STAT G8325
Gaussian Processes and Kernel Methods
Lecture Notes §04: Kernels

John P. Cunningham

Department of Statistics
Columbia University

Outline

Administrative interlude

Basic kernels

Kernel algebra [Gen02]

Spectral mixture kernels [WA13, WGNC14]

Transition

Graph kernels [VSKB10]

Ranking kernels [JV15]

Outline

Administrative interlude

Basic kernels

Kernel algebra [Gen02]

Spectral mixture kernels [WA13, WGNC14]

Transition

Graph kernels [VSKB10]

Ranking kernels [JV15]

Progress...

Week	Lectures	Content
3	Sep 23,28,30	Approximate inference
4	Oct 5,7	Kernels • Reading: [RW06, ch. 4-4.2; 4.4]; [Gen02]; [WA13, WGNC14]; [VSKB10]; [JV15]
5	Oct 12,14	Speed and scaling part 1: reduced-rank processes

- HW02 is due this Friday (see courseworks).
- Project brainstorming list available on courseworks.
- Make an appointment with me in the next two weeks.
- Anyone interested in Stan, talk to Daniel.

Comment

- In these lectures we will deal with kernels from an applied perspective, as that is the level of depth needed to use them in a gp context.

- Prior to discussing kernel statistical tests, we will return to kernels to deeply understand RKHS, Mercer's theorem, Moore-Aronsajn, etc., which will only be needed at that time.

Outline

Kernels, properly

Definition (kernel). Given a non-empty input set $\mathcal{X}$, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there is a Hilbert space $\mathcal{H}$, called the feature space, and a map $\phi : \mathcal{X} \to \mathcal{H}$, called the feature map, with:

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \quad , \qquad \forall x, x' \in \mathcal{X}.$$

- Warning: ϕ is not unique to k; $\mathcal{H}$ is often infinite dimensional.
- We'll return to Mercer's theorem, which is a set of necessary and sufficient conditions for this definition.
- Note this immediately implies kernel matrices are positive semidefinite:

$$\begin{aligned}
v^\top K v &= \sum_{i=1}^n \sum_{j=1}^n v_i k(x_i, x_j) v_j &&= \sum_{i=1}^n \sum_{j=1}^n v_i \left\langle \phi(x_i), \phi(x_j) \right\rangle_{\mathcal{H}} v_j \\
&= \sum_{i=1}^n \sum_{j=1}^n \left\langle v_i \phi(x_i), v_j \phi(x_j) \right\rangle_{\mathcal{H}} &&= \left\langle \sum_{i=1}^n v_i \phi(x_i), \sum_{j=1}^n v_j \phi(x_j) \right\rangle_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^n v_i \phi(x_i) \right\|_{\mathcal{H}}^2 &&\geq 0 \quad \forall v \in \mathbb{R}^n.
\end{aligned}$$

Recall mechanics

Example kernel (squared exponential or SE):

$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$$

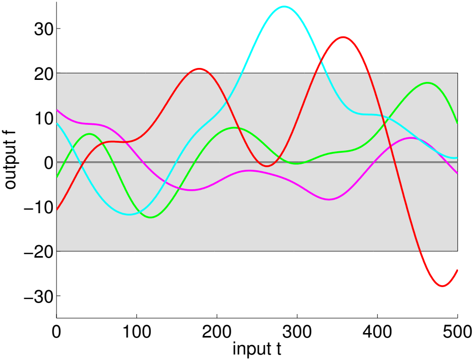
From kernel to covariance matrix

- Choose some *hyperparameters*: $\sigma_f = 7$, $\ell = 100$

$$t = \begin{bmatrix} 0700 \\ 0800 \\ 1029 \end{bmatrix} \qquad K(t,t) = \{k(t_i, t_j)\}_{i,j} = \begin{bmatrix} 49.0 & 29.7 & 00.2 \\ 29.7 & 49.0 & 03.6 \\ 00.2 & 03.6 & 49.0 \end{bmatrix}$$

