Advanced Topics in Statistical Machine Learning

Contents

Notation 0.1. $\lambda \succeq 0$ implies that $\forall i\ \lambda_i \geq 0$

Convexity - can also do via Hessian being positive semi-definite.

can do pos (semi) def by considering $z^\top Az$ for arbitrary z, seeing if everything turns out to be a square....

trick for computing distributions, e.g. posteriors - $p(w|\mathcal{D})$ must be a distribution, so we can lose all the scaling terms and just compare shape in w

trick: for a scalar λ, $\lambda = Trace(\lambda)$, and then if $\lambda = x^\top z$, say, $\lambda = Tr(x^\top z) = Tr(z^\top x)$

1 Basics/background

supervised learning, e.g. classification or regression

unsupervised learning

discriminative v. generative

1.1 ERM

[on discriminative supervised learning]

assume there is a joint distribution $\mathbb{P}[X, Y]$, have an iid sample from it, $\mathcal{D}$

learning a function $f : \mathcal{X} \to \mathcal{Y}$ that will approximate $Y|X = x$

loss function general form

$$L : \mathcal{Y} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$$

almost always doesn't depend on x, so generally $L(y, f(x))$

risk

$$R(f) := \mathbb{E}_{X,Y}[L(Y, f(X))]$$

hypothesis space $\mathcal{H}$, optimal f (in $\mathcal{H}$) is

$$f^* = \arg\min_{f \in \mathcal{H}} R(f)$$

(or optimising over parameters $\theta \in \mathcal{H}$

empirical risk

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i))$$

if f is chosen independently of the data used in the empirical risk, this is an unbiased estimate of the risk. . . . but if f is trained using this data it becomes a negatively biased because we have actively chosen f for which it will be smaller

loss functions

lots available

softmax/ mappings with sign/sigmoid to turn real line to probs

0/1 loss - Bayes classifer is optimal

hinge loss - in SVMs,

expo

1.2 Constrained optimisation

Primal problem:

$$\begin{aligned} \text{minimise: } & f(\boldsymbol{x}) & \\ \text{st } & f_i(\boldsymbol{x}) \leq 0 & i \leq m \\ & h_j(\boldsymbol{x}) = 0 & j \leq n \end{aligned}$$

the **(primal) optimum** value is $p^* = f_0(x^*)$. any x st $f_i(x) \leq 0, h_j(x) = 0$ for all i, j is a **primal feasible** point

Lagrangian:

$$L(x; \lambda, \nu) := f_0(x) + \sum_{i-1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{r} \nu_j h_j(x)$$

the **dual variables** are $\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^r$.

dual problem:

$$\begin{aligned} \text{maximise: } & \inf_{x \in \mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) & \\ \text{st } & \lambda_i \geq 0 & i \leq m \end{aligned}$$

to actually write this down: use differentiation/etc. to find the $\inf_{x \in \mathcal{D}}$

$d^* := \sup_{\lambda \succeq 0, \nu} \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$

weak duality: $d^* \leq p^*$

1.3 Lagrangian explanation

Given the primal problem f, we can consider

$$\tilde{f}(x) := f_0 \sum_{i=1}^{m} \infty_{f_i(x) > 0} + \sum_{j=1}^{r} \infty_{h_j(x) \neq 0},$$

and minimising $\tilde{f}$is equivalent to minimising f under the constraints. Then

$$\sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu) = \tilde{f(x)}$$

If we have no inequality constraints, considering the Lagrangian $L(x, \nu) := f(x) + \nu h(x)$ and differentiating, then $\nabla_{x,\nu} L(x, \nu) = 0 \iff h(x) = 0$ and $\nabla_x f(x) = -\nu \nabla_x h(x)$, which implies that we have the optimum value within the constraints, as the gradients of the objective and the constraints are opposite.

