Filter Design for Spectral Denoising

Sam Leone

November 30, 2022

1 Background

For our purposes, we will consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$. Without loss of generality, we will assume our graph G is finite and $\mathcal{V} = \{1, 2...n\}$. Then our weights w is a symmetric function $w : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ such that $w(a, b)$ is a measure of simmilarity between vertices a and b. We will also associate with G a number of algebraic objects. The first of which is the adjacency matrix $\mathbf{A}$ for which $\mathbf{A}(i, j) = w(i, j)$. From $\mathbf{A}$, we define a diagonal degree matrix $\mathbf{D}$ with $\mathbf{D}(a, b) = \sum_{a=1}^{n} w(a, b)$. The combinatorial Laplacian $\mathbf{L}$ is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, and the symmetric normalized Laplacian is defined as $\mathcal{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$. We will also define the random walk matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$.

1.1 Properties of the Laplacian

First, note that $\mathbf{L}$ is obviiously symmetric by the symmetry of $\mathbf{A}$. It also lends itself to a particular quadratic form. For any function $f : \mathcal{V} \to \mathbb{R}$, we have,

$$f^T \mathbf{L} f = \sum_{(a,b) \in E} (f(a) - f(b))^2$$

Since the above quantity is nonnegative for any f, it follows that $\mathbf{L}$ is positive semidefinite and is thus diagonalizable with a set of real eigenvalues $\lambda_1 \leq \lambda_2 ... \leq \lambda_n$ and corresponding mutually orhogonal eigenvectors $\psi_1 ... \psi_n$ ([3]). Let $\mathbf{U}$ be the matrix whose ith column is ψ_i and $\mathbf{\Lambda}$ be the diagonal matrix with $\mathbf{\Lambda}(i, i) = \lambda_i$. Since $\mathbf{U}$ is orthnormal, its transpose is its inverse, so $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.

These eigenvalues and eigenvectors contain valuable information about the graph G and are thus of great interest. The so-called graph frequencies are key to fields like graph signal processing, topological data analysis, and spectral graph theory. We can view $\mathbf{L}$ as a discretization as the usual Laplacian Δ which acts on functions in $\mathbb{R}^n$. Whereas,

$$\Delta f(x) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}$$

Measures local curvature of $f \in \mathcal{C}^2$. On the other hand, the combinatorial Laplacian $\mathbf{L}$ measures local differences:

$$\mathbf{L}f(i) = \sum_{j \sim i}(f(i) - f(j))$$

Thus, just as in traditional signal processing & analysis we are interested in the eigenfunctions of Δ, in graph signal processing & topological data analysis, we are interested in the eigenvectors of $\mathbf{L}$. Analogously, the eigenvectors corresponding to low eigenvalues are "low frequency" and those corresponding to high eigenvalues are "high frequency."

1.2 The Graph Fourier Transform & Graph Signal Processing

We define the *Graph Fourier Transform (GFT)* in much the same way as the usual Fourier Transform. Whereas in conventional signal processing, the fourier transform of a signal $\widehat{f}$ is defined via a convolution with eigenfunctions of Λ ([2]):

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(\omega)e^{-2\pi i\omega x}dx$$

The graph Fourier transform is defined as the convolution of f with the eigenvectors of $\mathbf{L}$:

$$\widehat{f} = \mathbf{U}^T f \qquad \widehat{f}(i) = \sum_{a=1}^{n} f(a)\psi_i(a)$$

From the Graph Fourier Transform, we also have a notion of Graph Signal Processing. For a function $h : \{\lambda_i\}_i \to \mathbb{R}$ on the spectrum of G, h induces a filter on the set of signals on G. This filter maps $f_{in} \mapsto f_{out}$ where:

$$f_{out} = h(\mathbf{L})f_{in} = \mathbf{U}h(\Lambda)\mathbf{U}^T f_{in}$$

Different h's are used for different tasks. For example, filters h which are decreasing in their argument comprise a class of low pass filters on the graph which map f to its low frequency components. This is often useful in denoising applications, where the true signal is assumed to vary smoothly over the graph, and thus removal of high frequency components corresponds to reconstruction of the ground truth signal. High pass filters are also useful for anomaly detection in networks & edge detection on image graphs.

1.3 Properties of the Random Walk Matrix

$\mathbf{P}$ can be viewed as the transition kernel defining a time inhomogeneous Markov Chain $\{X_t\}_t$ defined such that $\mathbf{P}(a, b) = \mathbb{P}(X_t = b|X_{t-1} = a)$. The left eigenvectors ν of $\mathbf{P}$ can be thought of eigen-initial-distributions. That is, ν is a left eigenvector of $\mathbf{P}$ of eigenvalue λ if $\mathbb{P}(X_t = a|X_0 \sim \nu) = \lambda\nu(a)$. It can be shown that the largest eigenvalue of $\mathbf{P}$ is 1, with multiplicity equal to the number of connected components of G. The corresponding left eigenvalue is denoted π, which is often referred to as the stationary distribution, since $\mathbb{P}(X_t = a|X_0 \sim \pi) = \pi(a)$. For an irreducible, aperiodic Markov Chain, the Strong Law of Large Numbers guarantees that for any initial distribution π_0, $\lim_{t\to\infty} \pi_0 \mathbf{P}^t = \pi$. Given our construction, it suffices to show that the graph G is fully connected and the lcm of its cycle lengths is 1. For fully connected graphs, this condition can be easily checked.

The right eigenvectors of $\mathbf{P}$ provide a set of eigenfunctions of the Markov Chain. That is, f is a right eigenvector of eigenvalue λ of $\mathbf{P}$ if $\mathbb{E}\left[f(X_t)|X_{t-1} = a\right] = \sum_b \mathbf{P}(a,b)f(b) = \lambda f(a)$ for all a. These eigenfunctions are of interest in data analysis. Intuitively, we may think of f as not only a function, but a coordinate map. If f is an informative set of coordinates, then it will remain roughly invariant upon diffusion. Therefore, the right eigenvectors of $\mathbf{P}$ can be used to obtain embeddings of the vertices themselves, which is the intuition underlying diffusion maps [1].

For computational purposes, we note the following relationship between the random walk matrix $\mathbf{P}$ and $\mathcal{L}$:

$$\mathcal{L} = \mathbf{I} - \mathbf{D}^{1/2}\mathbf{P}\mathbf{D}^{-1/2}$$
$$\mathbf{P} = \mathbf{I} - \mathbf{D}^{-1/2}\mathcal{L}\mathbf{D}^{1/2}$$

It can easily be checked ν is a right eigenvector of eigenvalue λ $\mathbf{P}$ if and only if $\mathbf{D}^{1/2}v$ is an eigenvector of $\mathcal{L}$ of eigenvalue $1-\lambda$. However, it is often the case that the eigenvectors of $\mathcal{L}$, being a symmetric matrix, are faster and more numerically stable to calculate. And so the eigenfunctions of $\mathbf{P}$ are calculated by first calculating the spectrum of $\mathcal{L}$.

1.4 Heat Diffusion for Denoising

Note that the operation $s(f) \triangleq \mathbf{P}^t f$ sets $f(a)$ equal to its expected value in a t-step random walk beginning from state a. That is,

$$s(f)(a) \leftarrow \sum_b \mathbf{P}(a,b)f(b) = \mathbb{E}\left[f(X_t)|X_0 = a\right]$$

For instance, if $t = 1$ and the underlying graph is simply a grid graph over which an image is defined, this will correspond to Gaussian blur. The idea is inherently this: if $\tilde{f}$ is an observed noisy signal of the form $f^\star + \epsilon$, where $f^\star$ is the true signal and ϵ is some mean-zero noise i.i.d. over the vertices of G, then the local averaging will "cancel" positive noise with negative noise, so that $P^t\tilde{f}$ more closely approximates the original $f^\star$ by "removing" ϵ. Of course, for larger and larger powers of t, we will also blur out real information from $\tilde{f}$, so there is a tradeoff.

In image processing, there is an established connection to filtering via heat diffusion and solving a Tikhonov & Sobolev regularized optimization.

2 Noise Models

To get started, we will consider four main models of noise for Graph Signals:

1. **Gaussian Noise**: In this case, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ is i.i.d. Gaussian noise of universal variance σ^2. Then the noisy signal is $\tilde{f} = f^\star + \epsilon$.

2. **Bernoulli-Downsampled Observations**: In this model, vertices are zeroed out with some probability $1 - p$ and preserved with probability p. That is, $\tilde{f}(i) = f^\star(i)$ with probability p, and $\tilde{f}(i) = 0$ with probability $(1 - p)$ (if $f^\star(i) \neq 0$). In other words, we can sample $\tilde{f}$ by first sampling $(z_1, z_2 ... z_n) \sim \text{Bern}(p, n)$ and letting $\tilde{f} = \vec{z} \odot f^\star$.

3. **Uniformly-Downsampled Observations**: This is similar to the previous noise model. In this, however, $\tilde{f}(i) \sim \text{Unif}(0, f^\star(i))$, so $\tilde{f}$ ranges uniformly between 0 and $f^\star$ for each vertex. We will represent this as $\tilde{f} = \vec{\alpha} \odot f^\star$, where $\alpha \sim^{i.i.d.} \text{Unif}([0, 1], n)$.

