Discrete-To-Continuum Limits in Graph-Based Semi-Supervised Learning

Algorithms & Computationally Intensive Inference Seminar
University of Warwick

Matthew Thorpe

Joint Work with Andrea Bertozzi (UCLA), Jeff Calder (Minnesota), Brendan Cook (Minnesota), Matt Dunlop (Courant Institute), Tan Nguyen (National University of Singapore), Stanley Osher (UCLA), Dejan Slepčev (CMU), Thomas Strohmer (UC Davis), Andrew Stuart (Caltech), Bao Wang (Utah), Adrien Weihs (Manchester) and Hedi Xia (UCLA)

Department of Statistics
University of Warwick

26th January 2023

Semi-Supervised Learning

- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i \in \mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1, \ldots, n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1,\dots,n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
- **'Supervised classical' approach:** minimise $\|\nabla u_n\|_{L^2}^2$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

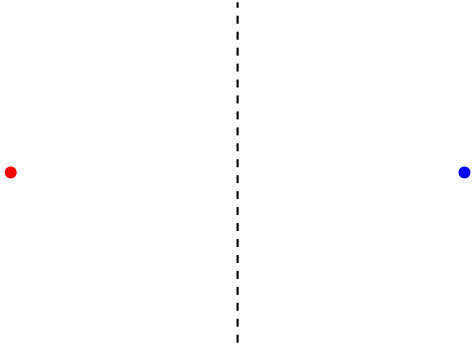
Semi-Supervised Learning

- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1,\dots,n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
- **'Supervised classical' approach:** minimise $\|\nabla u_n\|_{L^2}^2$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

Example by Jeff Calder.

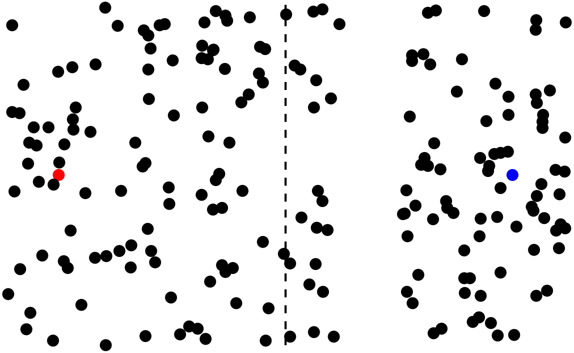
Semi-Supervised Learning

- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1,\dots,n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
- **'Supervised classical' approach:** minimise $\|\nabla u_n\|_{L^2}^2$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.



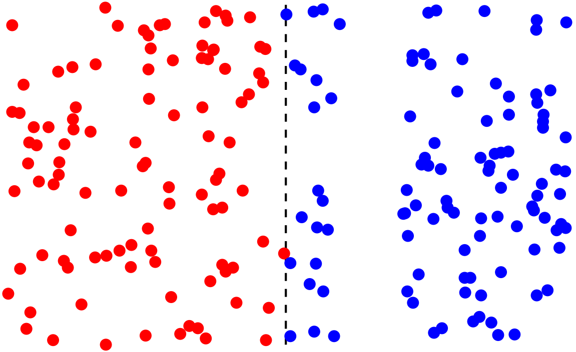
Example by Jeff Calder.

- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1,\dots,n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
- **'Supervised classical' approach:** minimise $\|\nabla u_n\|_{L^2}^2$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.



Example by Jeff Calder.

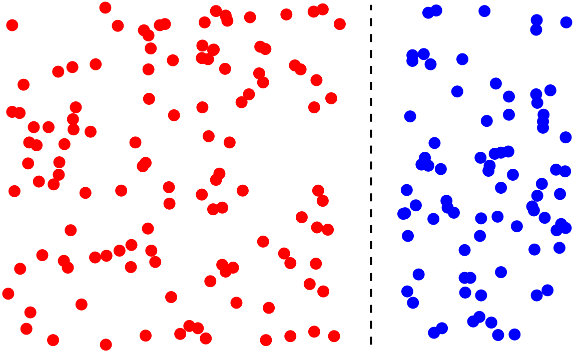
- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1,\ldots,n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
- **'Supervised classical' approach:** minimise $\|\nabla u_n\|_{L^2}^2$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.



Example by Jeff Calder.

Semi-Supervised Learning

- **Problem:** Given data $\Omega_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n} \subset \mathbb{R}$, where $\mathcal{I}_n \subseteq \{1,\dots,n\}$, find the 'best' $u_n : \Omega_n \to \mathbb{R}$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
- **'Supervised classical' approach:** minimise $\|\nabla u_n\|_{L^2}^2$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.



Example by Jeff Calder.

Semi-Supervised Learning

1. **Aim:** Given feature vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n}$ find labels of the unlabelled feature vectors $\{x_i\}_{i\notin\mathcal{I}_n}$.

1. **Aim:** Given feature vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n}$ find labels of the unlabelled feature vectors $\{x_i\}_{i\notin\mathcal{I}_n}$.
2. **Graph:** $G_n = (\Omega_n, W_n)$ where

1. **Aim:** Given feature vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n}$ find labels of the unlabelled feature vectors $\{x_i\}_{i\notin\mathcal{I}_n}$.
2. **Graph:** $G_n = (\Omega_n, W_n)$ where
    1. $\Omega_n = \{x_i\}_{i=1}^n$ are the vertices/nodes,

1. **Aim:** Given feature vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n}$ find labels of the unlabelled feature vectors $\{x_i\}_{i\notin\mathcal{I}_n}$.
2. **Graph:** $G_n = (\Omega_n, W_n)$ where
    1. $\Omega_n = \{x_i\}_{i=1}^n$ are the vertices/nodes,
    2. $W_n = (w_{ij})_{i,j=1}^n$ are nonnegative and symmetric edge weights with $w_{ij} \gg 1$ if x_i and x_j are similar and $w_{ij} \approx 0$ if dissimilar.

Semi-Supervised Learning

1. **Aim:** Given feature vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n}$ find labels of the unlabelled feature vectors $\{x_i\}_{i\notin\mathcal{I}_n}$.
2. **Graph:** $G_n = (\Omega_n, W_n)$ where
    1. $\Omega_n = \{x_i\}_{i=1}^n$ are the vertices/nodes,
    2. $W_n = (w_{ij})_{i,j=1}^n$ are nonnegative and symmetric edge weights with $w_{ij} \gg 1$ if x_i and x_j are similar and $w_{ij} \approx 0$ if dissimilar.
3. **Labels:** $\ell_j \in \{\mathrm{e}_1, \dots, \mathrm{e}_k\} \in \mathbb{R}^k$ or $\ell_j \in \mathbb{R}$ depending on the setting.

Semi-Supervised Learning

1. **Aim:** Given feature vectors $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{\ell_i\}_{i\in\mathcal{I}_n}$ find labels of the unlabelled feature vectors $\{x_i\}_{i\notin\mathcal{I}_n}$.
2. **Graph:** $G_n = (\Omega_n, W_n)$ where
   1. $\Omega_n = \{x_i\}_{i=1}^n$ are the vertices/nodes,
   2. $W_n = (w_{ij})_{i,j=1}^n$ are nonnegative and symmetric edge weights with $w_{ij} \gg 1$ if x_i and x_j are similar and $w_{ij} \approx 0$ if dissimilar.
3. **Labels:** $\ell_j \in \{\mathrm{e}_1, \ldots, \mathrm{e}_k\} \in \mathbb{R}^k$ or $\ell_j \in \mathbb{R}$ depending on the setting.
4. **Assumption:** Similar feature vectors should have similar labels.

1. **Laplacian Regularisation:** Zhu, Ghahramani and Lafferty (2003) or Zhou and Schölkopf (2005) define u_n^* as the minimiser of
$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij}|u_n(x_i) - u_n(x_j)|^p$$
over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

Laplace Learning

1. **Laplacian Regularisation:** Zhu, Ghahramani and Lafferty (2003) or Zhou and Schölkopf (2005) define u_n^* as the minimiser of

$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^p$$

over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

2. **Laplacian Classification:** We define

$$\ell_{u_n^*}(x_i) = \operatorname*{argmax}_{j \in \{1,\dots,k\}} u_{n,j}^*(x_i).$$

Laplace Learning

1. **Laplacian Regularisation:** Zhu, Ghahramani and Lafferty (2003) or Zhou and Schölkopf (2005) define u_n^* as the minimiser of
$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij}|u_n(x_i) - u_n(x_j)|^p$$
over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.
2. **Laplacian Classification:** We define
$$\ell_{u_n^*}(x_i) = \operatorname*{argmax}_{j \in \{1,\dots,k\}} u_{n,j}^*(x_i).$$
3. If $p = 2$ it follows that u_n^* satisfies the following Laplace equation
$$\begin{aligned} L_n u_n^*(x_i) &= 0 && \text{if } i \notin \mathcal{I}_n \\ u_n^*(x_i) &= \ell_i && \text{if } i \in \mathcal{I}_n \end{aligned}$$
where $L_n u(x_i) = \sum_{j=1}^n w_{ij}(u(x_i) - u(x_j))$ is the graph Laplacian.

Contents

Contents

The TL^p Topology

1. We want to compare $u_n : \Omega_n \to \mathbb{R}$ to $u : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^d$.

The TL^p Topology

1. We want to compare $u_n : \Omega_n \to \mathbb{R}$ to $u : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^d$.
2. Key idea: we extend each $u_n : \Omega_n \to \mathbb{R}$ to a piecewise constant function $\tilde{u}_n : \Omega \to \mathbb{R}$ and compute $\|\tilde{u}_n - u\|_{\mathrm{L}^p}$.

The TL^p Topology

1. We want to compare $u_n : \Omega_n \to \mathbb{R}$ to $u : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^d$.
2. Key idea: we extend each $u_n : \Omega_n \to \mathbb{R}$ to a piecewise constant function $\tilde{u}_n : \Omega \to \mathbb{R}$ and compute $\|\tilde{u}_n - u\|_{\mathrm{L}^p}$.
3. Let $\tilde{u}_n(x) = u_n(T_n(x))$ for some function $T_n : \Omega \to \Omega_n$.

The TL^p Topology

1. We want to compare $u_n : \Omega_n \to \mathbb{R}$ to $u : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}^d$.
2. Key idea: we extend each $u_n : \Omega_n \to \mathbb{R}$ to a piecewise constant function $\tilde{u}_n : \Omega \to \mathbb{R}$ and compute $\|\tilde{u}_n - u\|_{\mathrm{L}^p}$.
3. Let $\tilde{u}_n(x) = u_n(T_n(x))$ for some function $T_n : \Omega \to \Omega_n$.
4. We will choose T_n to be an optimal transport map.

The TL^p Metric

- As introduced by García Trillos and Slepčev (2016), let

$$\mathrm{TL}^p := \{(u, \mu) \,:\, u \in \mathrm{L}^p(\mu), \mu \in \mathcal{P}(\Omega)\}.$$

The TL^p Metric

- As introduced by García Trillos and Slepčev (2016), let

$$\mathrm{TL}^p := \{(u,\mu)\,:\, u \in \mathrm{L}^p(\mu), \mu \in \mathcal{P}(\Omega)\}.$$

- The TL^p metric is defined by $d_{\mathrm{TL}^p} : \mathrm{TL}^p \times \mathrm{TL}^p \to [0,\infty)$,

$$d^p_{\mathrm{TL}^p}((u,\mu),(v,\nu)) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} |x-y|^p + |u(x)-v(y)|^p \,\mathrm{d}\pi(x,y)$$

The TL^p Metric

- As introduced by García Trillos and Slepčev (2016), let
$$\mathrm{TL}^p := \{(u,\mu) \,:\, u \in \mathrm{L}^p(\mu), \mu \in \mathcal{P}(\Omega)\}\,.$$

- The TL^p metric is defined by $d_{\mathrm{TL}^p} : \mathrm{TL}^p \times \mathrm{TL}^p \to [0,\infty)$,
$$d^p_{\mathrm{TL}^p}((u,\mu),(v,\nu)) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} |x-y|^p + |u(x)-v(y)|^p \,\mathrm{d}\pi(x,y)$$

- Theoretically it is convenient to write:
$$d_{\mathrm{TL}^p}((u,\mu),(v,\nu)) = d_{\mathrm{W}^p}(\tilde{\mu},\tilde{\nu}) = \inf_{\tilde{\pi}\in\Pi(\tilde{\mu},\tilde{\nu})} \sqrt[p]{\int_{(\Omega\times\mathbb{R})\times(\Omega\times\mathbb{R})} |x-y|^p \,\mathrm{d}\tilde{\pi}(x,y)}.$$
where $\tilde{\mu} = (\mathrm{Id}\times u)_{\#}\mu$ and $\tilde{\nu} = (\mathrm{Id}\times v)_{\#}\nu$.

