Support Vector Machines

Nishant Mehta

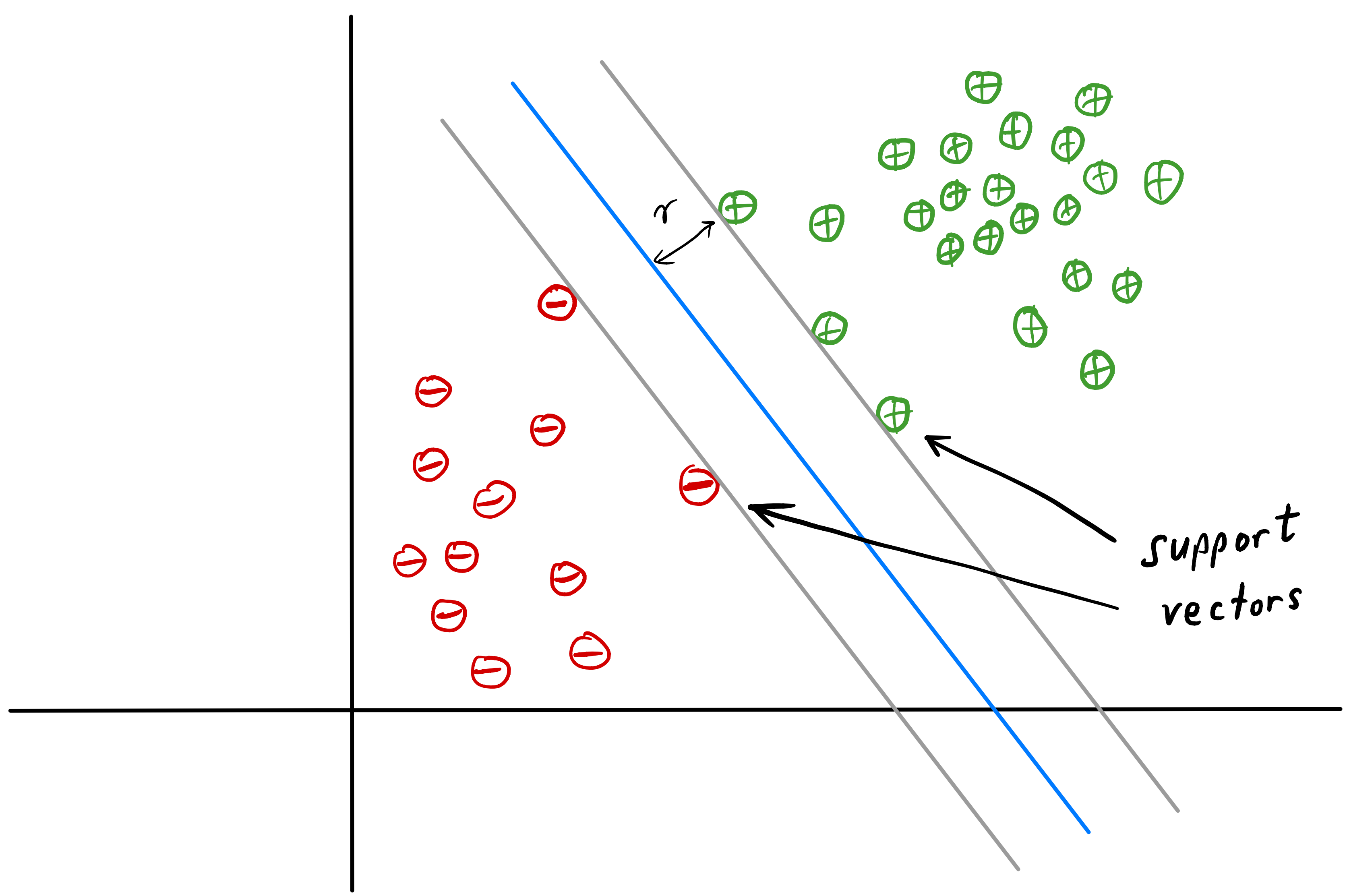
Lectures 12–14

Hard-margin SVM

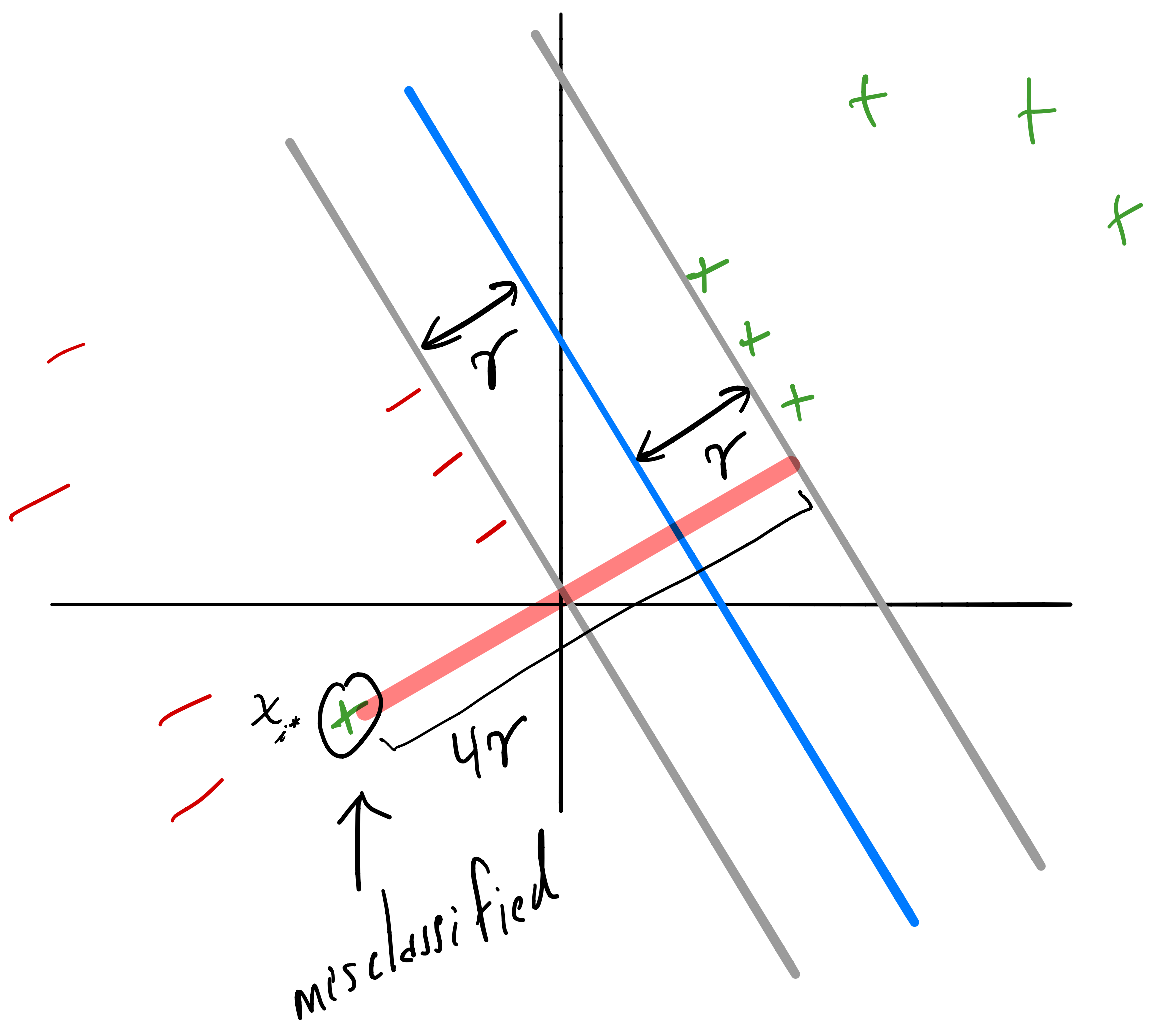
Margin $r = \frac{1}{\|w\|}$

Hard margin SVM problem

$$\begin{aligned} &\underset{w,b}{\text{minimize}} && \|w\|^2 \\ &\text{subject to} && y_i\big(\langle w, x_i\rangle + b\big) \geq 1,\ i = 1, \ldots, n. \end{aligned}$$



Soft-margin SVM



What if data isn't linearly separable?