4. **Exponentially-Downsampled Observations**: Continuing with our pattern, we will consider a noise model in which equality with $f^\star$ is favored. To do this, we will let $\gamma \sim \text{Exp}(\lambda, n)$ and let $\beta(i) \triangleq 1/\gamma(i)$ so that $\beta \in (0, 1]$ with a mode of 1. Then just as before, $\tilde{f}$ is obtained by sampling β then letting $\tilde{f} = \beta \odot f^\star$.

For the purposes of Bayesian inference, we will often need to construct a set of priors for the signal $f^\star$. Intuitively, we will assume functions which are smoother on the graph are more likely than those which are higher frequency.

1. **Uniform by Bandwidth**: In this setting, we choose a cutoff frequency ω. Then,

$$p_\omega(f) \propto \mathbb{I}\{\widehat{f}(\omega') = 0, \forall \omega' \geq \omega\}$$

   In other words, all signals f which are bandlimited to ω are equally likely under $p_\omega(f)$. Note that this distribution is not actually well defined. Since $\mathbf{U}$ is simply a rotation:

$$\text{Vol}(\{f : p_\omega(f) > 0\}) = \text{Vol}\left(\{g : g(\omega') = 0, \forall \omega' \geq \omega\}\right)$$

   Where the right hand side is the volume of an infinite rectangular prism. Thus, the volume of this set is either 0 or ∞, so no such uniform measure may exist. However, we will nevertheless go on as if this is well defined.

2. **Energy Gaussian Random Field**: If we would like smooth level curves of our prior, we can use the Gaussian Random Field described by:

$$p_\eta(f) = \frac{\exp(-\eta \sum_{(a,b)\in\mathcal{E}} (f(a) - f(b))^2)}{N_\eta} = \frac{e^{-\eta f^T \mathbf{L} f}}{Z_\eta}$$

   Where N_η is the normalizing constant. And so the prior likelihood of our initial signal f is monotonically decreasing in the smoothness of the signal f over G. This is an established prior for semi-supervised learning models in which we would like to impute continuous labels via a low-Bandwidth continuation of the existing labels ([4]). Note that, as provided, p_η could not be a valid probability distribution, since the level of curves of p_η are infinite, high-dimensional cylinders. There are two naive ways to handle this:

   (a) Restrict ourselves to an $n-1$ dimensional subspace $\mathcal{M}$ such that the function $\exp(-\eta x^t \mathbf{L} x \mathbb{I}\{\mathcal{M}\})$ is integrable.

   (b) Consider a positive-definite matrix L_γ such that $\lim_{\gamma\to 0} \mathbf{L}_\gamma = \mathbf{L}$ and the density proportional to $\exp(-\eta x^T \mathbf{L}_\gamma x)$, perform calculations for arbitrary γ, and take the limit of the results as $\gamma \to 0$.

   But let us consider a slightly more rigorous approach. Note that we cannot integrate with respect to *n-dimensional Lebesgue measure*. If we rather consider a measure on sets modulo their projection to the null space of $\mathbf{L}$, we can make rigorous the density of interest.

3. **Markov Gaussian Random Field**: We consider a modification of the Energy Gaussian Random field in which we measure the difference between the signal and its diffusion. That is, we consider

$$p_\eta(f) = \frac{\exp(-\eta \|(\mathbf{I} - \mathbf{P})f\|^2)}{Z_\eta}$$

   Where Z_η is again a normalizing factor. Where,

$$(\mathbf{I} - \mathbf{P})f(a) = f(a) - \sum_{b\sim a} p_t(b|a) f(b)$$

   Note that basic rearrangment yields:

$$\mathbf{I} - \mathbf{P} = \mathbf{I} - (\mathbf{I} - \mathbf{D}^{-1/2} \mathcal{L} \mathbf{D}^{1/2}) = \mathbf{D}^{-1}\mathbf{L}$$

2.1 Axiomatic Construction of Energy Gaussian Random Field

From a signal distribution standpoint, we would like to capture rigorously the notion of our signal of interest gravitating towards its neighbors. A natural model for a signal f is one in which $f(i)$ is most likely to be the average of its neighbors, but allowing for some variation. We might lead with the following assumptions:

1. The "influence" of a particular neighbor of a on $f(a)$ is proportional to its affinity to a
2. The mode of the distribution of $f(a)$ is the average of its neighbors, weighted by affinity
3. If a has more neighbors, the variance of $f(a)$ is lower. If a has fewer neighbors, $f(a)$ has higher variance.

To handle this rigorously, we can regard the signal f as observations from a Markov Random Field on $\mathcal{G}$. In particular, let us assume the MRF is dictated so that our signal f is taken so that $f(a) = X_a$, where $X \triangleq \{X_a\}_{a \in \mathcal{V}}$ is a Markov Random Field on $\mathcal{G}$. We will let X_a b normally distributed about the mean of its neighbors. Now, for a particular vertex a, let $\mathcal{N}(a)$ be the neighbors of a and, for some $\sigma \in \mathbb{R}$

$$\mu_{\mathcal{N}(a)} = \sum_{b \in \mathcal{N}(a)} p(b|a) X_b \qquad \sigma_a = \frac{\sigma}{\sqrt{\deg(a)}}$$

And assume that the conditional distribution of X is given by the map $\mathcal{N}(a) \mapsto \mathbb{P}_{\mathcal{N}(a)}$ where $\mathbb{P}_{\mathcal{N}(X_a)}$ is the $\mathcal{N}(\mu_{\mathcal{N}(a)}, \sigma_a)$. The intuition here is that X_a normally distributed, with a mean equal to a weighted average of the value of its neighbors. By choosing a variance monotonically decreasing in degree, we are modeling that a's neighbors enforce greater determinism in the distribution of X_a.

2.1.1 M.R.F. as a model of Communication

Suppose each neighbor $b \in \mathcal{N}(a)$ of a, sends a message to a. b tries to send its own value X_b, however this noise is corrupted by Gaussian noise. Furthermore, we assume this noise has standard deviation inversely proportional to the affinity; the idea being that the lower the affinity, the "longer" the message travels through space, and thus the greater the level of corruption of the message. so b sends a the message $m_{a,b} = X_b + \frac{\sigma}{w(a,b)} Z_{a,b}$, where $Z_{a,b}$ is some standard normal white noise. Then we set X_a equal to the weighted average of the messages it receives, so $X_a = \sum_{b \in \mathcal{N}(a)} p(b|a) m_{a,b}$. Obviously, X_a is a linear combination of randomly distributed random variables and is thus itself normal. It remains to provide its distribution by its mean & variance:

$$\mathbb{E}[X_a] = \mathbb{E}\left[\sum_{b\in\mathcal{N}(a)} p(b|a)m_{a,b}\right] = \sum_{b\in\mathcal{N}(a)} p(b|a)X_b = \mu_{\mathcal{N}(a)}$$

We now apply $p(b|a) = w(a,b)/\deg(a)$ to calculate the variance:

$$\begin{aligned}
\mathrm{Var}[X_a] = \mathrm{Var}\left[\sum_{b\in\mathcal{N}(a)} p(b|a)m_{a,b}\right] &= \sum_{b\in\mathcal{N}(a)} p(b|a)^2\mathrm{Var}[m_{a,b}] \\
= \sum_{b\in\mathcal{N}(a)} p(b|a)^2\frac{\sigma^2}{w(a,b)^2} &= \sum_{b\mathcal{N}(a)} \frac{w(a,b)^2}{d(a,b)^2}\cdot\frac{\sigma^2}{w(a,b)^2} \\
= \sum_{b\in\mathcal{N}(a)} \frac{\sigma^2}{\deg(a)^2} = \deg(a)\cdot\frac{\sigma^2}{\deg(a)^2} &= \frac{\sigma^2}{\deg(a)}
\end{aligned}$$

And so the standard deviation of X_a is $\frac{\sigma}{\sqrt{\deg(a)}}$. And thus the conditional distribution of X_a is consistent with that which we have modeled.

2.1.2 MRF Leads to Gaussian Random Field

Recall that a Gibbs distribution on X is given by a density $Q(x)$ which factorizes as a product of functions defined on the cliques of the graph $\mathcal{G}$:

$$Q(x) = \prod_{K \text{ a clique}} V_K(x)$$

Where V_K are functions of the coordinates of x indexed by vertices in the clique K. The Hammersley-Clifford Theorem guarantees that if $Q(x) > 0$ everywhere, then Q is also a Markov Random Field on X.

3 GSP on Gaussian Noise

To begin, we will establish a useful lemma which states that an i.i.d. normal distributed random variable has the same distribution as its Fourier Transform.

Lemma 1. *If $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, then $\widehat{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ as well.*

Proof. Recall that $\widehat{\epsilon} \triangleq \mathbf{U}^T\epsilon$. So then, $\widehat{\epsilon}(i) = \langle\psi_i, \epsilon\rangle$, and so for any point $x \in \mathbb{R}^n$, x has probability:

$$p(x) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left(-\frac{x_i^2}{2\sigma^2}\right)$$

$$p(x) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2}\right)$$

$$= \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\|x^2\|\right)$$

Thus, the probability of x only depends on its magnitude. But since $\mathbf{U}$ is orthonormal, $\|\mathbf{U}^T x\| = \|x\|$. Therefore, the likelihood of a normally distributed random variable is the same as the likelihood of its Fourier transform. And so, $\widehat{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$.

□

3.1 Filter Design for GRF Priors

Now, we will assume the noise model in which,

$$p_\eta(f) \propto \exp(-\eta f^T \mathbf{L} f)$$

And we have i.i.d. Gaussian noise in the spatial / vertex domain:

$$\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

We would now like to find a class of filters s which maximize signal likelihood, given this prior distribution. Suppose again that our true model is $f^\star$ and our observed signal is $\tilde{f} + \epsilon$ per the prescribed model. What is the maximum likelihood estimate of $\tilde{f}$, for fixed σ^2?

