Kernels

CPSC 532S: Modern Statistical Learning Theory
7 March 2022
cs.ubc.ca/~dsuth/532S/22/

$f(x) = w^T x$

$\text{sigh}(w^T x) \in \{\pm 1\}$

fiddle at 0

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1$$

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i)$$

$$\geq \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality;**

$\geq$ for **any** problem

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality;**

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i)$$

w optimization problem is differentiable + unconstrained

but here we have

strong duality:

it's equal

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality;**

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^{n} (-\alpha_i y_i x_i) = 0$$

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^{n} (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^{n} (-\alpha_i y_i x_i) = 0$$

$$\|w\| = \left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)^2$$

$$= \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i x_i^\top x_j y_j \alpha_j$$

$$\sum_i \alpha_i y_i x_i^\top \left(\sum_j \alpha_j y_j x_j\right)$$

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^{n} (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) \underbrace{X}_{n \times d} \underbrace{X^\top}_{d \times n} \operatorname{diag}(y) \alpha$$

Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^{n} (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

$$w = X^\top \operatorname{diag}(y) \alpha$$

Hard SVM Duality

$$\min_w \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_w \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i)$$

$$= \max_{\alpha_i \geq 0} \min_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**;

$\geq$ for **any** problem

but here we have

strong duality:

it's equal

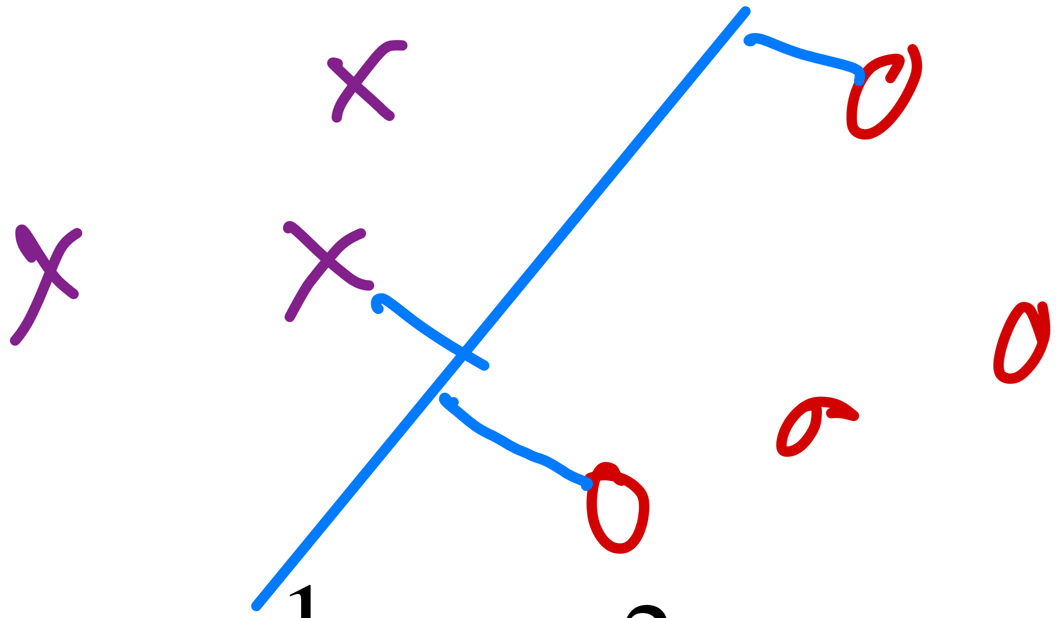
w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^n (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

$$w = X^\top \operatorname{diag}(y)\alpha \qquad w^\top x = \alpha^\top \operatorname{diag}(y) \overbrace{X}^{n\times d} \overbrace{x}^{d\times 1}$$



Hard SVM Duality

$$\min_{w} \frac{1}{2}\|w\|^2 \text{ s.t. } \forall i,\ y_i w^\top x_i \geq 1 \quad = \min_{w} \max_{\alpha_i \geq 0} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

called **weak (Lagrange) duality**; $\geq$ for **any** problem

$$= \max_{\alpha_i \geq 0} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i)$$

but here we have

strong duality:

it's equal

w optimization problem is differentiable + unconstrained

setting gradient to zero:

$$w + \sum_{i=1}^{n} (-\alpha_i y_i x_i) = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$= \max_{\alpha_i \geq 0} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i x_i^\top x_j y_j \alpha_j \quad = \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha$$

$$w = X^\top \operatorname{diag}(y) \alpha \qquad w^\top x = \alpha^\top \operatorname{diag}(y) X x$$

α_i is zero if $y_i w^\top x_i > 1$

Soft SVM Duality

$$\min_{w,\xi} \lambda \|w\|^2 + \frac{1}{n} \sum_i \xi_i \quad \text{s.t. } \forall i, \; y_i w^\top x_i \geq 1 - \xi_i, \;\; \xi_i \geq 0$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ \ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \mathrm{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \mathrm{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \mathrm{diag}(y)\alpha$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha$$

Soft SVM Duality

$$\min_{w,\xi} \lambda \|w\|^2 + \frac{1}{n} \sum_i \xi_i \quad \text{s.t. } \forall i,\; y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i, \beta_i \geq 0} \lambda \|w\|^2 + \frac{1}{n} \sum_i \xi_i + \sum_{i=1}^{n} \alpha_i (1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \max_{\alpha_i, \beta_i \geq 0} \min_{w,\xi} \lambda \|w\|^2 + \frac{1}{n} \mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y) \alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y) \alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda} \alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y) \alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

change variables: $2\lambda\tilde{\alpha} = \alpha$

$$= (2\lambda) \max_{0 \leq \tilde{\alpha}_i \leq \frac{1}{2\lambda n}} \mathbf{1}^\top \tilde{\alpha} - \frac{1}{2}\tilde{\alpha}^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\tilde{\alpha}$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i,\quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^{n} \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^{n} \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

