SENDNet (Supplementary material)

Anonymous submission

Spline Regression

Choice of Penalty

Following the traditional work on penalized B-Splines of (), we choose to impose an L2 penalty on the second order finite difference of the spline coefficients. We choose the Neumann boundary conditions, that is, we set the finite difference at the boundaries to $\alpha_0 - \alpha_1 = \alpha_K - \alpha_{K+1} = 0$. This is preferred over Dirichlet boundary conditions (setting $\alpha_0 = \alpha_{K+1} = 0$) because we argue that information near the boundaries is lost. The effect is worsened when we consider that we are computing many sets of coefficients (one for each channel) and later concatenating them. Then, the readout graph representation consists of a vector of length (# channels) $\times$ # splines) with about 2(# channels) entries that are almost redundant (forced to be near zero).

Lower order difference penalties (Tikhonov regularization and the first order finite difference) are too severe and may cause overly smooth regressions which do not capture the characteristics of a pectral energy distribution. An L1 penalty on the coefficients which enforces sparsity is attractive because the regression enjoys enhanced locality. However, it prohibits the offline computation of basis expansion matrices (the L1 optimization problem does not admit a closed form solution and must be solved by an iterative algorithm), which drastically increases the training time.

Estimates of the Smallest Singular Value

In practice, the normalized Laplacian eigenvalues of real-world graphs exhibit a high degree of repulsion, in the sense that a sample spectrum of size N would usually be favored to N i.i.d. points sampled uniformly from $[0, 2]$ were we to compare them by how well they cover the interval when balls of radius $r > 0$ (r arbitrary) are centered about the points.

It follows that the same holds when we compare a sample spectrum of size N with N copies sampled i.i.d. from P' (the marginal distribution of one eigenvalue when sampling from the true model) because the uniform distribution maximizes covering volume for any radius r. Thus, if the knots $\{t_i\}_{i=1}^{K+4}$ are the $(K + 4)$-quantiles of P', we may be justified in using i.i.d. samples from P' to underestimate statistics of a sample spectrum if that statistic favors repulsion between points. In particular, we may make the ansatz that, with high probability, we have $\nu_{\min}(B^T_{\sigma(\mathcal{L})}B_{\sigma(\mathcal{L})}) \geq \nu_{\min}(B^T_X B_X)$, where $X = \{x_1, ..., x_n\}$ with $x_i \sim P'$ i.i.d. In our setting we only consider finitely many possible sizes of graphs. Then, the true model can only yield finitely many possible Laplacian spectra, and so only finitely many possible eigenvalues. That is to say, P' is of the form $P' = (1/N_{\text{total}}) \sum_{i=1}^{N_{\text{total}}} \delta_{\lambda_i}$, where $N_{\text{total}} = (\text{total } \# \text{ eigenvalues })$ and δ_{λ_i} is the Dirac measure at λ_i.

With that said, we may apply classical Chernoff bounds to obtain that for suitable $\epsilon > 0$, with high probability,

$$\begin{aligned}\nu_{\min}(B^T_{\sigma(\mathcal{L})}B_{\sigma(\mathcal{L})}) &\geq \nu_{\min}(B^T_X B_X) \\ &> (1-\epsilon)N\nu_{\min}\left(\frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}} B^T_{\lambda_i}B_{\lambda_i}\right).\end{aligned}$$

Here, $B_{\lambda_i}(k) = B_k(\lambda_i), 1 \leq k \leq K$ is a row vector. To see this, we make the observation that $B^T_X B_X = \sum_{j=1}^{N} B^T_{x_i} B_{x_i}$ is a sum of random matrices with common expectation:

$$\mathbb{E}[B^T_{x_i}B_{x_i}] = \frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}} B^T_{\lambda_i}B_{\lambda_i}.$$

Moreover, $\nu_{\min}(B^T_{x_i}B_{x_i}) \geq 0$. The result follows from a direct application of Theorem 5.1.1. of ().

Why Locality of Spline Regression?

It is possible that two graphs differ in size but have similar spectra otherwise. For instance, the likeness between the spectrum of a graph and that of a subgraph is described by an interlacing theorem (). Then in cases like these the similarity is pronounced when restricting the SEDs to a subset of [0,2]. Hence, we would hope that the extra information in the SED of the larger graph does not influence too much the regression over the subset of interest.

To further address this problem beyond ensuring locality with a good conditioning number for $A = B^T_{\sigma(\mathcal{L})}B_{\sigma(\mathcal{L})} + \mu D_2^T D_2$, we can modulate the Fourier coefficients with learnable filter kernels defined over $[0, 2]$ for each channel before running spline regression. The filters can be conveniently parametrized by the very same B-spline we use for

regression, with backpropagation applied to the spline coefficients $(a_1, ..., a_K)$ defining the kernels like so:

$$g(x) = \sum_{i=1}^{K} a_k B_k(x).$$

This is done with the intention that the filter specific to one channel will be learnt to have support only over bands of $[0, 2]$ of interest for the comparison of spline coefficients in that channel.