Theorem 1. *Given that $\tilde{f} = f^\star + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $f^\star$ has prior density $p(f^\star) \propto \exp(-\eta {f^\star}^T \mathbf{L} f^\star)$, the maximum likelihood estimate of $f^\star$ is given by:*

$$\check{f} = (\eta\sigma^2\mathbf{L} + \mathbf{I})^{-1}\tilde{f}$$

Proof. Let us write for some estimate $\check{f}$ of $f^\star$ that:

$$p(f^\star = \check{f}|\tilde{f}) = \frac{p(\tilde{f}|\check{f})p_\eta(\check{f})}{p(\tilde{f})} = \frac{p(\epsilon = \check{f} - \tilde{f})\exp(-\eta\check{f}^T\mathbf{L}\check{f})}{p(\tilde{f})N_\eta}$$

$$= \frac{1}{(2\sigma^2\pi)^{n/2} p(\tilde{f}) N_\eta} \cdot \exp(-\eta\check{f}^T\mathbf{L}\check{f})\exp(-(\check{f} - \tilde{f})^2(\sigma^2\mathbf{I})^{-1}(\check{f} - \tilde{f}))$$

$$\propto \exp(-\eta\check{f}^T\mathbf{L}\check{f})\exp(-(\check{f} - \tilde{f})^2(\sigma^2\mathbf{I})^{-1}(\check{f} - \tilde{f}))$$

$$= \exp\left(-\eta\check{f}^T(\mathbf{L} + \frac{1}{\eta\sigma^2}\mathbf{I})\check{f} + \frac{2}{\sigma^2}\langle\check{f}, \tilde{f}\rangle - \frac{1}{\sigma^2}\|\tilde{f}\|^2\right)$$

$$\propto \exp\left(-\eta\check{f}^T(\mathbf{L} + \frac{1}{\eta\sigma^2}\mathbf{I})\check{f} + \frac{2}{\sigma^2}\langle\check{f}, \tilde{f}\rangle\right)$$

So maximizing likelihood is equivalent to minimizing the function F given by:

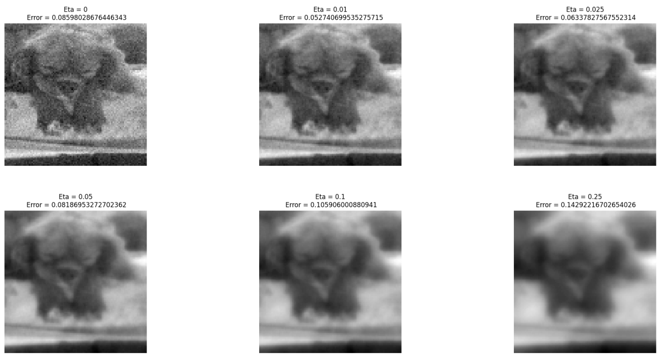
Eta = 0
Error = 0.08598028676446343
Eta = 0.01
Error = 0.05274069953527515
Eta = 0.025
Error = 0.06337827567552314
Eta = 0.05
Error = 0.08186953272702362
Eta = 0.1
Error = 0.105906000880941
Eta = 0.25
Error = 0.14292216702654026

Figure 1: Denoising using a noise standard deviation of 15 and increasing values of $\eta = 0, 0.01, 0.025, 0.05, 0.1, 0.2$. Thus, when $\eta = 0$, we are looking at the original noisy signal.

$$F(\tilde{f}) \triangleq \eta \check{f}^T(\mathbf{L} + \frac{1}{\eta\sigma^2}\mathbf{I})\check{f} - \frac{2}{\sigma^2}\langle \check{f}, \tilde{f}\rangle$$

To optimize, we do the routine setting of the gradient to zero. First, calculating the derivative, we obtain:

$$\frac{\partial F}{\partial \check{f}} = 2\eta(\mathbf{L} + \frac{1}{\eta\sigma^2}\mathbf{I})\check{f} - \frac{2}{\sigma^2}\tilde{f}$$

And setting the gradient equal to zero & doing the necessary rearrangements, we obtain the desired formula for $\check{f}$

$$\check{f} = (\eta\sigma^2\mathbf{L} + \mathbf{I})^{-1}\tilde{f}$$

Note $\mathbf{L}$ is positive semi-definite, so $\eta\sigma^2\mathbf{L} + \mathbf{I}$ is positive-definite and thus invertible.

□

Note that this only depends on $\eta\sigma^2$. So if we let $c = \eta\sigma^2$ be this parameter, we find $\check{f} = (c\mathbf{L}+\mathbf{I})^{-1}\tilde{f}$. Note that the higher the value of σ^2, $1/c \to 0$, meaning our estimate prefers very smooth functions. On the other hand, if $\sigma = 0$, then our optimization will let $\check{f} = \tilde{f}$, which is the sane thing. One final way of expressing this is via a graph filter. Letting $h_{\eta,\sigma^2}(\lambda) = \frac{1}{\eta\sigma^2\lambda+1}$, we find:

$$\check{f} = \mathbf{U}h_{\eta,\sigma^2}(\mathbf{\Lambda})\mathbf{U}^T\tilde{f}$$

The Class of Bandlimited Filters

Suppose instead that $f^\star$ is a known signal of maximum bandwidth ω. Without loss of generality, say $\omega = \lambda_j$ for some $j \in \{1, 2..n\}$. So we assume that for all $k \geq j$ that $\widehat{f^\star}(k) = 0$. If we suppose our first noise model holds true, in which $\tilde{f}$, the observed signal is equal to $f^\star + \epsilon$ for i.i.d. Gaussian noise, what is the best estimate of $\tilde{f}$?

3.2 An Unbiased Estimate

The naive guess is to simply bandlimit $\tilde{f}$ to the first j Fourier components. In other words, we let $h(\lambda_i) = \mathbb{I}\{i \leq j\}$ and set:

$$\check{f} = h(\mathbf{L})\tilde{f} = \sum_{i=1}^{n} \psi_i \mathbb{I}\{i \leq j\}\widehat{\tilde{f}}(i) = \sum_{i=1}^{j} \psi_i \widehat{\tilde{f}}(i)$$

Which we can separate into two terms:

$$= \sum_{i=1}^{j} \psi_i \widehat{f^\star}(i) + \sum_{i=1}^{j} \psi_i \widehat{\epsilon}(i)$$

So by assumption, $f^\star$ can be recovered exactly from its first j Fourier components. Meaning we may write $\check{f} = f^\star + \sum_{i=1}^{j} \psi_i \widehat{\epsilon}(i)$. Observe that each $\widehat{\epsilon}(i)$ is $\mathcal{N}(0, \sigma^2)$. Therefore, we can define the error of the estimate to be equal to $\delta \triangleq \sum_{i=1}^{j} \psi_i \hat{\epsilon}(i)$. Which can be represented as the matrix multiplication of $\mathbf{U}_j$ by $\hat{\epsilon}$, where $\mathbf{U}_j$ is the matrix whose columns are the first j eignvectors of $\mathbf{L}$ and last $n - j$ columns are all 0. Thus, we have,

$$\check{f} = f^\star + \delta = f^\star + \mathbf{U}_j \hat{\epsilon}$$

Which is obviously normally distributed with mean $f^\star$ and covariance matrix,

$$\Sigma = \mathbb{E}[\mathbf{U}_j \widehat{\epsilon}\widehat{\epsilon}^T \mathbf{U}_j^T] = \mathbf{U}_j \mathrm{cov}(\hat{\epsilon}) \mathbf{U}_j^T = \mathbf{U}_j \sigma^2 \mathbf{I} \mathbf{U}_j^T = \sigma^2 \mathbf{U}_j \mathbf{U}_j^T$$

Intuitively, for all ψ_i with $i \leq j$, Σ has variance σ^2 in direction ψ_i, and has variance 0 in the direction spanned by $\psi_k, \psi_{k+1} ... \psi_n$. So we conclude that, $\check{f} \sim \mathcal{N}(f^\star, \sigma^2 \mathbf{U}_j \mathbf{U}_j^T)$.

Lemma 2. $\mathbb{E}\|\check{f} - f^\star\|^2 = j\sigma^2$

Proof. By the Bias Variance decomposition,

$$\mathbb{E}\|\check{f} - f^\star\|^2 = \|f^\star - \mathbb{E}\check{f}\|^2 + \mathrm{Var}(\check{f})$$

And the cyclic property of trace:

$$= 0 + \mathrm{tr}(\Sigma) = \mathrm{tr}(\sigma^2 \mathbf{U}_j \mathbf{U}_j^T) = \sigma^2 \mathrm{tr}(\mathbf{U}_j^T \mathbf{U}_j) = \mathrm{tr}(\mathbf{I}_j) = j\sigma^2$$

□

Note that $\mathbb{E}\|\tilde{f} - f^\star\|^2 = \mathbb{E}\|\epsilon\|^2 = \sigma^2 n$, so we have reduced our error rate by a factor of $\frac{j}{n}$.

3.3 A Biased Estimate

Note that in the above calculation, we picked up a $j\sigma^2$, where j was the assumed index of the bandlimiting frequency λ_j. We may be lead to think that by reducing the number of Fourier Components onto which we project, we are reducing the variance of our estimate. This is, of course, true. But we assumed that $f^\star$ is recovered by its first k Fourier Components! So such an estimate $\hat{f}$ will be biased.

