Bayesian Posterior Consistency in Graph Based Semi-Supervised Learning

Kevin Miller
University of California, Los Angeles

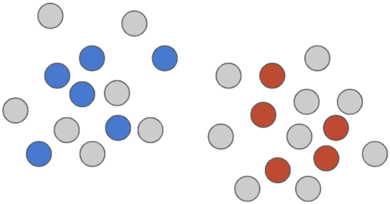
joint work with Hao Li (UCLA), Bamdad Hosseini (Caltech), Andrew Stuart (Caltech), and Andrea Bertozzi (UCLA).

December 13, 2019

Motivation - Semi Supervised Learning (SSL)

Given dataset that we know the classification (labeling) of **only some** of the datapoints.

- Can we infer the labeling of the rest of the **unlabeled** datapoints?



Setup – Semi-Supervised Learning (SSL)

Given $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subset \mathbb{R}^d$ (*unlabeled data*), with indexing set $Z = \{1, 2, \ldots, N\}$. Assume every point in Z belongs to one of M classes

- That is, assume there exists function $\ell : Z \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ ($\mathbf{e}_j \in \mathbb{R}^M$ are standard basis)

Setup – Semi-Supervised Learning (SSL)

Given $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subset \mathbb{R}^d$ (*unlabeled data*), with indexing set $Z = \{1, 2, \ldots, N\}$. Assume every point in Z belongs to one of M classes

- That is, assume there exists function $\ell : Z \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ ($\mathbf{e}_j \in \mathbb{R}^M$ are standard basis)

Let $Z' \subseteq Z$ be subset $J \leq N$ nodes, with $Y : Z' \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ the *noisily observed labels* of the points in Z'.

- refer to Y as *labeled data*

Given $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \subset \mathbb{R}^d$ (*unlabeled data*), with indexing set $Z = \{1, 2, \ldots, N\}$. Assume every point in Z belongs to one of M classes

- That is, assume there exists function $\ell : Z \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ ($\mathbf{e}_j \in \mathbb{R}^M$ are standard basis)

Let $Z' \subseteq Z$ be subset $J \leq N$ nodes, with $Y : Z' \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ the *noisily observed labels* of the points in Z'.

- refer to Y as *labeled data*

Semi-Supervised Learning (SSL) Problem:

- Can we “recover” labeling ℓ from X, Y, Z, Z'?

Setup – Semi-Supervised Regression

Cast SSL problem as inverse problem to infer a "ground-truth" latent variable $U^\dagger \in \mathbb{R}^{M \times N}$ under *regression model*:

$$Y = U^\dagger H^T + \gamma\eta, \qquad \eta \in \mathbb{R}^{M \times J}, \;\; \eta_{mj} \sim \mathcal{N}(0,1)$$

where $H \in \mathbb{R}^{J \times N}$ is matrix obtained by removing $Z - Z'$ rows of identity, I_N.

Setup – Semi-Supervised Regression

Cast SSL problem as inverse problem to infer a "ground-truth" latent variable $U^\dagger \in \mathbb{R}^{M\times N}$ under *regression model*:

$$Y = U^\dagger H^T + \gamma\eta, \qquad \eta \in \mathbb{R}^{M\times J}, \;\; \eta_{mj} \sim \mathcal{N}(0,1)$$

where $H \in \mathbb{R}^{J\times N}$ is matrix obtained by removing $Z - Z'$ rows of identity, I_N.

Semi-Supervised Regression (SSR) Problem:

- Can we infer ground-truth $U^\dagger$ from X, Y, Z, Z'?

Setup – BIP Semi-Supervised Regression

Previous problem is still ill-posed, so we regularize with prior μ_0 on $U^\dagger$. Obtain a Bayesian Inverse Problem (BIP) for our SSR problem:

BIP Semi-Supervised Regression Problem :

- Given X, Y, Z, Z' and prior measure μ_0 on U, we identify posterior probability measure μ^Y via Radon-Nikodym derivative

$$\frac{\mathrm{d}\mu^Y}{\mathrm{d}\mu_0}(U) \propto \exp\big(- \frac{1}{\gamma^2}\|UH^T - Y\|_F^2\big),$$

  per our regression model.