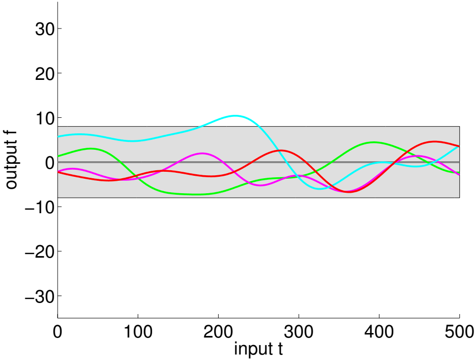
Impact of hyperparameters

- Squared exponential with $\sigma_f = 10$, $\ell = 50$



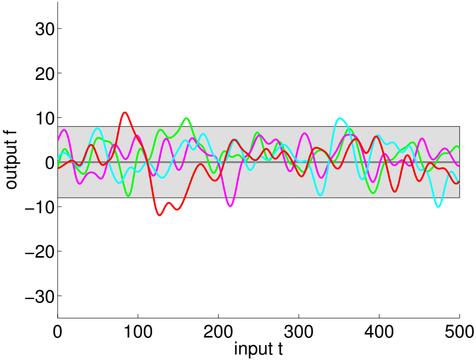
Impact of hyperparameters

- Squared exponential with $\sigma_f = 4$, $\ell = 50$



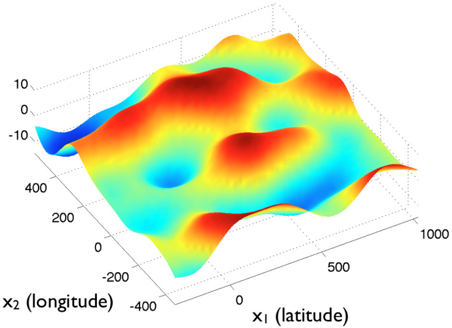
Impact of hyperparameters

- Squared exponential with $\sigma_f = 4$, $\ell = 10$



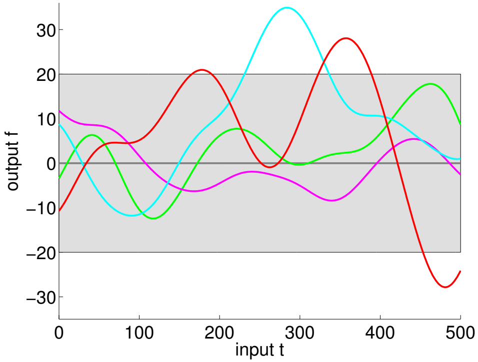
Multidimensional input

- Inputs $x \in \mathbb{R}^D$ (here $D = 2$, *e.g.* lat and long)
- $f \sim \mathcal{GP}(0, k_{ff})$, where $k_{ff}(x^{(i)}, x^{(j)}) = \sigma_f^2 \exp\left\{-\sum_d \frac{1}{2\ell_d^2}(x_d^{(i)} - x_d^{(j)})^2\right\}$



Squared exponential (exponentiated quadratic)

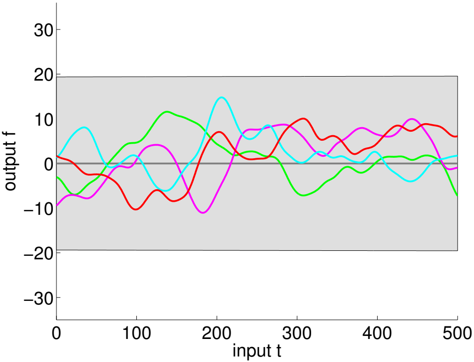
$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$$



Rational Quadratic

$$k(t_i, t_j) = \sigma_f^2 \left(1 + \frac{1}{2\alpha\ell^2}(t_i - t_j)^2\right)^{-\alpha}$$

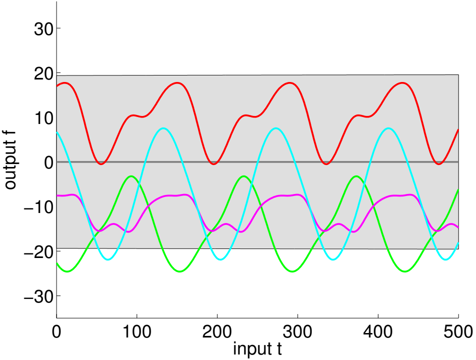
$$\propto \quad \sigma_f^2 \int z^{\alpha-1} \exp\left(-\frac{\alpha z}{\beta}\right) \exp\left(-\frac{z(t_i - t_j)^2}{2}\right) dz$$



Why does this look like an unnormalized t distribution?