Of course, we could develop a theory in which we may filter out any arbitrary subset of the spectrum. Let us broaden our framework. Suppose that $\hat{f}^\star$ has support on $\mathcal{S}$ in the spectral domain. We would like to construct an ideal graph filter h (such that h is binary on the spectrum) which minimizes the mean squared error. Let us suppose we have an index set $\mathcal{J}$ over which h includes the Fourier component. that is, for all $j \in \mathcal{J}$, $h(\lambda_j) = 1$ and for all $j \in \mathcal{J}^c$, $h(\lambda_j) = 0$. Without loss of generality, let us assume that $\mathcal{J} \subseteq \mathcal{S}$; certainly, we are not interested in including noise in the spectral components we assume not to be supported. We find that:

$$\check{f}_{\mathcal{J}} = \sum_{j\in\mathcal{J}} \psi_i \widehat{\tilde{f}}(i) = \sum_{j\in\mathcal{J}} \psi_i \widehat{f^\star}(j) + \sum_{j\in\mathcal{J}} f^\star(j)\psi_i \widehat{\epsilon}(j)$$

$$= f^\star - \sum_{j\in\mathcal{S}-\mathcal{J}} f^\star(j) + \sum_{j\in\mathcal{J}} \psi_i \widehat{\epsilon}(j)$$

For brevity, let $f^\star_{\mathcal{J}} = f^\star - \sum_{j\in\mathcal{S}-\mathcal{J}} f^\star(j)$ and $\mathbf{U}_{\mathcal{J}}$ be the matrix whose jth column is ψ_j if $j \in \mathcal{J}$ and a vector of 0's otherwise. Likewise, let $\Sigma_{\mathcal{J}} = \sigma^2 U_{\mathcal{J}} U_{\mathcal{J}}^T$ It can be shown just as before that, $\check{f}_{\mathcal{J}} \sim \mathcal{N}(f^\star_{\mathcal{J}}, \Sigma_{\mathcal{J}})$. And the mean squared error will be equal to:

$$\mathbb{E}\|\check{f}_{\mathcal{J}} - f^\star\|^2 = \|f^\star - f^\star_{\mathcal{J}}\|^2 + \sigma^2|\mathcal{J}| = \sum_{s\in\mathcal{S}-\mathcal{J}} \widehat{f^\star}(s)^2 + \sigma^2|\mathcal{J}|$$

So obviously, if we had access to the original $f^\star$, we could trivially let $\mathcal{J} = \{j : |\widehat{f^\star}(s)| > \sigma|\}$ and solve the optimization: the resulting $\check{f}_{\mathcal{J}}$ will be the binary filter which minimizes mean squared error. Since for these fourier components, including an element has $\sum_{s\in\mathcal{S}-\mathcal{J}} \widehat{f^\star}(s)^2$ decreasing by something larger than σ^2, while $\sigma^2|\mathcal{J}|$ increases by σ^2. Note it would also suffice to know only whether the expected magnitude of $|f^\star(s)|$ exceeds σ. What are the consequences of this? For sufficiently high variance noise, it will become practical to include nothing!! That is, if you assume $f^\star << \sigma^2$, you're best bet is simply letting $f^\star = 0$. In practice, however, these assumptions are not likely to be met, and we have to develop more complicated machinery.

Ideal Binary Filter Under GRF Prior

Now, we shall suppose the same prior distribution of our ground truth signals $f^\star$ according to our Gaussian Random Field, but again restrict ourselves to filters which cap at a certain frequency. In this scenario, we fix η and σ, and would like to determine which frequencies are worth containing. In other words, we would like to select,

$$\arg\min_{S\subseteq[n]} \mathbb{E}[|f^\star - h_S(\tilde{f})|^2]$$

Where $h_S(\tilde{f}) = \sum_{i=1}^n \mathbb{I}\{i \in S\}\langle\psi_i, \tilde{f}\rangle\psi_i$ is the filtered version of $\tilde{f}$ corresponding to the frequencies within S. Note that, for fixed S, since h_S is a linear operator, we may write:

$$\mathbb{E}[|f^\star - h_S(\tilde{f})|^2] = \mathbb{E}[|f^\star - h_S(f^\star + \epsilon)|^2] = \mathbb{E}[|f^\star - h_S(f^\star) - h_S(\epsilon)|^2]$$

$$= \mathbb{E}\|\sum_{i=1}^n \psi_i\langle b\psi_i\psi_i\rangle - \sum_{i=1}^n \mathbb{I}\{s\in S\}\psi_i\langle\psi_i\psi_i\rangle - \sum_{i=1}^n \mathbb{I}\{s\in S\}\psi_i\hat{\epsilon}(i)\|^2$$

$$= \mathbb{E}\|\sum_{i=1}^n \psi_i\left(\langle\psi_i, f^\star\rangle - \mathbb{I}\{s\in S\}(\langle\psi_i, f^\star\rangle + \hat{\epsilon}(i))\right)\|^2$$

By Parseval's Identity, since $\{\psi_i\}_i$ is an orthonormal basis,

$$= \sum_{i=1}^n \mathbb{E}\left(\langle\psi_i, f^\star\rangle - \mathbb{I}\{s\in S\}(\langle\psi_i, f^\star\rangle + \hat{\epsilon}(i))\right)^2$$

$$= \sum_{i\in S} \mathbb{E}\left(\langle\psi_i, f^\star\rangle - (1)(\langle\psi_i, f^\star\rangle + \hat{\epsilon}(i))\right)^2$$

$$+ \sum_{i\in S^c} \mathbb{E}\left(\langle\psi_i, f^\star\rangle - (0)(\langle\psi_i, f^\star\rangle + \hat{\epsilon}(i))\right)^2$$

Since $\hat{\epsilon}$ is i.i.d. normal with variance σ^2 in each coordinate, we may write:

$$= \sum_{i\in S}\mathbb{E}\left(\hat{\epsilon}(i)^2\right) + \sum_{i\in S^c}\mathbb{E}\langle\psi_i, f^\star\rangle^2 = |S|\sigma^2 + \sum_{i\in S^c}\mathbb{E}\langle\psi_i, f^\star\rangle^2$$

Now, what is the expected coordinate in each basis vector? Note that our distribution is actually poorly defined, since $\mathbf{L}$ is not invertible. That is, since the level curves of $\exp(-\eta x^T\mathbf{L}x)$ have infinite $n-1$-dimensional volume, we cannot define a probability distribution proportional to $\exp(-\eta x^T\mathbf{L}x)$. Let us instead tweak $\mathbf{L}$ so that it is invertible and look at the limiting behavior as we approach $\mathbf{L}$. Let $\zeta = \{i : \lambda_i = 0\}$. So momentarily, consider the matrix $\mathbf{L}_\gamma = \sum_{i=1}^n(\lambda_i + \gamma\mathbb{I}\{i \in \zeta\})\psi_i\psi_i^T$. Note $\mathbf{L}_\gamma$ has the spectral decomposition

$\Psi\Lambda_\gamma\Psi^T$, where the eigenvectors of $\mathbf{L}_\gamma$ are the same as $\mathbf{L}$, but the eigenvalues otherwise corresponding to $\lambda = 0$ are now set to γ. Suppose the density of $f^\star$ is proportional to $\exp(-\eta f^\star \mathbf{L}_\gamma f^\star)$. Note that any vector $f^\star$ can be written as $\sum_{i=1}^n \langle \psi_i, f^\star\rangle\psi_i$. Note that $p_{\eta,\gamma}$ is equivalent to the $\mathcal{N}(0, \mathbf{L}_\gamma^{-1})$ distribution. Notice that rather than drawing $f^\star \sim \mathcal{N}(0, \mathbf{L}_\gamma^{-1})$, we could draw $x \sim \mathcal{N}(0, I)$ and then let $f^\star = \Psi\Lambda_\gamma^{-1/2}x$. And clearly, $f^\star$ has mean 0 and covariance $\Psi(\Lambda_\gamma^{-1/2})^T\mathrm{cov}(x)\Lambda_\gamma^{-1/2}\Psi^T = \mathbf{L}_\gamma^{-1}$. But then more clearly, since Ψ has orthonormal columns:

$$\langle \psi_i, f^\star\rangle = \langle \psi_i, \Psi\Lambda_\gamma^{-1/2}x\rangle = \psi_i^T\Psi\Lambda_\gamma^{-1/2}x = (\lambda_i')^{-1/2}x(i)$$

And so $\mathbb{E}[\langle \psi_i, f^\star\rangle^2] = \mathbb{E}[(\lambda_i')^{-1}x(i)^2] = (\lambda_i')^{-1}\mathrm{Var}(x(i))$. Since the covariance matrix of x is I, this value is therefore $1/(\lambda_i')$ Where,

$$\lambda_i' = \begin{cases} \eta\lambda_i & i > 1 \\ \gamma & i = 1 \end{cases}$$

And thus,

$$\mathbb{E}[|f^\star - h_S(\tilde{f})|^2] = \sigma^2|S| + \frac{1}{\gamma}|S^c \cap \zeta| + \sum_{s \in S^c/\zeta} \frac{1}{\eta\lambda_i}$$

Taking $\gamma \to 0$, we obtain:

$$\mathbb{E}[|f^\star - h_S(\tilde{f})|^2] = \begin{cases} \infty & S^c \cap \zeta = \emptyset \\ \sigma^2|S| + \sum_{s\in S^c} \frac{1}{\eta\lambda_i} & \text{otherwise} \end{cases}$$

So now to minimize the above value, it follows that we must first set $S \supseteq \zeta$. Then, for each i, we include $i \in S$ if and only if $\sigma^2 \leq \frac{1}{\eta\lambda_i}$, so we incur a lesser penalty by including $i \in S$, (σ^2), than if we had not, $\left(\frac{1}{\eta\lambda_i}\right)$.

$$S^\star = \arg\min_{S\subseteq[n]} \mathbb{E}[|f^\star - h_S(\tilde{f})|^2] = \zeta \cup \{i : \lambda_i \leq \frac{1}{\eta\sigma^2}\}$$

In other wards, as our distribution stretches the level curves in the direction of the null space of $\mathbf{L}$, the ideal decision rule approaches the above. Note that we could alternatively suppose that $\hat{\epsilon}(i) = 0$ for $i \in \zeta$ (so we have no noise in the components of eigenvalue 0). Then the optimal $S^\star$ remains the same, but the resulting penalty differs by an additive factor of $\sigma^2|\zeta|$.

