昆山杜克大学
DUKE KUNSHAN UNIVERSITY

Sampling

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 13

RKHS (cont'd)

reproducing kernel Hilbert space (RKHS)

- In general, a Hilbert space $\mathcal{H}$ is said to be an RKHS if there exists a function $K(\cdot, \cdot): \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

1. $K(\mathbf{x}, \cdot) \in \mathcal{H}$ for any $\mathbf{x} \in \mathcal{X}$
2. $f(\mathbf{x}) = \langle f(\cdot), K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$

- Nevertheless, if $\mathcal{H}$ is an RKHS, then $K(\cdot, \cdot)$ must be PDS (why?). That is, there is no need to distinguish RKHS from PDS kernels.

uniqueness of RKHS

- An RKHS contains a unique reproducing kernel.

Suppose $K, \tilde{K}$ are two reproducing kernels of $\mathcal{H}$
then for any $f \in \mathcal{H}$, $x \in \mathcal{X}$,

$$\left.\begin{aligned} f(x) &= \langle f(\cdot), K(x,\cdot)\rangle_{\mathcal{H}} \\ f(x) &= \langle f(\cdot), \tilde{K}(x,\cdot)\rangle_{\mathcal{H}} \end{aligned}\right\} \Rightarrow 0 = \langle f(\cdot), K(x,\cdot) - \tilde{K}(x,\cdot)\rangle_{\mathcal{H}}.$$

Therefore, $(K - \tilde{K})(x,\cdot) = 0$. Since this holds for any x, $K = \tilde{K}$.

- Conversely, a reproducing kernel defines a unique RKHS.

feature space

- Given a RKHS with a PDS kernel K, any Hilbert space (inner-product space) $\mathbb{F}$ such that there exists $\phi: \mathcal{X} \to \mathbb{F}$ with $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathbb{F}}$ for any pair of $\mathbf{x}, \mathbf{x}'$ in $\mathcal{X}$ is called a feature space and this ϕ is called a feature map.
- In general, a PDS kernel K does not uniquely determines the feature space. It does not uniquely determine the dimensionality of feature either.

review of Mercer's Theorem

Theorem [Mercer's Theorem]

Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous and symmetric function. Then K admits the following expansion

$$K(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{x}')$$

ψ_j : "eigenfunction"

with nonnegative numbers $\{\lambda_j\}_{j=1}^{\infty}$ and orthonormal functions $\psi_j(\cdot): \mathcal{X} \to \mathbb{R}$ if and only if it satisfies the Mercer condition:

$$\int K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) f(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0$$

for any integrable function f.

This K is a reproducing kernel for $\mathcal{H} = \left\{ f \middle| \sum_{j=1}^{\infty} \frac{\langle f, \psi_j \rangle^2}{\lambda_j} < \infty \right\}$ such that

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{j=1}^{\infty} \frac{\langle f, \psi_j \rangle \langle g, \psi_j \rangle}{\lambda_j}$$

← regular inner products $\int f(x) \psi_j(x) dx$, ...

understanding Mercer's theorem

- Let $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$, $\mathbf{K} = [K(\mathbf{x}_n, \mathbf{x}_m)]_{n,m=1}^N$ and $\mathbf{f}: \mathcal{X} \to \mathbb{R}^N$ with $\mathrm{f}_n = f(\mathbf{x}_n)$.

- Then $\mathbf{f}^{\mathrm{T}}\mathbf{K}\mathbf{f} \geq 0$ and $\mathbf{K} = \sum_{j=1}^{N} \lambda_j \mathbf{v}_j \mathbf{v}_j^{\mathrm{T}} = V \Lambda V^{\mathrm{T}}$ where $V = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_N \\ | & & | \end{pmatrix}$

$\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$

- $K(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{K}_{nm} = (\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}})_{nm} = \sum_j \lambda_j v_{jn} v_{mj}$

~~$= \sum_j \lambda_j \phi_j(\mathbf{x}_n)\phi_k(\mathbf{x}_m)$.~~

$= \sum_j \lambda_j \psi_j(x_n) \psi_j(x_m)$

where $\psi_j(x_n) = v_{jn}$ for each j, n.

general regularization problem

$$\min_{f \in \mathcal{H}} H[f] = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2$$

Search for the best map f from a RKHS $\mathcal{H}$ which contains all possible solutions.

general regularization problem

$$\min_{f\in\mathcal{H}} H[f] = \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n)) + \lambda \|f\|_{\mathcal{H}}^2$$

Search for the best map f from a RKHS $\mathcal{H}$ which contains all possible solutions.