Only difference from hard SVM is upper bound on $\tilde{\alpha}_i$

change variables: $2\lambda\tilde{\alpha} = \alpha$

$$= (2\lambda) \max_{0 \leq \tilde{\alpha}_i \leq \frac{1}{2\lambda n}} \mathbf{1}^\top \tilde{\alpha} - \frac{1}{2}\tilde{\alpha}^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\tilde{\alpha}$$

Soft SVM Duality

$$\min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i \quad \text{s.t. } \forall i,\ y_i w^\top x_i \geq 1 - \xi_i,\quad \xi_i \geq 0$$

$$= \min_{w,\xi} \max_{\alpha_i,\beta_i \geq 0} \lambda\|w\|^2 + \frac{1}{n}\sum_i \xi_i + \sum_{i=1}^n \alpha_i(1 - y_i w^\top x_i - \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

$$= \max_{\alpha_i,\beta_i \geq 0} \min_{w,\xi} \lambda\|w\|^2 + \frac{1}{n}\mathbf{1}^\top \xi + \mathbf{1}^\top \alpha - \alpha^\top \operatorname{diag}(y) X w - \alpha^\top \xi - \beta^\top \xi$$

$2\lambda w - X^\top \operatorname{diag}(y)\alpha = 0 \quad w = \frac{1}{2\lambda} X^\top \operatorname{diag}(y)\alpha \qquad \frac{1}{n}\mathbf{1} - \alpha - \beta = 0 \quad \beta = \frac{1}{n}\mathbf{1} - \alpha$

$$= \max_{\alpha_i \geq 0} \mathbf{1}^\top \alpha - \frac{1}{4\lambda}\alpha^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\alpha \quad \text{s.t. } \frac{1}{n} \geq \alpha_i$$

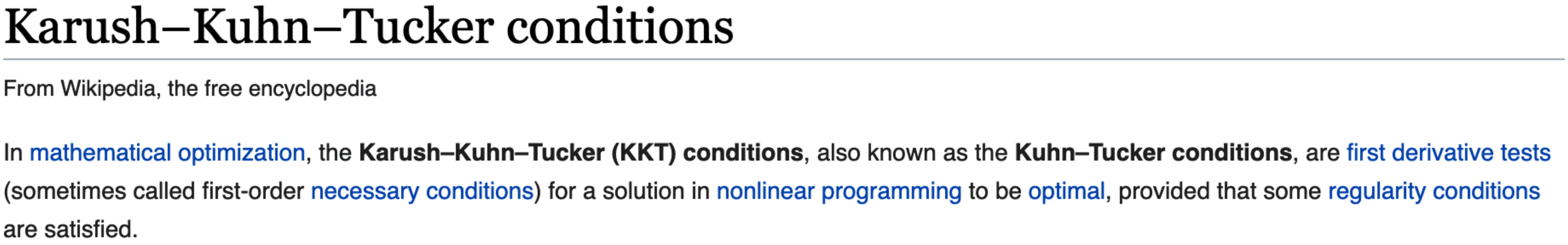
Only difference from hard SVM is upper bound on $\tilde{\alpha}_i$

change variables: $2\lambda\tilde{\alpha} = \alpha$

$$= (2\lambda) \max_{0 \leq \tilde{\alpha}_i \leq \frac{1}{2\lambda n}} \mathbf{1}^\top \tilde{\alpha} - \frac{1}{2}\tilde{\alpha}^\top \operatorname{diag}(y) X X^\top \operatorname{diag}(y)\tilde{\alpha}$$

Can do with b also:
add $\alpha^\top y = 0$ constraint,
set $b = w^\top x_i - y_i$ for any SV

FYI



Karush–Kuhn–Tucker conditions

From Wikipedia, the free encyclopedia

In mathematical optimization, the **Karush–Kuhn–Tucker (KKT) conditions**, also known as the **Kuhn–Tucker conditions**, are first derivative tests (sometimes called first-order necessary conditions) for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied.

- Summarize the process of going through Lagrange duality for you
    - Like Lagrange multipliers, but allow inequality constraints
- Make things a lot faster once you're familiar with them
- Related conditions for when strong duality holds
    - Especially important: "Slater's condition"

Motivation

- Machine learning! ...but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \textbf{sign}(f(x))$

Motivation

- Machine learning! ...but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \textbf{sign}(f(x))$

Motivation

- Machine learning! …but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \text{sign}(f(x))$