2 SVMs

two varieties: separable and inseparable data

margin for given weights w:

$$M(w) : +2 \min_i \frac{1}{\|w\|} \|w^\top x_i + b\|$$

so the aim is to solve

$$\begin{aligned} \text{maximise: } & M(w) \\ \text{over } & w \in \mathbb{R}^d \end{aligned}$$

derivation:

- can rescale the weights however we like
- for any two points x_+, x_i st $w^\top x_+ + b = 1, w^\top x_- + b = -1$, so $\|w\|\|x_+ - x_i\| = 2$, so the margin is $2/\|w\|$
- when all the points are classified correctly - i.e. $\min_i y_i(w^\top x_i + b) = 1$, relaxed to ≥ 1
- **check start of proof, wasn't paying attention**

2.0.1 Inseparable data

add hinge loss

which rescales to regularised ERM

add new variables ξ_i, which are constrained to $\xi_i \geq h(y_i(w^\top x_i + b))$ by $\xi_i \geq 0$, $\xi_i \geq 1 - y_i(w^\top x_i + b)$

note we need to prove that it equals it at the optimum: simple logic based on the constraints used, fact we can decrease ξ_i if > 0 and not equal to $1 - y_i(w^\top x_i + b)$

standard form:

$$\begin{aligned} \text{minimise: } & f_0(w, b, \xi) := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{st } & f_i(w, b, \xi) := 1 - \xi_i - y_i(w^\top x_i + b) \leq 0 && i \leq m \\ & f_{n+i}(w, b, \xi) := -\xi_i \leq 0 && j \leq n \end{aligned}$$

Strong duality will hold, as the problem is convex with afffine constraints and a feasible solution will exist

So the Lagrangian is

$$\begin{aligned} L(w, b, \xi; \alpha, \lambda) := \quad & \tfrac{1}{2}\|w\|^2 + C\textstyle\sum_{i=1}^{n} \xi_i \\ & + \textstyle\sum_{i=1}^{n} \alpha_i(1 - \xi_i - y_i(w^\top x_i + b)) + \sum_{i=1}^{n} \lambda_i(-\xi_i) \end{aligned}$$

which we differentiate wrt the primal variables to get the dual $g(\alpha, \lambda) := \inf_{w,b,\xi} L(w, b, \xi; \alpha, \lambda)$

- to get $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, $\sum_{i=1}^{n} y_i \alpha_i = 0$ and $\alpha_i = C - \lambda_i$,
- so we get rid of λ and have $\alpha_i \in [0, C]$ by the constraint on λ.

why we have equality constraints on the dual:

The result of the derivatives gives us $\sum_{i=1}^{n} y_i \alpha_i = 0$ and $\alpha_i = C - \lambda_i$, which do not include the primal variables, so they must hold, or the dual function $g = -\infty$, as we can always change b/ξ_i

dual program:

$$\begin{aligned} \max_\alpha & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{st } & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & \&0 \preceq \alpha \preceq C \end{aligned}$$

calculating vars

w is obvious, b is $y_{i_{margin}} - w^\top x_{i_{\text{margin}}}$ for any i st $0 < \alpha_i < C$

types of datapoint:

- non support vectors, if $\alpha_i = 0$

  - so $\lambda_i = C - \alpha_i > 0$, so $\xi_i = 0$
  - and so $y_i(w^\top x_i + b) \geq 1$ (almost always > 1)
- margin support vectors, if $\alpha_i \in (0, C)$:
  - so must have that $y_i(w^\top x_i + b) = 1 - \xi_i$, and since $\lambda_i > 0, \xi_i = 0$, so $y_i(w^\top x_i + b) = 1 =$ on the boundary
- margin errors/ non-margin Support Vectors: $\alpha_i = C > 0$
  - so $y_i(w^\top x_i + b) = 1 - \xi_i$, and since $\lambda_i = 0, \xi_i \geq 0$, which means $y_i(w^\top x_i + b) \leq 1$, which is a margin error (as within the margin (even if classified correctly

Insights in the solution:

what a support vector is

bounded influence

weights are in the span of the datapoints

multi-class

1-vs-all

1-v-1, pick class that wins the most

3 Kernel methods

feature map $x \mapsto \varphi(x)$, where $\varphi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is often infinite dimensional

in SVMs, data only appears in an inner product, so we just replace $x_i^\top x_j$ with $k(x_i, x_j)$, where k is the inner product in the space $\mathcal{H}$.