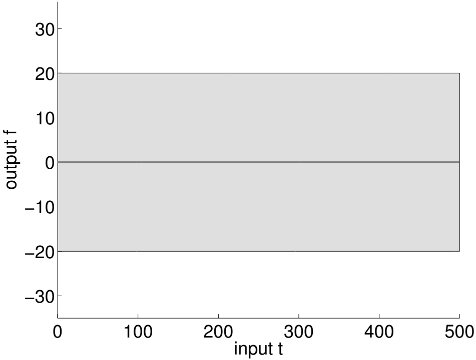
Periodic

$$k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{2}{\ell^2}\sin^2\left(\frac{\pi}{p}|t_i - t_j|\right)\right\}$$



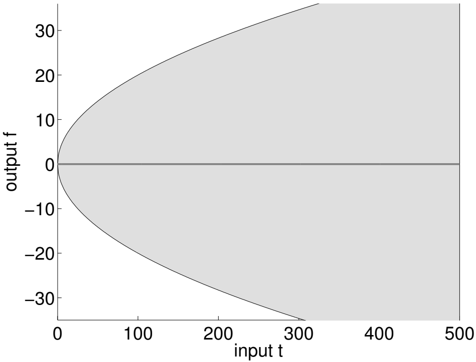
From Stationary to Nonstationary Kernels

- $k(t_i, t_j) = k(t_i - t_j) = k(\tau)$
- $k(t_i, t_j) = \sigma_f^2 \exp\left\{-\frac{1}{2\ell^2}(t_i - t_j)^2\right\}$



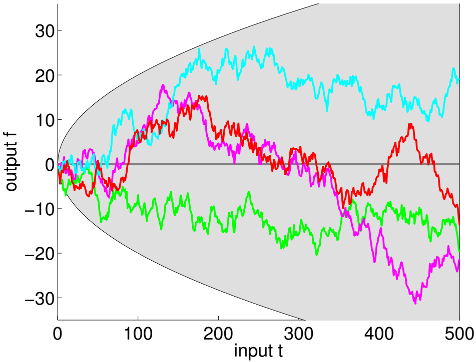
Brownian Motion

- $k(t_i, t_j) = \min(t_i, t_j)$
- Simple nonstationary gp



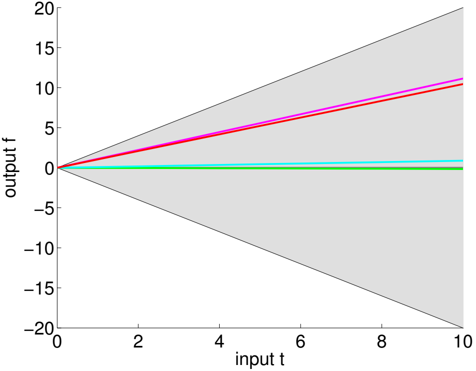
Brownian Motion

- $k(t_i, t_j) = \min(t_i, t_j)$
- Draws from this gp



Bayesian Linear Regression

- $f(t) = wt$ with $w \sim \mathcal{N}(0, 1)$
- $k(t_i, t_j) = E[f(t_i)f(t_j)] = t_i t_j$



Other popular kernels

- Polynomial kernels:

$$k(x_i, x_j) = \left(x_i^\top x_j + c\right)^d$$

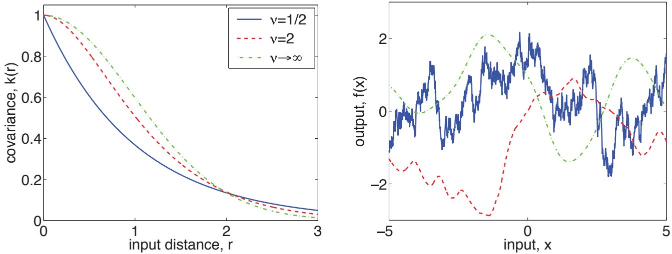
  - Constraints on c, d?

$c \geq 0$, $d \in \mathbb{N}_{+1}$.

- Matérn kernels:

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu r}}{\ell}\right)^\nu B_\nu \left(\frac{\sqrt{2\nu r}}{\ell}\right)$$

  - $\nu, \ell > 0$, modified Bessel function B_ν.



String kernels

$$k(x,x') = \sum_{s \in \mathcal{P}(\mathcal{A})} w_s \phi_s(x) \phi_s(x'),$$

- where $\mathcal{P}(\mathcal{A})$ denotes the powerset of an alphabet $\mathcal{A}$, and $\phi_s(x)$ is the number of occurrences of s in x.
- What constraint should their be on w_s?