The TL^p Metric

- As introduced by García Trillos and Slepčev (2016), let

$$\mathrm{TL}^p := \left\{ (u,\mu) \,:\, u \in \mathrm{L}^p(\mu), \mu \in \mathcal{P}(\Omega) \right\}.$$

- The TL^p metric is defined by $d_{\mathrm{TL}^p} : \mathrm{TL}^p \times \mathrm{TL}^p \to [0,\infty)$,

$$d^p_{\mathrm{TL}^p}((u,\mu),(v,\nu)) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} |x-y|^p + |u(x)-v(y)|^p \,\mathrm{d}\pi(x,y)$$

- Theoretically it is convenient to write:

$$d_{\mathrm{TL}^p}((u,\mu),(v,\nu)) = d_{\mathrm{W}^p}(\tilde{\mu},\tilde{\nu}) = \inf_{\tilde{\pi}\in\Pi(\tilde{\mu},\tilde{\nu})} \sqrt[p]{\int_{(\Omega\times\mathbb{R})\times(\Omega\times\mathbb{R})} |x-y|^p \,\mathrm{d}\tilde{\pi}(x,y)}.$$

  where $\tilde{\mu} = (\mathrm{Id}\times u)_{\#}\mu$ and $\tilde{\nu} = (\mathrm{Id}\times v)_{\#}\nu$.

- Numerically it is convenient to write:

$$d_{\mathrm{TL}^p}((u,\mu),(v,\nu)) = \inf_{\pi\in\Pi(\mu,\nu)} \sqrt[p]{\int_{\Omega\times\Omega} c(x,y;u,v)\,\mathrm{d}\pi(x,y)}.$$

  where $c(x,y;u,v) = |x-y|^p + |u(x)-v(y)|^p$.

Aside: A TL^p Approach to Histogram Specification

(a) Exemplar images.

(b) Original image to be shaded.

(c) The TL^p colour transfer solution.

Figure: More details and other applications in T., Park, Kolouri, Rohde and Slepčev (2017).

TL^p In Practice

Theorem (García Trillos and Slepčev (2016))

If μ is absolutely continuous, then $(u_n, \mu_n) \to (u, \mu)$ in TL^p if and only if $\mu_n \rightharpoonup^* \mu$ and there exists a sequence of maps $T_n : \Omega \to \Omega$ such that $(T_n)_{\#}\mu = \mu_n$, $T_n \to \mathrm{Id}$ in $\mathrm{L}^p(\mu)$ and

$$\|u_n \circ T_n - u\|_{\mathrm{L}^p(\mu)} \to 0.$$

TL^p In Practice

Theorem (García Trillos and Slepčev (2016))

If μ is absolutely continuous, then $(u_n, \mu_n) \to (u, \mu)$ in TL^p if and only if $\mu_n \rightharpoonup^* \mu$ and there exists a sequence of maps $T_n : \Omega \to \Omega$ such that $(T_n)_{\#}\mu = \mu_n$, $T_n \to \mathrm{Id}$ in $\mathrm{L}^p(\mu)$ and

$$\|u_n \circ T_n - u\|_{\mathrm{L}^p(\mu)} \to 0.$$

Theorem (García Trillos and Slepčev (2015))

Assume $x_i \overset{\text{iid}}{\sim} \mu$ and $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. With probability one, there exists $T_n : \Omega \to \Omega_n$ such that $(T_n)_{\#}\mu = \mu_n$ and

$$\|T_n - \mathrm{Id}\|_{\mathrm{L}^\infty} \lesssim \begin{cases} \frac{(\log n)^{\frac{3}{4}}}{\sqrt{n}} & \text{if } d = 2 \\ \left(\frac{\log n}{n}\right)^{\frac{1}{d}} & \text{if } d \geq 3. \end{cases}$$

TL^p In Practice

Theorem (García Trillos and Slepčev (2016))

If μ is absolutely continuous, then $(u_n, \mu_n) \to (u, \mu)$ in TL^p if and only if $\mu_n \rightharpoonup^* \mu$ and there exists a sequence of maps $T_n : \Omega \to \Omega$ such that $(T_n)_\# \mu = \mu_n$, $T_n \to \mathrm{Id}$ in $\mathrm{L}^p(\mu)$ and

$$\|u_n \circ T_n - u\|_{\mathrm{L}^p(\mu)} \to 0.$$

Theorem (García Trillos and Slepčev (2015))

Assume $x_i \overset{\mathrm{iid}}{\sim} \mu$ and $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. With probability one, there exists $T_n : \Omega \to \Omega_n$ such that $(T_n)_\# \mu = \mu_n$ and

$$\|T_n - \mathrm{Id}\|_{\mathrm{L}^\infty} \lesssim \begin{cases} \frac{(\log n)^{\frac{3}{4}}}{\sqrt{n}} & \text{if } d = 2 \\ \left(\frac{\log n}{n}\right)^{\frac{1}{d}} & \text{if } d \geq 3. \end{cases}$$

Remark: by (for example) Penrose (2003) the connectivity radius of the geometric random graph scales as $\left(\frac{\log n}{n}\right)^{\frac{1}{d}}$ for all $d \in \mathbb{N}$.

Contents

Asymptotic Setting

- **Aim:** minimise

$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^p$$

over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

Asymptotic Setting

- **Aim:** minimise

$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^p$$

over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

- Assume

$$w_{ij} = \eta_\varepsilon(|x_i - x_j|) = \frac{1}{\varepsilon^d} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right).$$

Asymptotic Setting

- **Aim:** minimise

$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^p$$

over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

- Assume

$$w_{ij} = \eta_\varepsilon(|x_i - x_j|) = \frac{1}{\varepsilon^d} \eta \left(\frac{|x_i - x_j|}{\varepsilon} \right).$$

- **Formal Definition:** The Laplacian regression problem is **asymptotically well-posed** if constrained minimisers of $\mathcal{E}_n^{(p)}$ converge to constrained minimisers of some $\mathcal{E}_\infty^{(p)}$ (to be defined).

Asymptotic Setting

- **Aim:** minimise

$$\mathcal{E}_n^{(p)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^p$$

over all $u_n : \Omega_n \to \mathbb{R}^k$ such that $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$.

- Assume

$$w_{ij} = \eta_\varepsilon(|x_i - x_j|) = \frac{1}{\varepsilon^d} \eta \left(\frac{|x_i - x_j|}{\varepsilon} \right).$$

- **Formal Definition:** The Laplacian regression problem is **asymptotically well-posed** if constrained minimisers of $\mathcal{E}_n^{(p)}$ converge to constrained minimisers of some $\mathcal{E}_\infty^{(p)}$ (to be defined).
- **Formal Definition:** The Laplacian regression problem is **asymptotically ill-posed** if constrained minimisers of $\mathcal{E}_n^{(p)}$ converge to constants.

Formal Derivation of Limit

The following (formal) calculation gives intuition as to what we should expect.

Formal Derivation of Limit

The following (formal) calculation gives intuition as to what we should expect.

$$\frac{1}{n^2\varepsilon^p}\mathcal{E}_n^{(p)}(u) = \frac{1}{n^2\varepsilon^{p+d}}\sum_{i,j=1}^n \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)|u(x_i) - u(x_j)|^p$$

The following (formal) calculation gives intuition as to what we should expect.

$$
\begin{aligned}
\frac{1}{n^2\varepsilon^p}\mathcal{E}_n^{(p)}(u) &= \frac{1}{n^2\varepsilon^{p+d}}\sum_{i,j=1}^n \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)|u(x_i) - u(x_j)|^p \\
&\approx \frac{1}{\varepsilon^{p+d}}\int\int \eta\left(\frac{|x-y|}{\varepsilon}\right)|u(x) - u(y)|^p\,\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y
\end{aligned}
$$

Formal Derivation of Limit

The following (formal) calculation gives intuition as to what we should expect.

$$
\begin{aligned}
\frac{1}{n^2\varepsilon^p}\mathcal{E}_n^{(p)}(u) &= \frac{1}{n^2\varepsilon^{p+d}}\sum_{i,j=1}^n \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)|u(x_i) - u(x_j)|^p \\
&\approx \frac{1}{\varepsilon^{p+d}}\int\int \eta\left(\frac{|x-y|}{\varepsilon}\right)|u(x) - u(y)|^p \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{\varepsilon^p}\int\int \eta(|z|)|u(y+\varepsilon z) - u(y)|^p\rho(y+\varepsilon z)\rho(y)\,\mathrm{d}y\,\mathrm{d}z
\end{aligned}
$$

Formal Derivation of Limit

The following (formal) calculation gives intuition as to what we should expect.

$$
\begin{aligned}
\frac{1}{n^2\varepsilon^p}\mathcal{E}_n^{(p)}(u) &= \frac{1}{n^2\varepsilon^{p+d}}\sum_{i,j=1}^{n}\eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)|u(x_i) - u(x_j)|^p \\
&\approx \frac{1}{\varepsilon^{p+d}}\int\int \eta\left(\frac{|x-y|}{\varepsilon}\right)|u(x) - u(y)|^p \rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{\varepsilon^p}\int\int \eta(|z|)|u(y+\varepsilon z) - u(y)|^p\rho(y+\varepsilon z)\rho(y)\,\mathrm{d}y\,\mathrm{d}z \\
&\approx \int\int \eta(|z|)|\nabla u(y)\cdot z|^p\rho^2(y)\,\mathrm{d}y\,\mathrm{d}z
\end{aligned}
$$

Formal Derivation of Limit

The following (formal) calculation gives intuition as to what we should expect.

$$
\begin{aligned}
\frac{1}{n^2\varepsilon^p}\mathcal{E}_n^{(p)}(u) &= \frac{1}{n^2\varepsilon^{p+d}}\sum_{i,j=1}^n \eta\left(\frac{|x_i-x_j|}{\varepsilon}\right)|u(x_i)-u(x_j)|^p \\
&\approx \frac{1}{\varepsilon^{p+d}}\int\int \eta\left(\frac{|x-y|}{\varepsilon}\right)|u(x)-u(y)|^p\,\rho(x)\rho(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{\varepsilon^p}\int\int \eta(|z|)|u(y+\varepsilon z)-u(y)|^p\rho(y+\varepsilon z)\rho(y)\,\mathrm{d}y\,\mathrm{d}z \\
&\approx \int\int \eta(|z|)|\nabla u(y)\cdot z|^p\rho^2(y)\,\mathrm{d}y\,\mathrm{d}z \\
&= \sigma_\eta\int |\nabla u(y)|^p\rho^2(y)\,\mathrm{d}y =: \mathcal{E}_\infty^{(p)}(u)
\end{aligned}
$$

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n \in \mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No!

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n \in \mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n \in \mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.
- Consider the function $u_n(x_1) = 1$ and $u_n(x_i) = 0$ for all $i \geq 2$.

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.
- Consider the function $u_n(x_1) = 1$ and $u_n(x_i) = 0$ for all $i \geq 2$.
- $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) = \frac{2}{\varepsilon_n^{p+d}n^2}\sum_{j=2}^n \eta\left(\frac{|x_1-x_j|}{\varepsilon_n}\right)$

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.
- Consider the function $u_n(x_1) = 1$ and $u_n(x_i) = 0$ for all $i \geq 2$.
- $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) = \frac{2}{\varepsilon_n^{p+d}n^2}\sum_{j=2}^n \eta\left(\frac{|x_1-x_j|}{\varepsilon_n}\right)$
  $= \left(\frac{2}{\varepsilon_n^p n}\right) \times \left(\frac{1}{n\varepsilon_n^d}\#\{\Omega_n \cap B(x_1,\varepsilon_n)\}\right).$

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.
- Consider the function $u_n(x_1) = 1$ and $u_n(x_i) = 0$ for all $i \geq 2$.
- $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) = \frac{2}{\varepsilon_n^{p+d}n^2}\sum_{j=2}^n \eta\left(\frac{|x_1 - x_j|}{\varepsilon_n}\right)$
  $= \left(\frac{2}{\varepsilon_n^p n}\right) \times \left(\frac{1}{n\varepsilon_n^d}\#\{\Omega_n \cap B(x_1,\varepsilon_n)\}\right)$.
- If $\varepsilon_n^p n \to \infty$ then $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) \to 0$ and the spike pays no cost in the limit!

Limiting Constraints - Intuition

We intuitively see that $p > d$ is necessary if the constrain set is finite, i.e. $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$, is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.
- Consider the function $u_n(x_1) = 1$ and $u_n(x_i) = 0$ for all $i \geq 2$.
- $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) = \frac{2}{\varepsilon_n^{p+d}n^2}\sum_{j=2}^n \eta\left(\frac{|x_1 - x_j|}{\varepsilon_n}\right)$
  $= \left(\frac{2}{\varepsilon_n^p n}\right) \times \left(\frac{1}{n\varepsilon_n^d}\#\{\Omega_n \cap B(x_1, \varepsilon_n)\}\right)$.
- If $\varepsilon_n^p n \to \infty$ then $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) \to 0$ and the spike pays no cost in the limit!

