A variational approach to consistency of graph-based methods for data clustering and dimensionality reduction.

Nicolás García Trillos
Brown University

Siam Annual Meeting,
July 11th 2017

Continuum Limit of Posteriors in Graph Bayesian Inverse Problems.

joint work with Daniel Sanz-Alonso (Brown University).

Inverse Problems

$$u \in X \longmapsto \mathcal{F}(u) \in Z \longmapsto \mathcal{O} \circ \mathcal{F}(u) \in \mathbb{R}^p$$

- Forward map $\mathcal{F} : X \to Z$.
- Observation map $\mathcal{O} : Z \to \mathbb{R}^p$.
- Spaces X, Z are spaces of functions on $\mathcal{M} \subseteq \mathbb{R}^d$. For example: $L^2(\mathcal{M}), C(\mathcal{M}), H^1(\mathcal{M}), \dots$

Observations are contaminated by noise.
For example, additive noise:

$$y_i = [\mathcal{O} \circ \mathcal{F}(u)]_i + \eta_i, \quad i = 1, \dots, p.$$

In general, we use a **negative log-likelihood function** to describe noise model:

$$\phi(u; y).$$

Goal: Given observations y learn input u.

Goal: Learn input from observations.
How?: Use Bayesian approach. Need a **prior** distribution:

$$u \sim \pi.$$

We can then obtain the **posterior** distribution of $u|y$:

$$d\mu^y(u) \propto \exp(-\phi(u; y))d\pi(u).$$

Example 1: Semi-supervised learning

- Input space: $u \in C(\mathcal{M})$ with $\int_{\mathcal{M}} u(x) d\gamma(x) = 0$.
- Forward map: $\mathcal{F} : u \mapsto u$.
- Observations:

$$y_i = S(u(x_i) + \eta_i), \quad i = 1, \dots, p.$$

$$\eta_i \sim N(0, \sigma^2).$$

Negative log-likelihood :

$$\phi(u; y) = -\sum_{i=1}^{p} \log(\Psi_\gamma(u(x_j) \cdot y_j))$$

- Prior: $\pi = N(0, (-\Delta_{\mathcal{M}})^{-s})$.

Example 2: Learning the initial condition of the heat equation.

- Input space: $u \in L^2(\mathcal{M})$.
- Forward map: $\mathcal{F} : u \mapsto e^{\Delta} u$.
- Observations:

$$y_i = \fint_{B(x_i,\delta)\cap\mathcal{M}} u(x)dx + \eta_i, \quad i = 1, \dots, p.$$

$$\eta_i \sim N(0, \sigma^2).$$

Negative log-likelihood:

$$\phi(u; y) = \frac{1}{\sigma^2} \|y - \mathcal{O} \circ \mathcal{F}(u)\|^2$$

- Prior: $\pi = N(0, (-\Delta_{\mathcal{M}})^{-s})$.

What do we do if the domain $\mathcal{M}$ is unknown?
Only access to:

- $y_1, \ldots, y_p$.
- $\mathcal{M}_n = \{x_1, \ldots, x_p, \ldots, x_n\} \subseteq \mathcal{M}$. (say i.i.d. samples from γ).

What do we do if the domain $\mathcal{M}$ is unknown?
Only access to:

- $y_1, \dots, y_p$.
- $\mathcal{M}_n = \{x_1, \dots, x_p, \dots, x_n\} \subseteq \mathcal{M}$. (say i.i.d. samples from γ).

Need surrogates for $\mathcal{F}, \mathcal{O}, \pi$.

First: Construct a geometric graph on $\mathcal{M}_n$

$$x_i \sim x_j \quad \text{if} \quad |x_i - x_j| < \varepsilon.$$

Then, produce graph Laplacian Δ_n.

- Forward map: $\mathcal{F}_n : u_n \in L^2(\gamma_n) \mapsto L^2(\gamma_n)$

$$\mathcal{F}_n u_n = e^{-\Delta_n} u_n.$$

- Observation map: $\mathcal{O}_n : v_n \in L^2(\gamma_n) \mapsto \mathbb{R}^p$

$$[\mathcal{O}_n v_n]_i = \fint_{B(x_i,\delta)\cap\mathcal{M}_n} v_n(x) d\gamma_n(x), \quad i = 1, \dots, p.$$

- Prior: $\pi_n = N(0, \Delta_n^{-s})$

Graph posterior:

$$\mu_n^y(u_n) \propto \exp(-\phi_n(u_n; y)) d\pi_n(u_n), \quad u_n \in L^2(\gamma_n).$$