Note the weights can't be expressed with just k, as $w = \sum_{i=1}^n \alpha_i y_i \varphi(x_i)$, though the intercept b can:

$$b = y_{\text{margin}} - \sum_{i=1}^n \alpha_i y_i \underbrace{\phi(x_i)^\top \phi(x_{\text{margin}})}_{k(x_i, x_{\text{margin}})}$$

however the decision function is still good:

$$\hat{y}(x) = \text{sign}(w^\top \phi(x) + b) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i k(x, x_i) + b\right)$$

Standard kernels:

- **polynomial kernel** $k(x_i, x_j) = (1 + x_i^\top x_j)^d$ introduces d-order polynomial interactions, φ is a horrible map
- **exponential/RBF kernel:**

  $$k(x_i, x_j) = \exp\left(-\frac{1}{2\gamma^2}\|x_i - x_j\|_2^2\right)$$

  where the feature mapping is

  $$\varphi(x) = \exp\left(-\frac{1}{2}x^2\right)\left[1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, ..., \frac{x^r}{\sqrt{r!}}, ...\right]^\top$$

3.1 RKHS

standard inner products (on spaces over $\mathbb{R}$), standard Hilbert space

Definition 3.1 (Kernel)**.** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if $\exists$a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ st $\forall x, x' \in \mathcal{X}$

$$k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$$

Definition 3.2 (Positive definite function)**.** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is postive definite if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0 \ \ \forall n \geq 1, \forall a_i \in \mathbb{R}, \forall x_i \in \mathcal{X}$$

which is equivalent to saying the matrix $(K_{i,j})_{i,j}$ defined by $K_{i,j} = k(x_i, x_j)$ is positive **semi-definite**

Notation 3.3. For kernels, we have strictly pos-def /pos def, for matrices pos-def/ pos-semi-def.

Lemma 3.4. *All kernels are positive definite (proof by rearranging sums*

3.2 RKHSs

Definition 3.5 (RKHS, reproducing kernel)**.** Given $\mathcal{H}$ is a Hilbert space $\subseteq \{f : X \to \mathbb{R}\}$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** of $\mathcal{H}$ if:

- $\forall x \in \mathcal{X} : k_x = y \mapsto k(y, x)$ is in $\mathcal{H}$
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$ $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property)

If $\mathcal{H}$ has such a k, then it is a RKHS.

Lemma 3.6 (Reproducing kernels are kernels)**.** *A reproducing kernel k is a normal kernel with the feature map $\varphi : x \mapsto k(\cdot, x)$, which is the* **canonical feature map**

Proof. Given $f = k(\cdot, x')$, we have [using defs of k, φ, fact $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is symmetric]
$k(x, x') = f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \langle k(\cdot, x'), k(\cdot, x) \rangle_{\mathcal{H}} = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ □

Theorem 3.7 (Moore-Aronszajn)**.** *every pos def function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is also a reproducing kernel with a unique corresponding RKHS.*

Proof. assume k is symmetric, as well as pos def [**prove this!**]

So we want to specify a RKHS associated with k.

Consider functions of the form $f = \sum_{i=1}^{r} \alpha_i k(\cdot, x_i)$ for $r \geq 1, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X}$.

Let $\mathcal{H}_0 = \text{span}\{k(\cdot, x) : x \in \mathcal{X}\}$, so $\mathcal{H}_0$ is the set of these f.

if $\mathcal{H} \supset \mathcal{H}_0$, then $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$

Define a function h st $h(k(\cdot, x), k(\cdot, x')) = k(x, x')$ of the form $h : (\mathcal{X} \to \mathbb{R}) \times (\mathcal{X} \to \mathbb{R}) \to \mathbb{R}$

define an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ by $\left\langle \sum_{i=1}^{r} \alpha_i k(\cdot, x_i), \sum_{j=1}^{s} \beta_j k(\cdot, x'_j) \right\rangle_{\mathcal{H}} := \sum_{i=1}^{r} \sum_{j=1}^{s} \alpha_i \beta_j h(k(\cdot, x_i), k(\cdot, x'_j)) = \sum_{i=1}^{r} \sum_{j=1}^{s} \alpha_i \beta_j k(x_i, x'_j)$

This is an inner product [linear, symmetric, non-neg norm, 0 iff f=0]

finally, take $\mathcal{H}$ to be the completion of $\mathcal{H}_0$, and extend $\langle \rangle_{\mathcal{H}}$ by completion.