Setup – BIP Semi-Supervised Regression

Previous problem is still ill-posed, so we regularize with prior μ_0 on $U^\dagger$. Obtain a Bayesian Inverse Problem (BIP) for our SSR problem:

BIP Semi-Supervised Regression Problem :

- Given X, Y, Z, Z' and prior measure μ_0 on U, we identify posterior probability measure μ^Y via Radon-Nikodym derivative

$$\frac{\mathrm{d}\mu^Y}{\mathrm{d}\mu_0}(U) \propto \exp\left(-\frac{1}{\gamma^2}\|UH^T - Y\|_F^2\right),$$

  per our regression model.

Our prior will capture unlabeled data's inherent geometry via similarity graph and associated graph Laplacian matrix.

Similarity Graph

Assume our data in X can be represented by similarity graph $G(Z, W)$, where

- W : self-adjoint matrix, with $w_{ij} \geq 0$
- $w_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$: "similarity kernel"

Similarity Graph

Assume our data in X can be represented by similarity graph $G(Z, W)$, where

- W : self-adjoint matrix, with $w_{ij} \geq 0$
- $w_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$: "similarity kernel"

Symmetric Graph Laplacian Matrix

$$L = D^{-p}(D - W)D^{-p}, \quad p \in \mathbb{R}$$

where $D = diag(d_i), d_i = \sum_{j \in Z} w_{ij}$ is degree matrix.

Similarity Graph

Assume our data in X can be represented by similarity graph $G(Z, W)$, where

- W : self-adjoint matrix, with $w_{ij} \geq 0$
- $w_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$: "similarity kernel"

Symmetric Graph Laplacian Matrix

$$L = D^{-p}(D - W)D^{-p}, \quad p \in \mathbb{R}$$

where $D = diag(d_i), d_i = \sum_{j \in Z} w_{ij}$ is degree matrix.

- $p = 0 \rightarrow$ *unnormalized* Graph Laplacian matrix
- $p = 1/2 \rightarrow$ *normalized* Graph Laplacian matrix

Gaussian Prior measure

With $G(Z, W)$ and L, we can define a *covariance operator*:

$$C_\tau = \tau^{2\alpha}\left(L + \tau^2 \mathrm{I}_N\right)^{-\alpha}$$

Well known that $L \geq 0$, so then $C_\tau > 0$ for $\alpha, \tau^2 > 0$.

Gaussian Prior measure

With $G(Z, W)$ and L, we can define a *covariance operator*:

$$C_\tau = \tau^{2\alpha}\left(L + \tau^2 \mathrm{I}_N\right)^{-\alpha}$$

Well known that $L \geq 0$, so then $C_\tau > 0$ for $\alpha, \tau^2 > 0$.

Gaussian Prior measure:

$$\begin{aligned}\mu_0(dU) &\sim \mathcal{N}(0, I_M \otimes C_\tau) \\ &\propto \prod_{\ell=1}^{M} \exp\left(-\frac{1}{2}\langle \mathbf{u}_\ell^T, \tau^{-2\alpha}\left(L + \tau^2 \mathrm{I}_N\right)^{\alpha} \mathbf{u}_\ell^T\rangle\right) dU\end{aligned}$$

Posterior Gaussian Measure

Can now identify posterior measure from our regression model (Gaussian likelihood) and Gaussian prior:

$$\mu^Y(dU) \propto \exp\left(-\frac{1}{2}\left[\underbrace{\langle U^T, C_\tau^{-1}U^T\rangle_F}_{prior} + \underbrace{\frac{1}{\gamma^2}\|UH^T - Y\|_F^2}_{likelihood}\right]\right) dU$$

Gaussian likelihood and Gaussian prior $\implies$ posterior μ^Y Gaussian

$$\mu^Y \sim \mathcal{N}(U^*, C^*)$$

where $C^* = \left(C_\tau^{-1} + \frac{1}{\gamma^2}H^T H\right)^{-1}, \quad U^* = \frac{1}{\gamma^2}Y^T HC^*$

Bayesian Posterior Consistency

Given a "ground-truth" $U^\dagger$, from which Y is observed, we want to show under what conditions the posterior $\mu^Y(dU)$ "contracts" onto $U^\dagger$ in the limit of model parameters.

Still Need:

- How to measure posterior contraction?
- Restrictions on data geometry (i.e. similarity graph properties)?
- Valid choices of possible $U^\dagger$ for this consistency?