Motivation

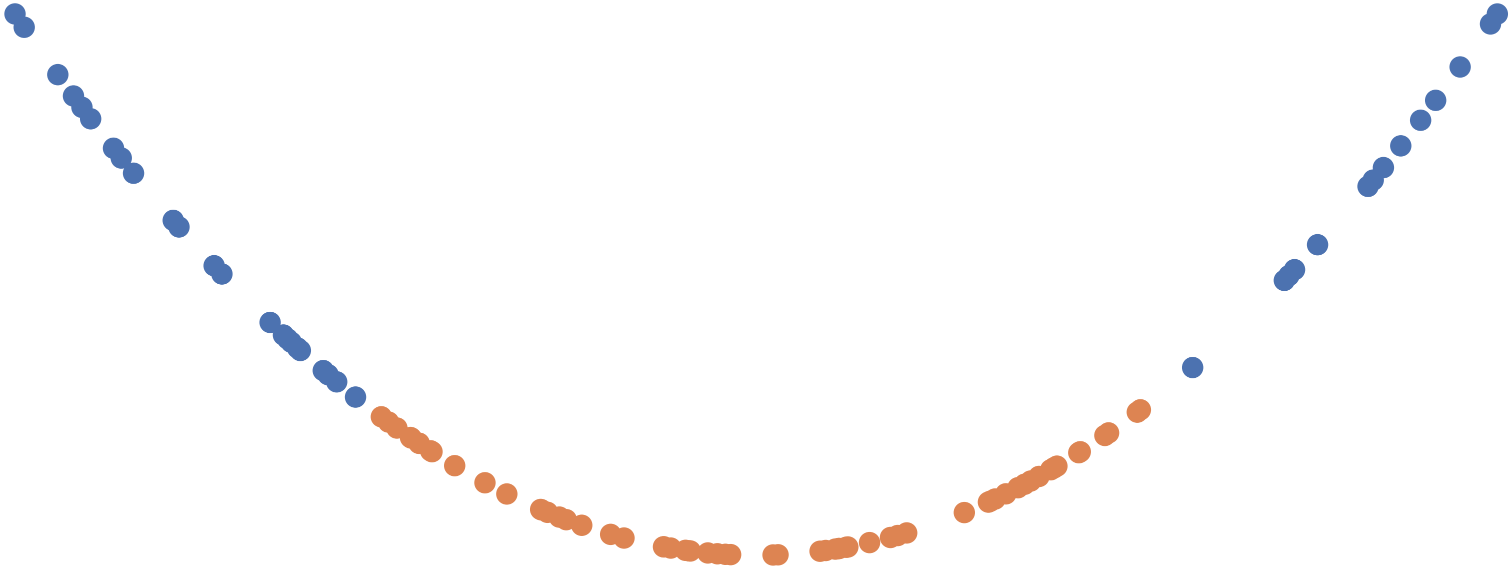
- Machine learning! …but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \mathbf{sign}(f(x))$
- Extend x…

$$f(x) = w^{\mathsf{T}}(1, x, x^2) = w^{\mathsf{T}}\phi(x)$$

Motivation

- Machine learning! ...but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \mathbf{sign}(f(x))$
- Extend x...

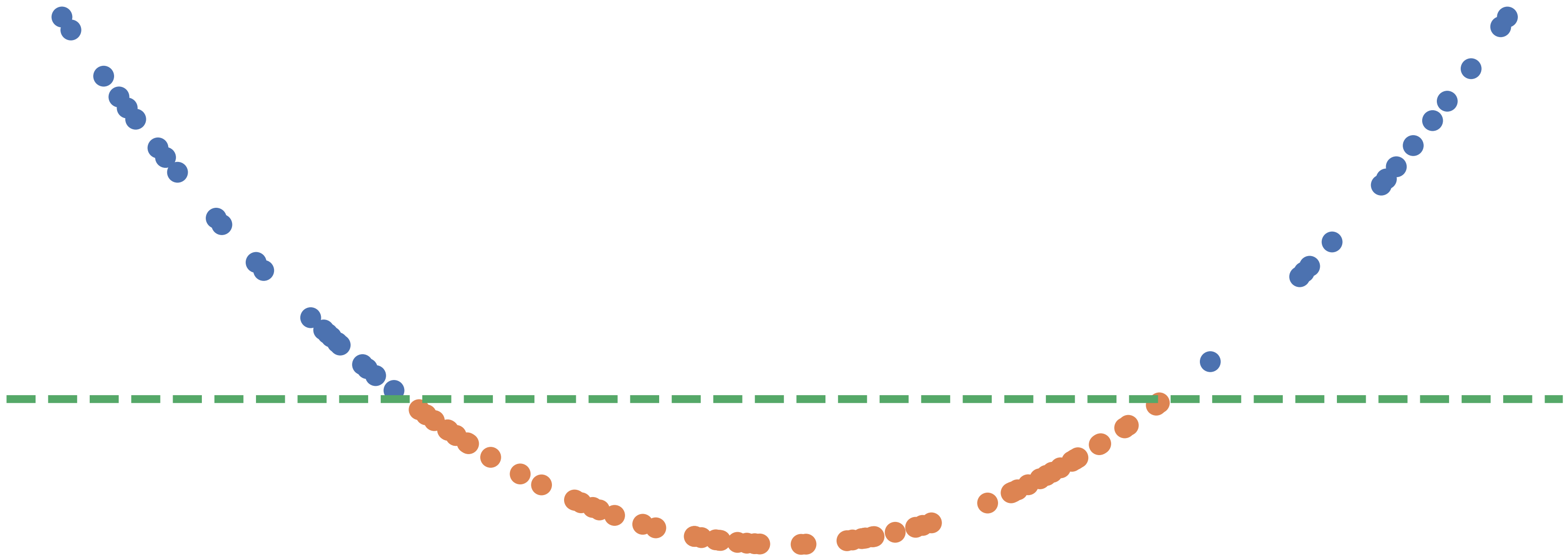
$$f(x) = w^\top(1, x, x^2) = w^\top\phi(x)$$



Motivation

- Machine learning! ...but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \mathbf{sign}(f(x))$
- Extend x...

$$f(x) = w^{\mathsf{T}}(1, x, x^2) = w^{\mathsf{T}}\phi(x)$$



Motivation

- Machine learning! ...but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \mathbf{sign}(f(x))$
- Extend x...