If Mercer's conditions are met, then we can solve

$$\{c_j\}_{j=1}^{\infty} \Leftrightarrow \min_{f\in\mathcal{H}} H[f] = \frac{1}{N}\sum_{n=1}^{N} L\left(y_n, \sum_{j=1}^{\infty} c_j \psi_j(x_n)\right) + \lambda \sum_{j=1}^{\infty} \frac{c_j^2}{\gamma_j}$$

which gives an infinite-dimensional representation of f

representer theorem

Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel whose RKHS is $\mathcal{H}$. Then, for any non-decreasing function $G: \mathbb{R} \to \mathbb{R}$ and any loss function L, the optimization problem

$$\min_{f \in \mathcal{H}} H[f] = \frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n)) + G(\|f\|_{\mathcal{H}})$$

admits a solution of the form $f^* = \sum_{n=1}^{N} c_n \, K(x_n, \cdot)$.

representer theorem

- For instance, in SVM, the optimizer is given by

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n K(\mathbf{x}_n, \mathbf{x}) + b$$

why sampling?

recall: difficulties we met in previous lectures

- Bayesian density estimation

$$\hat{p}_{\text{Bayes}}(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)[\text{redacted}]p(\theta|\mathcal{X})d\theta$$

- GP for classification

$$p(t_{N+1} = 1|\mathbf{t}_N) = \int p(t_{N+1} = 1|a_{N+1})p(a_{N+1}|\mathbf{t}_N)\,\mathrm{d}a_{N+1}$$

- Numerical integration methods do not work if we are in high dimension.

recall: difficulties we met in previous lectures

- In general, either

$$\hat{p}_{\text{Bayes}}(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\boldsymbol{x}|\theta)p(\theta|\mathcal{X})d\theta$$

or

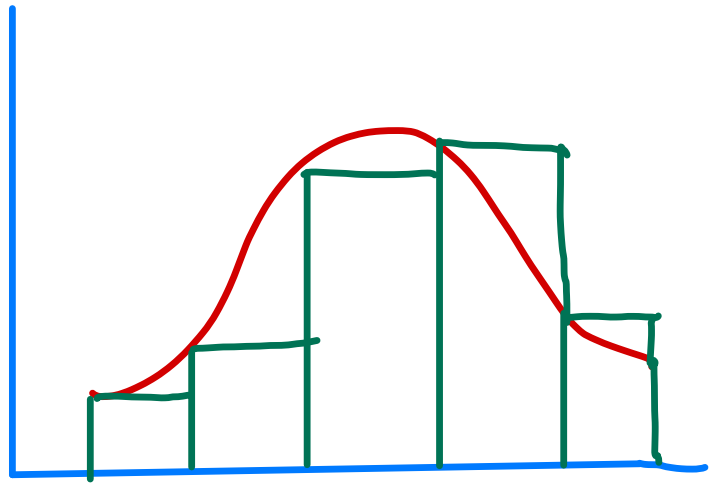
$$p(t_{N+1} = 1|\mathbf{t}_N) = \int p(t_{N+1} = 1|a_{N+1})p(a_{N+1}|\mathbf{t}_N)\,\mathrm{d}a_{N+1}$$

can be put in the more general form of

$$s = \int f(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

$$s = \int f(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

- If $\boldsymbol{x}$ is low-dimensional, we can apply numerical integration: evaluate the integrand at grid points (quadrature) and then take the sum.



$$s = \int f(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

- However, if $\boldsymbol{x}$ is high-dimensional, we would need a lot of grid points.
- e.g., 10 grid points in 1D $\rightarrow 10^{10}$ grid points in 10-dimensional spaces.