This example turns out to be sharp: $\varepsilon_n^p n \to \infty$ implies ill-posedness and $\varepsilon_n^p n \to 0$ implies well-posedness.

Continuum Limit of p-Laplace Learning

Theorem (Slepčev and T., 17)

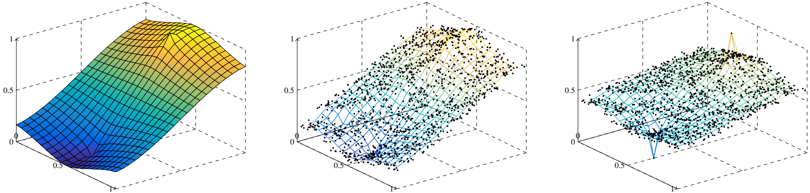
Let $p > 1$. Let u_n^* be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the $u_n^*(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$ where $\max_{n\in\mathbb{N}} |\mathcal{I}_n| < +\infty$. Then, almost surely, the sequence (u_n^*, μ_n) is precompact in TL^p. The TL^p limit of any convergent subsequence, $(u_{n_m}^*, \mu_{n_m})$, is of the form (u, μ) where $u \in W^{1,p}(\Omega)$. Furthermore,

- (i) *if $n\varepsilon_n^p \to 0$ as $n \to \infty$ then u is continuous and*
  - (a) *the whole sequence u_n^* converges to u both in TL^p and locally uniformly, meaning that for any Ω' with $\overline{\Omega'} \subset \Omega$*

$$\lim_{n\to\infty} \max_{\{k\in\{1,\dots,n\} \,:\, x_k\in\Omega'\}} |u(x_k) - u_n^*(x_k)| = 0,$$

  - (b) *u is a minimizer of $\mathcal{E}_\infty^{(p)}$ with constraints;*
- (ii) *if $n\varepsilon_n^p \to \infty$ as $n \to \infty$ then u is constant.*

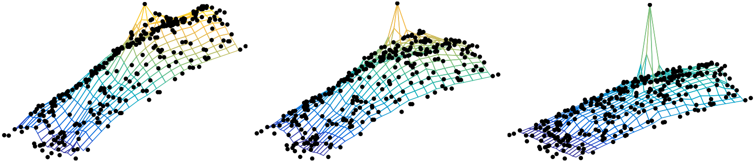
Numerical Comparisons



(a) $p = 4$ continuum limit minimiser.

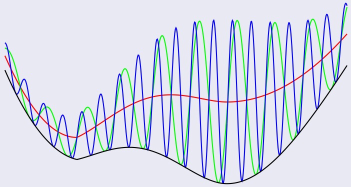
(b) $p = 4$ minimiser ($\varepsilon = 0.06$, $n = 1280$).

(c) $p = 2$ minimiser ($\varepsilon = 0.06$, $n = 1280$).



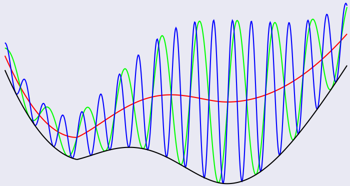
(a) $\varepsilon = 0.05$. (b) $\varepsilon = 0.1$. (c) $\varepsilon = 0.2$.

Variational Convergence



Green - $\mathcal{E}_n$, Blue - $\mathcal{E}_m$ for $m > n$, Red - weak limit, Black - Γ-limit.

Variational Convergence



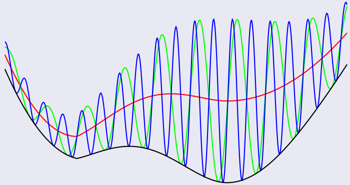
Green - $\mathcal{E}_n$, Blue - $\mathcal{E}_m$ for $m > n$, Red - weak limit, Black - Γ-limit.

We say $\mathcal{E}_\infty = \Gamma\text{-}\lim_n \mathcal{E}_n$, if for all u we have

(i) $\forall\, u_n \to u$,
$\mathcal{E}_\infty(u) \leq \liminf_{n\to\infty} \mathcal{E}_n(u_n)$;

(ii) $\exists\, u_n \to u$,
$\mathcal{E}_\infty(u) \geq \limsup_{n\to\infty} \mathcal{E}_n(u_n)$.

Variational Convergence



Green - $\mathcal{E}_n$, Blue - $\mathcal{E}_m$ for $m > n$, Red - weak limit, Black - Γ-limit.

We say $\mathcal{E}_\infty = \Gamma\text{-}\lim_n \mathcal{E}_n$, if for all u we have

(i) $\forall\, u_n \to u$,
$\mathcal{E}_\infty(u) \leq \liminf_{n\to\infty} \mathcal{E}_n(u_n)$;

(ii) $\exists\, u_n \to u$,
$\mathcal{E}_\infty(u) \geq \limsup_{n\to\infty} \mathcal{E}_n(u_n)$.

Theorem

Let u_n be a sequence of almost minimizers of $\mathcal{E}_n$. If u_n are precompact and $\mathcal{E}_\infty = \Gamma\text{-}\lim_n \mathcal{E}_n$ where $\mathcal{E}_\infty$ is not identically $+\infty$ then

$$\min \mathcal{E}_\infty = \lim_{n\to\infty} \inf \mathcal{E}_n.$$

Furthermore any cluster point of $\{u_n\}_{n=1}^\infty$ minimizes $\mathcal{E}_\infty$.

1. **Step 1:** We show $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{u}_n)$ where $\tilde{u}_n = u_n \circ T_n$ and J is a mollifier.

Intuition on the Proof

1. **Step 1:** We show $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{u}_n)$ where $\tilde{u}_n = u_n \circ T_n$ and J is a mollifier.
2. **Step 2:** We show $\mathrm{osc}_{\varepsilon_n}^{(n)}(u_n) \leq C\sqrt[p]{n\varepsilon_n^p\left(\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n)\right)}$ where

$$\mathrm{osc}_{\varepsilon}^{(n)}(u_n)(x_k) = \max_{z\in B(x_k,\varepsilon)\cap\Omega_n} u_n(z) - \min_{z\in B(x_k,\varepsilon)\cap\Omega_n} u_n(z).$$

Intuition on the Proof

1. **Step 1:** We show $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{u}_n)$ where $\tilde{u}_n = u_n \circ T_n$ and J is a mollifier.
2. **Step 2:** We show $\operatorname{osc}_{\varepsilon_n}^{(n)}(u_n) \leq C\sqrt[p]{n\varepsilon_n^p\left(\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n)\right)}$ where

$$\operatorname{osc}_{\varepsilon}^{(n)}(u_n)(x_k) = \max_{z\in B(x_k,\varepsilon)\cap\Omega_n} u_n(z) - \min_{z\in B(x_k,\varepsilon)\cap\Omega_n} u_n(z).$$

3. **Step 3:** Sobolev embedding of $J_{\varepsilon_n} * \tilde{u}_n$ plus the control over oscillations is enough to infer uniform convergence:

$$\lim_{n\to\infty} \max_{\{k\in\{1,\dots,n\}\,:\,x_k\in\Omega'\}} |u(x_k) - u_n(x_k)| = 0.$$

Intuition on the Proof

1. **Step 1:** We show $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{u}_n)$ where $\tilde{u}_n = u_n \circ T_n$ and J is a mollifier.
2. **Step 2:** We show $\mathrm{osc}_{\varepsilon_n}^{(n)}(u_n) \leq C \sqrt[p]{n\varepsilon_n^p \left(\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}(u_n)\right)}$ where
$$\mathrm{osc}_\varepsilon^{(n)}(u_n)(x_k) = \max_{z\in B(x_k,\varepsilon)\cap\Omega_n} u_n(z) - \min_{z\in B(x_k,\varepsilon)\cap\Omega_n} u_n(z).$$
3. **Step 3:** Sobolev embedding of $J_{\varepsilon_n} * \tilde{u}_n$ plus the control over oscillations is enough to infer uniform convergence:
$$\lim_{n\to\infty} \max_{\{k\in\{1,\dots,n\}\,:\,x_k\in\Omega'\}} |u(x_k) - u_n(x_k)| = 0.$$
4. **Step 4:** Γ-convergence of $\frac{1}{n^2\varepsilon_n^p}\mathcal{E}_n^{(p)}$ to $\mathcal{E}_\infty^{(p)}$ plus a TL^p compactness result is now enough to get convergence of constrained minimizers.

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_{\infty}^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.
- **Questions:** What happens in between?

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.
- **Questions:** What happens in between?
- **Model:** Recall $\{(x_i, \ell_i)\}_{i\in\mathcal{I}_n}$ is the labelled data.

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.
- **Questions:** What happens in between?
- **Model:** Recall $\{(x_i, \ell_i)\}_{i\in\mathcal{I}_n}$ is the labelled data.
    1. Assume $\mathbb{P}(i \in \mathcal{I}_n) = \beta_n \mathbb{1}_{\tilde{\Omega}}$, $\tilde{\Omega} \subset \Omega$ is open and bounded.

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.
- **Questions:** What happens in between?
- **Model:** Recall $\{(x_i, \ell_i)\}_{i \in \mathcal{I}_n}$ is the labelled data.
  1. Assume $\mathbb{P}(i \in \mathcal{I}_n) = \beta_n \mathbb{1}_{\tilde{\Omega}}$, $\tilde{\Omega} \subset \Omega$ is open and bounded.
  2. If $i \in \mathcal{I}_n$ we set $\ell_i = g^\dagger(x_i)$.

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.
- **Questions:** What happens in between?
- **Model:** Recall $\{(x_i, \ell_i)\}_{i\in\mathcal{I}_n}$ is the labelled data.
  1. Assume $\mathbb{P}(i \in \mathcal{I}_n) = \beta_n \mathbb{1}_{\tilde{\Omega}}$, $\tilde{\Omega} \subset \Omega$ is open and bounded.
  2. If $i \in \mathcal{I}_n$ we set $\ell_i = g^\dagger(x_i)$.
- **Well-posed case:** Minimisers of $\mathcal{E}_n^{(p)}$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$ converge to minimisers of $\mathcal{E}_\infty^{(p)}$ subject to $u(x) = g^\dagger(x)$ for all $x \in \tilde{\Omega}$.

Minimal Number of Labels

- Sobolev regularity implies that we can't minimise $\mathcal{E}_\infty^{(p)}$ with finite constraints, but the problem is well-posed if labels are imposed on an open subset.
- We expect asymptotic ill-posedness if $|\mathcal{I}_n| \sim 1$ and asymptotic well-posedness if $|\mathcal{I}_n| \sim n$.
- **Questions:** What happens in between?
- **Model:** Recall $\{(x_i, \ell_i)\}_{i \in \mathcal{I}_n}$ is the labelled data.
  1. Assume $\mathbb{P}(i \in \mathcal{I}_n) = \beta_n \mathbb{1}_{\tilde{\Omega}}$, $\tilde{\Omega} \subset \Omega$ is open and bounded.
  2. If $i \in \mathcal{I}_n$ we set $\ell_i = g^\dagger(x_i)$.
- **Well-posed case:** Minimisers of $\mathcal{E}_n^{(p)}$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$ converge to minimisers of $\mathcal{E}_\infty^{(p)}$ subject to $u(x) = g^\dagger(x)$ for all $x \in \tilde{\Omega}$.
- **Ill-posed case:** Minimisers of $\mathcal{E}_n^{(p)}$ subject to $u_n(x_i) = \ell_i$ for all $i \in \mathcal{I}_n$ converge to constants.

Random Walks on Graphs

- Let $G_n = (\Omega_n, W)$ be the graph with edge weights $W = (w_{ij})$.

Random Walks on Graphs

- Let $G_n = (\Omega_n, W)$ be the graph with edge weights $W = (w_{ij})$.
- Let B_t^x be the random walk on Ω_n starting from $B_0^x = x \in \Omega_n$ and transitioning with probability

$$\mathbb{P}(B_{t+1}^x = x_j \mid B_t^x = x_i) = \frac{w_{ij}}{d_i}$$

where $d_i = \sum_{j=1}^n w_{ij}$.