$$f(x) = w^\mathsf{T}(1, x, x^2) = w^\mathsf{T}\phi(x)$$

- Kernels are basically a way to study doing this with any, potentially very complicated, ϕ

Motivation

- Machine learning! …but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \mathbf{sign}(f(x))$
- Extend x…

$$f(x) = w^\mathsf{T}(1, x, x^2) = w^\mathsf{T}\phi(x)$$

- Kernels are basically a way to study doing this with any, potentially very complicated, ϕ
- Convenient way to make models on documents, graphs, videos, datasets, …

Motivation

- Machine learning! …but how do we actually do it?
- Linear models! $f(x) = w_0 + wx$, $\hat{y}(x) = \mathbf{sign}(f(x))$
- Extend x…

$$f(x) = w^{\mathsf{T}}(1, x, x^2) = w^{\mathsf{T}}\phi(x)$$

- Kernels are basically a way to study doing this with any, potentially very complicated, ϕ
- Convenient way to make models on documents, graphs, videos, datasets, …
- ϕ will live in a *reproducing kernel Hilbert space*

Hilbert spaces

- A complete (real or complex) inner product space.

Hilbert spaces

- A complete (real ~~or complex~~) inner product space.

Hilbert spaces

- A complete (real ~~or complex~~) inner product space.
- Inner product space: a vector space with an **inner product**:
  - $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
  - $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
  - $\langle f, f \rangle_{\mathcal{H}} > 0$ for $f \neq 0$, $\langle 0, 0 \rangle_{\mathcal{H}} = 0$

$\mathbb{R}^d \quad \langle a, b \rangle_{\mathbb{R}^d} = a^T b$

Hilbert spaces

- A complete (real ~~or complex~~) inner product space.
- Inner product space: a vector space with an **inner product**:
    - $\langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g\rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g\rangle_{\mathcal{H}}$
    - $\langle f, g\rangle_{\mathcal{H}} = \langle g, f\rangle_{\mathcal{H}}$
    - $\langle f, f\rangle_{\mathcal{H}} > 0$ for $f \neq 0$, $\langle 0, 0\rangle_{\mathcal{H}} = 0$

Induces a **norm**: $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$

$\|f-g\|_{\mathcal{H}}$

$\|x\|_{\mathbb{R}^d} = \sqrt{x^T x}$

Hilbert spaces

- A complete (real ~~or complex~~) inner product space.
- Inner product space: a vector space with an **inner product**:
  - $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
  - $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
  - $\langle f, f \rangle_{\mathcal{H}} > 0$ for $f \neq 0$, $\langle 0, 0 \rangle_{\mathcal{H}} = 0$

  Induces a **norm**: $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$
- Complete: "well-behaved" (Cauchy sequences have limits in $\mathcal{H}$)

$x_1, x_2, x_3, x_4, \cdots$

Kernel: an inner product between feature maps

- Call our domain $\mathcal{X}$, some set
    - $\mathbb{R}^d$, functions, distributions of graphs of images, …

Kernel: an inner product between feature maps

- Call our domain $\mathcal{X}$, some set
    - $\mathbb{R}^d$, functions, distributions of graphs of images, …
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ so that

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

Kernel: an inner product between feature maps

- Call our domain $\mathcal{X}$, some set
  - $\mathbb{R}^d$, functions, distributions of graphs of images, …
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ so that

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

- Roughly, k is a notion of “similarity” between inputs

Kernel: an inner product between feature maps

- Call our domain $\mathcal{X}$, some set
    - $\mathbb{R}^d$, functions, distributions of graphs of images, …
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel on $\mathcal{X}$ if there exists a Hilbert space $\mathcal{H}$ and a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$ so that

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$

- Roughly, k is a notion of “similarity” between inputs
- *Linear kernel* on $\mathbb{R}^d$: $k(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$

Building kernels from other kernels

- Scaling: if $\gamma \geq 0$, $k_\gamma(x, y) = \gamma k(x, y)$ is a kernel

Building kernels from other kernels

- Scaling: if $\gamma \geq 0$, $k_\gamma(x, y) = \gamma k(x, y)$ is a kernel
  - $k_\gamma(x, y) = \gamma \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \sqrt{\gamma}\phi(x), \sqrt{\gamma}\phi(y) \rangle_{\mathcal{H}}$

Building kernels from other kernels

- Scaling: if $\gamma \geq 0$, $k_\gamma(x, y) = \gamma k(x, y)$ is a kernel
    - $k_\gamma(x, y) = \gamma\langle\phi(x), \phi(y)\rangle_\mathcal{H} = \langle\sqrt{\gamma}\phi(x), \sqrt{\gamma}\phi(y)\rangle_\mathcal{H}$
- Sum: $k_+(x, y) = k_1(x, y) + k_2(x, y)$ is a kernel

Building kernels from other kernels

- Scaling: if $\gamma \geq 0$, $k_\gamma(x, y) = \gamma k(x, y)$ is a kernel
    - $k_\gamma(x, y) = \gamma\langle\phi(x), \phi(y)\rangle_\mathcal{H} = \langle\sqrt{\gamma}\phi(x), \sqrt{\gamma}\phi(y)\rangle_\mathcal{H}$
- Sum: $k_+(x, y) = k_1(x, y) + k_2(x, y)$ is a kernel
    - $k_+(x, y) = \left\langle \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \end{bmatrix}, \begin{bmatrix} \phi_1(y) \\ \phi_2(y) \end{bmatrix} \right\rangle_{\mathcal{H}_1 \oplus \mathcal{H}_2}$

$$= \langle \varphi_1(x), \varphi_1(y) \rangle_{\mathcal{H}_1} + \langle \varphi_2(x), \varphi_2(y) \rangle_{\mathcal{H}_2}$$