4 Bernoulli Downsampled Observations

Now, we assume that $\tilde{f} = f^\star \odot z$, where $z \sim \mathrm{Bern}(p, n)$. Consider the subspace $\mathcal{M}$ of functions equal to $\tilde{f}$ over its nonzero observations. Then, there exists a density proportional over $\mathcal{M}$ proportional to $\exp(-\eta x\mathbf{L}x)$ for $x \in \mathcal{M}$. Then, we may write:

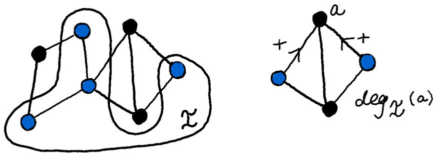

Figure 2: Visualization of $\mathcal{I}$ - the vertices over which $\tilde{f}$ is nonzero is shown in blue. Note that $\mathcal{I}^c$ has three vertices. The subgraph induced by them, H, has two connected components. On the right, a depiction of $\deg_{\mathcal{I}}(a)$ for a particular vertex $a \in H$.

$$p(\check{f}|\tilde{f}) \propto p(\check{f})p(\tilde{f}|\check{f}) \propto \exp(-\eta \check{f}\mathbf{L}\check{f})p(\tilde{f}|\check{f})$$

For simplicity, let us assume that $f^\star \neq 0$ everywhere so that $\tilde{f} = 0$ if and only if $z = 0$.

Theorem 2. *Let G be a graph and $\tilde{f}$ an observed signal such that, for all connected components C_k of G, $\tilde{f} \neq 0$ somewhere on G. Let $\mathcal{I}$ be the indices over which $\tilde{f} \neq 0$. Let $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}$ be the submatrix of $\mathbf{L}$ with rows and columns indexed by $\mathcal{I}^c$. Similarly define $\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}$ to be a submatrix of $\mathbf{A}$. Then for $p \neq 0$, the maximum likelihood estimate of $f^\star$, up to permutation, is given by:*

$$\check{f} = \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tilde{f}$$

4.1 Algorithmic Implementation

Algorithm 1 Interpolation of Signals Over Dropout Vertices

```
$\mathbf{L} \leftarrow \mathbf{D} - \mathbf{A}$
Form surjections $\phi : \mathcal{I}^c \to \{1, 2...|\mathcal{I}^c|\}$ and $\psi : \mathcal{I} \to \{1, 2...|\mathcal{I}^c|\}$.
for $i, j \in \{1, 2...|\mathcal{I}|$ do
    $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}(i, j) \leftarrow \mathbf{L}(\phi^{-1}(i), \phi^{-1}(j))$
end for
for $i \in \{1, 2..|\mathcal{I}^c|\}, j \in \{1, 2...|\mathcal{I}|\}$ do
    $\tilde{f}_{\mathcal{I}}(j) \leftarrow \tilde{f}(\psi^{-1}(j))$
    $\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}(i, j) \leftarrow \mathbf{A}(\phi^{-1}(i), \psi^{-1}(j))$
end for
$\delta_{\mathcal{I}^c} \leftarrow \mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$
$\check{f} \leftarrow \tilde{f} + \phi(\delta_{\mathcal{I}^c})$
return $\check{f}$
```

4.2 Runtime Analysis

Note that the surjections can be calculated in time $\mathcal{O}(n)$ by sweeping through $\tilde{f}$ once. Likewise, the necessary submatrices can be generated in time $\mathcal{O}(n)$. The most expensive step is calculation of $\mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}$, which occurs in time roughly equal to $\mathcal{O}(|\mathcal{I}^c|^3)$. Then the multiplication $\mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$ is done in time $\mathcal{O}(|\mathcal{I}^c||\mathcal{I}| + |\mathcal{I}^c|^2)$. So overall runtime is $\mathcal{O}(|\mathcal{I}^c|^3 + |\mathcal{I}^c||\mathcal{I}|)$. Then for dropout probability $1-p$ and a graph of size n, the expected runtime of this algorithm is $\mathcal{O}\left(n^3(1-p)^3 + n^2(1-p)^2 + n(1-p^2)\right)$, which we will now prove.

Theorem 3. *For dropout probability p and graph size n, algorithm 5 has expected runtime on the order of $\mathcal{O}\left(n^3(1-p)^3 + n^2(1-p)^2 + n(1-p^2)\right)$*

Proof. Note that the algorithm is deterministic and the input is random, so this is *not* to be confused with the expected runtime of a randomized algorithm. First, let T be the random variable denoting the runtime of the algorithm. Furthermore, note that $|\mathcal{I}^c| \sim \text{Bin}(n, 1-p)$. So for shorthand, let $X \sim \text{Bin}(n, 1-p)$. We may also write $X = X_1 + X_2 .. + X_n$ as a sum of i.i.d. Bernoulli $p-1$ random variables. Likewise, let $Y = n - X$. So we have that, as we have already argued, $T = \mathcal{O}(X^3 + XY)$. And thus, $\mathbb{E}[T] = \mathcal{O}(\mathbb{E}[X^3] + \mathbb{E}[XY])$. So let us calculate each of these expectations individually. For $\mathbb{E}[X^3]$, we will use the moment generating function of a binomial distribution:

$$m_X(t) = \mathbb{E}[e^{tX}] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] = \prod_{i=1}^{n} (e^t(1-p) + 1(p)) = (e^t(1-p) + p)^n$$

And we do the usual with the moment-generating function to find the non-central moments. Basic computation yields that,

$$\mathbb{E}[X^3] = \frac{\partial^3}{\partial t^3} m_X(0) = n(n-1)(n-2)(1-p)^3 + 3n(n-1)(1-p)^2 + n(1-p)$$

$$\leq n^3(1-p)^3 + 3n^2(1-p)^2 + n(1-p)$$

Now, we examine the distribution of XY. Note that, $\mathbb{E}[XY] = \mathbb{E}[X(n-X)] = n\mathbb{E}[X] - \mathbb{E}[X^2]$. Where $\mathbb{E}[X] = n(1-p)$ and $\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 = np(1-p) + (n(1-p))^2 = np(1-p) + n^2(1-p)^2$. And so, $\mathbb{E}[XY] = 3n^2(1-p) + np(1-p)$. And thus,

$$\mathbb{E}[T] \leq C(\mathbb{E}[X^3] + \mathbb{E}[XY]) \leq C\left(n^3(1-p) + 6n^2(1-p)^2 + n(1-p)(1+p)\right)$$

Therefore, $\mathbb{E}[T] = \mathcal{O}\left(n^3(1-p)^3 + n^2(1-p)^2 + n(1-p^2)\right)$, which is the desired result.

□

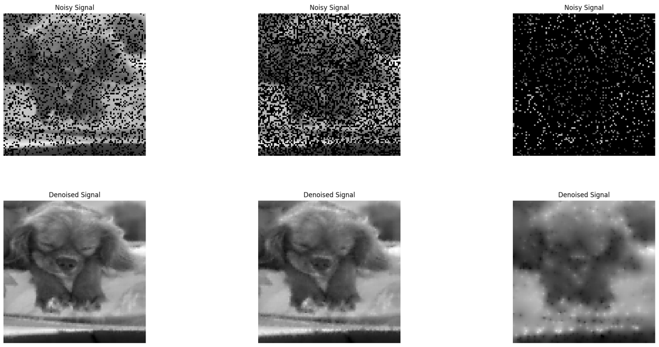

Figure 3: Denoising a puppy signal when dropout probabilities are $p = 0.25, 0.5$, and 0.9

4.3 Special Case

If H is disconnected - that is, all zeros are non-adjacent to any other zero, the optimization can be computed more simply. L_H becomes $\mathbf{I}$, and so $\delta_{\mathcal{I}^c}$ becomes $D_{\mathcal{I}^c,\mathcal{I}}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$. In other words, the optimization is achieved by setting each zeroed vertex equal to the average of its neighbors, which is intuitively the correct thing to do.