$w_s \geq 0$

- A standard example (e.g., [RW06, ch. 4.4.1]) is size-biased $w_s = \lambda^{|s|}$ for $\lambda \in (0,1)$.
- $w_s = 0 \quad \forall s : |s| > 1$ is bag of characters.
- $w_s = 0 \quad \forall s : s = \backslash \textvisiblespace * \textvisiblespace \backslash$ is bag of words.
- and so on...

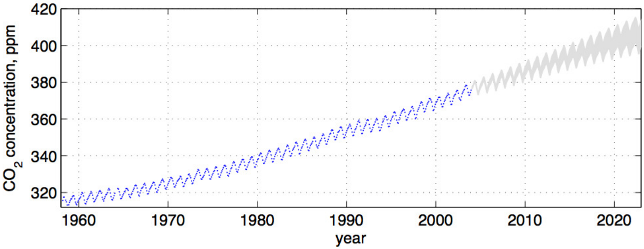
Outline

Linearity

- For kernels k_a and k_b and constants $\alpha, \beta \geq 0$, their sum is also a kernel:

$$k(x, x') = \alpha k_a(x, x') + \beta k_b(x, x').$$

- Note this implies that kernels form a convex cone.
- This is useful for modeling additive trends, e.g.:



See [RW06, §5.4.3].

Multiplicity

- For kernels k_a and k_b, their product is also a kernel:

$$k(x, x') = k_a(x, x')k_b(x, x').$$

- To do this in full generality requires some more machinery (soon).
- For now we can prove the finite case $\phi : \mathcal{X} \to \mathbb{R}^q$, for $q < \infty$.
- Hint: use $\langle A, B \rangle_{\mathbb{R}^{q_a \times q_b}} = tr(A^\top B)$.
- Along with linearity, any positive polynomial is also a kernel:

$$p_+ (k(x, x')) = \sum_{k=1}^{K} \alpha_k k(x, x')^k \qquad \forall \alpha_k \geq 0.$$

- In particular then, $\exp(k(x, x'))$ is a kernel (Taylor expansion).

More properties

- Integration:

$$
\begin{aligned}
z(t) &= \int g(u,t) f(u) du \quad \leftrightarrow \\
k_z(t_i, t_j) &= \int\int g(u, t_i) k_f(t_i, t_j) g(v, t_j) du dv.
\end{aligned}
$$

- Differentiation:

$$
\begin{aligned}
z(t) &= \frac{\partial}{\partial t} f(t) \quad \leftrightarrow \\
k_z(t_i, t_j) &= \frac{\partial^2}{\partial t_i \partial t_j} k_f(t_i, t_j).
\end{aligned}
$$

- Warping:

$$
\begin{aligned}
z(t) &= f\left(h(t)\right) \quad \leftrightarrow \\
k_z(t_i, t_j) &= k_f\left(h(t_i), h(t_j)\right).
\end{aligned}
$$

Variance function trick [Gen02]

- Let $h : \mathcal{X} \to \mathbb{R}_+$ be a positive function with minimum at $x = 0$.

- Then the following is a kernel:

$$k(x, x') = \frac{1}{4} \left(h(x + x') - h(x - x') \right).$$

- This exploits the covariance identity:

$$cov(w_1, w_2) = \frac{1}{4} \left(var(w_1 + w_2) - var(w_1 - w_2) \right).$$

Bochner's theorem

- A stationary kernel $k(t_i, t_j) = k(t_i - t_j) = k(\tau)$ is positive semidefinite iff:

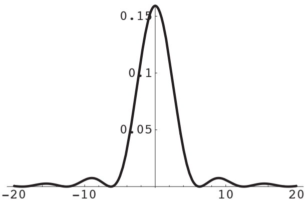
$$S(\omega) = \mathcal{F}\{k\}(\omega) \geq 0 \;\; \forall \;\; \omega.$$

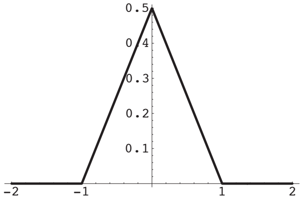
- In other words, the power spectral density is nonnegative everywhere.
- Sometimes also handy: $k(0) = \int S(\omega) d\omega$, Parseval's, and other fourier facts.
- $\bar{S}(\omega) = \frac{1}{k(0)} S(\omega)$ is sometimes called spectral probability density.
- This theorem opens the door for easier and more general constructions.