Building kernels from other kernels

- Scaling: if $\gamma \geq 0$, $k_\gamma(x, y) = \gamma k(x, y)$ is a kernel
    - $k_\gamma(x, y) = \gamma \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \sqrt{\gamma}\phi(x), \sqrt{\gamma}\phi(y) \rangle_{\mathcal{H}}$
- Sum: $k_+(x, y) = k_1(x, y) + k_2(x, y)$ is a kernel
    - $k_+(x, y) = \left\langle \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \end{bmatrix}, \begin{bmatrix} \phi_1(y) \\ \phi_2(y) \end{bmatrix} \right\rangle_{\mathcal{H}_1 \oplus \mathcal{H}_2}$
- Is $k_1(x, y) - k_2(x, y)$ necessarily a kernel?

Building kernels from other kernels

- Scaling: if $\gamma \geq 0$, $k_\gamma(x, y) = \gamma k(x, y)$ is a kernel
    - $k_\gamma(x, y) = \gamma \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \sqrt{\gamma}\phi(x), \sqrt{\gamma}\phi(y) \rangle_{\mathcal{H}}$
- Sum: $k_+(x, y) = k_1(x, y) + k_2(x, y)$ is a kernel
    - $k_+(x, y) = \left\langle \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \end{bmatrix}, \begin{bmatrix} \phi_1(y) \\ \phi_2(y) \end{bmatrix} \right\rangle_{\mathcal{H}_1 \oplus \mathcal{H}_2}$
- Is $k_1(x, y) - k_2(x, y)$ necessarily a kernel?
    - Take $k_1(x, y) = 0$, $k_2(x, y) = xy$, $x \neq 0$.
    - Then $k_1(x, x) - k_2(x, x) = -x^2 < 0$
    - But $k(x, x) = \|\phi(x)\|^2_{\mathcal{H}} \geq 0$.

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

Positive definiteness

$k(x,y) = k(y,x)$

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Equivalently: *kernel matrix* K is PSD

$$K := \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \ldots & k(x_n, x_n) \end{bmatrix}$$

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}}$$

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{H}}$$

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \end{aligned}$$

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd

$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{j=1}^{n} a_j \phi(x_j) \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0\end{aligned}$$

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd

Positive definiteness

- A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive semi-definite (psd)* if for all $n \geq 1$, $a_1, \ldots, a_n \in \mathbb{R}^n$, $x_1, \ldots, x_n \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

- Hilbert space kernels are psd
- psd functions are Hilbert space kernels
  - Moore-Aronszajn Theorem; we'll come back to this

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
    - $\displaystyle \lim_{n \to \infty} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k_n(x_i, x_j) \geq 0$

Some more ways to build kernels

- Limits: if $k_\infty(x,y) = \lim_{n\to\infty} k_n(x,y)$ exists, k_∞ is psd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
  - Let $V \sim \mathcal{N}(0, K_1)$, $W \sim \mathcal{N}(0, K_2)$ be independent
  - $\operatorname{Cov}(V_i W_i, V_j W_j) = \operatorname{Cov}(V_i, V_j) \operatorname{Cov}(W_i, W_j) = k_\times(x_i, x_j)$
  - Covariance matrices are psd, so $k_\times$ is too

Schur's Theorem

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y) k_2(x, y)$ is psd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y) k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$

$$x^\mathsf{T} y$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$

$$x^\mathsf{T} y + c$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$

$$\left(x^\mathsf{T} y + c\right)^n$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y) k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$

  $\left(x^\mathsf{T} y + c\right)^n$, the **polynomial kernel**

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
    - $k_{\exp}(x, y) = \lim_{N\to\infty} \sum_{n=0}^{N} \frac{1}{n!} k(x, y)^n$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y) k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd
    - Use the feature map $x \mapsto f(x)\phi(x)$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y) k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x) k(x, y) f(y)$ is pd

$$x^\mathsf{T} y$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd

$$\frac{1}{\sigma^2} x^\mathsf{T} y$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd

$$\exp\left(\frac{1}{\sigma^2} x^\mathsf{T} y\right)$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd

$$\exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right)\exp\left(\frac{1}{\sigma^2}x^\mathsf{T}y\right)\exp\left(-\frac{1}{2\sigma^2}\|y\|^2\right)$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n \to \infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y) k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x) k(x, y) f(y)$ is pd

$$\exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right) \exp\left(\frac{1}{\sigma^2} x^\mathsf{T} y\right) \exp\left(-\frac{1}{2\sigma^2}\|y\|^2\right)$$

$$= \exp\left(-\tfrac{1}{2\sigma^2}\left[\|x\|^2 - 2x^\mathsf{T} y + \|y\|^2\right]\right)$$

Some more ways to build kernels

- Limits: if $k_\infty(x, y) = \lim_{n\to\infty} k_n(x, y)$ exists, k_∞ is psd
- Products: $k_\times(x, y) = k_1(x, y)k_2(x, y)$ is psd
- Powers: $k_n(x, y) = k(x, y)^n$ is pd for any integer $n \geq 0$
- Exponents: $k_{\exp}(x, y) = \exp(k(x, y))$ is pd
- If $f : \mathcal{X} \to \mathbb{R}$, $k_f(x, y) = f(x)k(x, y)f(y)$ is pd

$$\exp\left(-\frac{1}{2\sigma^2}\|x\|^2\right)\exp\left(\frac{1}{\sigma^2}x^\mathsf{T}y\right)\exp\left(-\frac{1}{2\sigma^2}\|y\|^2\right)$$