Approximation

Note that our optimization also requires the solution $\check{f}$ must be harmonic over $\mathcal{I}^c$ in the sense that,

Claim 1. *$\check{f}$ is harmonic in $\mathcal{I}^c$ in the sense that,* $\mathbf{L}\check{f}\big|_{\mathcal{I}^c} = 0$

Proof. Observe that we can rewrite the Laplacian Quadratic form like so:

$$\check{f}^T\mathbf{L}\check{f} = \sum_{(a,b)\in\mathcal{E}} w(a,b)(\check{f}(a) - \check{f}(b))^2 = \frac{1}{2}\sum_{a\in V}\sum_{b\sim a} w(a,b)(\check{f}(a) - \check{f}(b))^2$$

$$= \frac{1}{2}\sum_{a\in\mathcal{I}}\sum_{b\sim a} w(a,b)(\check{f}(a) - \check{f}(b))^2 + \frac{1}{2}\sum_{a\in\mathcal{I}^c}\sum_{b\sim a} w(a,b)(\check{f}(a) - \check{f}(b))^2$$

If we differentiate with respect to $\check{f}(a)$ for $a \in \mathcal{I}^c$, we find that,

$$0 = \frac{\partial}{\partial\check{f}(a)}\check{f}^T\mathbf{L}\check{f} = \frac{\partial}{\partial\check{f}(a)}\sum_{b\sim a}(\check{f}(a) - \check{f}(b))^2 = 2\sum_{b\sim a}(\check{f}(a) - \check{f}(b))$$

Which implies that,

$$\deg(a)\check{f}(a) = \sum_{b\sim a} w(a,b)\check{f}(b), \check{f}(a) = \sum_{b\sim a} p(a,b)\check{f}(b)$$

And thus $\check{f}$ is harmonic at vertex $a \in \mathcal{I}^c$.

□

We can interpret this in terms of heat diffusion. Suppose we are to "diffuse" the function $\tilde{f}$ over the rest of $\mathcal{G}$. In other words, we consider the discrete equation. Let $\partial\mathcal{I} = \{b : b \in \mathcal{I}, (a,b) \in \mathcal{E} \text{ for some } a \in \mathcal{I}\}$. We then seek the solution to the Dirichlet problem:

$$\mathbf{L}\check{f} = 0 \text{ in } \mathcal{I}^c \qquad \check{f}(a) = \tilde{f}(a) \text{ in } \partial\mathcal{I}(a)$$

But if $\check{f}$ is harmonic, this becomes $\frac{\partial}{\partial t}\check{f}(a) = 0$. In other words, $\check{f}$ is at equilibrium under diffusion. Note that we could represent this diffusion at time t as the vector $x \in \mathbb{R}^n \times \mathbb{N}$ obeying,

$$x(a,t) = \sum_{b\sim a} p(a,b)x(b,t-1) \text{ for } a \in \mathcal{I}^c \qquad x(a,t) = \tilde{f}(a) \text{ for } a \in \partial\mathcal{I}$$

Diffusion over $\mathcal{G}$ is provided by the matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$, then we develop the system of equations,

$$x(a,0) = \tilde{f} \qquad x(a,t) = \mathbf{W}x(a,t-1) \qquad \text{where } \mathbf{W} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{\mathcal{I}^c,\mathcal{I}} & \mathbf{P}_{\mathcal{I}^c,\mathcal{I}^c} \end{pmatrix}$$

Thus, if $x(a,t) = x(a,t-1)$, then x is harmonic in $\mathcal{I}^c$. Thus, we find that the solution is also given by,

$$\check{f} = \lim_{t\to\infty} x(a,\infty) = \lim_{t\to\infty} \mathbf{W}^t\tilde{f}$$

Where, for finite t, $\mathbf{W}^t\tilde{f}$ can be calculated in time $\mathcal{O}(tn^2)$.

5 Uniformly Downsampled Observations

Now consider the model in which, for each vertex i, we assume $\tilde{f}(i)$ is uniformly sampled between 0 and $f^\star$. For now, let us assume that $\tilde{f}$ is all positive. Thus,

$$p(\tilde{f}|x) = \prod_{i=1}^{n} \frac{1}{x(i)} \mathbb{I}\{0 \le \tilde{f}(i) \le x(i)\}$$

And so,

$$p(x|\tilde{f}) \propto p(\tilde{f}|x)p(x) \propto \exp(-\eta x^T \mathbf{L} x) \prod_{i=1}^{n} \frac{1}{|x(i)|} \mathbb{I}\{0 \le \tilde{f}(i) \le x(i)\}$$

$$= \mathbb{I}\{\tilde{f}/x \in [0,1]\} \exp\left(-\eta x^T \mathbf{L} x\right) \exp\left(-\sum_{i=1}^{n} \log(|x(i)|)\right)$$

$$= \mathbb{I}\{\tilde{f}/x \in [0,1]\} \exp\left(-\eta x^T \mathbf{L} x - \sum_{i=1}^{n} \log(|x(i)|)\right)$$

So our optimization becomes equivalent to:

$$\check{f} = \arg\min_{x} \left(\eta x^T \mathbf{L} x + \sum_{i=1}^{n} \log(x(i))\right) \qquad \text{such that} \qquad \tilde{f}(i)/x \in [0,1]$$

Note that this problem does not lend itself to a closed form solution. However, we can consider estimation of this constrained optimization via gradient descent.

Algorithm 2 Recovery of Signals from Uniform Dropout

```
$\mathbf{L} \leftarrow \mathbf{D} - \mathbf{A}$
$x \leftarrow \tilde{f}$
while $\|x_{prev} - x\| > \epsilon$ do
    $x_{prev} \leftarrow x$
    $C(x) \leftarrow \eta x^T \mathbf{L} x + \sum_{i=1}^{n} \log(|x(i)|)$
    $x \leftarrow x - \nu \nabla_x C(x)$
    for $i \in \{1, 2...n\}$ do
        if $\tilde{f}(i) > 0$ then
            $x(i) \leftarrow \max(x(i), \tilde{f(i)})$
        else
            $x(i) \leftarrow \min(x(i), \tilde{f(i)})$
        end if
    end for
end while
```

References

[1] Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[2] Antonio Ortega, Pascal Frossard, Jelena Kovačević, José MF Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5):808–828, 2018.

[3] Daniel Spielman. Spectral graph theory. *Combinatorial scientific computing*, 18, 2012.

[4] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

.1 Energy Gaussian Random Field Construction - Measures on Quotient Groups

Theorem 4. *Let ker($\mathbf{L}$) be the null space of $\mathbf{L}$, $\mathcal{N} = ker(\mathbf{L})$, $\mathcal{S} = \mathcal{N}^{\perp}$, and $\mathbf{H}$ be the projection matrix onto $\mathcal{S}$. Now let $\mathcal{A}$ be the σ-algebra generated by $\mathbf{H}$. We may define a probability distribution $\mathbb{P} : \mathcal{A} \to \mathbb{R}$ in which, for $A \in \mathcal{A}$,*

$$\mathbb{P}(A) \propto \int_{\mathbf{H}(A)} \exp(-\eta x^T \mathbf{L} x) dx$$

Proof. Intuitively, we are just constructing a measure on the quotient space $\mathbb{R}^d/\mathcal{N} \cong \mathcal{S}$ (since $\mathcal{S} = \text{im}(\mathbf{H}), \mathcal{N} = \ker(\mathbf{H})$, this statement holds by the first isomorphism theorem). Suppose that $\dim(\mathcal{N}) = d$ so that $\dim(\mathcal{S}) = n - d$.

Lemma 3. *For all $A \in \mathcal{A}$, A posesses the property that for all $a \in A$, $x \in \mathcal{N}$, $x + a \in A$. And therefore, $A = \{x + a : a \in \mathbf{H}(A), x \in \mathcal{N}\}$.*

Proof. If $A = \emptyset$, this holds vacuously. Otherwise, suppose $a \in A$, $x \in \mathcal{N}$. Then since $\mathcal{A}$ is the σ-algebra generated by the map $\mathbf{H}$, $\mathcal{A} \triangleq \{\mathbf{H}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^n)\}$. And thus, for our particular A, there exists a B such that $A = \mathbf{H}^{-1}(B)$. But now note that if $a \in A$, there exists a $b \in B$ such that $\mathbf{H}(a) = b$. And therefore, $\mathbf{H}(a+x) = \mathbf{H}(a) + \mathbf{H}(x) = b + 0 = b \in B$. Furthermore, since $x \in \mathcal{N}, \mathbf{H}(x) = 0$. Therefore, $\mathbf{H}(a + x) \in B$, so $a + x \in \mathbf{H}^{-1}(B) = A$. It follows that A is closed under addition of elements of $\mathcal{N}$.