- Let $G_n = (\Omega_n, W)$ be the graph with edge weights $W = (w_{ij})$.
- Let B_t^x be the random walk on Ω_n starting from $B_0^x = x \in \Omega_n$ and transitioning with probability

$$\mathbb{P}(B_{t+1}^x = x_j \mid B_t^x = x_i) = \frac{w_{ij}}{d_i}$$

  where $d_i = \sum_{j=1}^n w_{ij}$.
- Define the stopping time

$$S(x) = \min\left\{t \in \mathbb{N} \,:\, B_t^x \in \{x_i\}_{i \in \mathcal{I}_n}\right\}.$$

Random Walks on Graphs

- Let $G_n = (\Omega_n, W)$ be the graph with edge weights $W = (w_{ij})$.
- Let B_t^x be the random walk on Ω_n starting from $B_0^x = x \in \Omega_n$ and transitioning with probability

$$\mathbb{P}(B_{t+1}^x = x_j \mid B_t^x = x_i) = \frac{w_{ij}}{d_i}$$

  where $d_i = \sum_{j=1}^n w_{ij}$.
- Define the stopping time

$$S(x) = \min \left\{ t \in \mathbb{N} \, : \, B_t^x \in \{x_i\}_{i \in \mathcal{I}_n} \right\}.$$

Proposition

Define $u_n^*(x) = \mathbb{E}[g^\dagger(B_{S(x)}^x)]$. Then u_n^* minimises $\mathcal{E}_n^{(2)}$ subject to the constraints.

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\dots,k} |B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\ldots,k}|B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$
2. **Step 2:** Within the labelled domain we have a probability β of stopping and so
$$\mathbb{P}(S(x) > k) \leq (1-\beta)^k \leq e^{-ck\beta} \qquad \forall x \in \tilde{\Omega}.$$

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\dots,k} |B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$
2. **Step 2:** Within the labelled domain we have a probability β of stopping and so
$$\mathbb{P}(S(x) > k) \leq (1-\beta)^k \leq e^{-ck\beta} \qquad \forall x \in \tilde{\Omega}.$$
3. **Step 3:** Combining the previous results, for all $x \in \tilde{\Omega}$,

$|u_n^*(x) - g^\dagger(x)|$

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\ldots,k}|B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$
2. **Step 2:** Within the labelled domain we have a probability β of stopping and so
$$\mathbb{P}(S(x) > k) \leq (1-\beta)^k \leq e^{-ck\beta} \qquad \forall x \in \tilde{\Omega}.$$
3. **Step 3:** Combining the previous results, for all $x \in \tilde{\Omega}$,

$$|u_n^*(x) - g^\dagger(x)| \leq \mathbb{E}\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right|$$

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\dots,k} |B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$
2. **Step 2:** Within the labelled domain we have a probability β of stopping and so
$$\mathbb{P}(S(x) > k) \leq (1-\beta)^k \leq e^{-ck\beta} \qquad \forall x \in \tilde{\Omega}.$$
3. **Step 3:** Combining the previous results, for all $x \in \tilde{\Omega}$,

$$\begin{aligned} |u_n^*(x) - g^\dagger(x)| &\leq \mathbb{E}\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \\ &= \mathbb{E}\left[\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \,\middle|\, S(x) \leq k\right]\mathbb{P}(S(x) \leq k) \\ &\quad + \mathbb{E}\left[\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \,\middle|\, S(x) > k\right]\mathbb{P}(S(x) > k) \end{aligned}$$

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\dots,k} |B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$
2. **Step 2:** Within the labelled domain we have a probability β of stopping and so
$$\mathbb{P}(S(x) > k) \leq (1-\beta)^k \leq e^{-ck\beta} \qquad \forall x \in \tilde{\Omega}.$$
3. **Step 3:** Combining the previous results, for all $x \in \tilde{\Omega}$,

$$\begin{aligned}
|u_n^*(x) - g^\dagger(x)| &\leq \mathbb{E}\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \\
&= \mathbb{E}\left[\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \,\middle|\, S(x) \leq k\right] \mathbb{P}(S(x) \leq k) \\
&\quad + \mathbb{E}\left[\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \,\middle|\, S(x) > k\right] \mathbb{P}(S(x) > k) \\
&\leq \alpha \mathrm{Lip}(g^\dagger)\sqrt{k}\varepsilon + 2\|g^\dagger\|_{L^\infty} e^{-ck\beta}.
\end{aligned}$$

Intuition on the Minimal Number of Labels Proof I

1. **Step 1:** We show B_t^x behaves approximately as a Brownian motion and therefore
$$\mathbb{P}\left(\max_{t=1,\dots,k} |B_t^x - x| > \alpha\sqrt{k}\varepsilon\right) \leq e^{-c\alpha^2}.$$
2. **Step 2:** Within the labelled domain we have a probability β of stopping and so
$$\mathbb{P}(S(x) > k) \leq (1-\beta)^k \leq e^{-ck\beta} \qquad \forall x \in \tilde{\Omega}.$$
3. **Step 3:** Combining the previous results, for all $x \in \tilde{\Omega}$,

$$\begin{aligned}
|u_n^*(x) - g^\dagger(x)| &\leq \mathbb{E}\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \\
&= \mathbb{E}\left[\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \,\middle|\, S(x) \leq k\right] \mathbb{P}(S(x) \leq k) \\
&\quad + \mathbb{E}\left[\left|g^\dagger(B_{S(x)}^x) - g^\dagger(x)\right| \,\middle|\, S(x) > k\right] \mathbb{P}(S(x) > k) \\
&\leq \alpha \mathrm{Lip}(g^\dagger)\sqrt{k}\varepsilon + 2\|g^\dagger\|_{L^\infty} e^{-ck\beta}.
\end{aligned}$$

Choosing $k = \frac{C}{\beta}\log\frac{\sqrt{\beta}}{\varepsilon}$ implies (with high probability)
$$|u_n^*(x) - g^\dagger(x)| \leq C\frac{\varepsilon}{\sqrt{\beta}}\log\frac{\sqrt{\beta}}{\varepsilon}.$$

Euler-Lagrange Equations

1. Discrete variational problem: minimise
$$\mathcal{E}_n^{(2)}(u_n) = \sum_{i,j=1}^{n} w_{ij}|u_n(x_i) - u_n(x_j)|^2$$
s.t. $u_n(x_i) = \ell_i \,\forall\, i \in \mathcal{I}_n$.

Euler-Lagrange Equations

1. Discrete variational problem:
   minimise
$$\mathcal{E}_n^{(2)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^2$$
   s.t. $u_n(x_i) = \ell_i \, \forall \, i \in \mathcal{I}_n$.

2. Euler-Lagrange equation:
$$\begin{aligned} L_n u_n^*(x_i) &= 0 && \text{for } i \notin \mathcal{I}_n \\ u_n^*(x_i) &= \ell_i && \text{for } i \in \mathcal{I}_n \end{aligned}$$
   where
$$L_n u_n(x_i) = \sum_{j=1}^{n} w_{ij} \left(u_n(x_i) - u_n(x_j) \right).$$

Euler-Lagrange Equations

1. Discrete variational problem: minimise

$$\mathcal{E}_n^{(2)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^2$$

s.t. $u_n(x_i) = \ell_i \, \forall \, i \in \mathcal{I}_n$.

2. Euler-Lagrange equation:

$$\begin{aligned} L_n u_n^*(x_i) &= 0 \quad \text{for } i \notin \mathcal{I}_n \\ u_n^*(x_i) &= \ell_i \quad \text{for } i \in \mathcal{I}_n \end{aligned}$$

where

$$L_n u_n(x_i) = \sum_{j=1}^{n} w_{ij} \left(u_n(x_i) - u_n(x_j) \right).$$

3. Continuum variational problem: minimise

$$\mathcal{E}_\infty^{(2)}(u) = \sigma_\eta \int_\Omega \|\nabla u(x)\|^2 \rho^2(x) \, \mathrm{d}x$$

s.t. $u(x) = g^\dagger(x) \, \forall \, x \in \tilde{\Omega}$.

Euler-Lagrange Equations

1. Discrete variational problem: minimise
$$\mathcal{E}_n^{(2)}(u_n) = \sum_{i,j=1}^{n} w_{ij} |u_n(x_i) - u_n(x_j)|^2$$
s.t. $u_n(x_i) = \ell_i \, \forall\, i \in \mathcal{I}_n$.

2. Euler-Lagrange equation:
$$\begin{aligned} L_n u_n^*(x_i) &= 0 \quad \text{for } i \notin \mathcal{I}_n \\ u_n^*(x_i) &= \ell_i \quad \text{for } i \in \mathcal{I}_n \end{aligned}$$
where
$$L_n u_n(x_i) = \sum_{j=1}^{n} w_{ij} \left(u_n(x_i) - u_n(x_j) \right).$$

3. Continuum variational problem: minimise
$$\mathcal{E}_\infty^{(2)}(u) = \sigma_\eta \int_\Omega \|\nabla u(x)\|^2 \rho^2(x) \, \mathrm{d}x$$
s.t. $u(x) = g^\dagger(x) \, \forall\, x \in \tilde{\Omega}$.

4. Euler-Lagrange equation:
$$\begin{aligned} \mathcal{L} u^*(x) &= 0 && \text{for } x \in \Omega \setminus \tilde{\Omega} \\ u^*(x) &= g^\dagger(x) && \text{for } x \in \tilde{\Omega} \\ \frac{\partial u^*}{\partial \mathrm{n}}(x) &= 0 && \text{for } x \in \partial\Omega \end{aligned}$$
where
$$\mathcal{L} u(x) = -\frac{1}{\rho(x)} \mathrm{div}(\rho^2 \nabla u)(x).$$

Intuition on the Minimal Number of Labels Proof II

From Step 3, we have

$$\max_{x_i \in \tilde{\Omega}} |u_n^*(x_i) - g^\dagger(x_i)| \le C \frac{\varepsilon}{\sqrt{\beta}} \log \frac{\sqrt{\beta}}{\varepsilon}$$

and now we need to extend the convergence to the whole domain.

Intuition on the Minimal Number of Labels Proof II

From Step 3, we have

$$\max_{x_i \in \tilde{\Omega}} |u_n^*(x_i) - g^\dagger(x_i)| \leq C \frac{\varepsilon}{\sqrt{\beta}} \log \frac{\sqrt{\beta}}{\varepsilon}$$

and now we need to extend the convergence to the whole domain.

4. **Step 4:** Pointwise convergence of the graph Laplacian.

Theorem (Calder, Slepčev and T. (2020))

There exists $C > c > 0$ such that for any $\varphi \in C^3(\overline{\Omega})$ and any $\varepsilon \leq \vartheta \leq \frac{1}{\varepsilon}$,

$$\sup_{x \in \Omega_n} \left| L_n\varphi(x) - \mathcal{L}\varphi(x) + b.c.\text{'s} \right| \leq C\|\varphi\|_{C^3(\overline{\Omega})}(\varepsilon + \vartheta)$$

with probability at least $1 - Cne^{-cn\varepsilon^{d+2}\vartheta^2}$.

Intuition on the Minimal Number of Labels Proof III

5. **Step 5:** u^* solves

$$\begin{cases} \mathcal{L}u^* = 0 & \text{in } \Omega \setminus \tilde{\Omega} \\ u^* = g^\dagger & \text{in } \tilde{\Omega} \\ \dfrac{\partial u^*}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

5. **Step 5:** u^* solves

$$\begin{cases} \mathcal{L}u^* = 0 & \text{in } \Omega \setminus \tilde{\Omega} \\ u^* = g^\dagger & \text{in } \tilde{\Omega} \\ \dfrac{\partial u^*}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Let φ solve

$$\begin{cases} \mathcal{L}\varphi = 1 & \text{in } \Omega \setminus \tilde{\Omega} \\ \varphi = 0 & \text{in } \tilde{\Omega} \\ \dfrac{\partial \varphi}{\partial n} = 1 & \text{on } \partial\Omega. \end{cases}$$

Intuition on the Minimal Number of Labels Proof III

5. **Step 5:** u^* solves

$$\begin{cases} \mathcal{L}u^* = 0 & \text{in } \Omega \setminus \tilde{\Omega} \\ u^* = g^\dagger & \text{in } \tilde{\Omega} \\ \dfrac{\partial u^*}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Let φ solve

$$\begin{cases} \mathcal{L}\varphi = 1 & \text{in } \Omega \setminus \tilde{\Omega} \\ \varphi = 0 & \text{in } \tilde{\Omega} \\ \dfrac{\partial \varphi}{\partial n} = 1 & \text{on } \partial\Omega. \end{cases}$$

Then let

$$v = \begin{cases} u^* + M\vartheta\varphi & \text{in } \Omega \setminus \tilde{\Omega} \\ g^\dagger & \text{on } \tilde{\Omega}. \end{cases}$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$L_n v =$$

6. **Step 6:** Choosing M large enough we have

$$L_n v = \mathcal{L} u^* + M\vartheta \mathcal{L}\varphi + O(\varepsilon + \vartheta)$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L}u^* + M\vartheta\mathcal{L}\varphi + O(\varepsilon+\vartheta) \\ &= M\vartheta + O(\varepsilon+\vartheta) > 0. \end{aligned}$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L}u^* + M\vartheta\mathcal{L}\varphi + O(\varepsilon + \vartheta) \\ &= M\vartheta + O(\varepsilon + \vartheta) > 0. \end{aligned}$$