Posterior Contraction Measure

Define the following measure of posterior contraction

$$\mathcal{I} := \mathbb{E}_{Y|U^\dagger}\mathbb{E}_{U|Y}\left\|U - U^\dagger\right\|_F^2$$

- inner expectation $\rightarrow$ w.r.t. the posterior measure $\mu^Y(\mathrm{d}U)$
- outer expectation $\rightarrow$ w.r.t. the measure of Y conditioned on $U^\dagger$ following SSR model

.

Goal: to show that $\mathcal{I} \rightarrow 0$ with the noise standard deviation γ and other the prior hyperparameters such as τ, α for certain *weakly connected* graphs.

Disconnected Graph

(a) $W_0 \in \mathbb{R}^{N\times N}$ is block diagonal

$$W_0 = \mathrm{diag}(\widetilde{W}_1, \widetilde{W}_2, \cdots, \widetilde{W}_K),$$

with $\widetilde{W}_k \in \mathbb{R}^{N_k \times N_k}$ denoting the weight matrices of the subgraphs G_k.

(b) $\tilde{L}_k$ graph Laplacian matrices of G_k, i.e.,

$$\widetilde{L}_k := \widetilde{D}_k^{-p}(\widetilde{D}_k - \widetilde{W}_k)\widetilde{D}_k^{-p}$$

There exists uniform $\theta > 0$ so that the submatrices $\widetilde{L}_k$ have a uniform spectral gap, i.e.,

$$\langle \mathbf{x}, \widetilde{L}_k \mathbf{x} \rangle \geq \theta \langle \mathbf{x}, \mathbf{x} \rangle, \tag{1}$$

for all vectors $\mathbf{x} \in \mathbb{R}^{N_k}$ and $\mathbf{x} \perp \widetilde{D}_k^p \mathbf{1}$.

Disconnected to Weakly Connected Graph

Now, we perturb this disconnected graph G_0 to obtain $G_\epsilon(Z, W_\epsilon)$:

$$W_\epsilon = W_0 + \sum_{h=1}^{\infty} \epsilon^h W_h,$$

- W_h are self-adjoint and $\{\|W_h\|_2\}_{h=1}^{\infty} \in \ell^\infty$.
- Let $w_{ij}^{(0)}$ and $w_{ij}^{(h)}$ denote the entries of W_0 and W_h respectively. Then

$$\begin{cases} w_{ij}^{(h)} \geq 0, \quad \text{if} \quad w_{ij}^{(0)} = 0 \quad \text{for} \quad i, j \in Z, i \neq j. \\ w_{ii}^{(h)} = 0. \end{cases}$$

Disconnected to Weakly Connected Graph

Therefore, we have

$$L_\epsilon := D_\epsilon^{-p}(D_\epsilon - W_\epsilon)D_\epsilon^{-p}, \quad \text{and} \quad C_{\tau,\epsilon} := \tau^{2\alpha}(L_\epsilon + \tau^2 \mathrm{I}_N)^{-\alpha}$$

where D_ϵ corresponds to the diagonal degree matrix of W_ϵ.

Disconnected to Weakly Connected Graph

Therefore, we have

$$L_\epsilon := D_\epsilon^{-p}(D_\epsilon - W_\epsilon)D_\epsilon^{-p}, \quad \text{and} \quad C_{\tau,\epsilon} := \tau^{2\alpha}(L_\epsilon + \tau^2 \mathrm{I}_N)^{-\alpha}$$

where D_ϵ corresponds to the diagonal degree matrix of W_ϵ.

Remind Goal: Given a weakly connected graph representation of X, can we recover a "ground-truth" function $U^\dagger$ with some observations Y from $U^\dagger$?

- Need some restrictions on $U^\dagger$!

Assumptions about Ground-Truth $U^\dagger$

Let $(\mathbf{u}_\ell^\dagger)^T$ for $\ell = 1, \ldots, M$ denote the rows of $U^\dagger$. Then

$$\mathbf{u}_\ell^\dagger \in \operatorname{span}\{\bar{\boldsymbol{\chi}}_1, \ldots, \bar{\boldsymbol{\chi}}_K\},$$

where the weighted set functions

$$\bar{\boldsymbol{\chi}}_k := \frac{D_0^p \mathbf{1}_k}{\left|D_0^p \mathbf{1}_k\right|},$$

with $\mathbf{1}_k \in \mathbb{R}^N$ denoting indicator of the clusters Z_k (subgraph $\tilde{G}_k$).