As a Corollary, note that since

□

Consider the matrix $\mathbf{L}_\gamma = \sum_{i=1}^n (\lambda_i + \gamma \mathbb{I}_{\lambda_i = 0}) \psi_i \psi_i^T$ and $\mathcal{B}(\mathbb{R}^n)$-measurable distribution $\mathbb{P}_\gamma$ with the density

$$p_\gamma(x) \propto \exp(-\eta x^T \mathbf{L}_\gamma x \eta) = \exp(-\eta x^T \mathbf{L} x) \exp\left(-\eta x^T \sum_{i:\lambda_i = 0} \gamma \psi_i \psi_i^T x\right)$$

Note the particular form of $\mathbb{P}_\gamma$ is not important. All that's important is that we require it be integrable, which we do by setting some non-degenerate covariance. Then the desired measure $\mathbb{P}$ can be constructed by simply setting, for all $A \in \mathcal{A}$, $\mathbb{P}(\mathcal{A}) = \mathbb{P}_\gamma(\mathcal{A})$. That is, $\mathbb{P}$ is simply the restriction of $\mathbb{P}_\gamma$ with respect to $\mathcal{A} \subset \mathcal{B}(\mathbb{R}^n)$.

$$\mathbb{P}(A) = \mathbb{P}_\gamma(A) \propto \int_A \exp(-\eta x^T \mathbf{L} x) \exp\left(-\eta x^T \sum_{i:\lambda_i = 0} \gamma \psi_i \psi_i^T x\right) dx$$

For some $A \in \mathcal{A}$, it is clear that A is closed under linear combinations in $\mathcal{N}$, and thus we can write $A = \{a + t : a \in \mathbf{H}(A), t \in \mathcal{N}\}$. Additionally, we can write a change of variables $\varphi(x) = \Psi^T x$ so that,

$$\int_A \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i \varphi(x)(i)^2\right) \exp\left(-\eta \sum_{i:\lambda_i = 0} \gamma \varphi(x)(i)^2 x\right) dx$$

Since $D\varphi = \Psi^T$ is an orthonormal matrix, it has determinant 1, and so:

$$= \int_A \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i \varphi(x)(i)^2\right) \exp\left(-\eta \sum_{i:\lambda_i = 0} \gamma \varphi(x)(i)^2 x\right) |\det D(\varphi)| dx$$

By the change of variables formula,

$$= \int_{\varphi(A)} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i x(i)^2\right) \exp\left(-\eta \sum_{i:\lambda_i = 0} \gamma x(i)^2 x\right) dx$$

Where $\varphi(A) = \{\Psi^T a : a \in A\} = \mathbb{R}^d \times R$, where $R = \{(\langle \psi_{n-d+1}, a\rangle ... \langle \psi_n, a\rangle) : a \in A\} \subseteq \mathbb{R}^{n-d}$.

$$= \int_{y \in \mathbb{R}^d} \int_{x \in R} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i x(i)^2\right) \exp\left(-\eta \sum_{i:\lambda_i = 0} \gamma y(i)^2\right) dx dy$$

$$= \left(\int_{y \in \mathbb{R}^d} \exp\left(-\eta \sum_{i:\lambda_i = 0} \gamma y(i)^2 x\right) dy\right) \left(\int_{x \in R} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i x(i)^2\right) dx\right)$$

Since the leftmost integral has no dependence on A,

$$\propto \int_{x \in R} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i x(i)^2\right) dx$$

And another change of variables $\phi : \mathbf{H}(A) \to R$ with $\phi(a) = (\langle \psi_{n-d}, a\rangle ... \langle \psi_n, a\rangle)$ gives that, with $R = \phi(\mathbf{H}(A))$

$$\int_{\phi(\mathbf{H}(A))} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i x(i)^2\right) dx = \int_{\mathbf{H}(A)} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i \phi(x)(i)^2\right) dx =$$

$$= \int_{\mathbf{H}(A)} \exp\left(-\eta \sum_{i:\lambda_i \neq 0} \lambda_i \langle \phi_i, x\rangle^2\right) dx = \int_{\mathbf{H}(A)} \exp\left(-\eta x^T \mathbf{L} x\right) dx$$

Which gives the desired result. Note that the density at $\bar{f} \in \mathbb{R}^n/\mathcal{N}$ is given by what you would expect: $\frac{1}{Z_\eta} \exp\left(-\eta x^T \mathbf{L} x\right)$. The key insight is that we essentially choose, without loss of generality, $\bar{f} \in \mathcal{S}$ and perform the necessary calculations in this regime, while implicity performing calculations for the quotient space.

□

.2 Gibbs Distribution From Appropriate Markov Random Field

Theorem 5. *The Markov Random Field X with conditional probability kernel $\mathcal{N}(a) \mapsto \mathbb{P}_{\mathcal{N}(a)}$ as defined induces a Gibbs Distribution equal to the Energy Gaussian Random Field with density $Q(x) \propto \exp(-\eta x^T \mathbf{L} x)$ for some η.*

Proof. First, given the measure $\mathbb{P}_{\mathcal{N}(a)}$, we can let the corresponding density function be $P_{\mathcal{N}(a)}$. Fix $X_b = x_b$ for each $j \in \mathcal{N}(a)$. As a first observation, if we let Z be a random variable taking on values in $\mathcal{N}(a)$ such that $\mathbb{P}(Z = X(b)) = p(b|a)$, then the bias-variance decomposition guarantees us that,

$$\mathbb{E}_Z[(x_a - Z)^2] = (x_a - \mathbb{E}[Z])^2 + \mathrm{Var}(Z)$$

And thus,

$$\sum_{b \in \mathcal{N}(a)} p(b|a)(x_a - x_b)^2 = (x_a - \mu_{\mathcal{N}(a)})^2 + \mathrm{Var}(Z)$$

And similarly, multiplying through by $\deg(a)$,

$$\sum_{b \in \mathcal{N}(a)} w(a,b)(x_a - x_b)^2 = \deg(a)(x_a - \mu_{\mathcal{N}(a)})^2 + \deg(a)\mathrm{Var}(Z)$$

Then the density of X_a is given by:

$$P_{\mathcal{N}(a)}(x_a) \propto \exp\left(-\frac{(x_a - \mu_{\mathcal{N}(a)})^2}{2(\sigma_a)^2}\right)$$

By the Bias Variance Decomposition,

$$= \exp\left(-\frac{1}{2(\sigma)^2/\deg(a)}\left(\frac{1}{\deg(a)} \sum_{b \in \mathcal{N}(a)} w(a,b)(x_a - x_b)^2 - \frac{1}{\deg(a)}\mathrm{Var}(Z)\right)\right)$$

$$= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{b\in\mathcal{N}(a)} w(a,b)(x_a - x_b)^2\right)\right)\exp\left(-\frac{1}{2\sigma^2}\mathrm{Var}(Z)\right)$$

Since $\mathrm{Var}(Z)$ has no dependence on x_a, we conclude that,

$$P_{\mathcal{N}(a)}(x_a) \propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{b\in\mathcal{N}(a)} w(a,b)(x_a - x_b)^2\right)\right)$$

Note that for some vector $x = (x_1\, x_2 ... \, x_n)$, our distribution Q factorizes as a product of functions defined on the edges of $\mathcal{G}$. While technically speaking, x should be regarded as an element of $\mathbb{R}^n/\mathcal{N}$, since Q is $\mathcal{A}/\mathcal{B}(\mathbb{R}^n)$ measurable by construct, Q is constant over x's equivalence class, so it suffices to consider the representative $(x_1\, x_2 ... \, x_n)$ of $\bar{x}$

$$Q(x) = \frac{1}{Z_\eta}\exp(-\eta x^T \mathbf{L} x) = \frac{1}{Z_\eta}\exp\left(-\eta \sum_{(a,b)\in E}(x_a - x_b)^2\right)$$

$$= \frac{1}{Z_\eta}\prod_{(a,b)\in E}\exp\left(-\eta w(a,b)(x_a - x_b)^2\right)$$

Now fix $a \in V$. Let $x_{\mathcal{N}(a)}$ be the entries of x corresponding to $\mathcal{N}(a)$ and x_C correspond to entries of $V - \{a \cup \mathcal{N}(a)\}$. Observe that we can write $Q(x)$ as:

$$Q(x) = \frac{1}{Z_\eta}\prod_{b\in\mathcal{N}(a')}\exp\left(-\eta w(a',b)(x_a - x_b)^2\right)\prod_{\substack{(c,d)\in E\\ c,d\in C}}\exp\left(-\eta w(a,b)(x_c - x_d)^2\right)$$

So then $Q(x)$ factorizes into the product of two functions $g(x_C)$, which has no dependence on x_a or $x_{\mathcal{N}(a)}$ and a function $f(a, \mathcal{N}(a))$, given explicitly by:

$$g(x_C) \triangleq \frac{1}{Z_\eta}\prod_{\substack{(c,d)\in E\\ c,d\in C}}\exp\left(-\eta w(a,b)(x_c - x_d)^2\right)$$

$$f(x_a, x_{\mathcal{N}(a)}) \triangleq \prod_{b\in\mathcal{N}(a')}\exp\left(-\eta w(a',b)(x_a - x_b)^2\right)$$

Where g is some function of $\{X_a\}_a$ taken so that it does not depend on a or its neighbors. From this, we can deduce the desired result. So then, since p is a Gibbs distribution, it is a Markov Random Field, and so we may write:

$$\frac{Q(X_a = x_a | X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)})}{Q(X_a = 0 | X_{\mathcal{N}(a)})} = \frac{Q(X_a = x_a | X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)}{Q(X_a = 0 | X_{\mathcal{N}(a)}, X_C = x_C)}$$

$$= \frac{Q(X_a = x_a, X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)/Q(X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)}{p(X_a = 0, X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)/Q(X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)}$$

$$= \frac{Q(X_a = x_a, X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)}{Q(X_a = 0, X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}, X_C = x_C)} = \frac{g(x_C)f(x_a, x_{\mathcal{N}(a)})}{g(x_C)f(0, x_{\mathcal{N}(a)})}$$

$$= \frac{1}{f(0, x_{\mathcal{N}(a)})} \prod_{j \in \mathcal{N}(a)} \exp\left(-\eta w(a,b)(x_a - x_b)^2\right)$$

$$= \frac{1}{f(0, x_{\mathcal{N}(a)})} \exp\left(-\eta \sum_{j \in \mathcal{N}(a)} w(a,b)(x_a - x_b)^2\right)$$

Note $p(X_a = 0|X_{\mathcal{N}(a)})$ and $f(0, x_{\mathcal{N}(a)})$ have no dependence on the value of x_a. And letting $\eta = \frac{1}{2\sigma^2}$, we have:

$$Q(X_a = x_a | X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{j \in \mathcal{N}(a)} w(a,b)(x_a - x_b)^2\right)$$

And we conclude that $Q(X_a = x_a | X_{\mathcal{N}(a)} = x_{\mathcal{N}(a)}) \propto P_{\mathcal{N}(a)}(x_a)$. And since probability measures are normalized, it must be that they are in fact the same. Since a was selected arbitrarily, this holds for all vertices $a \in \mathcal{V}$. Again, it is important to remember that we are truly working mod $\mathcal{N}$, but the calculations do not change. This concludes the proof.

