||h||_{\mathcal{H}}^2 + ||h^{\perp}||_{\mathcal{H}}^2$$

- h' has higher (or =) value of the norm than h
  - h is a better solution w.r.t. ||h||^2 term
- And has the same loss (prev. slide)

No need to search through whole space H

If we stick to just Gaussians centered at training points, we get optimal solution (risk + ||h||^2)