By the max principle, and since $L_n(u_n^* - v) < 0$ on $\Omega \setminus \tilde{\Omega}$,

$$\max_{\Omega_n}(u_n^* - v)$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L}u^* + M\vartheta\mathcal{L}\varphi + O(\varepsilon + \vartheta) \\ &= M\vartheta + O(\varepsilon + \vartheta) > 0. \end{aligned}$$

By the max principle, and since $L_n(u_n^* - v) < 0$ on $\Omega \setminus \tilde{\Omega}$,

$$\max_{\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - v)$$

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L}u^* + M\vartheta\mathcal{L}\varphi + O(\varepsilon + \vartheta) \\ &= M\vartheta + O(\varepsilon + \vartheta) > 0. \end{aligned}$$

By the max principle, and since $L_n(u_n^* - v) < 0$ on $\Omega \setminus \tilde{\Omega}$,

$$\max_{\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - g^\dagger)$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L}u^* + M\vartheta\mathcal{L}\varphi + O(\varepsilon + \vartheta) \\ &= M\vartheta + O(\varepsilon + \vartheta) > 0. \end{aligned}$$

By the max principle, and since $L_n(u_n^* - v) < 0$ on $\Omega \setminus \tilde{\Omega}$,

$$\max_{\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - g^\dagger) \leq \frac{C\varepsilon}{\sqrt{\beta}} \log \frac{\sqrt{\beta}}{\varepsilon}.$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L} u^* + M\vartheta \mathcal{L}\varphi + O(\varepsilon + \vartheta) \\ &= M\vartheta + O(\varepsilon + \vartheta) > 0. \end{aligned}$$

By the max principle, and since $L_n(u_n^* - v) < 0$ on $\Omega \setminus \tilde{\Omega}$,

$$\max_{\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - g^\dagger) \leq \frac{C\varepsilon}{\sqrt{\beta}} \log \frac{\sqrt{\beta}}{\varepsilon}.$$

Using the same argument on $v - u_n^*$ we have

$$\|u_n^* - v\|_{\mathrm{L}^\infty(\Omega_n)} \leq \frac{C\varepsilon}{\sqrt{\beta}} \log \frac{\sqrt{\beta}}{\varepsilon}.$$

Intuition on the Minimal Number of Labels Proof IV

6. **Step 6:** Choosing M large enough we have

$$\begin{aligned} L_n v &= \mathcal{L}u^* + M\vartheta\mathcal{L}\varphi + O(\varepsilon + \vartheta) \\ &= M\vartheta + O(\varepsilon + \vartheta) > 0. \end{aligned}$$

By the max principle, and since $L_n(u_n^* - v) < 0$ on $\Omega \setminus \tilde{\Omega}$,

$$\max_{\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - v) = \max_{\tilde{\Omega}\cap\Omega_n}(u_n^* - g^\dagger) \leq \frac{C\varepsilon}{\sqrt{\beta}} \log\frac{\sqrt{\beta}}{\varepsilon}.$$

Using the same argument on $v - u_n^*$ we have

$$\|u_n^* - v\|_{\mathrm{L}^\infty(\Omega_n)} \leq \frac{C\varepsilon}{\sqrt{\beta}} \log\frac{\sqrt{\beta}}{\varepsilon}.$$

7. **Step 7:** Since $\|\varphi\|_{\mathrm{L}^\infty} \leq C$ then

$$\|u_n^* - u^*\|_{\mathrm{L}^\infty(\Omega_n)} \leq \frac{C\varepsilon}{\sqrt{\beta}} \log\frac{\sqrt{\beta}}{\varepsilon}.$$

Large Data Limits for $|\mathcal{I}_n| \to \infty$

Theorem (Calder, Slepčev and T. (2020))

Ill-Posed Regime. *Let* ε_n *satisfy a lower bound. Let* u_n^* *be a sequence of minimizers of* $\mathcal{E}_n^{(2)}$ *satisfying the constraints. Assume* $\beta_n \ll \varepsilon_n^2$*. Then, almost surely,* $\{u_n^*\}_{n\in\mathbb{N}}$ *is precompact and any convergent subsequence converges to a constant.*

Large Data Limits for $|\mathcal{I}_n| \to \infty$

Theorem (Calder, Slepčev and T. (2020))

Ill-Posed Regime. *Let ε_n satisfy a lower bound. Let u_n^* be a sequence of minimizers of $\mathcal{E}_n^{(2)}$ satisfying the constraints. Assume $\beta_n \ll \varepsilon_n^2$. Then, almost surely, $\{u_n^*\}_{n\in\mathbb{N}}$ is precompact and any convergent subsequence converges to a constant.*

Theorem (Calder, Slepčev and T. (2020))

Well-Posed Regime. *Let ε_n satisfy a lower bound. Let u_n^* be a sequence of minimizers of $\mathcal{E}_n^{(2)}$ satisfying the constraints and u^* be the minimiser of $\mathcal{E}_\infty^{(2)}$ with constraints. Assume $\beta_n \gg \varepsilon_n^2$. Then, almost surely, u_n^* converges to u^* uniformly, in particular*

$$\max_{i=1,\dots,n} |u_n^*(x_i) - u^*(x_i)| \lesssim \frac{\varepsilon_n}{\sqrt{\beta_n}} \log \frac{\sqrt{\beta_n}}{\varepsilon_n}.$$

Contents

Finite Constraint Degeneracy

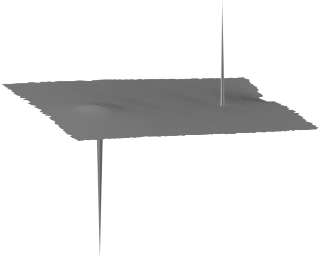


Figure: A toy example with two labels which are seen as spikes.

1. Let us assume $\mathcal{I}_n = \{1, \ldots, m\}$.

[1] Nadler, Srebro and Zhou, *Statistical Analysis of Semi-Supervised Learning*, NeurIPS, 2009, pp. 1330–1338

Finite Constraint Degeneracy

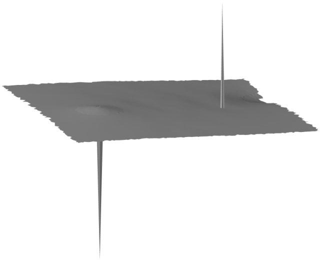


Figure: A toy example with two labels which are seen as spikes.

1. Let us assume $\mathcal{I}_n = \{1, \ldots, m\}$.
2. We have that $u_n^*(x_i) \approx c$ for $i \notin \mathcal{I}_n$.[1]

[1] Nadler, Srebro and Zhou, *Statistical Analysis of Semi-Supervised Learning*, NeurIPS, 2009, pp. 1330–1338

Finite Constraint Degeneracy

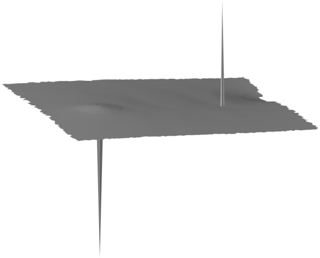


Figure: A toy example with two labels which are seen as spikes.

1. Let us assume $\mathcal{I}_n = \{1, \ldots, m\}$.
2. We have that $u_n^*(x_i) \approx c$ for $i \notin \mathcal{I}_n$.[1]
3. Say $c > 0$, then this means the majority of labels, classified using $\ell_{u_n^*}(x_i) = \mathrm{sign}(u_n^*(x_i))$, will be classed as $\ell_{u_n^*}(x_i) = 1$.

[1] Nadler, Srebro and Zhou, *Statistical Analysis of Semi-Supervised Learning*, NeurIPS, 2009, pp. 1330–1338

Finite Constraint Degeneracy

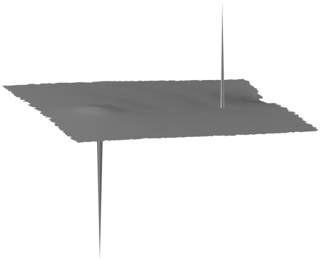


Figure: A toy example with two labels which are seen as spikes.

1. Let us assume $\mathcal{I}_n = \{1, \ldots, m\}$.
2. We have that $u_n^*(x_i) \approx c$ for $i \notin \mathcal{I}_n$.[1]
3. Say $c > 0$, then this means the majority of labels, classified using $\ell_{u_n^*}(x_i) = \mathrm{sign}(u_n^*(x_i))$, will be classed as $\ell_{u_n^*}(x_i) = 1$.
4. One way to correct this bias would be to consider $u_n^* - c$, but this is just the solution Laplace Learning with the labels $\ell_i - c$, **why would we expect to do better with the the wrong label?**

[1] Nadler, Srebro and Zhou, *Statistical Analysis of Semi-Supervised Learning*, NeurIPS, 2009, pp. 1330–1338

Laplace Learning on MNIST

# Labels/class	**1**	**2**	**3**	**4**
Laplace	16.1 (6.2)	28.2 (10)	42.0 (12)	57.8 (12)
Graph NN	58.8 (5.6)	66.6 (2.8)	70.2 (4)	71.3 (2.6)

# Labels/class	**5**	**10**	**50**	**100**
Laplace	69.5 (12)	93.2 (2.3)	96.9 (0.1)	97.1 (0.1)
Graph NN	73.4 (1.9)	82.3 (1.0)	89.0 (0.5)	90.6 (0.4)

Average accuracy over 10 trials with standard deviation in brackets.

C.f. for 1 label per class the shifted Laplacian method achieves 85.9% accuracy.

Graph NN: 1-nearest neighbour using graph geodesic distance.

1. The random walk interpretation of Laplace learning explains the ill-posedness...

1. The random walk interpretation of Laplace learning explains the ill-posedness...
2. If $|\mathcal{I}_n| \ll n$ then the probability of hitting a label is low.

Random Walks at Low Labelling Rates

1. The random walk interpretation of Laplace learning explains the ill-posedness...
2. If $|\mathcal{I}_n| \ll n$ then the probability of hitting a label is low.
3. Hence, we expect $S(x) \gg 1$, and in particular $S(x)$ may be greater than the mixing time of the random walk.

Random Walks at Low Labelling Rates

1. The random walk interpretation of Laplace learning explains the ill-posedness...
2. If $|\mathcal{I}_n| \ll n$ then the probability of hitting a label is low.
3. Hence, we expect $S(x) \gg 1$, and in particular $S(x)$ may be greater than the mixing time of the random walk.
4. This means $B^x_{S(x)}$ is distributionally independent of x.

Random Walks at Low Labelling Rates

1. The random walk interpretation of Laplace learning explains the ill-posedness...
2. If $|\mathcal{I}_n| \ll n$ then the probability of hitting a label is low.
3. Hence, we expect $S(x) \gg 1$, and in particular $S(x)$ may be greater than the mixing time of the random walk.
4. This means $B^x_{S(x)}$ is distributionally independent of x.
5. This implies u_n^* is approximately a constant on $\{x_i\}_{i \notin \mathcal{I}_n}$.

Random Walks at Low Labelling Rates

1. The random walk interpretation of Laplace learning explains the ill-posedness...
2. If $|\mathcal{I}_n| \ll n$ then the probability of hitting a label is low.
3. Hence, we expect $S(x) \gg 1$, and in particular $S(x)$ may be greater than the mixing time of the random walk.
4. This means $B^x_{S(x)}$ is distributionally independent of x.
5. This implies u_n^* is approximately a constant on $\{x_i\}_{i \notin \mathcal{I}_n}$.
6. The stationary distribution of B_t^x is $\pi(x_i) = \frac{d_i}{\sum_{j=1}^n d_j}$, so it follows that

$$u_n^*(x_i) = \mathbb{E}[\ell(B^x_{S(x)})] \approx \frac{\sum_{i\in\mathcal{I}_n} d_i \ell_i}{\sum_{i\in\mathcal{I}_n} d_i} =: c$$

for all $i \notin \mathcal{I}_n$.