□

Theorem 6. *Let G be a graph and $\tilde{f}$ an observed signal such that, for all connected components C_k of G, $\tilde{f} \neq 0$ somewhere on G. Let $\mathcal{I}$ be the indices over which $\tilde{f} \neq 0$. Let $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}$ be the submatrix of $\mathbf{L}$ with rows and columns indexed by $\mathcal{I}^c$. Similarly define $\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}$ to be a submatrix of $\mathbf{A}$. Then for $p \neq 0$, the maximum likelihood estimate of $f^\star$, up to permutation, is given by:*

$$\check{f} = \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tilde{f}$$

Proof. We assume that $\tilde{f} = 0$ only when $z = 0$. In which case, $p(\tilde{f}|\check{f}) = (1-p)^{m_z}\mathbb{I}\{\tilde{f}(i) \in \{\check{f}(i), 0\}, \forall i\}$, where m_z is the number of zero entries of $\tilde{f}$, and $\mathbb{I}\{\tilde{f}(i) \in \{\check{f}(i), 0\}, \forall i\}$ asserts that $\tilde{f} = \check{f}$ whenever $\tilde{f} \neq 0$. Note that $(1-p)^m$ is independent of the particular $\check{f}$, therefore, the maximum likelihood estimate is equivalent to the following optimization:

$$\min_{\check{f}} \check{f}^T \mathbf{L} \check{f} \text{ subject to } \tilde{f}(i) \in \{\check{f}(i), 0\}, \forall i$$

So this becomes a problem of interpolating an optimally smooth signal which respects our observations over their support. Let $\check{f} = \tilde{f} + \delta$, so δ represents a vector of allowable changes. That is, $\delta = 0$ whenever $\tilde{f} > 0$. Then the new optimization becomes:

$$\min_{\delta}(\tilde{f}+\delta)^T\mathbf{L}(\tilde{f}+\delta) \text{ subject to } f(i) \neq 0 \implies \delta(i) = 0$$

Where the objective function is equal to:

$$\delta^T\mathbf{L}\delta + 2\delta^T\mathbf{L}\tilde{f} + \tilde{f}^T\mathbf{L}\tilde{f}$$

We introduce a Lagrangian. Let $\mathcal{I}$ be the nonzero indices of $\tilde{f}$. Suppose without loss of generality that $I = \{1...k\}$. So then, with our constraint, $\delta^T\mathbf{L}\delta$ becomes,

$$\begin{pmatrix}\delta_{\mathcal{I}^c} & \delta_{\mathcal{I}}\end{pmatrix}\begin{bmatrix}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c} & \mathbf{L}_{\mathcal{I}^c,\mathcal{I}}\\ \mathbf{L}_{\mathcal{I},\mathcal{I}^c} & \mathbf{L}_{\mathcal{I},\mathcal{I}}\end{bmatrix}\begin{pmatrix}\delta_{\mathcal{I}^c}\\ \delta_{\mathcal{I}}\end{pmatrix} = \delta_{\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}\delta_{\mathcal{I}^c} + 2\delta_{\mathcal{I}}\mathbf{L}_{\mathcal{I},\mathcal{I}^c}\delta_{\mathcal{I}^c} + \delta_{\mathcal{I}}\mathbf{L}_{\mathcal{I},\mathcal{I}}\delta_{\mathcal{I}}$$

$$= \delta_{\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}\delta_{\mathcal{I}^c} + 0 + 0 = \delta_{\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}\delta_{\mathcal{I}^c}$$

Likewise,

$$2\delta^T\mathbf{L}\tilde{f} = 2\begin{pmatrix}\delta_{\mathcal{I}^c} & \delta_{\mathcal{I}}\end{pmatrix}\begin{bmatrix}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c} & \mathbf{L}_{\mathcal{I}^c,\mathcal{I}}\\ \mathbf{L}_{\mathcal{I},\mathcal{I}^c} & \mathbf{L}_{\mathcal{I},\mathcal{I}}\end{bmatrix}\begin{bmatrix}\tilde{f}_{\mathcal{I}^c}\\ \tilde{f}_{\mathcal{I}}\end{bmatrix} = 2\delta_{\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$$

We now see that our desired optimization is equal to a constrained optimization over $\mathcal{I}^c$. We find that this is equivalent to optimizing,

$$\min_{\delta_{\mathcal{I}^c}} F(\delta_{\mathcal{I}^c}) \text{ where } F(\delta_{\mathcal{I}^c}) = \delta_{\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}\delta_{\mathcal{I}^c} + 2\delta_{\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$$

Setting the gradient to zero:

$$0 = \frac{\partial F}{\partial \delta_{\mathcal{I}^c}} = 2\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}\delta_{\mathcal{I}^c} + 2\mathbf{L}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$$

So a solution would be, provided the invertibility of $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}$,

$$\delta_{\mathcal{I}^c} = -\mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{L}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$$

And this invertibility exists. Let H be the subgraph induced on G by $\mathcal{I}^c$. Let $\mathbf{L}_H$ be the Laplacian of H, indexed the same way as $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}$. Furthermore, let $\mathbf{D}_{\mathcal{I}^c,\mathcal{I}}$ be a sort of flux degree matrix in which, for some $a \in \mathcal{I}^c$,$\mathbf{D}_{\mathcal{I}^c,\mathcal{I}}(a,a) = \sum_{b:b\sim a, b\in\mathcal{I}} w(a,b)$. For brevity, we will refer to $\mathbf{D}_{\mathcal{I}^c,\mathcal{I}}(a,a)$ as $\deg_{\mathcal{I}}(a)$. Note $\deg_{\mathcal{I}}(a)$ is positive if a is adjacent to anything in $\mathcal{I}$. So then, $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c} = \mathbf{L}_H + \mathbf{D}_{\mathcal{I}^c,\mathcal{I}}$. So then for any $x \in \mathbb{R}^{|\mathcal{I}^c|}$ with $x \neq 0$, we have,

$$x^T\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}x = x^T\mathbf{D}_{\mathcal{I}^c,\mathcal{I}}x + x^T\mathbf{L}_H x$$

Since $\mathbf{L}_H$ is a Laplacian, it is postive semi-definite, so $x^T\mathbf{L}_H x \geq 0$. Furthermore, since we assume that there is at least one edge between $\mathcal{I}$ and $\mathcal{I}^c$, there exists an a for which $\deg_{\mathcal{I}}(a) > 0$. Meaning,

$$x^T\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}x = \sum_{a\in\mathcal{I}^c}\deg_{\mathcal{I}}(a)x(a)^2 + x^T Lx$$

Now there are two scenarios. First, if $x(a) \neq 0$ for at least one a adjacent to $\mathcal{I}$, then the above quantity will be > 0. Now assume otherwise that for all a such that a is adjacent to some $b \in \mathcal{I}^c$ that $x(a) = 0$. Note that, by assumption, every connected component of H is adjacent to some vertex in $\mathcal{I}$. So then letting $\mathcal{J} = \{a : x(a) = 0\}$ be the vertices over which x is zero, it follows that for all connected components $\{B_k\}_k$ of H, $B_k \cap \mathcal{J} \neq \emptyset$. Now letting $\{\mathbf{L}_{B_k}\}_k$ be the respective Laplacians of the connected components of H and $x(B_k)$ be x restricted to B_k, it follows that,

$$x^T\mathbf{L}_H x = \sum_k x(B_k)^T\mathbf{L}_{B_k}x(B_k)$$

Note the above is strictly nonnegative, and is zero if and only if $x(B_k)$ is constant for all k. But since $x(a) = 0$ for some a in each B_k, this means x is identically zero over all the connected components of H, so x itself must be the zero vector. Since we assumed $x \neq 0$, then, it follows that $x^T\mathbf{L}_H x > 0$. We conclude that, for any x, $x^T\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}x > 0$. Since $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}^c}$ is obviously symmetric, it follows that it is positive definite and thus invertible.

As one final observation, observe that $\mathbf{L}_{\mathcal{I}^c,\mathcal{I}}$, containing no diagonals is equal to the negative adjacency over the same index set: $-\mathbf{L}_{\mathcal{I}^c,\mathcal{I}} = \mathbf{A}_{\mathcal{I}^c,\mathcal{I}}$. So our solution is:

$$\delta_{\mathcal{I}^c} = \mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}$$

And thus the best interpolation is:

$$\check{f} = \tilde{f} + \begin{pmatrix}\delta_{\mathcal{I}^c}\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}\tilde{f}_{\mathcal{I}}\\ \tilde{f}_{\mathcal{I}}\end{pmatrix} = \begin{pmatrix}\mathbf{0} & \mathbf{L}^{-1}_{\mathcal{I}^c,\mathcal{I}^c}\mathbf{A}_{\mathcal{I}^c,\mathcal{I}}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}\tilde{f}$$

□