Assumptions about Ground-Truth $U^\dagger$

Let $(\mathbf{u}_\ell^\dagger)^T$ for $\ell = 1, \ldots, M$ denote the rows of $U^\dagger$. Then

$$\mathbf{u}_\ell^\dagger \in \mathrm{span}\{\bar{\boldsymbol{\chi}}_1, \ldots, \bar{\boldsymbol{\chi}}_K\},$$

where the weighted set functions

$$\bar{\boldsymbol{\chi}}_k := \frac{D_0^p \mathbf{1}_k}{\left|D_0^p \mathbf{1}_k\right|},$$

with $\mathbf{1}_k \in \mathbb{R}^N$ denoting indicator of the clusters Z_k (subgraph $\tilde{G}_k$).

And... at least one label is observed in each cluster Z_k

$$|Z' \cap Z_k| > 0 \qquad \forall k = 1, \ldots, K.$$

Main Result

All together – want to show that:

$$\mathcal{I}(\gamma, \alpha, \tau, \epsilon) = \mathbb{E}_{Y|U^\dagger}\mathbb{E}_{U|Y}\left\|U - U^\dagger\right\|_F^2 \to 0$$

in the limit of model parameters γ, τ, ϵ.

Main Result – $\epsilon = 0$ Case

Theorem ($\epsilon = 0$ Case)

Suppose have G_0, $U^\dagger$, and Z' that satisfy all Assumptions presented. Then there exists a constant $\Xi > 0$, such that $\forall(\tau, \alpha, \gamma) \in \mathbb{R}^3_+$ it holds that

$$\mathcal{I}(\gamma, \alpha, \tau) \leq \Xi \max\left\{\gamma^2, \tau^{2\alpha}\right\}\left(1 + \max\left\{\gamma^2, \tau^{2\alpha}\right\}\sum_{m=1}^{M} |\mathbf{u}_m^\dagger|^2\right).$$

Main Result – $\epsilon = 0$ Case

Theorem ($\epsilon = 0$ Case)

Suppose have G_0, $U^\dagger$, and Z' that satisfy all Assumptions presented. Then there exists a constant $\Xi > 0$, such that $\forall(\tau, \alpha, \gamma) \in \mathbb{R}^3_+$ it holds that

$$\mathcal{I}(\gamma, \alpha, \tau) \leq \Xi \max\left\{\gamma^2, \tau^{2\alpha}\right\}\left(1 + \max\left\{\gamma^2, \tau^{2\alpha}\right\}\sum_{m=1}^{M}|\mathbf{u}_m^\dagger|^2\right).$$

Note if we fix α and set $\tau = \gamma^{1/\alpha}$, we can simplify

$$\begin{aligned}\mathcal{I}(\gamma, \alpha, \tau) &\leq \Xi\gamma^2\left(1 + \gamma^2\|U^\dagger\|_F^2\right)\\ &\to 0, \quad \text{as } \gamma \to 0\end{aligned}$$

Main Result

Main Theorem

Suppose have G_0, $U^\dagger$, Z', and G_ϵ that satisfy all Assumptions presented. Then there exist constants $\epsilon_0 \in (0,1)$, and $\Xi, \Xi_1 > 0$, such that $\forall(\epsilon,\tau,\alpha,\gamma) \in (0,\epsilon_0) \times \mathbb{R}^3_+$ it holds that

$$\mathcal{I}(\gamma,\alpha,\tau,\epsilon) \leq \Xi \max\left\{\gamma^2, \left(\frac{\tau^2}{1-\Xi_1\epsilon/\tau^2}\right)^\alpha\right\} \times \left(1 + A(\tau,\epsilon)\max\left\{\gamma^2, \left(\frac{\tau^2}{1-\Xi_1\epsilon/\tau^2}\right)^\alpha\right\}\sum_{m=1}^{M}|\mathbf{u}_m^\dagger|^2\right).$$

where $A(\epsilon,\tau) = \left(\epsilon + \frac{\epsilon}{\tau^{2\alpha}} + \left(1+\frac{\epsilon}{\tau^2}\right)^\alpha\right)^2$