$= \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$, the **Gaussian kernel**

Reproducing property

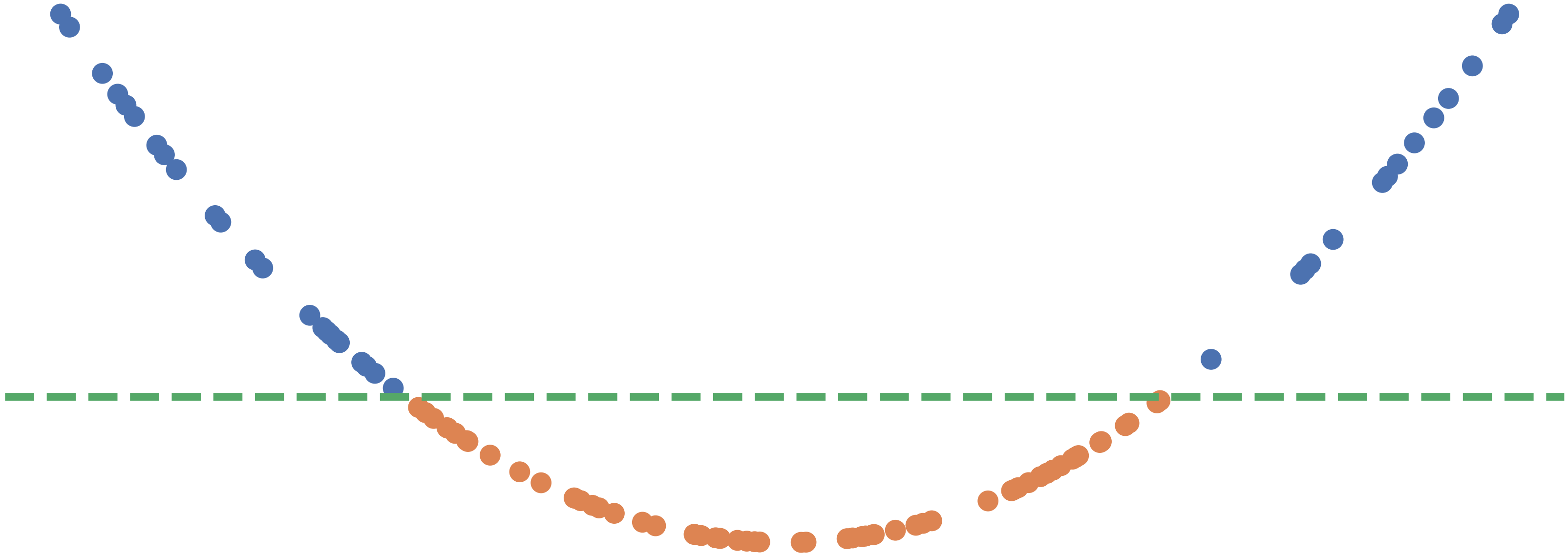
- Recall original motivating example with

$$\boldsymbol{\mathcal{X}} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

Reproducing property

- Recall original motivating example with

$$\mathcal{X} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

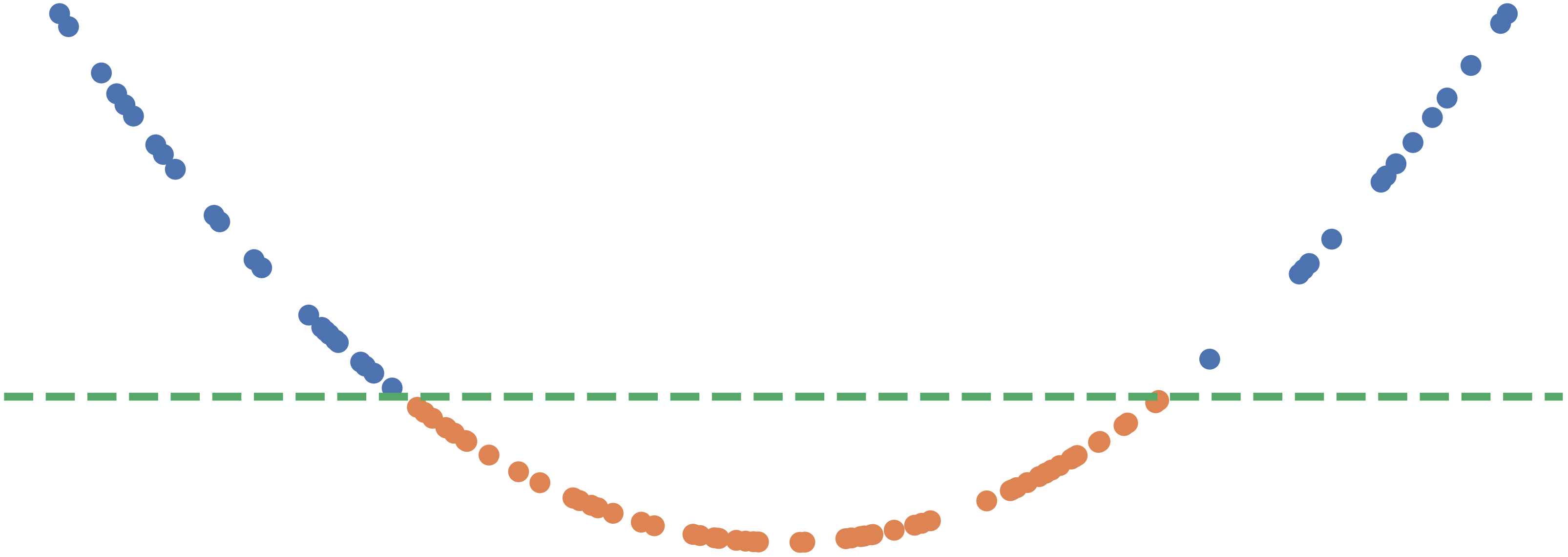


Reproducing property

- Recall original motivating example with

$$\mathcal{X} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

- Kernel is $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = 1 + xy + x^2 y^2$