Or, most of the data is separable with large margin, and some only with very low margin?

Soft-margin SVM problem

hyperparameter $C > 0$

$$\underset{\substack{w\in\mathbb{R}^n, b\in\mathbb{R}\\ \xi\in\mathbb{R}^n_+}}{\text{minimize}} \quad \|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

nonnegative vector in $\mathbb{R}^n$

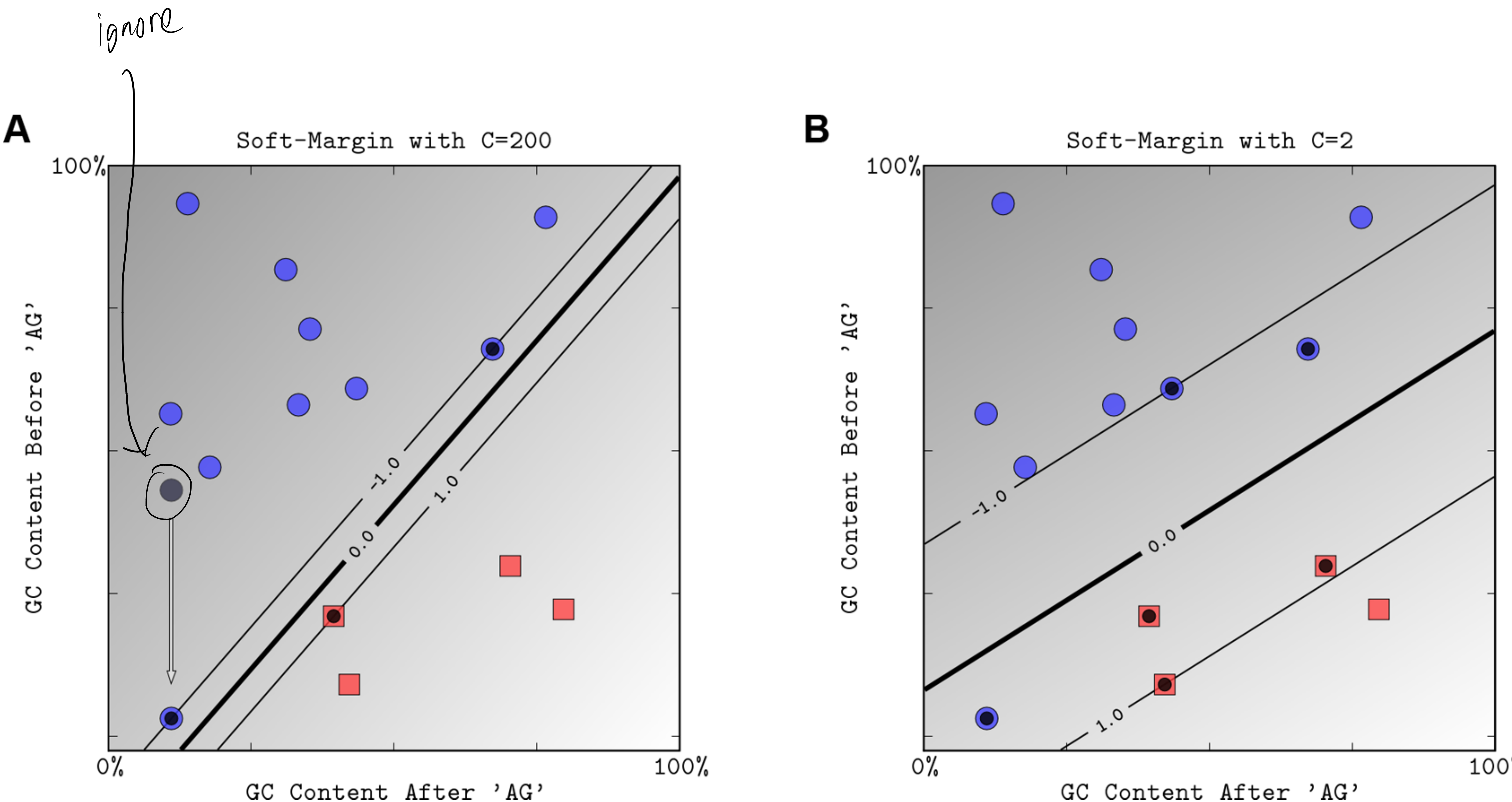
$$\text{subject to} \quad y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \; i = 1, \ldots, n$$