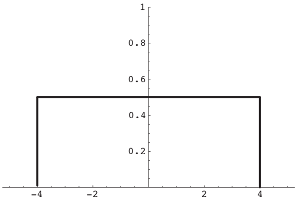
Why?

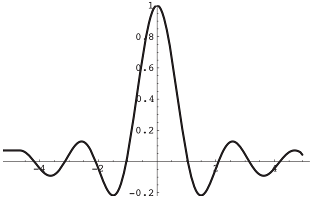
One of these things is not like the others

- Identify which of the following are kernels:









Bochner's theorem for isotropic kernels

- Warning: do not confuse the applicability of Bochner's theorem when working with seemingly one-dimensional isotropic kernels:

$$k(x, x') = k_I(||x - x'||).$$

- Here Bochner's theorem must be extended to this larger volume case.
- See details in [Gen02, §2].
- A surprising and quite cool bit of trivia:

$$\frac{k_I(||x - x'||)}{k_I(0)} \geq -1 \text{ when } x, x' \in \mathbb{R}$$

$$\frac{k_I(||x - x'||)}{k_I(0)} \geq -0.403 \text{ when } x, x' \in \mathbb{R}^2$$

$$\frac{k_I(||x - x'||)}{k_I(0)} \geq -0.218 \text{ when } x, x' \in \mathbb{R}^3$$

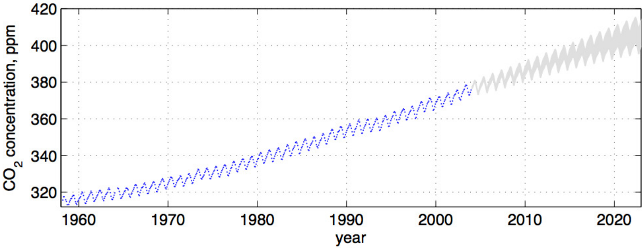
$$\frac{k_I(||x - x'||)}{k_I(0)} \geq 0 \text{ when } x, x' \in \mathbb{R}^\infty$$

- I don't know of any great applications of this fact, but it's still cool.

Outline

Learning kernels

- Recall the exercise of [RW06, §5.4.3]:



- Is there a kernel class generic enough to include this structure?

Starting with Bochner's theorem

- Approximate *any* stationary kernel with a mixture of gaussians *in the spectral domain*.
- Gaussian mixture components in pairs to preserve spectral symmetry:

$$S_a(\omega) = \frac{\alpha_a}{2}\left(\mathcal{N}(\omega; \mu_a, \sigma_a^2) + \mathcal{N}(\omega; -\mu_a, \sigma_a^2)\right).$$

Why is symmetry necessary? Why real?

- We can then take the (inverse) fourier transform to see:

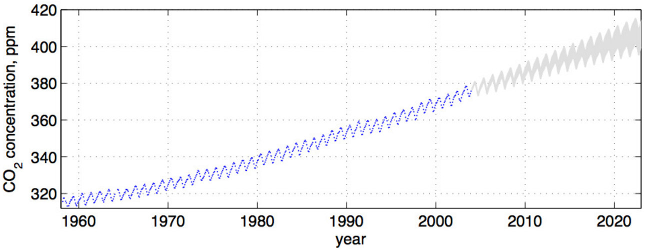
$$k_a(\tau) = \alpha_a \exp\left\{-2\pi^2\tau^2\sigma_a^2\right\}\cos\left(2\pi\tau\mu_a\right).$$

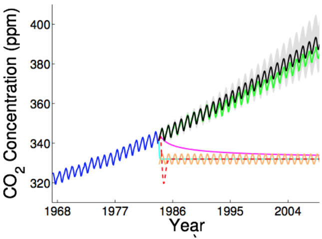
- Leading to the *spectral mixture kernel*:

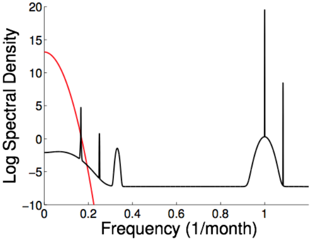
$$k(\tau) = \sum_{a=1}^{A}\alpha_a \exp\left\{-2\pi^2\tau^2\sigma_a^2\right\}\cos\left(2\pi\tau\mu_a\right).$$

- Caution: more hyperparameters to learn!

