Transductive Kernels for Gaussian Processes on Graphs

Oxford-Man Institute
Department of Engineering

Yin-Cong Zhi
Felix L. Opolka, Yin Cheng Ng, Pietro Liò, Xiaowen Dong

March 9, 2023
Semi-Supervised Classification with Graph Gaussian Processes

What is $k(x_i, x_j)$?
Kernels and Regularization

Many machine learning problems can be presented as a minimization task

$$\min_f L(f, y) + \Omega(\|f\|^2).$$
Kernels and Regularization

Many machine learning problems can be presented as a minimization task

$$\min_{f} L(f, y) + \Omega(\|f\|^2).$$

From a kernel methods point of view, we are interested in the regularizer $\Omega(\|f\|^2)$:

$$\Omega(\|f\|^2) = \langle Pf, Pf \rangle = \langle f, P. Pf \rangle = \langle f, r(\Delta)f \rangle = \langle f, f \rangle_\mathcal{H}.$$

The Hilbert space $\mathcal{H}$ w.r.t. inner product with $r(\Delta)$.
Kernels and Regularization

Many machine learning problems can be presented as a minimization task

$$\min_{f} L(f, y) + \Omega(\|f\|^2).$$

From a kernel methods point of view, we are interested in the regularizer $$\Omega(\|f\|^2):$$

$$\Omega(\|f\|^2) = \langle P f, P f \rangle = \langle f, P. P f \rangle = \langle f, r(\Delta) f \rangle = \langle f, f \rangle_\mathcal{H}.$$

The Hilbert space $$\mathcal{H}$$ w.r.t. inner product with $$r(\Delta).$$ Examples include $$P = 1, \nabla, (1, \nabla)^T, \ldots;$$ and $$\nabla.\nabla = \Delta$$

$$\nabla = \left( \frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right) \quad \Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}.$$
Reproducing Kernel Hilbert Space

A function $k(\cdot, \cdot)$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies

$$\forall x, f, \langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \quad (= \langle f, r(\Delta)k(x, \cdot) \rangle),$$

$\mathcal{H}$ is then a RKHS. This is called the reproducing property.

The kernel function is derived by computing $K = r^{-1}(\Delta)$ (more details later).
Reproducing Kernel Hilbert Space

A function $k(\cdot, \cdot)$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies

$$\forall x, f, \langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \quad (= \langle f, r(\Delta)k(x, \cdot) \rangle),$$

$\mathcal{H}$ is then a RKHS. This is called the reproducing property.

The kernel function is derived by computing $K = r^{-1}(\Delta)$ (more details later).

**Representer Theorem:**
If $\mathcal{H}$ is a reproducing kernel hilbert space (RKHS), then there is a solution to

$$\min_f L(f, y) + \Omega(\|f\|_{\mathcal{H}}^2)$$

that takes the form $f^*(\cdot) = \sum_i \alpha_i k(\cdot, x_i)$. 

Kernel Functions

Examples of regularization functions $r(\Delta)$, and their kernels (Smola. A. 2003):

- **RBF:**
  
  \[
  r(\Delta) = \sum_i \frac{\sigma^{2i}}{2^i i!} \Delta^i \implies k(x, x') = \exp \left\{ -\frac{1}{2\sigma^2} ||x - x'||^2 \right\}
  \]

- **Laplacian kernel:**
  
  \[
  r(\Delta) = 1 + \sigma^2 \Delta \implies k(x, x') = \exp \left\{ -\frac{1}{\sigma} ||x - x'|| \right\}
  \]
Graphs Laplacians

The graph Laplacian is defined as

$$L = D - A$$
Graphs Laplacians

The graph Laplacian is defined as

\[ L = D - A \]

This is the finite difference of \( \Delta \) on a discrete space

\[ \Delta \equiv L \]
Graphs Laplacians

The graph Laplacian is defined as

$$L = D - A$$

This is the finite difference of $\Delta$ on a discrete space

$$\Delta \equiv L$$

To define kernels on graphs, the regularization function $\Omega(||f||^2_{\mathcal{H}})$ becomes (Smola. A. 2003):

$$\langle f, r(\Delta)f \rangle \rightarrow \langle f, r(L)f \rangle$$
Kernels on Graphs

If $f$ is a function on a graph $G$, reproducing property is

$$\langle f, r(L)K(i,:) \rangle_i = f(i)$$
$$\implies f^\top r(L)K = f^\top$$
$$\implies K = r^{-1}(L)$$

Kernel between two nodes is

$$k(v_i, v_j) = r^{-1}(L)_{i,j}$$

Note, this kernel depends on the graph only, and not any node data.
Kernels on Attributed Graphs

What is $k(x_i, x_j)$?
Kernels on Attributed Graphs

What if we have a graph with feature data on the nodes (node attributes)?