Examples: $\xi_i = 0$ $\quad y_i(\langle w, x_i\rangle + b) \geq 1$ margin of (w,b) on (x_i, y_i) is at least $\frac{1}{\|w\|}$

$\xi_i = \frac{2}{3}$ $\quad y_i(\langle w, x_i\rangle + b) = 1 - \xi_i = 1 - \frac{2}{3} = \frac{1}{3}$

$$\Rightarrow \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} = \frac{1/3}{\|w\|} = \frac{1-\xi_i}{\|w\|}$$

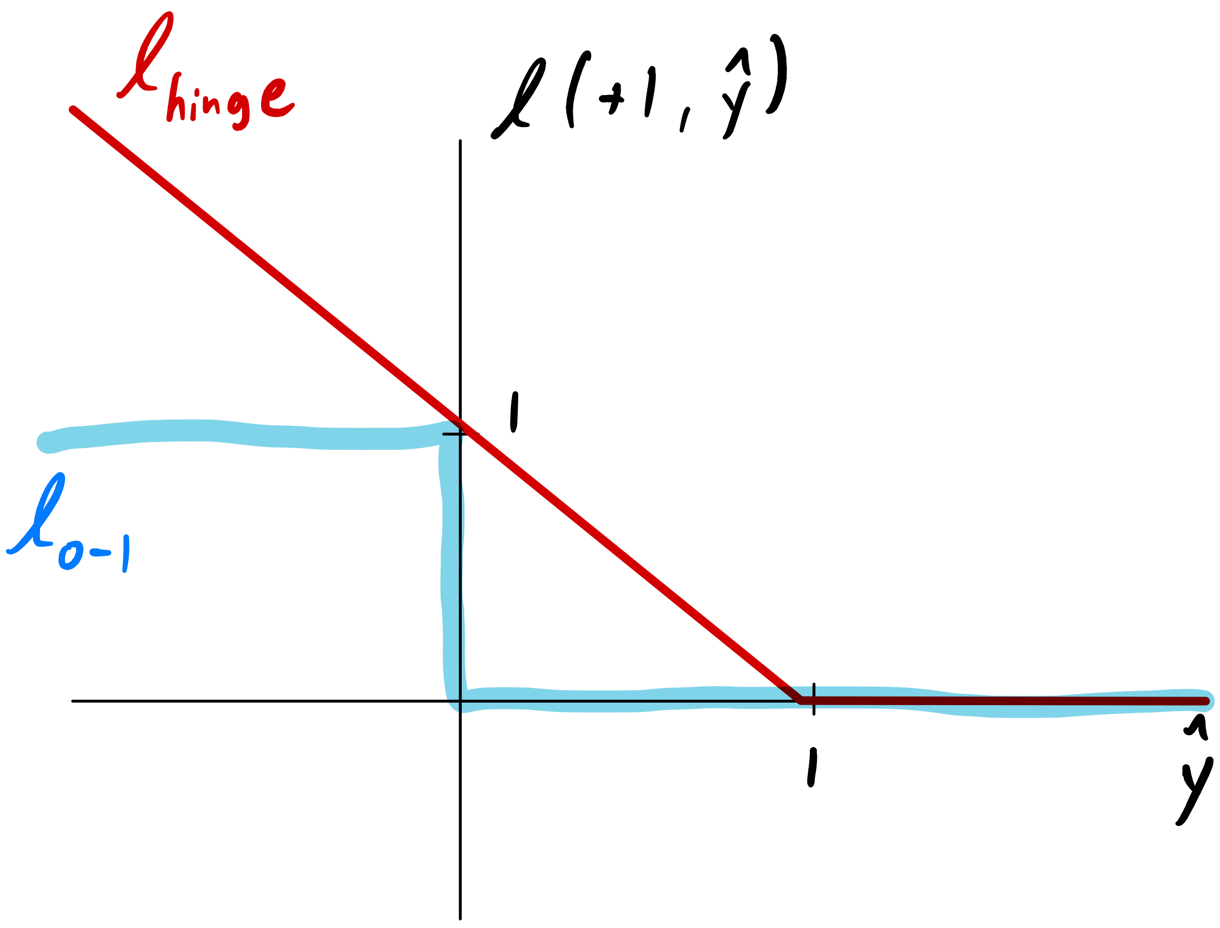
Varying C (linear kernel)



From "Support Vector Machines and Kernels for Computational Biology" (Ben-Hur et al., 2008)

Soft-margin SVM - Hinge Loss

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \quad \|w\|^2 + C \sum_{i=1}^{n} \max\left\{0, 1 - y_i \underbrace{\left(\langle w, x_i \rangle + b\right)}_{\in \mathbb{R}}\right\}$$