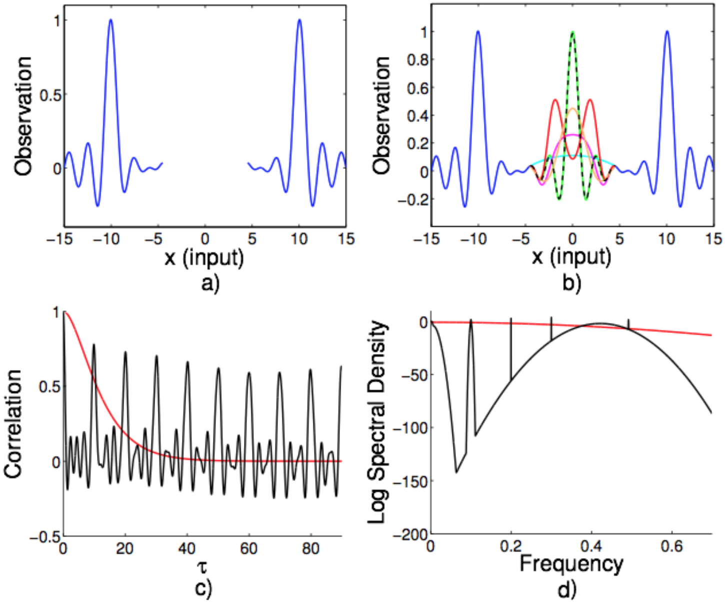
Spectral mixture kernels



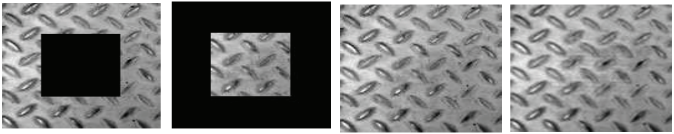




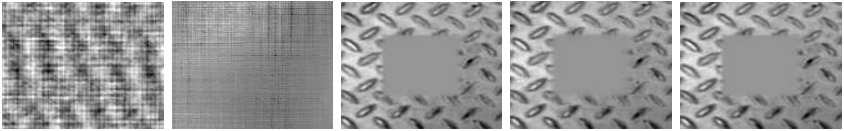
Spectral mixture kernels



Spectral mixture kernels in multiple dimensions [WGNC14]



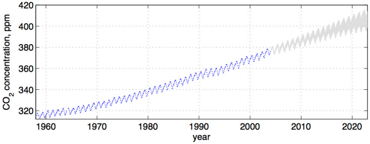
(a) Train (b) Test (c) Full (d) Spectral Mixture



(e) SSGP (f) FITC (g) GP-SE (h) GP-MA (i) GP-RQ

Outline

Kernels on more interesting spaces



- To now we have mostly considered kernels on $\mathbb{R}^D$ (very often $\mathbb{R}$).
- But we know $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for very general sets $\mathcal{X}$.
- We'll now discuss a few examples of kernels on more interesting spaces:
  - (strings)
  - (the Grassmann manifold $\mathcal{G}(\mathbb{R}^d, r)$ [HL08]; [HSJ$^+$14]; [HSH14]; [CG15])
  - graphs [VSKB10]
  - permutations and rankings [JV15]

Outline

Preliminaries

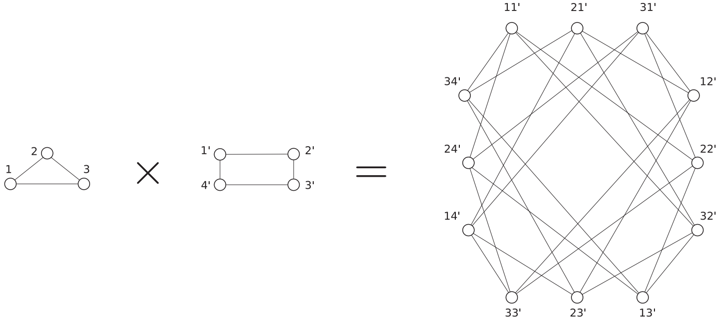
- Define a graph $G = (V, E, Z)$:
  - vertex set $V = \{v_1, ..., v_n\}$
  - edge set $E = \{(v_i, v_j)\}_{i,j} \subset V \times V$
  - edge labels $Z \in \mathcal{Z}^{n \times n}$.
  - feature map for labels $\phi : \mathcal{Z} \to \mathcal{H}$
- Note this generalizes some more conventional definitions:
  - undirected, unweighted, simple: $\mathcal{Z} = \{0, 1\}, Z_{ij} = Z_{ji}, Z_{ii} = 0, \phi(Z) = Z$.
  - undirected, weighted, simple: $\mathcal{Z} = \mathbb{R}_+, Z_{ij} = Z_{ji}, Z_{ii} = 0, \phi(Z) = Z$.
- We then have a feature "matrix" $\Phi(Z) = \{\phi(Z_{ij})\}_{i,j}$.
- Note this is general; e.g., zero edges require only $\phi(Z_{ij}) = 0$ [VSKB10, §2.2].
- Important: if nothing else, understand that we have just created a feature map $\Phi(Z)$ for a graph G...