Monte Carlo sampling

- Suppose we are able to draw i.i.d. samples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from the density $p(\boldsymbol{x})$.
- Then calculate

$$\hat{s}_N = \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{x}_n)$$

as an estimator for $s = \int f(\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$.

Monte Carlo sampling

- The estimator $\hat{s}_N = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}_n)$ is unbiased:

$$\mathbb{E}[\hat{s}_N] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[f(x_n)] = \frac{1}{N}\sum_{n=1}^{N} s = s$$

- Also, by the Law of Large Numbers (lol#), $\lim_{N\to\infty} \hat{s}_N = s$.

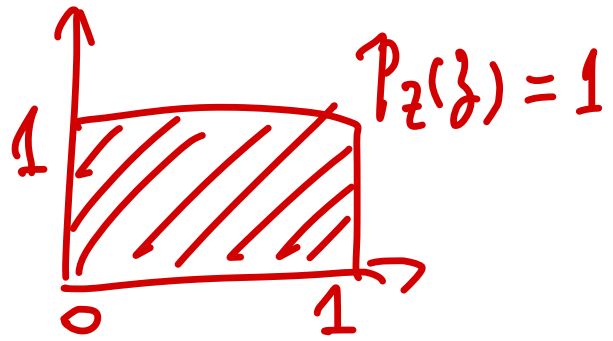
Monte Carlo sampling

- By the Central Limit Theorem (CLT), the estimator $\hat{s}_N = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x}_n)$ converges in distribution to a normal distribution with mean s and variance $N^{-1}\mathrm{Var}(f(\boldsymbol{x}))$.

- By sampling a large number of data points, we will approach the true s with a small variance.

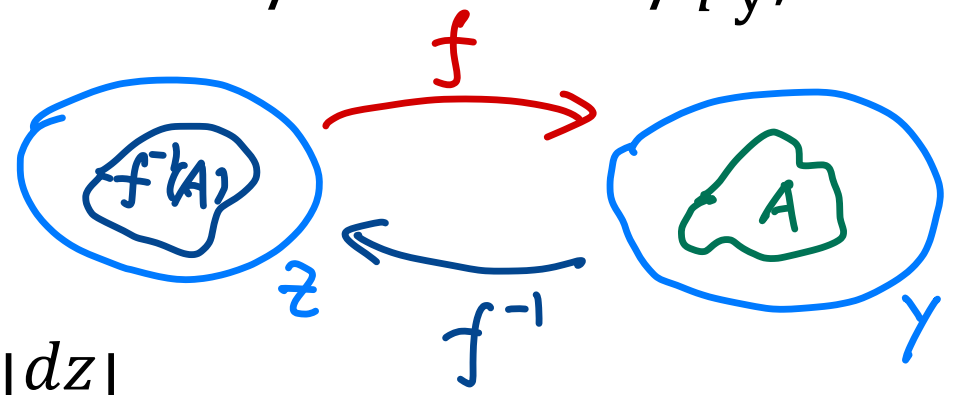
simple sampling methods

sample from univariate distribution



- Assume we can generate (pseudo-)random numbers uniformly distributed in $(0, 1)$.
- Suppose $y = f(z)$ where $z \sim \mathrm{Unif}(0,1)$. Then z is "pushed forward" by f to produces y. The *uniform density* is also "pushed forward" to produce the *corresponding density* of y, so that the probability of any event $\mathcal{A}$, measured by the density p_y, should satisfy $\mathbb{P}_y(\mathcal{A}) = \mathbb{P}_z(f^{-1}(\mathcal{A}))$.

$$\int_A p_y(y)\,dy = \int_{f^{-1}(A)} p_z(z)\,dz$$



- This implies

"$p_y(y)\,dy = p_z(z)\,dz$" $\Rightarrow$ $p_y(y) = p_z(z)\left|\frac{dz}{dy}\right| = \left|\frac{dz}{dy}\right|$

sample from univariate distribution

- Now that we have $p_{\mathrm{Y}}(y) = p_{\mathrm{Z}}(z)\left|\frac{dz}{dy}\right| = \left|\frac{dz}{dy}\right|$.