hinge loss

$$\ell_{\text{hinge}}(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$$

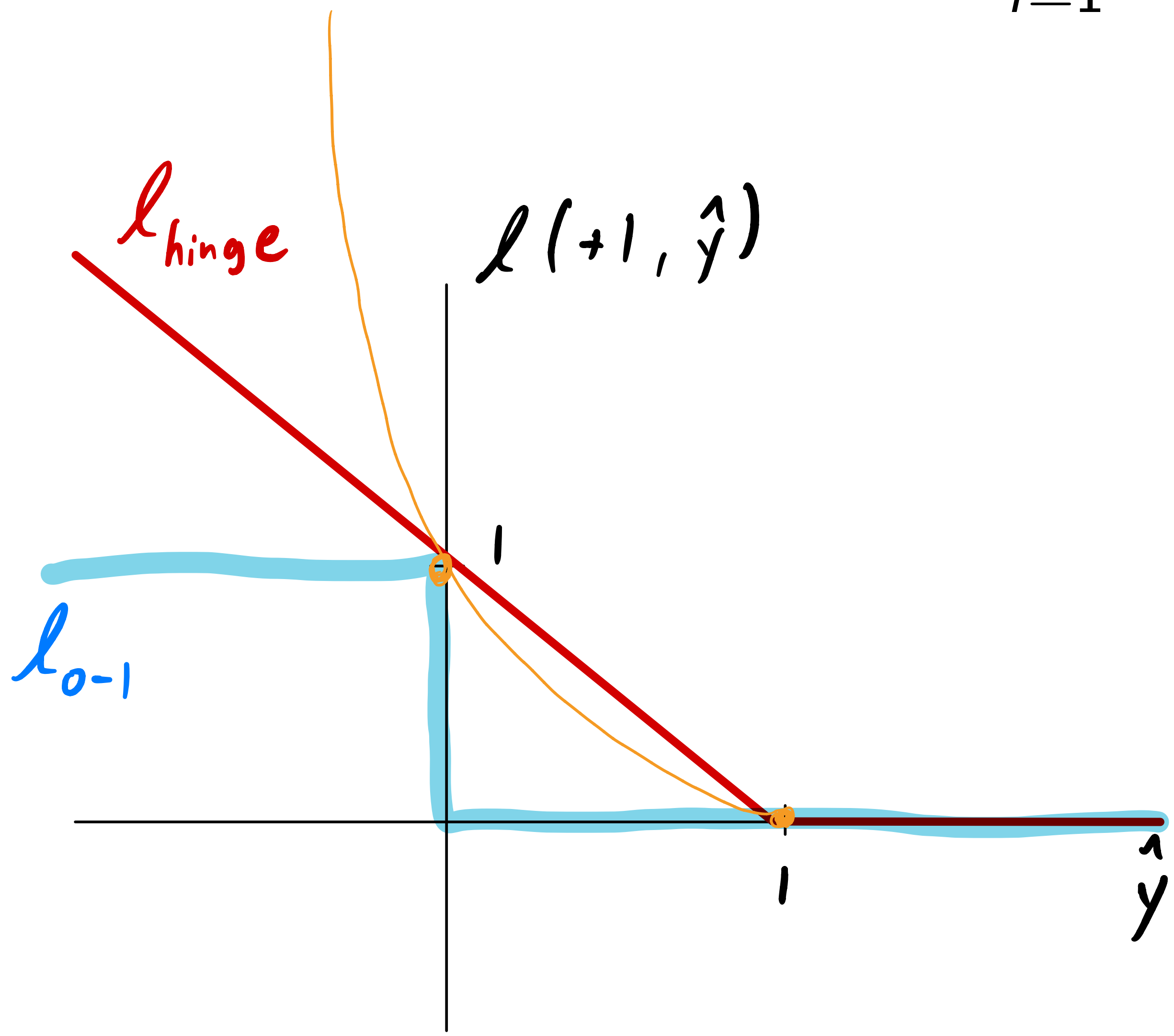
$\hat{y} \in \mathbb{R}$

$\ell_{hinge}(1, \hat{y})$

$= \max\{0, 1 - \hat{y}\}$

Soft-margin SVM - Hinge Loss

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \quad \|w\|^2 + C \sum_{i=1}^{n} \max\Big\{0, 1 - y_i\big(\langle w, x_i \rangle + b\big)\Big\}$$



hinge loss

$$\ell_{\text{hinge}}(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$$

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \quad \|w\|^2 + C \sum_{i=1}^{n} \ell_{\text{hinge}}\big(y_i, f_{w,b}(x_i)\big)$$

SVM - Regularization viewpoint

SVM can be viewed as minimizing regularized training error under hinge loss

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \ \frac{1}{C}\|w\|^2 + \frac{C}{C}\sum_{i=1}^{n} \ell_{\text{hinge}}\left(y_i, f_{w,b}(x_i)\right)$$

$\Updownarrow$ **Equivalent**

$\lambda \geq 0$

$$\lambda = \frac{1}{C}$$

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \ \sum_{i=1}^{n} \ell_{\text{hinge}}\left(y_i, f_{w,b}(x_i)\right) + \lambda\|w\|^2$$

λ = regularization parameter

SVM dual problem

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle \\ \text{subject to} \quad & \sum_{i=1}^{n} y_i \alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \; i = 1, \ldots, n \end{aligned}$$

How to get w and b from this?

$$w = \sum_{i=1}^{n} y_i \alpha_i x_i$$

$$b = y_i - \sum_{j=1}^{n} y_j \alpha_j \langle x_i, x_j \rangle \quad \text{for any } i \text{ satisfying } 0 < \alpha_i < C$$

How to predict?

$$f_{w,b}(x_{\text{test}}) = \langle w, x_{\text{test}} \rangle + b = \sum_{i=1}^{n} y_i \alpha_i \langle x_i, x_{\text{test}} \rangle + b$$

SVM dual problem - Inner products only

$$\underset{\alpha \in \mathbb{R}^n}{\text{maximize}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \overbrace{\langle x_i, x_j \rangle}^{\langle \varphi(x_i), \varphi(x_j) \rangle}$$