Laplace's Equation at Low Labelling Rates I

1. Assume no labels are connected, i.e. $w_{ij} = 0$ for all $i, j \in \mathcal{I}_n$.

Laplace's Equation at Low Labelling Rates I

1. Assume no labels are connected, i.e. $w_{ij} = 0$ for all $i, j \in \mathcal{I}_n$.
2. Then, for $i \in \mathcal{I}_n = \{1, \ldots, m\}$ and $c = \frac{\sum_{i \in \mathcal{I}_n} d_i \ell_i}{\sum_{i \in \mathcal{I}_n} d_i}$,

$$\begin{aligned} L_n u_n^*(x_i) &= \sum_{j=1}^{n} w_{ij}(u_n^*(x_i) - u_n^*(x_j)) \\ &\approx \sum_{j \notin \mathcal{I}_n} w_{ij}(\ell_i - c) \\ &= d_i(\ell_i - c). \end{aligned}$$

Laplace's Equation at Low Labelling Rates I

1. Assume no labels are connected, i.e. $w_{ij} = 0$ for all $i, j \in \mathcal{I}_n$.
2. Then, for $i \in \mathcal{I}_n = \{1, \ldots, m\}$ and $c = \frac{\sum_{i \in \mathcal{I}_n} d_i \ell_i}{\sum_{i \in \mathcal{I}_n} d_i}$,

$$\begin{aligned} L_n u_n^*(x_i) &= \sum_{j=1}^{n} w_{ij}(u_n^*(x_i) - u_n^*(x_j)) \\ &\approx \sum_{j \notin \mathcal{I}_n} w_{ij}(\ell_i - c) \\ &= d_i(\ell_i - c). \end{aligned}$$

3. We also have

$$\sum_{i=1}^{n} d_i u_n(x_i) \approx \sum_{i \in \mathcal{I}_n} d_i \ell_i + c \sum_{i \notin \mathcal{I}_n} d_i = c \sum_{i=1}^{n} d_i.$$

Laplace's Equation at Low Labelling Rates II

- For $|\mathcal{I}_n| \ll n$, u_n^* approximately satisfies

$$L_n u_n^*(x_i) \approx \sum_{j \in \mathcal{I}_n} d_j(\ell_j - c)\delta_{ij}, \qquad \frac{1}{\sum_{i=1}^n d_i} \sum_{i=1}^n d_i u_n^*(x_i) \approx c.$$

Laplace's Equation at Low Labelling Rates II

- For $|\mathcal{I}_n| \ll n$, u_n^* approximately satisfies

$$L_n u_n^*(x_i) \approx \sum_{j\in\mathcal{I}_n} d_j(\ell_j - c)\delta_{ij}, \qquad \frac{1}{\sum_{i=1}^n d_i} \sum_{i=1}^n d_i u_n^*(x_i) \approx c.$$

- Shifting by c we could define v_n^* by

$$L_n v_n^*(x_i) = \sum_{j\in\mathcal{I}_n} d_j(\ell_j - c)\delta_{ij}, \qquad \sum_{i=1}^n d_i v_n^*(x_i) = 0.$$

Laplace's Equation at Low Labelling Rates II

- For $|\mathcal{I}_n| \ll n$, u_n^* approximately satisfies

$$L_n u_n^*(x_i) \approx \sum_{j\in\mathcal{I}_n} d_j(\ell_j - c)\delta_{ij}, \qquad \frac{1}{\sum_{i=1}^n d_i} \sum_{i=1}^n d_i u_n^*(x_i) \approx c.$$

- Shifting by c we could define v_n^* by

$$L_n v_n^*(x_i) = \sum_{j\in\mathcal{I}_n} d_j(\ell_j - c)\delta_{ij}, \qquad \sum_{i=1}^n d_i v_n^*(x_i) = 0.$$

- However, we find a slight improvement in performance if we additionally normalise each node and therefore we define v_n^* to satisfy

$$L_n v_n^*(x_i) = \sum_{j\in\mathcal{I}_n} (\ell_j - \bar{c})\delta_{ij}, \qquad \sum_{i=1}^n v_n^*(x_i) = 0$$

where $\bar{c} = \frac{1}{|\mathcal{I}_n|}\sum_{i\in\mathcal{I}_n} \ell_i$.

Poisson Random Walk

Recall that B_t^x is the random walk starting from x and transitioning from x_i to x_j with probability proportional to w_{ij}.

Poisson Random Walk

Recall that B_t^x is the random walk starting from x and transitioning from x_i to x_j with probability proportional to w_{ij}.

Theorem (Calder, Cook, Slepčev and T. (2020))

Let

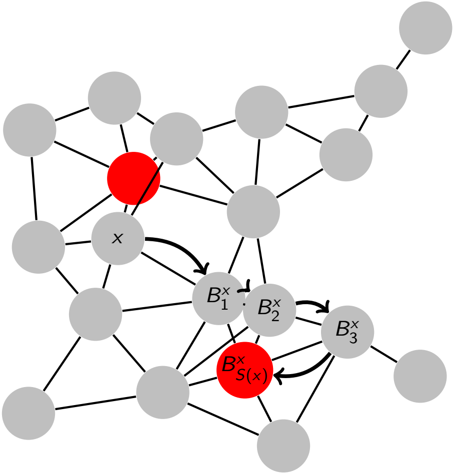
$$v_n^{(T)}(x_i) = \mathbb{E}\left[\frac{1}{d_i}\sum_{t=0}^{T}\sum_{j\in\mathcal{I}_n}(\ell_j - \bar{c})\mathbb{1}_{B_t^{x_j}=x_i}\right].$$

Then,

$$v_n^{(T+1)}(x_i) = v_n^{(T)}(x_i) + \frac{1}{d_i}\left(\sum_{j\in\mathcal{I}_n}(\ell_j - \bar{c})\delta_{ij} - L_n v_n^{(T)}(x_i)\right)$$

and moreover $v_n^{(T)} \to v_n^*$ *as* $T \to \infty$.

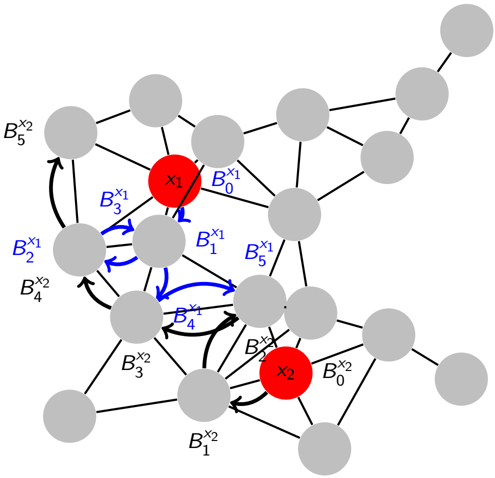
Laplace's Random Walk (Again)



Red - labelled nodes, grey - unlabelled nodes.

$$u_n^*(x) = \mathbb{E}\left[\sum_{j\in\mathcal{I}_n} \ell_j \mathbb{1}_{B^x_{S(x)}=x_j}\right]$$

Poisson's Random Walk



Red - labelled nodes, grey - unlabelled nodes.

Notice that

$$\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{d_i}\sum_{j\in\mathcal{I}_n} \ell_j \mathbb{1}_{B_t^{x_j}=x_i}\right] = \frac{m\bar{c}}{\sum_{j=1}^n d_j} = \lim_{T\to\infty} \mathbb{E}\left[\frac{\bar{c}}{d_i}\sum_{j\in\mathcal{I}_n} \mathbb{1}_{B_t^{x_j}=x_i}\right].$$

$$v_n^*(x) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{d_i}\sum_{t=0}^{T}\sum_{j\in\mathcal{I}_n} (\ell_j - \bar{c})\mathbb{1}_{B_t^{x_j}=x}\right].$$

MNIST Results

	# Labels per class				
	1	**2**	**3**	**4**	**5**
Laplace/LP	16.1 (6.2)	28.2 (10.3)	42.0 (12.4)	57.8 (12.3)	69.5 (12.2)
NN	55.8 (5.1)	65.0 (3.2)	68.9 (3.2)	72.1 (2.8)	74.1 (2.4)
Random Walk	66.4 (5.3)	76.2 (3.3)	80.0 (2.7)	82.8 (2.3)	84.5 (2.0)
MBO	19.4 (6.2)	29.3 (6.9)	40.2 (7.4)	50.7 (6.0)	59.2 (6.0)
VolumeMBO	89.9 (7.3)	95.6 (1.9)	96.2 (1.2)	96.6 (0.6)	96.7 (0.6)
WNLL	55.8 (15.2)	82.8 (7.6)	90.5 (3.3)	93.6 (1.5)	94.6 (1.1)
Centered Kernel	19.1 (1.9)	24.2 (2.3)	28.8 (3.4)	32.6 (4.1)	35.6 (4.6)
Sparse LP	14.0 (5.5)	14.0 (4.0)	14.5 (4.0)	18.0 (5.9)	16.2 (4.2)
p-Laplace	72.3 (9.1)	86.5 (3.9)	89.7 (1.6)	90.3 (1.6)	91.9 (1.0)
Poisson	90.2 (4.0)	93.6 (1.6)	94.5 (1.1)	94.9 (0.8)	95.3 (0.7)
PoissonMBO	**96.5 (2.6)**	**97.2 (0.1)**	**97.2 (0.1)**	**97.2 (0.1)**	**97.2 (0.1)**

Average (standard deviation) classification accuracy over 100 trials.

FashionMNIST Results

	# Labels per class				
	1	**2**	**3**	**4**	**5**
Laplace/LP	18.4 (7.3)	32.5 (8.2)	44.0 (8.6)	52.2 (6.2)	57.9 (6.7)
NN	44.5 (4.2)	50.8 (3.5)	54.6 (3.0)	56.6 (2.5)	58.3 (2.4)
Random Walk	49.0 (4.4)	55.6 (3.8)	59.4 (3.0)	61.6 (2.5)	63.4 (2.5)
MBO	15.7 (4.1)	20.1 (4.6)	25.7 (4.9)	30.7 (4.9)	34.8 (4.3)
VolumeMBO	54.7 (5.2)	61.7 (4.4)	66.1 (3.3)	68.5 (2.8)	70.1 (2.8)
WNLL	44.6 (7.1)	59.1 (4.7)	64.7 (3.5)	67.4 (3.3)	70.0 (2.8)
Centered Kernel	11.8 (0.4)	13.1 (0.7)	14.3 (0.8)	15.2 (0.9)	16.3 (1.1)
Sparse LP	14.1 (3.8)	16.5 (2.0)	13.7 (3.3)	13.8 (3.3)	16.1 (2.5)
p-Laplace	54.6 (4.0)	57.4 (3.8)	65.4 (2.8)	68.0 (2.9)	68.4 (0.5)
Poisson	60.8 (4.6)	66.1 (3.9)	69.6 (2.6)	71.2 (2.2)	72.4 (2.3)
PoissonMBO	**62.0 (5.7)**	**67.2 (4.8)**	**70.4 (2.9)**	**72.1 (2.5)**	**73.1 (2.7)**

Average (standard deviation) classification accuracy over 100 trials.

C.f. state-of-the-art clustering result of *67.2%* [McConville et al., 2019].

CIFAR-10 Results

	# Labels per class				
	1	**2**	**3**	**4**	**5**
Laplace/LP	10.5 (1.3)	12.5 (4.4)	13.1 (3.8)	14.5 (4.7)	18.0 (6.9)
NN	33.6 (4.4)	37.3 (3.3)	40.3 (3.0)	40.9 (2.7)	42.1 (2.4)
Random Walk	37.1 (5.0)	42.1 (3.7)	45.8 (3.4)	47.0 (2.8)	48.8 (2.5)
MBO	15.2 (4.1)	20.4 (4.8)	25.9 (4.1)	29.6 (4.3)	34.5 (4.2)
VolumeMBO	40.3 (8.0)	47.2 (7.1)	52.2 (5.3)	53.3 (4.7)	55.9 (4.0)
WNLL	20.8 (6.4)	34.5 (6.2)	42.1 (5.2)	46.1 (4.4)	50.2 (3.5)
Centered Kernel	13.8 (1.1)	15.5 (1.2)	17.3 (1.4)	18.8 (1.7)	20.4 (1.6)
Sparse LP	10.4 (2.1)	11.1 (1.4)	11.8 (2.1)	12.8 (4.4)	13.6 (3.3)
p-Laplace	28.7 (6.6)	39.8 (6.4)	45.7 (2.6)	46.8 (1.7)	50.4 (2.9)
Poisson	41.6 (5.4)	46.9 (4.2)	51.1 (3.4)	52.5 (3.0)	54.5 (3.0)
PoissonMBO	**42.1 (7.0)**	**49.1 (5.3)**	**53.8 (4.4)**	**55.6 (3.7)**	**57.4 (3.4)**

Average (standard deviation) classification accuracy over 100 trials.

C.f. state-of-the-art clustering result of *41.2%* [Mukherjee et al., ClusterGAN, CVPR 2019].

Contents

The Fractional Graph Laplacian

- Let $(\lambda_i^{(n)}, q_i^{(n)})$ be the eigenvalues and eigenvectors of the normalised graph Laplacian $\frac{1}{n\varepsilon_n^2\sigma_\eta} L_n$.