Main Result – Simplified

Note if we fix α, set $\tau = \gamma^{1/\alpha}$, and for $\beta \geq 2$ let $\epsilon = \tau^{\beta} = \gamma^{\beta/\alpha}$, we can simplify the bound in Main Theorem to be:

$$\begin{aligned}
\mathcal{I}(\gamma, \alpha, \tau, \epsilon) &\leq \Xi K \gamma^2 \left(1 + K' \gamma^2 \left[\gamma^{\beta/\alpha} + \frac{\gamma^{\beta/\alpha}}{\gamma^2} + \left(1 + \frac{\gamma^{\beta/\alpha}}{\gamma^{1/\alpha}} \right)^{\alpha} \right]^2 \right) \\
&\leq \Xi' \left(\gamma^2 + \gamma^4 \left[\gamma^{\beta/\alpha} + \frac{\gamma^{\beta/\alpha}}{\gamma^2} + 1 \right]^2 \right) \\
&\leq \tilde{\Xi} \left(\gamma^2 + \gamma^{2\beta/\alpha} \right).
\end{aligned}$$

where $K, K', \Xi', \tilde{\Xi}$ are constants that are derived from Ξ, Ξ_1 from the Theorem and bounds for the other terms.

Numerical Example

Synthetic Data:
Disconnected $G_0(Z, W_0)$ and ground-truth $U^\dagger$ created from:

- 3 clusters of 100 nodes each
  - each cluster is different class, Erdos-Renyi graph ($p = 0.8$)
- 5 nodes from each class labeled

Then, weakly-connected G_ϵ obtained by ϵ perturbations of G_0.

From theory, see desired relationship in the scaling τ, γ, and ϵ. We set $\gamma = \tau^\alpha$ for bounds.

- 3 regimes:
  - $\epsilon = \tau^2 = \mathcal{O}(\tau^2)(\beta = 2)$
  - $\epsilon = \tau^3 = o(\tau^2)(\beta = 3)$
  - $\epsilon = 0(\approx \beta \to \infty)$

Numerical Example

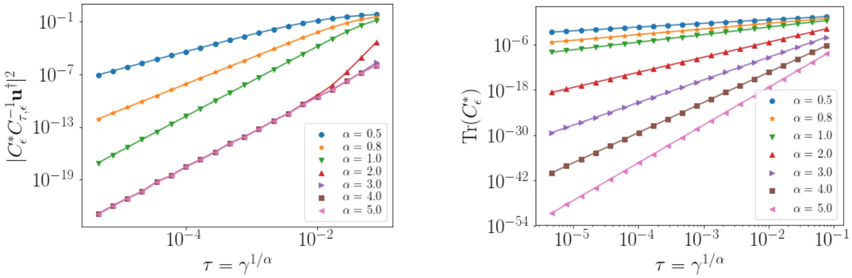
Calculation of $\mathcal{I}(\gamma, \tau, \epsilon, \alpha)$ found by 3 different terms derived in proof:

$$\mathcal{I}(\gamma, \alpha, \tau, \epsilon) = M\mathrm{Tr}(C_\epsilon^*) + \frac{M}{\gamma^2}\mathrm{Tr}(C_\epsilon^* B C_\epsilon^*) + \sum_{m=1}^{M} \left| \frac{1}{\gamma^2} C_\epsilon^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger \right|^2 .$$