And finally, $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \langle \sum_{i=1}^{r} \alpha_i k(\cdot, x_i), k(\cdot, x) \rangle = \sum_{i=1}^{r} \alpha_i k_i(x_i, x) = f(x)$ [by symmetry of k] □

3.2.1 More on RKHSs

Definition 3.8 (Alt def of RKHSs)**.** $\mathcal{H}$ is an RKHS if the evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, defined by $\delta_x f = f(x)$ are continuous for all $x \in \mathcal{X}$, or equivalently if δ_x is a bounded operator - i.e. $\|\delta_x\|_{\mathcal{H}^*} < \infty$.

Note this implies $\|f - g\|_{\mathcal{H}} = 0 \implies f(x) = g(x)$ for all $x \in \mathcal{X}$.

Proposition 3.9 (Equivalence of definitions)**.**

Proof. first def $\implies$ second: $|\delta_x f| = |f(x)| = |\langle f, k(\cdot, x)\rangle| \leq \|f\|_{\mathcal{H}}\|k(\cdot, x)\|_{\mathcal{H}} = \sqrt{k(x,x)} \cdot \|f\|_{\mathcal{H}}$

other way uses R-r theorem □

Proposition 3.10 (Uniqueness of reproducing kernels)**.** *Each RKHS has a unique corresponding reproducing kernel.*

Proof. assume there exists 2, $k_1 \neq k_2$. Then $\langle f, k_1(\cdot, x) - k_2(\cdot, x)\rangle_{\mathcal{H}} = \cdots = 0$, apply with $f = k_1(\cdot, x) - k_2(\cdot, x)$ □

Proposition 3.11 (Uniqueness of RKHS)**.** *For any given kernel/pos def func, the RKHS is unique.*

Theorem 3.12 (Representer theorem)**.** *There is always a solution to*

$$f^* = \arg\min_{f \in \mathcal{H}_k} \hat{R}(f) + g(\|f\|^2_{\mathcal{H}_k})$$

that takes the form

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$$

where the x_i are the datapoints. If g is a strictly increasing function, all solutions have this form.

Proof. let f_S be the projection of a function f onto span$\{k(\cdot, x_i) : i\}$, and let $f_\perp = f - f_S$. Obviously $f_S = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for some $\alpha_i \in \mathbb{R}$. Then $\|f\|^2_{\mathcal{H}_k} = \|f_S\|^2_{\mathcal{H}_k} + \|f_\perp\|^2_{\mathcal{H}_k} \geq \|f_S\|^2_{\mathcal{H}_k}$ by Pythagoras. Thus, $g(\|f\|^2_{\mathcal{H}_k}) \geq g(\|f_\perp\|^2_{\mathcal{H}_k})$.

$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^n L(y, f(x_i)) = \frac{1}{n}\sum_{i=1}^n L(y, \langle f_S, k(\cdot, x_i)\rangle_{\mathcal{H}_k} = \hat{R}(f_S)$ using def of $f \in$RKHS, orthogonality □

Remark 3.13 (on the implications of the Representer theorem).

- we can work with complex RKHS hypothesis classes, but solution is still fairly simple
- kernel directly affects solution
- complexity is limited to n - good for large d, bad that we need to keep all data.

Remark 3.14 (on kernels/RKHSs).

- RKHS are general and powerful
- can do ERM with solution in RKHS
- simple/analytic solutions to ERM as solutions end up being linear maps of kernels
- choosing kernel is v. important - some RKHSs too restrictive, others too broad
- bad scaling - at least $O(n^2)$ at training, $O(n)$ at test

3.3 Constructing kernels

3 methods we have:

- define the feature map, take the inner product
- as a positive definite function
- choosing the RKHS $\mathcal{H}$, and taking the unique reproducing kernel associated with it.