$$\text{subject to} \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

$$0 \leq \alpha_i \leq C, \; i = 1, \ldots, n$$

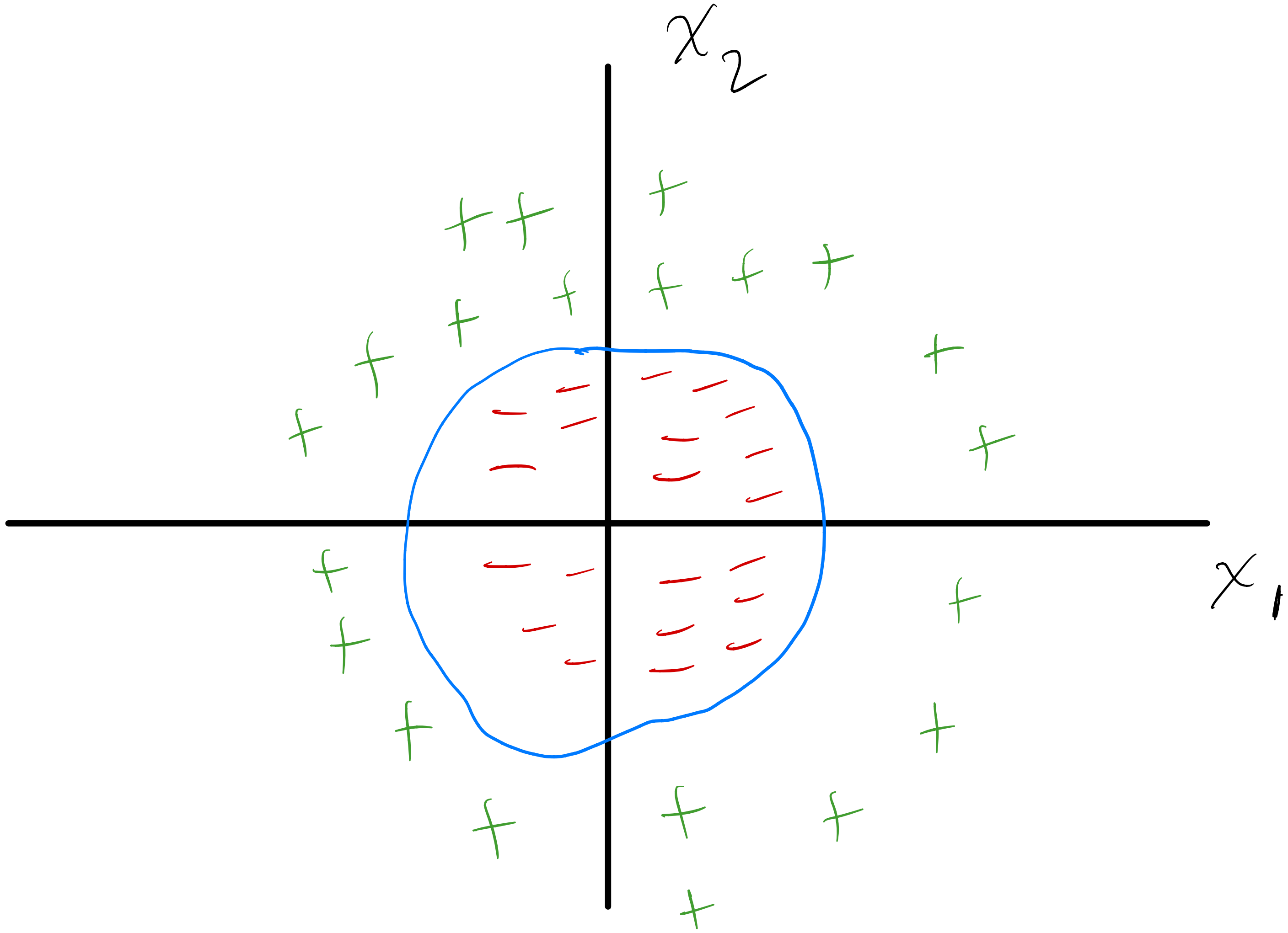
compute this without explicitly storing $\varphi(x_i)$, $\varphi(x_j)$

$\langle \varphi(x_i), \varphi(x_{test}) \rangle$

How to predict? $f_{w,b}(x_{\text{test}}) = \langle w, x_{\text{test}} \rangle + b = \sum_{i=1}^{n} y_i \alpha_i \langle x_i, x_{\text{test}} \rangle + b$

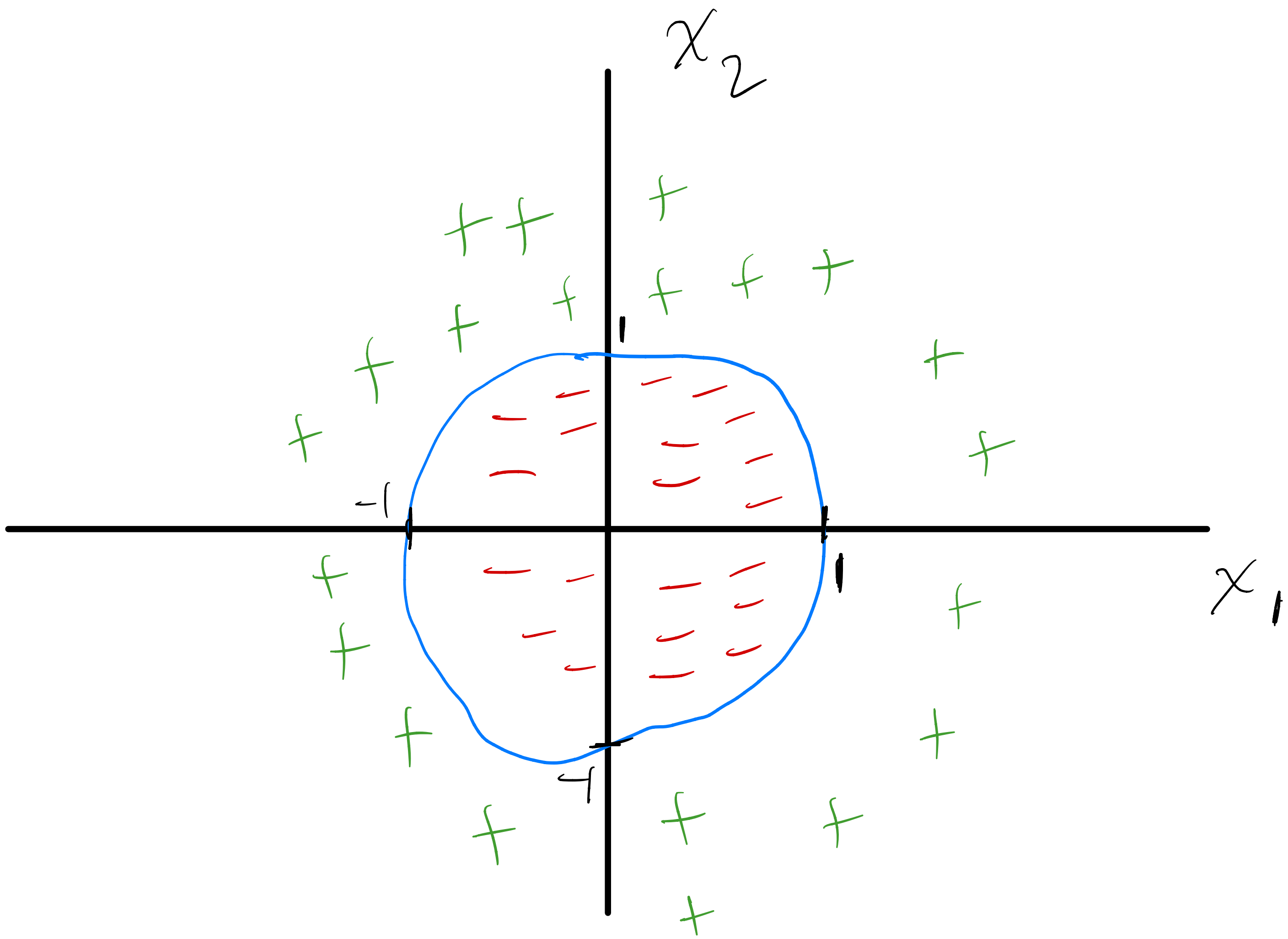
Dual SVM only needs inner products between input examples!