- Need: design such f for our desired distribution $p_{\mathrm{Y}}(y)$

sample $\longrightarrow$ $z = \int_0^z d\hat{z} = \int_{-\infty}^{y} p_{\mathrm{Y}}(\hat{y})\, d\hat{y} =: h(y)$ (cdf of y)

by drawing "random #" in (0,1)

- Take $f = h^{-1}$, the inverse function of the cdf of y.

sample from univariate distribution

$z \sim Unif(0,1)$

Geometrical interpretation of the transformation method for generating nonuniformly distributed random numbers. $h(y)$ is the indefinite integral of the desired distribution $p(y)$. If a uniformly distributed random variable z is transformed using $y = h^{-1}(z)$, then y will be distributed according to $p(y)$.

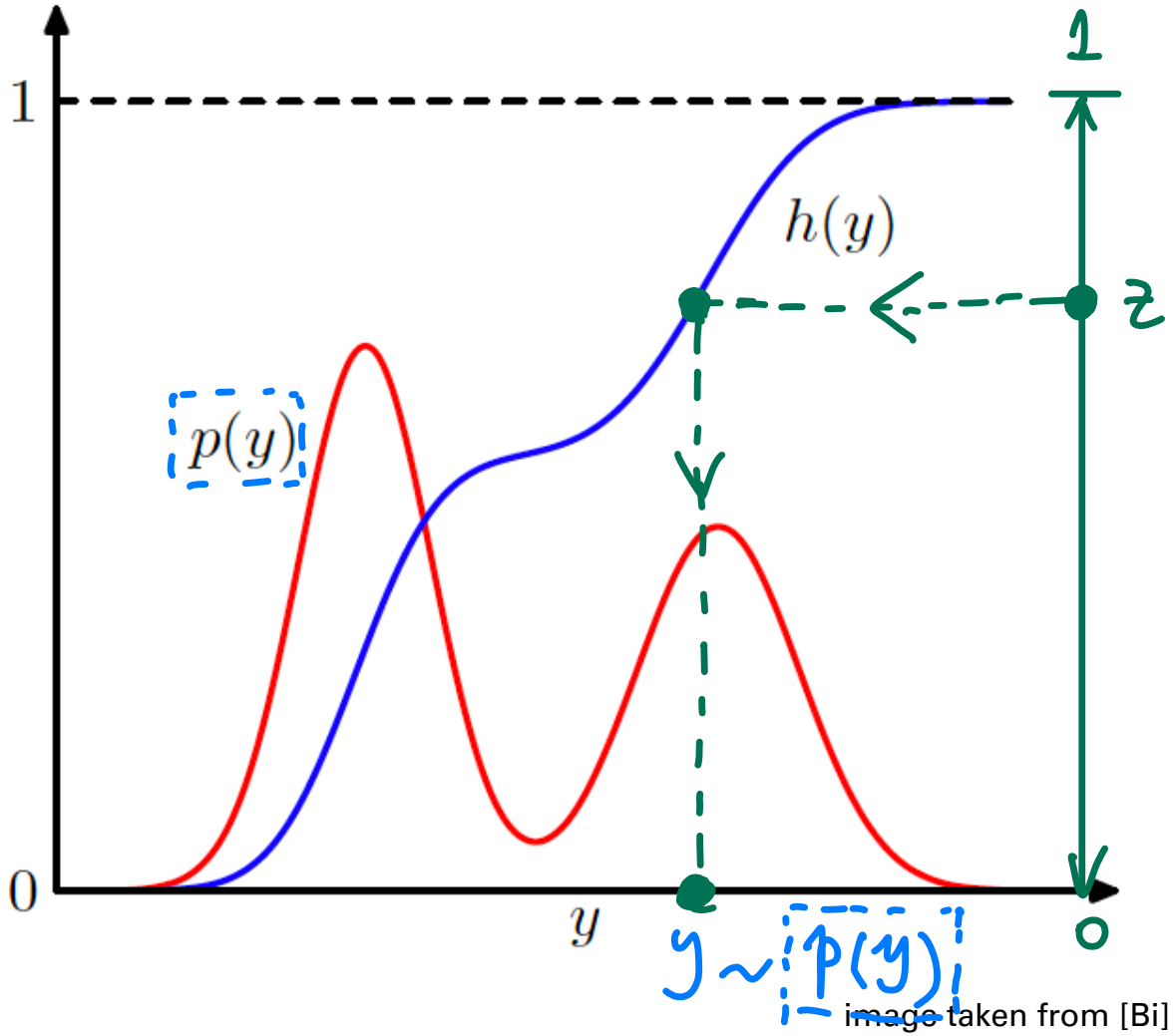


image taken from [Bi]

example: exponential distribution

$p_{\mathrm{y}}(y) = \lambda \exp(-\lambda y)$ where $0 \leq y < \infty$

$$h(y) = \int_{-\infty}^{y} p_{\mathrm{y}}(\hat{y})\, d\hat{y} = \int_{0}^{y} \lambda \exp(-\lambda \hat{y}) d\hat{y} = 1 - \exp(-\lambda y)$$

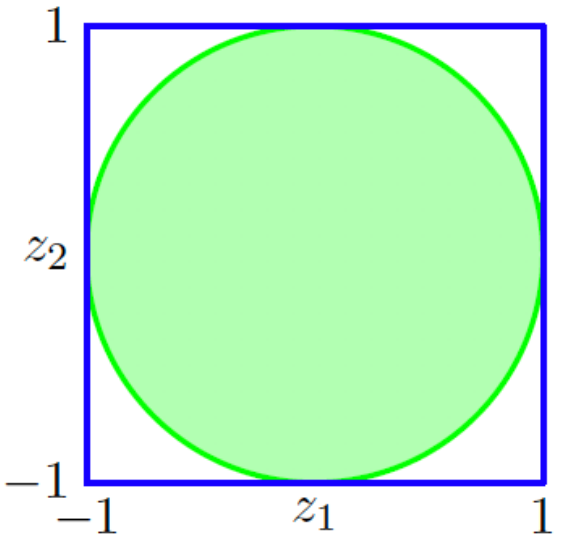
$$f(z) = h^{-1}(z) = -\lambda^{-1} \ln(1 - z)$$

sample from multivariate distribution