Comparing graphs

- We have a graph G implicitly defined by a feature matrix $\Phi(Z)$.
- Recall the usual kernel setup $k_0(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$.
- To define $k(G, G')$, first define a Kronecker product of feature matrices:

$$W_\times = \Phi(Z) \otimes \Phi(Z') = \left\{ \langle \phi(Z_{ij}), \phi(Z'_{k\ell}) \rangle_{\mathcal{H}} \right\}_{i,j=1,\ldots,n \ ; \ k,\ell=1,\ldots,n'} = \left\{ k_0(Z_{ij}, Z'_{k\ell}) \right\}_{ijk\ell}$$

- This is a weighted product graph $G \times G'$ with edge weights $W_\times^{(ik),(j\ell)}$:



Random walks

- Take a breath:
  - Two graphs G, G' of different size and different label sets $Z, Z' \in \mathcal{Z}$.
  - The feature map ϕ defines a kernel $k_0(Z_{ij}, Z'_{k\ell})$ between any two edges.
  - Important: $W_\times^{(ik),(j\ell)}$ is the similarity between edge $(i,j) \in G$ and $(k,\ell) \in G'$.

- Random walk on a graph G:
  - Start at a vertex according to p (a p.m.f.).
  - Quit at a vertex according to q (a p.m.f.).
  - Randomly move $v_i \to v_j$, accruing W_{ij}.
  - If properly normalized, $(W^r)_{ij}$ is $\mathbb{P}(\ v_j$ at step r $\mid v_i$ at step 0 $)$.
  - Thus $q^\top W^r p$ is the total expected weight of a length r random walk on G.

Graph kernel [VSKB10, §3.2]

- Take simultaneous random walks on G and G' and see how similar the experience is (in the total expected weight sense).
- Fact: random walk on $G \times G' \leftrightarrow$ simultaneous random walks on G and G'.
- Define the length r simultaneous random walk kernel as:

$$k^r(G, G') = q_\times^\top W_\times^r p_\times \quad , \qquad \text{where } p_\times = p \otimes p', W_\times^{(ik),(j\ell)} = k_0(Z_{i,j}, Z'_{k,\ell})$$

- Valid kernel! ...positive weighted sum of $(nn')^2$ elements $k_0(\cdot, \cdot)$.
- Consider all length r walks:

$$k(G, G') = \sum_{r=0}^{\infty} k^r(G, G') = \sum_{r=0}^{\infty} \alpha(r) q_\times^\top W_\times^r p_\times .$$

$\alpha(r) \to k(G, G') < \infty$ .... if time, [VSKB10, Lemma 2, Theorem 3].

Making this kernel practical

$$k(G, G') = \sum_{r=0}^{\infty} \alpha(r) q_\times^\top W_\times^r p_\times.$$

- Numerous (not all!) existing graph kernels are special cases of this kernel (unifying).
- This kernel is, at face value, $\mathcal{O}\left((nn')^3\right)$ (at least) to compute.
- [VSKB10, §4-5] is then efficient computation and demonstration thereof.
- A heavily exploited trick is $\text{vec}(ABC) = (C^\top \otimes A)\text{vec}(B)$.

...which is used all over GP, in multiple dimensions particularly.

Outline

Rankings

- Object of interest is a ranking of a set of items $\{x_1, ..., x_n\}$:

$$x_{i_1} \succ x_{i_2} \succ ... \succ x_{i_n},$$

- namely a permutation $\sigma : \{1, ..., n\} \to \{1, ..., n\}$ (distinctly).

Think of $\sigma(i) = j$ as "item i is my jth favorite".

- Call $\mathbb{S}$ the set of all permutations.
- We then seek a kernel $k : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$.
- Such a kernel likens individuals based on their polling/voting preferences.