How can we achieve nonlinear classifiers?

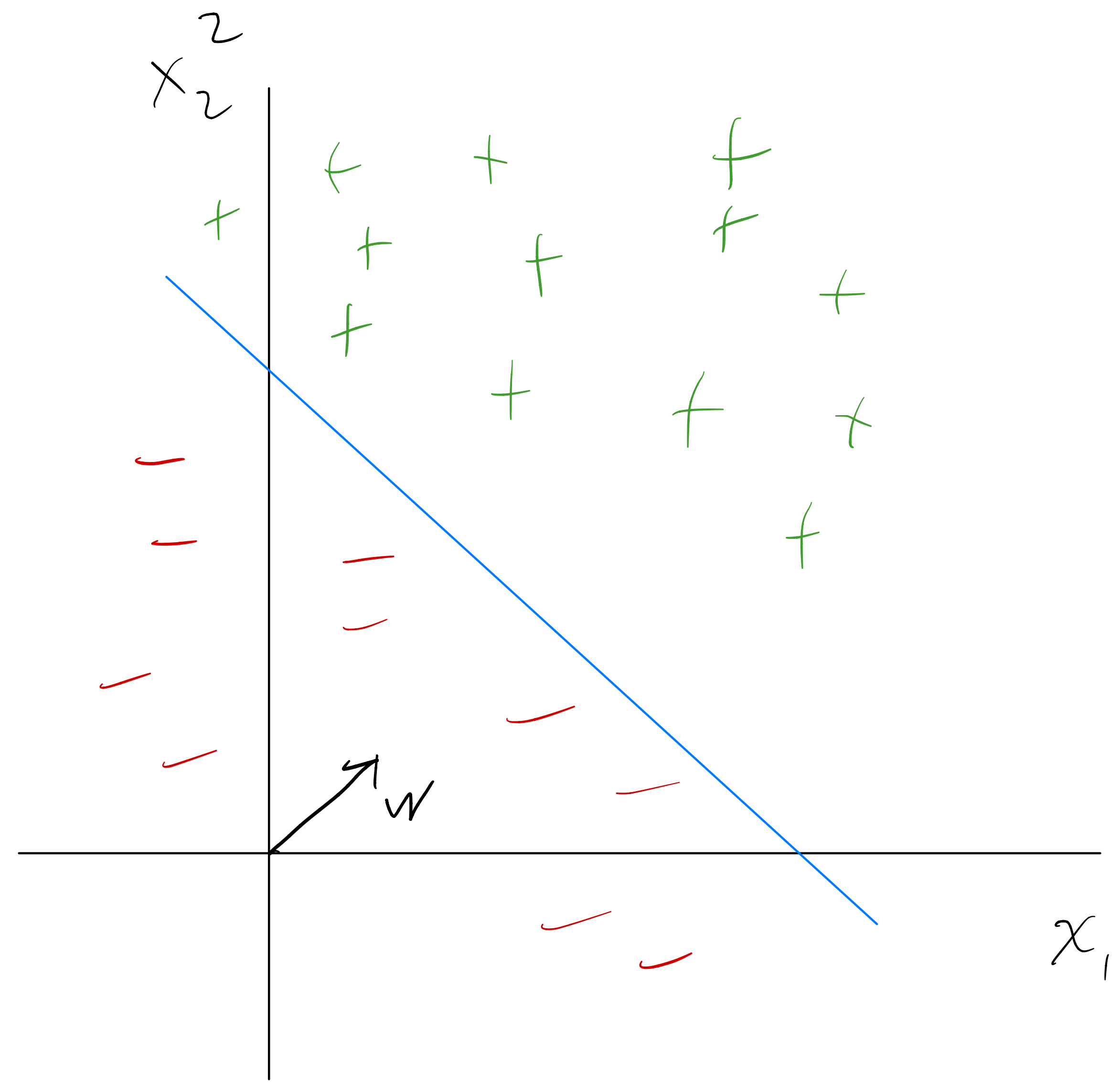


Idea: feature map

Classification in original space

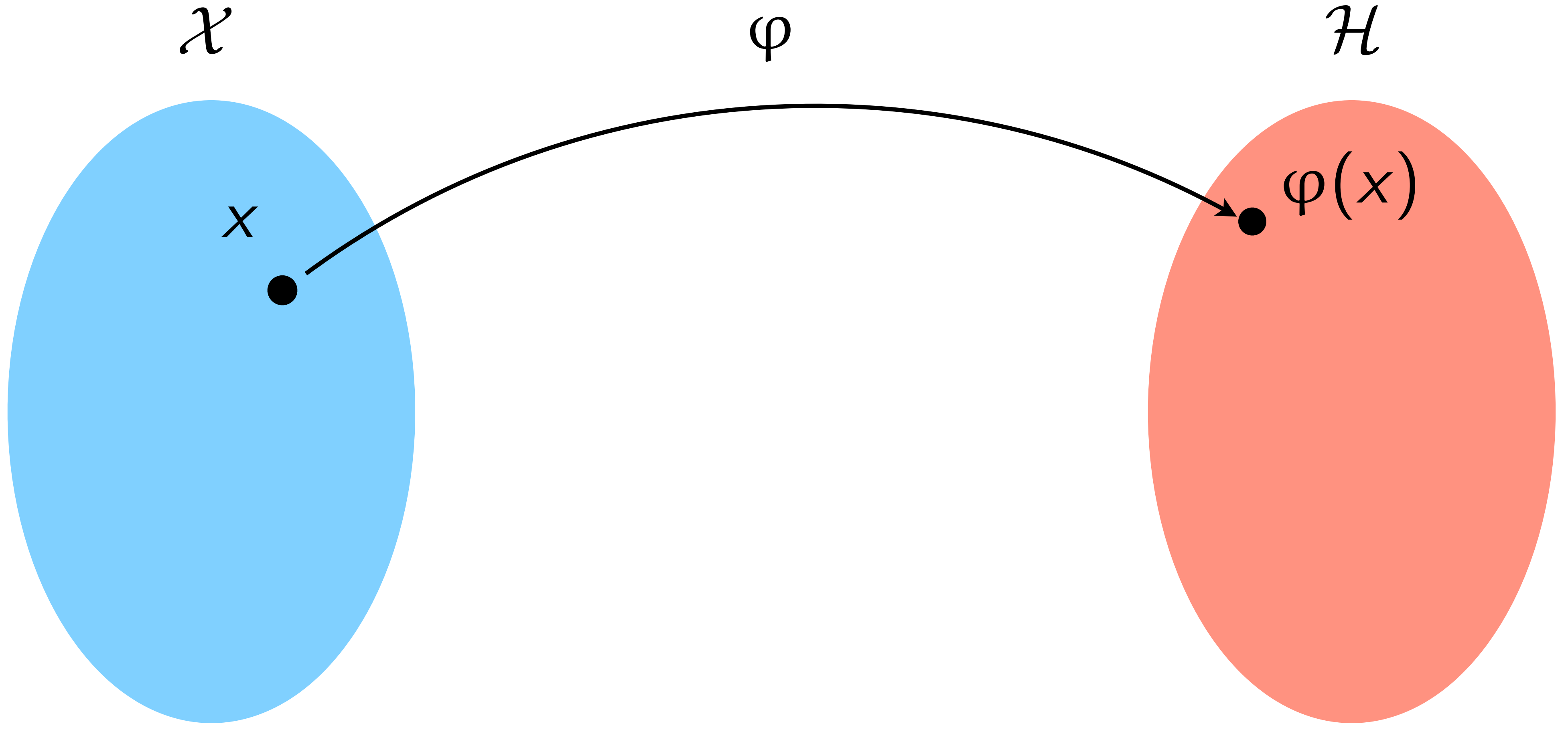


Classification in feature space



Idea: feature map

Use a feature map: $\varphi(x) : \mathcal{X} \to \mathcal{H}$



Kernel trick

Question: Can we compute inner product between input examples x and z in feature space without explicitly computing $\varphi(x)$ and $\varphi(z)$?

In many cases, yes! We use a *kernel function*:

$$k(x, z) = \langle \varphi(x), \varphi(\overset{z}{\cancel{x}}) \rangle$$

Equal to inner product… but we won't compute it this way!

Example 1: Warm-up exercise

(dimension $d=2$)

original space $(\mathcal{X})$: $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

(dimension $=3$)

feature space $(\mathcal{H})$: $\varphi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \end{pmatrix}$

kernel function

$$k(x,z) = \langle \varphi(x), \varphi(z) \rangle = \left\langle \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}\, x_1 x_2 \end{pmatrix}, \begin{pmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}\, z_1 z_2 \end{pmatrix} \right\rangle$$