Ground-truth posterior:

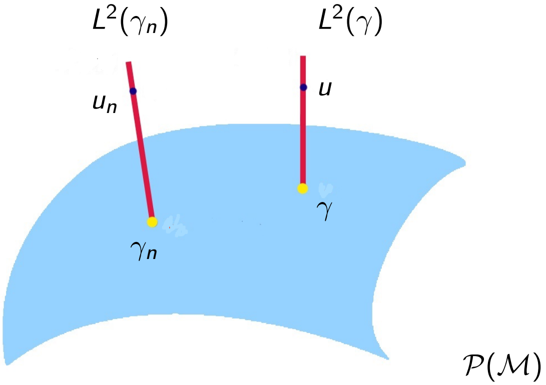
$$\mu^y(u) \propto \exp(-\phi(u; y)) d\pi(u), \quad u \in L^2(\gamma).$$

How and when do we recover μ^y from μ_n^y?

How and when do we recover μ^y from μ_n^y?
Note: μ_n^y is supported on $L^2(\gamma_n)$ whereas μ^y is supported on $L^2(\gamma)$.

How?

TL^2 space



$$TL^2 = \{(\theta, v) \,:\, \theta \in \mathcal{P}(\mathcal{M}), \quad v \in L^2(\theta)\}.$$

with distance between (θ_1, v_1) and (θ_2, v_2):

$$\inf_{\pi \in \Gamma(\theta_1, \theta_2)} \int_{\mathcal{M}\times\mathcal{M}} d_{\mathcal{M}}^2(x,y)d\pi(x,y) + \int_{\mathcal{M}\times\mathcal{M}} |v_1(x) - v_2(y)|^2 d\pi(x,y).$$

TL^2 space in previous works:

- *Continuum limit of total variation on point clouds.* ARMA. NGT and Slepčev.
- *A variational approach to the consistency of spectral clustering.* ACHA. NGT and Slepčev.
- *A new analytic approach to consistency and overfitting in regularized empirical risk minimization* EJAM. NGT and R. Murray.
- *A transportation L^p distance for signal analysis.* Preprint. Slepčev, Thorpe, et al.

$$L^2(\gamma_n) \hookrightarrow TL^2 \text{ induces } \mathcal{P}(L^2(\gamma_n)) \hookrightarrow \mathcal{P}(TL^2).$$

$$L^2(\gamma) \hookrightarrow TL^2 \text{ induces } \mathcal{P}(L^2(\gamma)) \hookrightarrow \mathcal{P}(TL^2).$$

When?

Theorem (NGT & D. Sanz-Alonso)

Suppose that

$$\frac{\log(n)^{1/m}}{n^{1/m}} \ll \varepsilon \ll \frac{1}{n^{1/s}},$$

where $s > 2m$. *Then,*

$$\mu_n^y \overset{\mathcal{P}(TL^2)}{\longrightarrow} \mu^y.$$

Moreover,

$$\mathcal{F}_{n\sharp}\mu_n^y \overset{\mathcal{P}(TL^2)}{\longrightarrow} \mathcal{F}_{\sharp}\mu^y.$$

$$\frac{\log(n)^{1/m}}{n^{1/m}} \ll \varepsilon \ll \frac{1}{n^{1/s}},$$

- Lower bound: ∞-OT distance between γ_n and γ.
- Upper bound: Needed to control high frequencies graph Laplacian.

Variational characterization of posteriors.

Graph:

$$J_n(\nu_n) := D_{KL}(\nu_n || \pi_n) + \int_{L^2(\gamma_n)} \phi_n(u_n; y) d\nu_n(u_n), \quad \nu_n \in \mathcal{P}(L^2(\gamma_n)).$$

$$\mu_n^y = \text{argmin}_{\nu_n} J_n(\nu_n).$$

Ground-Truth:

$$J(\nu) := D_{KL}(\nu || \pi) + \int_{L^2(\gamma)} \phi(u; y) d\nu(u), \quad \nu \in \mathcal{P}(L^2(\gamma)).$$

$$\mu^y = \text{argmin}_{\nu} J(\nu).$$

Note: The variational characterization of posteriors allows us to use variational techniques.

Conclusions

- We set forth formulation of Bayesian inverse problems in unknown domains.
- Contribute to the study of **robust** UQ in machine learning tasks such as Semi-supervised learning.

Thank you for your attention!