- $p_{\mathbf{y}}(y_1, \cdots, y_M) = p_{\mathbf{z}}(z_1, \cdots, z_M) \left| \frac{\partial(z_1, \cdots, z_M)}{\partial(y_1, \cdots, y_M)} \right|$

example: Box-Muller method for Gaussian

The Box-Muller method for generating Gaussian distributed random numbers starts by generating samples from a uniform distribution inside the unit circle.



- First, uniformly sample $(z_1, z_2)^{\mathrm{T}}$ from a unit disk.

How?

example: Box-Muller method for Gaussian

- Next, apply the transform: $y_1 = z_1\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$, $y_2 = z_2\left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$ where $r^2 = z_1^2 + z_2^2$.

Then it is easy to verify:

$$p(y_1, y_2) = p(z_1, z_2)\left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right| = \left[\frac{1}{\sqrt{2\pi}}\exp(-y_1^2/2)\right]\left[\frac{1}{\sqrt{2\pi}}\exp(-y_2^2/2)\right]$$

Questions?

Reference

- *Gaussian processes:*
  - *[Bi] Ch.6.4.1-6.4.2, 6.4.5*
  - *[HaTF] Ch.5.8.1-5.8.2*
- *RKHS:*
  - *[HaTF] Ch.5.8*
- *Sampling - Simple methods:*
  - *[Bi] Ch.11.1*