Ranking kernels

- Kendall's tau distance between σ and σ' is the number of discordant pairs:

$$n_d(\sigma,\sigma') = \sum_{j=1}^{n}\sum_{i=1}^{j-1}\left(\mathbb{1}(\sigma(i)<\sigma(j))\mathbb{1}(\sigma'(i)>\sigma'(j)) + \mathbb{1}(\sigma(i)>\sigma(j))\mathbb{1}(\sigma'(i)<\sigma'(j))\right)$$

- Similarly concordant pairs:

$$n_c(\sigma,\sigma') = \sum_{j=1}^{n}\sum_{i=1}^{j-1}\left(\mathbb{1}(\sigma(i)<\sigma(j))\mathbb{1}(\sigma'(i)<\sigma'(j)) + \mathbb{1}(\sigma(i)>\sigma(j))\mathbb{1}(\sigma'(i)>\sigma'(j))\right)$$

- Define the Mallows and Kendal kernels for any $\lambda \geq 0$ as:

$$\begin{aligned} k_M(\sigma,\sigma') &= \exp\{-\lambda n_d(\sigma,\sigma')\} \\ k_K(\sigma,\sigma') &= \frac{n_c(\sigma,\sigma') - n_d(\sigma,\sigma')}{\binom{n}{2}}. \end{aligned}$$

Working with these kernels

- [JV15, Thm. 1] prove these are kernels via the feature map $\phi : \mathbb{S} \to \mathcal{H}$:

$$\phi(\sigma) = \left(\frac{1}{\sqrt{\binom{n}{\ell}}} \left(\mathbb{1}(\sigma(i) > \sigma(j)) - \mathbb{1}(\sigma(i) < \sigma(j)) \right) \right)_{i=1,\ldots,j-1\ ;\ j=1,\ldots,n}$$

- Again, naively we must spend $\mathcal{O}(n^2)$ to compute a *single* $k(\sigma, \sigma')$.
- MergeSort can be used to reduce this cost to $\mathcal{O}(n \log n)$.
- [JV15] also consider partial rankings:

$$\sigma = x_{i_1} \succ x_{i_2} \succ \ldots \succ x_{i_\ell}, \qquad \text{where } |\sigma| = \ell < n$$

or

$$\sigma = x_{i_1} \succ x_{i_2} \succ \ldots \succ x_{i_\ell} \succ \text{the rest}, \qquad \text{where } |\sigma| = \ell < n$$

- which they show to be computable in $\mathcal{O}(\ell \log \ell + m \log m)$ time.

References

[CG15] John P Cunningham and Zoubin Ghahramani.
Linear dimensionality reduction: Survey, insights, and generalizations.
Journal of Machine Learning Research, 2015.

[Gen02] Marc G Genton.
Classes of kernels for machine learning: a statistics perspective.
The Journal of Machine Learning Research, 2:299–312, 2002.

[HL08] Jihun Hamm and Daniel D Lee.
Grassmann discriminant analysis: a unifying view on subspace-based learning.
In *Proceedings of the 25th international conference on Machine learning*, pages 376–383. ACM, 2008.

[HSH14] Mehrtash T Harandi, Mathieu Salzmann, and Richard Hartley.
From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices.
In *Computer Vision–ECCV 2014*, pages 17–32. Springer, 2014.

[HSJ+14] Mehrtash T Harandi, Mathieu Salzmann, Sadeep Jayasumana, Richard Hartley, and Hongdong Li.
Expanding the family of grassmannian kernels: An embedding perspective.
In *Computer Vision–ECCV 2014*, pages 408–423. Springer, 2014.

[JV15] Yunlong Jiao and Jean-Philippe Vert.
The kendall and mallows kernels for permutations.
In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1935–1944, 2015.

[RW06] C. E. Rasmussen and C.K.I. Williams.
Gaussian Processes for Machine Learning.
MIT Press, Cambridge, 2006.

[VSKB10] S Vichy N Vishwanathan, Nicol N Schraudolph, Risi Kondor, and Karsten M Borgwardt.
Graph kernels.
The Journal of Machine Learning Research, 11:1201–1242, 2010.

[WA13] Andrew Wilson and Ryan Adams.
Gaussian process kernels for pattern discovery and extrapolation.
In *Proceedings of The 30th International Conference on Machine Learning*, pages 1067–1075, 2013.

[WGNC14] Andrew Wilson, Elad Gilboa, Arye Nehorai, and John P Cunningham.
Fast kernel learning for multidimensional pattern extrapolation.
In *Advances in Neural Information Processing Systems*, pages 3626–3634, 2014.