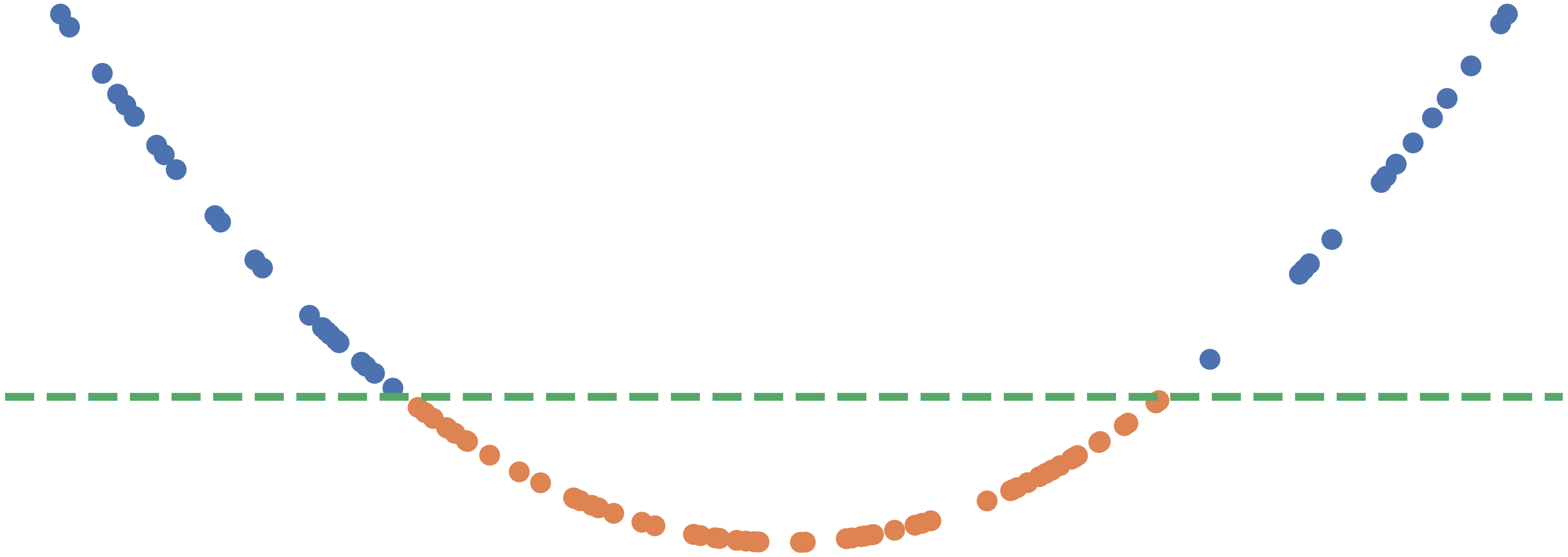


Reproducing property

- Recall original motivating example with

$$\mathcal{X} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

- Kernel is $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = 1 + xy + x^2y^2$
- Classifier based on linear $f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}}$



Reproducing property

- Recall original motivating example with

$$\mathcal{X} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

- Kernel is $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = 1 + xy + x^2 y^2$
- Classifier based on linear $f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}}$
- $f(\cdot)$ is the function f itself, represented by a vector in $\mathbb{R}^3$
  $f(x) \in \mathbb{R}$ is the function evaluated at a point x

Reproducing property

- Recall original motivating example with

$$\mathcal{X} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

- Kernel is $k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = 1 + xy + x^2y^2$
- Classifier based on linear $f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}}$
- $f(\cdot)$ is the function f itself, represented by a vector in $\mathbb{R}^3$
  $f(x) \in \mathbb{R}$ is the function evaluated at a point x
- Elements of $\mathcal{H}$ correspond to **functions**, $f : \mathcal{X} \to \mathbb{R}$

Reproducing property

- Recall original motivating example with

$$\mathcal{X} = \mathbb{R} \qquad \phi(x) = (1, x, x^2) \in \mathbb{R}^3$$

- Kernel is $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = 1 + xy + x^2 y^2$
- Classifier based on linear $f(x) = \langle w, \phi(x) \rangle_{\mathcal{H}}$
- $f(\cdot)$ is the function f itself, represented by a vector in $\mathbb{R}^3$
  $f(x) \in \mathbb{R}$ is the function evaluated at a point x
- Elements of $\mathcal{H}$ correspond to **functions**, $f : \mathcal{X} \to \mathbb{R}$
- **Reproducing prop.**: $f(x) = \langle f(\cdot), \phi(x) \rangle_{\mathcal{H}}$ for $f \in \mathcal{H}$

Reproducing kernel Hilbert space (RKHS)

- Every psd kernel k on $\mathcal{X}$ defines a (unique) Hilbert space, its RKHS $\mathcal{H}$, and a map $\phi : \mathcal{X} \to \mathcal{H}$ where
    - $$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$
    - Elements $f \in \mathcal{H}$ are functions on $\mathcal{X}$, with
      $$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$$
- Combining the two, we sometimes write $k(x, \cdot) = \phi(x)$

Reproducing kernel Hilbert space (RKHS)