Start with the minimization task again

\[
\min_f L(f, y) + \Omega(||f||^2)
\]

\[
\Omega(||f||^2) = \langle f, f \rangle_{\mathcal{H}}.
\]

What \( \mathcal{H} \) do we choose?
Kernels on Attributed Graphs

What if we have a graph with feature data on the nodes (node attributes)?

Start with the minimization task again

$$\min_{f} L(f, y) + \Omega(||f||^2)$$

$$\Omega(||f||^2) = \langle f, f \rangle_{\mathcal{H}}.$$  

What $\mathcal{H}$ do we choose?

- $\langle f, r(\Delta)f \rangle$ leads to $k(x_i, x_j)$ - depends on feature only
- $\langle f, r(L)f \rangle$ leads to $k(v_i, v_j)$ - depends on graph only
Kernels on Attributed Graphs

What if we have a graph with feature data on the nodes (node attributes)?

Start with the minimization task again

$$\min_{f} L(f, y) + \Omega(||f||^2)$$

$$\Omega(||f||^2) = \langle f, f \rangle_{\mathcal{H}}.$$

What $\mathcal{H}$ do we choose?

- $\langle f, r(\Delta)f \rangle$ leads to $k(x_i, x_j)$ - depends on feature only
- $\langle f, r(L)f \rangle$ leads to $k(v_i, v_j)$ - depends on graph only

Solution: $f$ depends on $x$ and $G$, so use $r(\Delta)$ and $r(L)$!
Kernels for Attributed Graphs

\[ \Omega(||f||^2) = \langle f, [r_1(\Delta) + r_2(L)]f \rangle = \langle f, r_1(\Delta)f \rangle + \langle f, r_2(L)f \rangle \]

- \( r_1(\Delta) \) deals with the feature data \((K_1 = r_1^{-1}(\Delta))\)
- \( r_2(L) \) deals with the graph

\[ K = [K_1^{-1} + r_2(L)]^{-1} \]
Kernels for Attributed Graphs

\[ \Omega(||f||^2) = \langle f, [r_1(\Delta) + r_2(L)]f \rangle = \langle f, r_1(\Delta)f \rangle + \langle f, r_2(L)f \rangle \]

- \( r_1(\Delta) \) deals with the feature data \( (K_1 = r_1^{-1}(\Delta)) \)
- \( r_2(L) \) deals with the graph

\[ K = [K_1^{-1} + r_2(L)]^{-1} \]

Now use Woodbury matrix identity:

\[ [K_1^{-1} + r_2(L)]^{-1} = K_1 - K_1[K_1 + r_2^{-1}(L)]^{-1}K_1 \]

or elementwise:

\[ k_g(x_1, x_2) = k_1(x_1, x_2) - k_1(x_1, X)^\top [K_1 + r_2^{-1}(L)]^{-1}k_1(X, x_2) \]
Kernels for Attributed Graphs

$$\Omega(||f||^2) = \langle f, [r_1(\Delta) + r_2(L)]f \rangle = \langle f, r_1(\Delta)f \rangle + \langle f, r_2(L)f \rangle$$

- $r_1(\Delta)$ deals with the feature data ($K_1 = r_1^{-1}(\Delta)$)
- $r_2(L)$ deals with the graph

$$K = [K_1^{-1} + r_2(L)]^{-1}$$

Now use Woodbury matrix identity:

$$[K_1^{-1} + r_2(L)]^{-1} = K_1 - K_1[K_1 + r_2^{-1}(L)]^{-1}K_1$$

or elementwise:

$$k_g(x_1, x_2) = k_1(x_1, x_2) - k_1(x_1, X)^\top[K_1 + r_2^{-1}(L)]^{-1}k_1(X, x_2)$$

Note this depends on all the data - the transductive property
Synthetic Experiments

Swiss Roll regression, 1000 points, 10 training, 990 testing:

training (red) and test points
Synthetic Experiments

Predictions:

(a) Ground truth
(b) GP with RBF
(c) Graph only
(d) TGGP (ours)
(e) GGP
(f) WGGP
## Semi-Supervised Classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Texas</th>
<th>Cornell</th>
<th>Wisconsin</th>
<th>Chameleon</th>
<th>Cora</th>
<th>Citeseer</th>
<th>Squirrel</th>
<th>Actor</th>
</tr>
</thead>
<tbody>
<tr>
<td># Nodes</td>
<td>183</td>
<td>183</td>
<td>251</td>
<td>2,277</td>
<td>2,708</td>
<td>3,327</td>
<td>5,201</td>
<td>7,600</td>
</tr>
<tr>
<td>Homo. Rat.</td>
<td>0.11</td>
<td>0.30</td>
<td>0.21</td>
<td>0.23</td>
<td>0.81</td>
<td>0.74</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>GCN</td>
<td>59.5 ±5.3</td>
<td>57.0 ±4.7</td>
<td>59.8 ±7.0</td>
<td>59.8 ±2.6</td>
<td>80.5 ±0.8</td>
<td>68.1 ±1.3</td>
<td>36.9 ±1.3</td>
<td>30.3 ±0.8</td>
</tr>
<tr>
<td>GAT</td>
<td>58.4 ±4.5</td>
<td>58.9 ±3.3</td>
<td>55.3 ±8.7</td>
<td>54.7 ±2.0</td>
<td>82.6 ±0.7</td>
<td>72.2 ±0.9</td>
<td>30.6 ±2.1</td>
<td>26.3 ±1.7</td>
</tr>
<tr>
<td>ChebNet</td>
<td>77.3 ±4.1</td>
<td><strong>74.3 ±7.5</strong></td>
<td>79.4 ±4.5</td>
<td>55.2 ±2.8</td>
<td>78.0 ±1.2</td>
<td>70.1 ±0.8</td>
<td>43.9 ±1.6</td>
<td>34.1 ±1.1</td>
</tr>
<tr>
<td>LP</td>
<td>37.8</td>
<td>21.6</td>
<td>23.5</td>
<td>44.5</td>
<td>71.3</td>
<td>49.9</td>
<td>32.7</td>
<td>22.4</td>
</tr>
<tr>
<td>GP</td>
<td>78.4</td>
<td>73.0</td>
<td>78.4</td>
<td>46.1</td>
<td>60.8</td>
<td>54.7</td>
<td>34.4</td>
<td>34.9</td>
</tr>
<tr>
<td>GGP</td>
<td>78.4</td>
<td>62.1</td>
<td>60.8</td>
<td>73.5</td>
<td>80.9</td>
<td>69.7</td>
<td>64.8</td>
<td>26.3</td>
</tr>
<tr>
<td>ChebGP</td>
<td><strong>81.1</strong></td>
<td>64.9</td>
<td><strong>82.4</strong></td>
<td>69.1</td>
<td>79.7</td>
<td>66.5</td>
<td>28.8</td>
<td>31.8</td>
</tr>
<tr>
<td>WGGP</td>
<td>78.4</td>
<td>67.6</td>
<td><strong>84.3</strong></td>
<td>64.5</td>
<td><strong>84.7</strong></td>
<td><strong>70.8</strong></td>
<td><strong>58.3</strong></td>
<td>32.6</td>
</tr>
<tr>
<td>TGGP (ours)</td>
<td><strong>81.1</strong></td>
<td><strong>75.7</strong></td>
<td><strong>82.4</strong></td>
<td>63.2</td>
<td>80.3</td>
<td>70.5</td>
<td>53.8</td>
<td><strong>34.9</strong></td>
</tr>
<tr>
<td>GGP-X</td>
<td>78.4</td>
<td>56.8</td>
<td>60.8</td>
<td>77.6</td>
<td>84.7</td>
<td>75.6</td>
<td><strong>71.9</strong></td>
<td>OOM</td>
</tr>
<tr>
<td>WGGP-X</td>
<td><strong>81.1</strong></td>
<td>75.7</td>
<td><strong>84.3</strong></td>
<td>65.6</td>
<td><strong>87.5</strong></td>
<td><strong>76.8</strong></td>
<td>61.3</td>
<td>OOM</td>
</tr>
<tr>
<td>TGGP-X (ours)</td>
<td><strong>86.5</strong></td>
<td><strong>81.1</strong></td>
<td><strong>86.3</strong></td>
<td>63.4</td>
<td>83.8</td>
<td>76.7</td>
<td>54.2</td>
<td><strong>36.9</strong></td>
</tr>
</tbody>
</table>
Supplementary
Solving Kernels

For finding $K = r^{-1}(\Delta)$:

- Firstly, note
  $\Delta e^{-i\omega x} = -\omega^2 e^{-i\omega x} \implies r(\Delta)e^{-i\omega x} = \hat{g}(\omega)e^{-i\omega x}$.