Lemma 3.15 (Mapping between spaces)**.** *given $A : \mathcal{X} \to \tilde{\mathcal{X}}$, and a kernel k on $\tilde{\mathcal{X}}$, $k(A(x), A(x'))$ is a kernel on $\mathcal{X}$*

Lemma 3.16 (sums of kernels)**.** *given k_1, k_2 on $\mathcal{X}$ and $\alpha_1, \alpha_2 > 0$, $k = \alpha_1 k_1 + \alpha_2 k_2$ is a kernel [prove with pos def]*

Lemma 3.17 (products of kernels). *k_1 on $\mathcal{X}$, k_2 on $\mathcal{Y}$, then $k((x,y),(x',y)) := k_1(x,x')k_2(y,y')$ is a kernel, and if $\mathcal{Y} = \mathcal{X}$, then $k(x,x') = k_1(x,x')k_2(x,x')$ is also a kernel.*

Definition 3.18 (Common kernels).

- **RBF**: $k(x,x') = \exp(-\frac{1}{2\gamma^2}\|x-x'\|_2^2)$. The RKHS contains functions which are infinitely differentiable
- **Matérn** kernels: less smooth, only s times differentiable
  - general [slightly useless] form includes the Bessel function...
  - for $\nu = s + 1/2$,
    - $\nu = 1/2$: $k(x,x') = \exp(-\frac{1}{\gamma}\|x-x'\|)$
    - $\nu = 3/2$: $k(x,x') = (1+\frac{\sqrt{3}}{\gamma}\|x-x'\|)\exp(-\frac{\sqrt{3}}{\gamma}\|x-x'\|)$
    - ....
  - as $\nu \to \infty$ this converges to the RBF kernel
  - $\|f\|_{\mathcal{H}_k}^2 \propto \int f''(x)^2 dx + \frac{6}{\gamma^2}\int f'(x)^2 dx + \frac{9}{\gamma^4}\int f(x)^2 dx$ for $\nu = 3/2$, demonstrating that $\|f\|_{\mathcal{H}_k}$ directly penalises derivatives
- **constant** $k(x,x') := c > 0$ [useful for sums!]
- **linear**: $k(x,x') = x^\top x'$
- **poly**: $k(x,x') = (c + x^\top x')^m$ for $c \in \mathbb{R}, m \in \mathbb{N}$ **what about** $m = 1, c < 0$**?**
- **periodic (1d)** $k(x,x') = \exp(-\frac{2\sin^2(\pi|x-x'|/p)}{\gamma^2})$ for $\gamma \neq 0$, which has period p,
- **Laplace**: Matern w/ $\nu = 1/2$
- **Rational quadratic** $k(x,x') = \left(1 + \|x-x'\|_2^2/(2\alpha\gamma^2)\right)^{-\alpha}$ for $\alpha, \gamma > 0$

Remark 3.19. In kernel (ridge) regression, we are doing linear regression in the hypothesis space, so we are comparing y_i to $\langle f, k(\cdot, x_i)\rangle_{\mathcal{H}}$, so the reproducing property is v. useful, as this simplifies to $f(x_i)$.

Remark 3.20. Confused about interpretation of sums of kernels etc - point about fitting linear reg, then doing stuff on residuals

Remark 3.21 (Issues with choosing kernels).

- many kernels use Euclidean distances, but in high-D everything may far away from each other
- challenge is more about deciding which points are similar/close than ensuring the predictor is powerful enough to discriminate/fit

4 Bayesian chapter?

5 Gaussian processes

parametric models collapse to the modal θ of the posterior distribution $p(\theta|\mathcal{D})$

non-parametric (in the sense that there's no w that we do $w^\top x$ with) don't have this problem, and if the model is set up correctly, θ may be infinite dimensional, but it marginalises out to something finite dimension.

instead of having a prior over θ and the model using some set of functions f_θ, we have a prior over f directly, and the posterior (assuming IID data, distributed as $y_i \sim f(x_i) + \sigma^2\varepsilon$) becomes

$$p(f|\mathcal{D}) \propto p(f)\prod_{i=1}^{N} p(y_i|f(x_i))$$

in practise, we want to only work with the set of RVs $\{f(x_i)\}_{i\in[N]}$....

