L6: Building Embeddings : Liftings

About data $X \in \mathbb{R}^d$ when d is small.
Want a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^D$ where D is large.
Result: - more expressive features - still use linear approaches (e.g., halfspaces) - higher dimensions (oh my!)

Parabolic Lifting

Halfspace to Balls: $D = d + 1$

$$\phi(x) = (x_1, x_2, \ldots, x_d, \sum_{j=1}^{d} x_j^2)$$

- includes halfspaces $h_{u,t} = \{x \in \mathbb{R}^d \mid \langle u, x\rangle - t > 0\}$
  since $b_{u',t} = \{x \in \mathbb{R}^D \mid \langle u', x\rangle - t > 0\}$ where $u' = (u, 0)$
- includes balls $b_{c,r} = \{x \in \mathbb{R}^d \mid \|x - c\|^2 \leq r^2\}$
  as halfspaces $h_{u',r'} = \{x \in \mathbb{R}^D \mid \langle u', x\rangle - r' > 0\}$
  where $u' = (-2c, 1)$ and $r' = \|c\|^2 - r^2$

  $\|x - c\|^2 \leq r^2$
  $\langle x, x\rangle + \langle c, c\rangle - 2\langle x, c\rangle \leq r^2$
  $2\langle x, c\rangle \geq \langle x, x\rangle + (\langle c, c\rangle - r^2)$
  $\langle x, 2c\rangle \geq \sum_{j=1}^{d} x_j^2 + (\|c\|^2 - r^2)$
  $\langle (x, \sum_{j=1}^{d} x_j^2), (2c, -1)\rangle \geq \|c\|^2 - r^2$

Note that the free variables in u' and r' are c and r.
Once c is set, we can pick any r. Some values of r' not feasible. These contain no points in X.
Only feasible regions are same as balls.

Polynomial Lifting

We can also generate **any** polynomial boundary of degree p, not just balls.
With $p = 2$ and $d = 2$ leads to $D = 5$

$$\phi(x_1, x_2) = (x_1, x_2, x_1^2, x_2^2, x_1 x_2)$$

Any set with polynomial boundary, with maximum degree $p = 2$, can be reduced to some

$$\langle \alpha, \phi(x)\rangle = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1^2 + \alpha_4 x_2^2 + \alpha_5 x_1 x_2 > \alpha_0$$

Often this is written with $D = 6$

$$\phi(x_1, x_2) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$$

then use $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_5)$
and only need halfspaces with origin on boundary:
e.g., $\langle \phi(x), \alpha\rangle \geq 0$

In general d and degree p we need $D = \binom{d+p}{d} = \binom{d+p}{d}$ which is both $O(d^p)$ and $O(p^d)$.

Reproducing Kernels

Bivariate kernels with scale parameter σ

- $K(x, p) = \exp(-\|x - p\|^2/\sigma^2)$, the Gaussian kernel.

- $K(x,p) = \exp(-\|x-p\|/\sigma)$, the Laplace kernel.
- $K(x,p) = \frac{\sigma}{\|x-p\|}\sin(-\|x-p\|/\sigma)$, the Sinc kernel.

These kernels are a notion of similarity between inputs $x, p \in \mathbb{R}^d$.
Within about $\sigma \to$ close. Otherwise $\to$ far.

Kernel Trick: use $K(p,x)$ in place of $\langle p, x\rangle$.
e.g., $\|x-p\|_K^2 = K(x,x) + K(p,p) - 2K(p,x)$
also useful for non-linear classification, regression, PCA, clustering
... but precomputes $K(x_i, x_j)$ for all x_i, x_j (in $O(dn^2)$ time)

But lifting exactly like polynomials needs infinite dimensions! (or n dimensions if we know X).
$\phi(x) = K(x,\cdot)$ is a point in a function space.
For *reproducing* kernels K, each $\phi(x)$ is linear independent of all others sets not containing x.
But for $x, p \in \mathbb{R}^d$ then $\langle \phi(x), \phi(p)\rangle_{H_K} = K(x,p)$.

Random Fourier Features:
For Gaussian kernels (and others) can cleverly approximate $\phi : \mathbb{R}^d \to H_K$ with $\hat{\phi} : \mathbb{R}^d \to \mathbb{R}^D$.

Generate $w_1, \ldots, w_D \sim N(0, 1/\sigma^2)$ and $t_1, \ldots, t_D \sim Unif(0, 2\pi)$
Let $\hat{\phi}_j(x) = \cos(\langle w_j, x\rangle + t_j)$.
Then $\hat{\phi}(x) = (\hat{\phi}_1(x), \hat{\phi}_2(x), \ldots, \hat{\phi}_D(x))/\sqrt{D}$

Or generate $w_1, \ldots, w_{D/2} \sim N(0, 1/\sigma^2)$.
Let $\tilde{\phi}_{2j-1}(x) = \cos(\langle w_j, x\rangle)$ and $\tilde{\phi}_{2j}(x) = \sin(\langle w_j, x\rangle)$.
Again $\tilde{\phi}(x) = (\tilde{\phi}_1(x), \tilde{\phi}_2(x), \ldots, \tilde{\phi}_D(x))/\sqrt{D/2}$.

With $D = O((1/\varepsilon^2)\log(n/\delta))$ then with probability at least $1-\delta$ - For all $x_1, x_2 \in X$: $|K(x_1,x_2) - \langle \hat{\phi}(x_1), \hat{\phi}(x_2)\rangle| \leq \varepsilon$.
- For all $x_1, x_2 \in X$: $|\|\phi(x_1) - \phi(x_2)\|_{H_K} - \|\hat{\phi}(x_1) - \hat{\phi}(x_2)\|| \leq \varepsilon \|\phi(x_1) - \phi(x_2)\|_{H_K}$.

Same for $\tilde{\phi}$.