$$= x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 x_2 z_1 z_2$$

$$= (x_1 z_1)^2 + (x_2 z_2)^2 + 2 (x_1 z_1)(x_2 z_2) = (x_1 z_1 + x_2 z_2)^2$$

$$= \langle x, z \rangle^2$$

Example 2: Polynomial kernel, one dimension

inner product $\langle x, z \rangle = xz$

The *polynomial kernel* (one dimension):

$$k(x,z) = (xz + a)^r = (xz)^r + \binom{r}{1}(xz)^{r-1}a + \dots + \binom{r}{r-1}(xz)a^{r-1} + a$$

hyperparameter

$$= x^r z^r + \sqrt{\binom{r}{1}a}\, x^{r-1} \sqrt{\binom{r}{1}a}\, z^{r-1} + \dots + \sqrt{\binom{r}{r-1}a^{r-1}}\, x \sqrt{\binom{r}{r-1}a^{r-1}}\, z + \sqrt{a}\cdot\sqrt{a}$$

What is the feature space?

$$\varphi(x) = \begin{pmatrix} x^r \\ \sqrt{\binom{r}{1}a}\, x^{r-1} \\ \vdots \\ \sqrt{\binom{r}{r-1}a^{r-1}}\, x \\ \sqrt{a} \end{pmatrix}$$

Example 3: Polynomial kernel, general dimension

$d \geq 1$

The *polynomial kernel* (general dimension):

$$k(x, z) = (\langle x, z \rangle + a)^r \qquad C \cdot x_1^{r-5} x_3^2 x_6^2 x_7$$

$\varphi(x)$ has one feature for each monomial up to degree *r*

How many features are there in the feature space?