Definition 5.1 (Gaussian process)**.** f is a GP if it is a stochastic process whos evaluations are jointly Gaussian - i.e. $[f(x_1), ..., f(x_N)]^\top$ is multivariate Gaussian for any N, x_i. It is specified by its mean function $m : \mathcal{X} \to \mathbb{R}$, and a covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, so that $m(x) = \mathbb{E}f(x)$ and $k(x, x') = \mathbb{E}\left[(f(x) - m(x))(f(x') - m(x'))\right]$, so that

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix} \sim \mathcal{N}\left(\underbrace{\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}}_{\boldsymbol{m}}, \underbrace{\begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix}}_{\boldsymbol{K}} \right)$$

Definition 5.2 (GP priors)**.** Define a prior over m and k, though typically $m = 0$, as GPs are linear in their mean

Lemma 5.3 (Gaussian marginals and conditionals)**.** *Given $\boldsymbol{z} \sim \mathcal{N}(\mu, \Sigma)$, we split its dimensions as*

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix}, \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

note that $\Sigma_{21} = \Sigma_{12}^\top$ because of symmetry - **why?**

Then $p(\boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_1; \mu_1, \Sigma_{11})$, and

$$p(\boldsymbol{z}_2, \boldsymbol{z}_1) = \mathcal{N}(\boldsymbol{z}_2; \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{z}_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

6 GPs

WRITE UP SLIDES 1 AND 2

6.1 Approximations

Want to reduce $O(n^3)$ cost. 2 approaches - approx K_{xx} with a rank m approximation, so inversion is only $O(m^3)$, or we summarise the dataset with m **inducing** datapoints, and fit the model on that dataset.

6.1.1 Low rank matrix approximations

If $\tilde{K}_{xx}$ is a symmetric rank m approx to K_{xx}, then $\tilde{K}_{xx} = QQ^\top$ for a $n \times m$ matrix Q.

Thus,

$$\left(\tilde{K}_{xx} + \sigma^2 I\right)^{-1} = \sigma^{-2} I - \sigma^{-2} Q \left(\sigma^2 I + Q^\top Q\right)^{-1} Q^\top,$$

where $Q^\top Q$ is $m \times m$, and can be caluclated in $O(m^2 n)$. Fully calculating the inverse still takes $O(n^2 m)$, but applying it to a vector β only takes $O(m^2 n)$.

Q represents a m-dimensional feature mapping of X

Theorem 6.1 (Bochner's)**.** *Given $k(x, x')$ is a stationary kernel $k(x, x') = \kappa(x - x')$, then*

$$k(x, x') = 2\kappa(0)\mathbb{E}\left[\cos(\omega^\top x + b)\cos(\omega^\top x' + b)\right],$$

where $b \sim$ Unif$(0, 2\pi)$, and ω has density given by

$$p(\omega) \propto \int_{\delta \in \mathbb{R}^p} \kappa(\delta)\cos(\omega^\top \delta)\,\mathrm{d}\delta$$

Definition 6.2 (Random Fourier Features)**.** For many common kernels, $p(\omega)$ is simple - e.g. for RBF, $p(\omega) \sim \mathcal{N}(0, \gamma^{-2} I)$

So we can actually form an unbiased Monte Carlo estimate of the kernel by sampling ω's and $b's$, and writing the approximation as an inner product of feature maps

$$\varphi_m(x) = \sqrt{\frac{2\kappa(0)}{m}}\left[\cos(\omega_1^\top x + b_1), ..., \cos(\omega_1^\top x + b_m)\right]^\top$$

Remark 6.3. On RFF: it works pretty well on any finite interval, as $m \to \infty$ (e.g. $m = 100$ works pretty well), but it produces periodic functions, which means uncertainty estimates repeat, which is not ideal.

6.1.2 Sparse Gaussian Processes

summarise the datasat with m "pseudo" datapoints. these often overpredict uncertainty.

7 Deep Learningn