- By Plancherel theorem,
  $\langle h, h \rangle_{\mathcal{H}} = \int |h(x)^2 g(x)^2| dx = \int |\hat{h}(\omega)^2 \hat{g}(\omega)^2| d\omega$.

- The inverse is
  $\left( r^{-1}(\Delta)_{x^*,x} \right) = k(x^*, x) = \int \frac{1}{\hat{g}(\omega)} e^{i\omega(x^*-x)} d\omega$. 

Example

Gaussian regularizer:

\[ r(\Delta) = \sum_{i=0}^{\infty} \frac{1}{i!} \left( \frac{\sigma^2 \Delta}{2} \right)^i \equiv \exp \left\{ \frac{\sigma^2 \Delta}{2} \right\}. \]

This leads to \( \hat{g}(\omega) = \exp\{\sigma^2 \omega^2 / 2\} \).

Compute the inverse Fourier transform of the reciprocal:

\[
\int_{-\infty}^{\infty} \exp \left\{ - \frac{\sigma^2 \omega^2}{2} \right\} e^{i\omega (x-x')} \, d\omega \propto \exp \left\{ - \frac{1}{2\sigma^2} (x - x')^2 \right\}.
\]

Last term corresponds to the popular Gaussian RBF kernel.
## Generalization

How are kernels, graph kernels, and graph GPs defined so far?

<table>
<thead>
<tr>
<th>Model</th>
<th>Name</th>
<th>( r_1(\Delta) )</th>
<th>( r_2(L) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label Propagation</td>
<td>Label Propagation</td>
<td>0</td>
<td>( \frac{1}{1-\alpha}(I + \alpha L) )</td>
</tr>
<tr>
<td>Kernels on graphs</td>
<td>Regularized Laplacian</td>
<td>0</td>
<td>( (I + \sigma^2 L) )</td>
</tr>
<tr>
<td></td>
<td>Diffusion</td>
<td>0</td>
<td>( \exp{\frac{\sigma^2}{2} L} )</td>
</tr>
<tr>
<td></td>
<td>( p )-step random walk</td>
<td>0</td>
<td>( (\alpha I - L)^{-p} )</td>
</tr>
<tr>
<td></td>
<td>Cosine</td>
<td>0</td>
<td>( (\cos(L\pi/4))^{-1} )</td>
</tr>
<tr>
<td></td>
<td>GP kernels</td>
<td></td>
<td>( (\frac{2\nu}{\kappa^2} - L)^{\nu/2+d/4} )</td>
</tr>
<tr>
<td></td>
<td>Matérn kernel on graphs</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Laplacian kernel</td>
<td>( 1 +</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gaussian kernel</td>
<td>( e^{\frac{\sigma^2}{2}</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Matérn kernel on manifolds</td>
<td>( (\frac{2\nu}{\kappa^2} - \Delta)^{\nu/2+d/4} )</td>
<td>0</td>
</tr>
<tr>
<td>Graph GP</td>
<td>GGP</td>
<td>( (P^T)^{-1} r(\Delta) P^{-1} )</td>
<td>0</td>
</tr>
<tr>
<td>Wavelet Graph GP</td>
<td>WGGP</td>
<td>( (W^T)^{-1} r(\Delta) W^{-1} )</td>
<td>0</td>
</tr>
<tr>
<td>Transductive kernel (ours)</td>
<td>TGGP (ours)</td>
<td>( (\frac{2\nu}{\kappa^2} - \Delta)^{\nu/2+d/4} )</td>
<td>( U[\text{softplus}(\sum_i (\beta_i \Lambda^i))] U^T )</td>
</tr>
</tbody>
</table>

(In our experiments, we chose the Matérn kernel w.r.t. \( r_1(\Delta) \) and a softplus polynomial for the graph kernel \( r_2(L) \))