The Fractional Graph Laplacian

- Let $(\lambda_i^{(n)}, q_i^{(n)})$ be the eigenvalues and eigenvectors of the normalised graph Laplacian $\frac{1}{n\varepsilon_n^2\sigma_\eta} L_n$.
- We define the fractional graph Laplacian energy $\mathcal{J}_n^{(\alpha,\tau)}$ by

$$\mathcal{J}_n^{(\alpha,\tau)}(u_n) = \sum_{i=1}^{n} \left(\lambda_i^{(n)} + \tau^2\right)^{\alpha} \langle u_n, q_i^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}^2.$$

The Fractional Graph Laplacian

- Let $(\lambda_i^{(n)}, q_i^{(n)})$ be the eigenvalues and eigenvectors of the normalised graph Laplacian $\frac{1}{n\varepsilon_n^2\sigma_\eta} L_n$.
- We define the fractional graph Laplacian energy $\mathcal{J}_n^{(\alpha,\tau)}$ by

$$\mathcal{J}_n^{(\alpha,\tau)}(u_n) = \sum_{i=1}^{n} \left(\lambda_i^{(n)} + \tau^2\right)^{\alpha} \langle u_n, q_i^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}^2.$$

- When $\alpha = 1$ and $\tau = 0$,

$$\begin{aligned}\mathcal{J}_n^{(1,0)}(u_n) &= \sum_{i=1}^{n} \lambda_i^{(n)} \langle u_n, q_i^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}^2 \\ &= \langle u_n, L_n u_n \rangle_{\mathrm{L}^2(\mu_n)} \\ &= \frac{1}{2} \mathcal{E}_n^{(2)}(u_n).\end{aligned}$$

- Let (λ_i, q_i) be the eigenvalues and eigenfunctions of the continuum operator $\mathcal{L}$.

- Let (λ_i, q_i) be the eigenvalues and eigenfunctions of the continuum operator $\mathcal{L}$.
- We define $\mathcal{J}_\infty^{(\alpha,\tau)}$ by

$$\mathcal{J}_\infty^{(\alpha,\tau)}(u) = \sum_{i=1}^{\infty} (\lambda_i + \tau^2)^\alpha \langle u, q_i \rangle_{\mathrm{L}^2(\mu)}^2.$$

Continuum Limit of the Graph Fractional Laplacian

- Let (λ_i, q_i) be the eigenvalues and eigenfunctions of the continuum operator $\mathcal{L}$.
- We define $\mathcal{J}_\infty^{(\alpha,\tau)}$ by

$$\mathcal{J}_\infty^{(\alpha,\tau)}(u) = \sum_{i=1}^{\infty} (\lambda_i + \tau^2)^\alpha \langle u, q_i \rangle_{\mathrm{L}^2(\mu)}^2.$$

- When $\alpha = 1$ and $\tau = 0$ we have

$$\mathcal{J}_\infty^{(1,0)}(u) = \int_\Omega |\nabla u(x)|^2 \rho^2(x) \, \mathrm{d}x = \frac{1}{\sigma_\eta} \mathcal{E}_\infty^{(2)}(u).$$

Convergence of the Fractional Graph Laplacian

Theorem (Dunlop, Slepčev, Stuart and T. (2017))

Under assumptions on η, Ω, μ and a lower bound on $\epsilon_n \to 0$ we have, with probability one,

1. $\Gamma\text{-}\lim_{n\to\infty} 2\sigma_\eta \mathcal{J}_n^{(\alpha,\tau)} = \mathcal{J}_\infty^{(\alpha,\tau)}$ *with respect to the* TL^2 *topology;*
2. *if* $\tau = 0$*, any sequence* $\{u_n\}$ *with* $u_n : \Omega_n \to \mathbb{R}$ *satisfying* $\sup_n \|u_n\|_{\mathrm{L}^2(\mu_n)} < \infty$ *and* $\sup_{n\in\mathbb{N}} \mathcal{J}_n^{(\alpha,0)}(u_n) < \infty$ *is pre-compact in the* TL^2 *topology;*
3. *if* $\tau > 0$*, any sequence* $\{u_n\}$ *with* $u_n : \Omega_n \to \mathbb{R}$ *satisfying* $\sup_{n\in\mathbb{N}} \mathcal{J}_n^{(\alpha,\tau)}(u_n) < \infty$ *is pre-compact in the* TL^2 *topology.*

Large Data Limits of Fractional Laplace Learning: Ill-Posed Case

Theorem (Dunlop, Slepčev, Stuart and T. (2017) and Weihs and T.(2023))

Assume $\varepsilon_n^{2\alpha} n \to \infty$ *and* $|\mathcal{I}_n| = m$ *is fixed. Let* $\{u_n^*\}_{n\in\mathbb{N}}$ *be constrained minimisers of* $\mathcal{J}_n^{(\alpha,\tau)}$. *Assume* $\sup_{n\in\mathbb{N}} \|u_n^*\|_{\mathrm{L}^2(\mu_n)} < +\infty$. *Then, with probability one,* $\{u_n^*\}_{n\in\mathbb{N}}$ *are precompact in* TL^2 *and any converging subsequence converges to a constant.*

Large Data Limits of Fractional Laplace Learning: Ill-Posed Case

Theorem (Dunlop, Slepčev, Stuart and T. (2017) and Weihs and T.(2023))

Assume $\varepsilon_n^{2\alpha} n \to \infty$ and $|\mathcal{I}_n| = m$ is fixed. Let $\{u_n^*\}_{n\in\mathbb{N}}$ be constrained minimisers of $\mathcal{J}_n^{(\alpha,\tau)}$. Assume $\sup_{n\in\mathbb{N}} \|u_n^*\|_{\mathrm{L}^2(\mu_n)} < +\infty$. Then, with probability one, $\{u_n^*\}_{n\in\mathbb{N}}$ are precompact in TL^2 and any converging subsequence converges to a constant.

- Remark 1: $\varepsilon_n^{2\alpha} n \to \infty$ is always true if $\alpha \leq \frac{d}{2}$ (due to the lower bound on ε_n).

Large Data Limits of Fractional Laplace Learning: Ill-Posed Case

Theorem (Dunlop, Slepčev, Stuart and T. (2017) and Weihs and T.(2023))

Assume $\varepsilon_n^{2\alpha} n \to \infty$ and $|\mathcal{I}_n| = m$ is fixed. Let $\{u_n^*\}_{n\in\mathbb{N}}$ be constrained minimisers of $\mathcal{J}_n^{(\alpha,\tau)}$. Assume $\sup_{n\in\mathbb{N}} \|u_n^*\|_{\mathrm{L}^2(\mu_n)} < +\infty$. Then, with probability one, $\{u_n^*\}_{n\in\mathbb{N}}$ are precompact in TL^2 and any converging subsequence converges to a constant.

- Remark 1: $\varepsilon_n^{2\alpha} n \to \infty$ is always true if $\alpha \leq \frac{d}{2}$ (due to the lower bound on ε_n).
- Remark 2: The idea behind the proof is the same as in the p-Laplacian: measure the cost of a spike $u_n(x_i) = 1$ for $i = 1$ and $u(x_i) = 0$ otherwise.

Large Data Limits of Fractional Laplace Learning: Well-Posed Case

Theorem (Weihs and T. (2023))

Let $\Omega = [0,1]^d$ be the torus. Assume $\varepsilon_n^{\frac{\alpha-1}{2}} n$ is bounded, $\alpha > \frac{5d}{2} + 4$ and $|\mathcal{I}_n| = m$ is fixed. Let $\{u_n^*\}_{n\in\mathbb{N}}$ be constrained minimisers of $\mathcal{J}_n^{(\alpha,\tau)}$. Then, with probability one, the sequence u_n^* converges uniformly to the constrained minimizer of $\mathcal{J}_\infty^{(\alpha,\tau)}$.

Intuition on the Proof I

- As in the p-Laplacian case we want to control

$$u_n(x) - u_n(y) = \sum_{k=1}^{n} \langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)} \left(q_k^{(n)}(x) - q_k^{(n)}(y) \right).$$

Intuition on the Proof I

- As in the p-Laplacian case we want to control

$$u_n(x) - u_n(y) = \sum_{k=1}^{n} \langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)} \left(q_k^{(n)}(x) - q_k^{(n)}(y) \right).$$

- We split the summation at $K_n \sim \varepsilon_n^{-\frac{d}{2}}$.

Intuition on the Proof I

- As in the p-Laplacian case we want to control

$$u_n(x) - u_n(y) = \sum_{k=1}^{n} \langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)} \left(q_k^{(n)}(x) - q_k^{(n)}(y) \right).$$

- We split the summation at $K_n \sim \varepsilon_n^{-\frac{d}{2}}$.
- For $k = K_n$ we use Weyl's law: $\lambda_{n,K_n}^{-1} \sim K_n^{-\frac{2}{d}} \sim \varepsilon_n$.

Intuition on the Proof I

- As in the p-Laplacian case we want to control

$$u_n(x) - u_n(y) = \sum_{k=1}^{n} \langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)} \left(q_k^{(n)}(x) - q_k^{(n)}(y) \right).$$

- We split the summation at $K_n \sim \varepsilon_n^{-\frac{d}{2}}$.
- For $k = K_n$ we use Weyl's law: $\lambda_{n,K_n}^{-1} \sim K_n^{-\frac{2}{d}} \sim \varepsilon_n$.
- For $k = K_n, \dots, n$ we use

$$\left| q_k^{(n)}(x) - q_k^{(n)}(y) \right| \lesssim \sqrt{n \mathcal{E}_n^{(2)}(q_k^{(n)})} |x - y| = \sqrt{n \lambda_k^{(n)}} |x - y|$$

to show that

$$\begin{aligned}
\sum_{k=K_n}^{n} |\langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}| \left| q_k^{(n)}(x) - q_k^{(n)}(y) \right| &\leq \sqrt{n} |x - y| \sum_{k=K_n}^{n} \sqrt{\lambda_k^{(n)}} |\langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}| \\
&\lesssim n |x - y| \left(\sum_{k=1}^{n} \lambda_k^{(n)} |\langle u_n, q_k^{(n)} \rangle_{\mathrm{L}^2(\mu_n)}|^2 \right)^{\frac{1}{2}} \\
&\lesssim n |x - y| \sqrt{\mathcal{J}_n^{(\alpha,0)}(u_n)} (\lambda_{K_n}^{(n)})^{\frac{1-\alpha}{2}} \lesssim n \varepsilon_n^{\frac{\alpha-1}{2}} |x - y| \sqrt{\mathcal{J}_n^{(\alpha,0)}}.
\end{aligned}$$

Intuition on the Proof II

- For $k = 1, \ldots, K_n$ we can control

$$|\lambda_{n,k} - \lambda_k| \lesssim \lambda_k \left(\sqrt{\lambda_k} \varepsilon_n + \frac{d_{\mathrm{W}^\infty}(\mu_n, \mu)}{\varepsilon_n} \right)$$
$$\|q_i^{(n)}\|_{\mathrm{L}^\infty} \lesssim \lambda_k^{d+1}$$

thanks to García Trillos, Gerlach, Hein and Slepčev (2020) and Calder, García Trillos and Lewicka (2022).

Intuition on the Proof II

- For $k = 1, \dots, K_n$ we can control

$$|\lambda_{n,k} - \lambda_k| \lesssim \lambda_k \left(\sqrt{\lambda_k} \varepsilon_n + \frac{d_{\mathrm{W}^\infty}(\mu_n, \mu)}{\varepsilon_n} \right)$$
$$\|q_i^{(n)}\|_{\mathrm{L}^\infty} \lesssim \lambda_k^{d+1}$$

  thanks to García Trillos, Gerlach, Hein and Slepčev (2020) and Calder, García Trillos and Lewicka (2022).

- Putting everything together implies

$$\frac{|u_n(x) - u_n(y)|}{|x-y| + d_{\mathrm{W}^\infty}(\mu_n, \mu)} \lesssim \sqrt{\mathcal{J}_n^{(\alpha,0)}(u_n)}(1 + n\varepsilon_n^{\frac{\alpha-1}{2}}) + \|u_n\|_{\mathrm{L}^2(\mu_n)}.$$

Intuition on the Proof II

- For $k = 1, \ldots, K_n$ we can control

$$|\lambda_{n,k} - \lambda_k| \lesssim \lambda_k \left(\sqrt{\lambda_k} \varepsilon_n + \frac{d_{\mathrm{W}^\infty}(\mu_n, \mu)}{\varepsilon_n} \right)$$
$$\|q_i^{(n)}\|_{\mathrm{L}^\infty} \lesssim \lambda_k^{d+1}$$

  thanks to García Trillos, Gerlach, Hein and Slepčev (2020) and Calder, García Trillos and Lewicka (2022).

- Putting everything together implies

$$\frac{|u_n(x) - u_n(y)|}{|x - y| + d_{\mathrm{W}^\infty}(\mu_n, \mu)} \lesssim \sqrt{\mathcal{J}_n^{(\alpha,0)}(u_n)}(1 + n\varepsilon_n^{\frac{\alpha-1}{2}}) + \|u_n\|_{\mathrm{L}^2(\mu_n)}.$$

- Compactness: after (piecewise) extension and mollification use the above Lipschitz bound and the Arzela-Ascoli theorem to infer the existence of a uniformly converging subsequence.