where

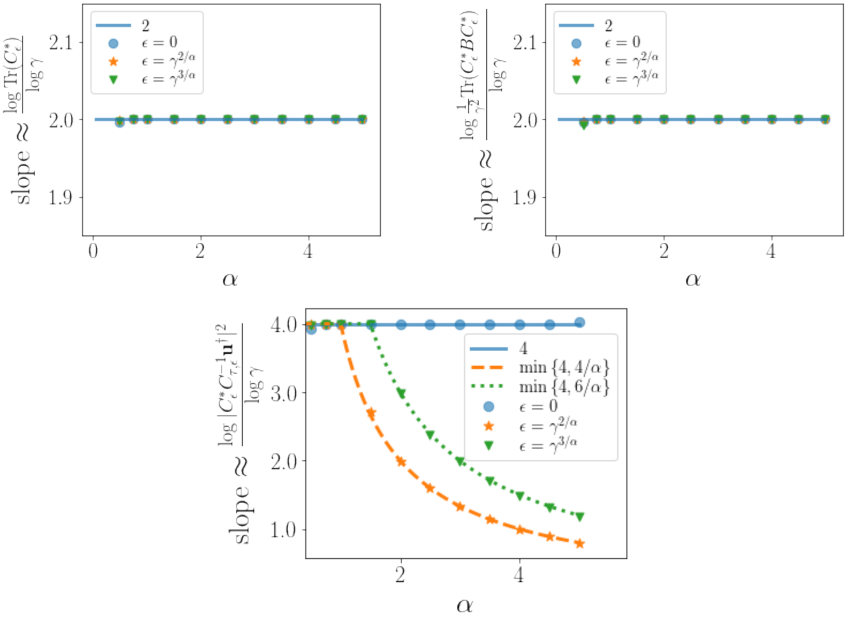
- C_ϵ^* : posterior measure's covariance matrix
- $B = H^T H \in \mathbb{R}^{N \times N}$: projection onto labeled nodes

Numerical Example – Convergence

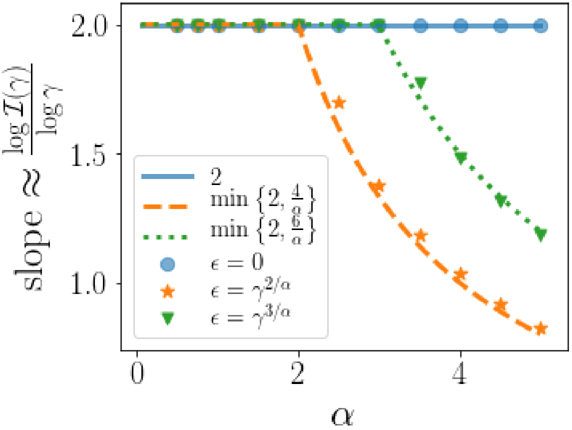


Bias and $\mathrm{Tr}(C_\epsilon^*)$ convergence plots for $\beta = 2, \ (\epsilon = \tau^2)$

Numerical Example – Convergence Rates of 3 terms



Bound seen for varying levels of $\beta \geq 2$:



Theoretical Bounds seem tight in testing!

Application Takeaway:

- Scaling needed in theory $\rightarrow$ need τ not to be too small compared to ϵ but also non-zero with relationship to γ

Future Directions:

- Apply to other likelihood choices
  - Regression not "natural" for underlying task of classification
  - Probit likelihood
- Try on real-world datasets – how to estimate ϵ?

References I

H. Li K. Miller A. L. Bertozzi, B. Hosseini and A. Stuart.
Posterior consistency of semi-supervised regression on graphs.
Preprint, 2019.

Sergios Agapiou, Stig Larsson, and Andrew M. Stuart.
Posterior contraction rates for the bayesian approach to linear ill-posed inverse problems.
Stochastic Processes and their Applications, 123, 03 2012.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani.
Manifold regularization: A geometric framework for learning from labeled and unlabeled examples.
Journal of machine learning research, 7(Nov):2399–2434, 2006.

Franca Hoffman, Bamdad Hosseini, Assad A. Oberai, and Andrew M. Stuart.
Spectral analysis of weighted laplacians arising in data clustering.
Preprint, 2019.

Franca Hoffman, Bamdad Hosseini, Zhi Ren, and Andrew M. Stuart.
Consistency of semi-supervised learning algorithms on graphs.
Preprint, 2019.

B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten.
Bayesian inverse problems with gaussian priors.
Ann. Statist., 39(5):2626–2657, 10 2011.

C. E. Rasmussen and C. K. I. Williams.
Gaussian Processes for Machine Learning.
The MIT press, Cambridge, 2006.

References II

A. M. Stuart.
Inverse problems: a Bayesian perspective.
Acta Numerica, 19:451–559, 2010.

Ambuj Tewari and Peter L Bartlett.
On the consistency of multiclass classification methods.
Journal of Machine Learning Research, 8(May):1007–1025, 2007.

Xiaojin Zhu.
Semi-supervised Learning with Graphs.
PhD thesis, Pittsburgh, PA, USA, 2005.
AAI3179046.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty.
Semi-supervised learning using Gaussian fields and harmonic functions.
In *ICML '03 Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003.