Example 3: Polynomial kernel, general dimension

hyperparameter

$d \geq 1$

The *polynomial kernel*:

$$k(x, z) = (\langle x, z \rangle + a)^r$$

$$C \cdot x_1^{r-5} \, x_3^2 \, x_6^2 \, x_7$$

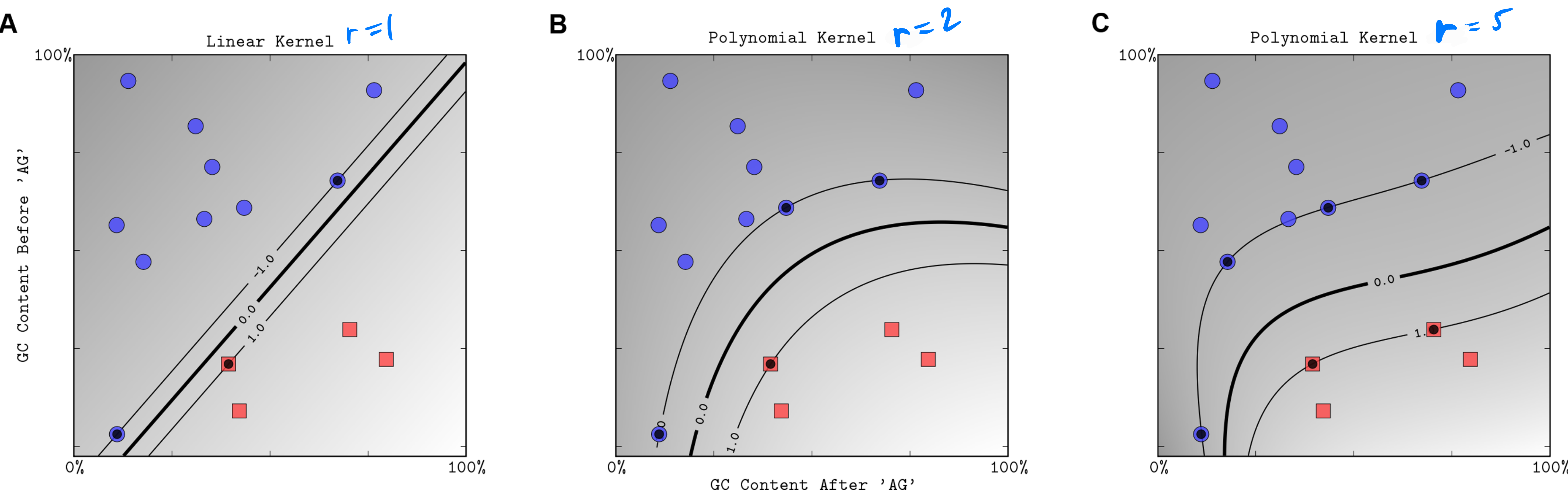
$\varphi(x)$ has one feature for each monomial up to degree r

How many features are there in the feature space?

$$\binom{r+d}{d} = \frac{(r+d)!}{r!\,d!} = \binom{r+d}{r} = O(d^r)$$

But the kernel can be computed in only $O(d)$

Polynomial kernels of increasing degree



Gaussian kernel

Suppose $Y \sim N(0, \sigma^2)$ $\mu \in \mathbb{R}$ $\sigma^2 > 0$

pdf of Y $p(Y=y) = \dfrac{\exp\left(-\frac{1}{2}\frac{y^2}{\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$

The *Gaussian kernel* is based on the distance between two examples

$$k(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

(hyperparameter)

bandwidth parameter

The Gaussian kernel is a type of similarity measure, taking values between 0 and 1

What is the corresponding feature map $\varphi(x)$?

Gaussian kernel

The *Gaussian kernel* is based on the distance between two examples

$$k(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

bandwidth parameter

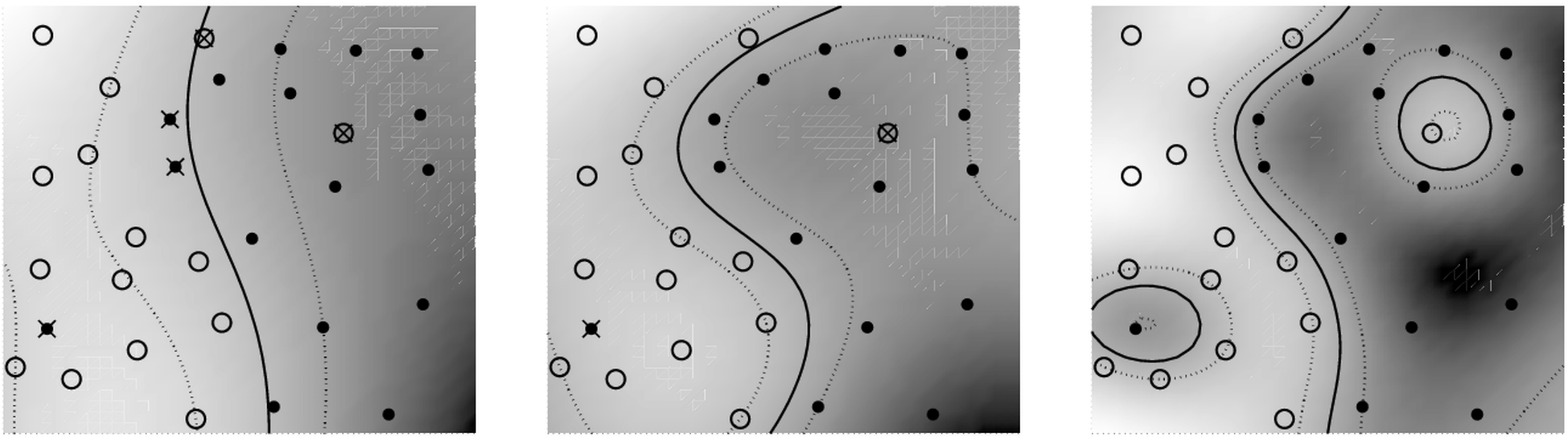
The Gaussian kernel is a type of similarity measure, taking values between 0 and 1

What is the corresponding feature map $\varphi(x)$?

It's infinite dimensional!

Varying Gaussian kernel bandwidth
(C kept constant)

Decreasing kernel bandwidth



From the book “Learning with Kernels” (Schölkopf and Smola, 2001)