- Every psd kernel k on $\mathcal{X}$ defines a (unique) Hilbert space, its RKHS $\mathcal{H}$, and a map $\phi : \mathcal{X} \to \mathcal{H}$ where
  - $$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$$
  - Elements $f \in \mathcal{H}$ are functions on $\mathcal{X}$, with

    $$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$$

- Combining the two, we sometimes write $k(x, \cdot) = \phi(x)$
- $k(x, \cdot)$ is the **evaluation functional**

  An RKHS is defined by it being *continuous*, or

  $$|f(x)| \leq M_x \|f\|_{\mathcal{H}}$$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
    - Start with $\mathcal{H}_0 = \mathbf{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
    - Start with $\mathcal{H}_0 = \mathrm{span}(\{k(x,\cdot) : x \in \mathcal{X}\})$
    - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x,\cdot), k(y,\cdot) \rangle_{\mathcal{H}_0} = k(x,y)$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \textbf{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
    - Start with $\mathcal{H}_0 = \mathbf{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$
    - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$
    - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$
    - Get that the reproducing property holds for $k(x, \cdot)$ in $\mathcal{H}$

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \textbf{span}(\{k(x, \cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$
  - Get that the reproducing property holds for $k(x, \cdot)$ in $\mathcal{H}$
  - Can also show uniqueness

Moore-Aronszajn Theorem

- Building $\mathcal{H}$ for a given psd k:
  - Start with $\mathcal{H}_0 = \operatorname{\mathbf{span}}(\{k(x, \cdot) : x \in \mathcal{X}\})$
  - Define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ from $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_0} = k(x, y)$
  - Take $\mathcal{H}$ to be completion of $\mathcal{H}_0$ in the metric from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$
  - Get that the reproducing property holds for $k(x, \cdot)$ in $\mathcal{H}$
  - Can also show uniqueness
- Theorem: k is psd iff it's the reproducing kernel of an RKHS

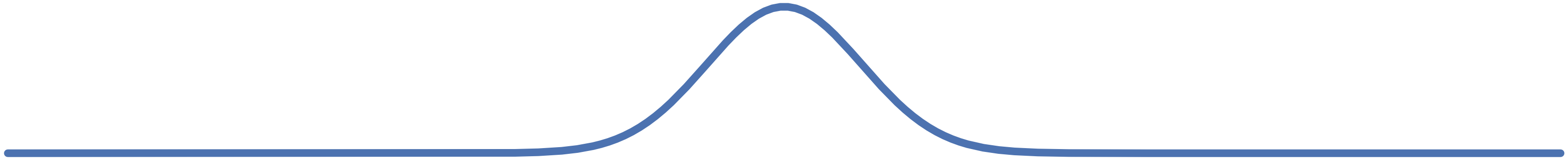
A quick check: linear kernels

- $k(x, y) = x^{\top} y$ on $\mathcal{X} = \mathbb{R}^d$
- If $f(y) = \sum_{i=1}^{n} a_i k(x_i, y)$, then $f(y) = \left[\sum_{i=1}^{n} a_i x_i\right]^{\top} y$
- Closure doesn't add anything here, since $\mathbb{R}^d$ is closed
- So, linear kernel gives you RKHS of linear functions
- $\|f\|_{\mathcal{H}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j)} = \|\sum_{i=1}^{n} a_i x_i\|$

More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2}\|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*



More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2} \|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*

More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2} \|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*

More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2} \|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*

More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2} \|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*

More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2}\|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*
- Functions in $\mathcal{H}$ are bounded:

$$f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}} \leq \sqrt{k(x, x)}\|f\|_{\mathcal{H}} = \|f\|_{\mathcal{H}}$$

$$\|k(x,\cdot)\|_{\mathcal{H}}^2 = \langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{H}} = k(x,x)$$

More complicated: Gaussian kernels

$$k(x,y) = \exp(\frac{1}{2\sigma^2}\|x-y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*
- Functions in $\mathcal{H}$ are bounded:

$$f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}} \leq \sqrt{k(x,x)}\|f\|_{\mathcal{H}} = \|f\|_{\mathcal{H}}$$

- Choice of σ controls how fast functions can vary:

$\langle k(x+\xi \cdot) - k(x;\cdot), f \rangle_{\mathcal{H}}$

$$f(x+t) - f(x) \leq \|k(x+t,\cdot) - k(x',\cdot)\|_{\mathcal{H}}\|f\|_{\mathcal{H}}$$

$$\|k(x+t,\cdot) - k(x,\cdot)\|^2_{\mathcal{H}} = 2 - 2k(x,x+t) = 2 - 2\exp\left(-\frac{\|t\|^2}{2\sigma^2}\right)$$

More complicated: Gaussian kernels

$$k(x, y) = \exp(\frac{1}{2\sigma^2}\|x - y\|^2)$$

- $\mathcal{H}$ is *infinite-dimensional*
- Functions in $\mathcal{H}$ are bounded:
  $$f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}} \leq \sqrt{k(x, x)}\|f\|_{\mathcal{H}} = \|f\|_{\mathcal{H}}$$
- Choice of σ controls how fast functions can vary:

$$f(x + t) - f(x) \leq \|k(x + t, \cdot) - k(x', \cdot)\|_{\mathcal{H}}\|f\|_{\mathcal{H}}$$

$$\|k(x + t, \cdot) - k(x, \cdot)\|_{\mathcal{H}}^2 = 2 - 2k(x, x + t) = 2 - 2\exp\left(-\frac{\|t\|^2}{2\sigma^2}\right)$$

- Can say lots more with Fourier properties