Intuition on the Proof II

- For $k = 1, \ldots, K_n$ we can control

$$|\lambda_{n,k} - \lambda_k| \lesssim \lambda_k \left(\sqrt{\lambda_k}\varepsilon_n + \frac{d_{\mathrm{W}^\infty}(\mu_n, \mu)}{\varepsilon_n} \right)$$

$$\|q_i^{(n)}\|_{\mathrm{L}^\infty} \lesssim \lambda_k^{d+1}$$

  thanks to García Trillos, Gerlach, Hein and Slepčev (2020) and Calder, García Trillos and Lewicka (2022).

- Putting everything together implies

$$\frac{|u_n(x) - u_n(y)|}{|x-y| + d_{\mathrm{W}^\infty}(\mu_n, \mu)} \lesssim \sqrt{\mathcal{J}_n^{(\alpha,0)}(u_n)}(1 + n\varepsilon_n^{\frac{\alpha-1}{2}}) + \|u_n\|_{\mathrm{L}^2(\mu_n)}.$$

- Compactness: after (piecewise) extension and mollification use the above Lipschitz bound and the Arzela-Ascoli theorem to infer the existence of a uniformly converging subsequence.
- Combined with the Γ-convergence result we can conclude the theorem.

Contents

Graph Diffusions

- Let $X(t) = [x_1(t)^\top, x_2(t)^\top, \ldots, x_n(t)^\top]^\top \in \mathbb{R}^{n \times d}$ satisfy

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = \mathrm{div}(F(X(t), t) \odot \nabla X(t))$$

where $F : \mathbb{R}^{n \times d} \times [0, \infty) \to \mathbb{R}^{n \times n}$ is a given function, ∇ is the graph gradient operator and div is the graph divergence operator.

Graph Diffusions

- Let $X(t) = [x_1(t)^\top, x_2(t)^\top, \dots, x_n(t)^\top]^\top \in \mathbb{R}^{n\times d}$ satisfy

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = \mathrm{div}(F(X(t),t) \odot \nabla X(t))$$

  where $F : \mathbb{R}^{n\times d} \times [0,\infty) \to \mathbb{R}^{n\times n}$ is a given function, ∇ is the graph gradient operator and div is the graph divergence operator.

- Special case: Assume $[F(X,t)]_{ij} = \frac{1}{d_i}$ where $d_i = \sum_{j=1}^n w_{ij}$ then

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = -\tilde{L}_n X(t)$$

  where $\tilde{L} = \mathrm{Id} - D^{-1}W = D^{-1}L_n$ is the random walk Laplacian.

Graph Diffusions

- Let $X(t) = [x_1(t)^\top, x_2(t)^\top, \ldots, x_n(t)^\top]^\top \in \mathbb{R}^{n \times d}$ satisfy

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = \mathrm{div}(F(X(t), t) \odot \nabla X(t))$$

  where $F : \mathbb{R}^{n \times d} \times [0, \infty) \to \mathbb{R}^{n \times n}$ is a given function, ∇ is the graph gradient operator and div is the graph divergence operator.

- Special case: Assume $[F(X, t)]_{ij} = \frac{1}{d_i}$ where $d_i = \sum_{j=1}^{n} w_{ij}$ then

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = -\tilde{L}_n X(t)$$

  where $\tilde{L} = \mathrm{Id} - D^{-1}W = D^{-1}L_n$ is the random walk Laplacian.

- Remark: This is the gradient flow corresponding to minimising a Dirichlet energy (without constraints). In particular, $x_i(t) \to c \in \mathbb{R}^d$, as $t \to \infty$, for all $i = 1, 2, \ldots, n$.

GRAph Neural Diffusion (GRAND) networks were proposed by Chamberlain et. al.[2] as a architecture for graph neural networks.

The architecture is based on

$$X(T) = X(0) + \int_0^T \frac{\mathrm{d}X}{\mathrm{d}t}(t)\,\mathrm{d}t$$

where

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = \mathrm{div}(F(X(t), t) \odot \nabla X(t))$$

and the parameter values that define F are to be learned.

[2] Chamberlain, Rowbottom, Gorinova, Bronstein, Webb and Rossi, *GRAND: Graph neural diffusion*, ICML, 2021, pp. 1407–1418.

Random Walk Viewpoint of GRAND

- We consider the (slightly modified) random walk B_t^x on $\{x_i(0)\}_{i=1}^n$

$$B_0^x = x \in \{x_i(0)\}_{i=1}^n$$
$$\mathbb{P}\left(B_{t+1}^x = x_j(0) | B_t^x = x_i(0)\right) = \begin{cases} 1 - \delta_t & \text{if } i = j \\ \frac{\delta_t W_{ij}}{d_i} & \text{if } i \neq j \end{cases}$$

Random Walk Viewpoint of GRAND

- We consider the (slightly modified) random walk B_t^x on $\{x_i(0)\}_{i=1}^n$

$$B_0^x = x \in \{x_i(0)\}_{i=1}^n$$

$$\mathbb{P}\left(B_{t+1}^x = x_j(0) | B_t^x = x_i(0)\right) = \begin{cases} 1-\delta_t & \text{if } i = j \\ \frac{\delta_t W_{ij}}{d_i} & \text{if } i \neq j \end{cases}$$

- **Result:** Let $X(t) = [x_1(t)^\top, x_2(t)^\top, \ldots, x_n(t)^\top]^\top$ solve

$$X(0) = [\overline{x}_1^\top, \overline{x}_2^\top, \ldots, \overline{x}_n^\top]^\top$$

$$X(k\delta_t) = X((k-1)\delta_t) - \delta_t \tilde{L}_n X((k-1)\delta_t).$$

Then,

$$x_i(k\delta_t) = \mathbb{E}[B_k^{\overline{x}_i}].$$

Random Walk Viewpoint of GRAND

- We consider the (slightly modified) random walk B_t^x on $\{x_i(0)\}_{i=1}^n$

$$B_0^x = x \in \{x_i(0)\}_{i=1}^n$$
$$\mathbb{P}\left(B_{t+1}^x = x_j(0) | B_t^x = x_i(0)\right) = \begin{cases} 1 - \delta_t & \text{if } i = j \\ \frac{\delta_t W_{ij}}{d_i} & \text{if } i \neq j \end{cases}$$

- **Result:** Let $X(t) = [x_1(t)^\top, x_2(t)^\top, \dots, x_n(t)^\top]^\top$ solve

$$X(0) = [\overline{x}_1^\top, \overline{x}_2^\top, \dots, \overline{x}_n^\top]^\top$$
$$X(k\delta_t) = X((k-1)\delta_t) - \delta_t \tilde{L}_n X((k-1)\delta_t).$$

Then,

$$x_i(k\delta_t) = \mathbb{E}[B_k^{\overline{x}_i}].$$

- **Result:** As $k \to \infty$

$$x_i(k\delta_t) \to \tilde{x} := \sum_{j=1}^n \overline{x}_j \pi_j, \qquad \pi_j = \frac{d_j}{\sum_{i=1}^n d_i}.$$

- **Problem:** In GRAND we suffer from the oversmoothing phenomena.

GRAND++

- **Problem:** In GRAND we suffer from the oversmoothing phenomena.
- **Solution:** Add a source term: let $Z(t) = [z_1(t)^\top, z_2(t)^\top, \ldots, z_n(t)^\top]^\top$ solve

$$\frac{\mathrm{d}z_i}{\mathrm{d}t}(t) = [\mathrm{div}(F(Z(t), t) \odot \nabla Z(t))]_i + \sum_{j \in \mathcal{I}_n} \delta_{ij} C_j$$

where C_j is the source added at nodes $j \in \mathcal{I}_n$.

GRAND++

- **Problem:** In GRAND we suffer from the oversmoothing phenomena.
- **Solution:** Add a source term: let $Z(t) = [z_1(t)^\top, z_2(t)^\top, \ldots, z_n(t)^\top]^\top$ solve

$$\frac{\mathrm{d}z_i}{\mathrm{d}t}(t) = [\mathrm{div}(F(Z(t), t) \odot \nabla Z(t))]_i + \sum_{j \in \mathcal{I}_n} \delta_{ij} C_j$$

  where C_j is the source added at nodes $j \in \mathcal{I}_n$.
- We choose

$$C_j = \overline{x}_j - \hat{x}, \qquad \hat{x} = \frac{1}{|\mathcal{I}_n|} \sum_{j \in \mathcal{I}_n} \overline{x}_j.$$

GRAND++

- **Problem:** In GRAND we suffer from the oversmoothing phenomena.
- **Solution:** Add a source term: let $Z(t) = [z_1(t)^\top, z_2(t)^\top, \ldots, z_n(t)^\top]^\top$ solve

$$\frac{\mathrm{d}z_i}{\mathrm{d}t}(t) = [\mathrm{div}(F(Z(t), t) \odot \nabla Z(t))]_i + \sum_{j \in \mathcal{I}_n} \delta_{ij} C_j$$

  where C_j is the source added at nodes $j \in \mathcal{I}_n$.
- We choose

$$C_j = \overline{x}_j - \hat{x}, \qquad \hat{x} = \frac{1}{|\mathcal{I}_n|} \sum_{j \in \mathcal{I}_n} \overline{x}_j.$$

- GRAph Neural Diffusion with source (GRAND++) is based on this architecture.

Random Walk Viewpoint of GRAND++

Assume that the initial condition satisfies

$$\sum_{i=1}^{n} z_i(0) = \sum_{i \in \mathcal{I}_n} \frac{1}{d_i} \left(\overline{x}_i - \hat{x} \right).$$

Random Walk Viewpoint of GRAND++

Assume that the initial condition satisfies

$$\sum_{i=1}^{n} z_i(0) = \sum_{i \in \mathcal{I}_n} \frac{1}{d_i} \left(\overline{x}_i - \hat{x}\right).$$

Theorem (T., Nguyen, Xia, Strohmer, Bertozzi, Osher and Wang (2021))

Let $Z(t) = [z_1(t)^\top, z_2(t)^\top, \ldots, z_n(t)^\top]^\top$ *solve*

$$z_i(k\delta_t) = z_i((k-1)\delta_t) - \delta_t[\tilde{L}_n Z((k-1)\delta_t)]_i + \sum_{j \in \mathcal{I}_n} \delta_{ij} \left(\overline{x}_j - \hat{x}\right).$$

Then,

$$\left| z_i(k\delta_t) - \mathbb{E}\left[\sum_{s=0}^{k} \frac{1}{d_i} \sum_{j \in \mathcal{I}_n} \left(\overline{x}_j - \hat{x}\right) \mathbb{1}_{B_s^{\overline{x}_j} = \overline{x}_i} \right] \right| \to 0$$

as $k \to \infty$.

"Deep" Layer Results

	Depth	GRAND-nl	GRAND-nl-rw	GRAND++-nl	GRAND++-nl-rw
CORA	1	**79.70 (1.88)**	79.07 (3.05)	79.24 (1.48)	79.24 (1.48)
	4	82.31 (0.91)	82.47 (1.32)	**82.64 (0.89)**	82.23 (1.14)
	16	82.11 (1.42)	82.05 (1.31)	**83.24 (0.20)**	81.48 (1.07)
	32	79.42 (0.64)	81.01 (0.81)	81.21 (0.37)	**82.20 (1.15)**
CiteSeer	1	71.84 (2.98)	**71.84 (2.66)**	70.45 (2.12)	71.74 (1.37)
	16	72.65 (2.42)	73.06 (2.98)	72.48 (1.10)	**73.29 (1.37)**
	64	70.29 (2.58)	69.65 (2.50)	72.64 (0.93)	**73.38 (0.95)**
	128	65.19 (6.77)	65.45 (7.18)	**74.24 (0.70)**	74.23 (0.70)
PubMed	1	77.93 (1.27)	77.93 (1.26)	**78.01 (0.68)**	78.01 (0.68)
	4	77.95 (1.28)	78.02 (1.14)	**78.41 (0.88)**	78.17 (0.93)
	16	76.51 (2.73)	76.88 (2.57)	**78.43 (0.78)**	78.12 (0.87)

Table: Classification accuracy of GRAND and GRAND++ variants of different depth trained with 20 labels per class. (Unit: %)

Low Label Rate Results

Model	Labels/Class	CORA	CiteSeer	PubMed	CoauthorCS	Computer	Photo
GRAND++-l	1	**54.94 (16.09)**	**58.95 (9.59)**	**65.94 (4.87)**	60.30 (1.50)	**67.65 (0.37)**	**83.12 (0.78)**
	2	**66.92 (10.04)**	**64.98 (8.31)**	**69.31 (4.87)**	76.53 (1.85)	76.47 (1.48)	**83.71 (0.90)**
	5	**77.80 (4.46)**	**70.03 (3.63)**	71.99 (1.91)	84.83 (0.84)	**82.64 (0.56)**	88.33 (1.21)
	10	**80.86 (2.99)**	**72.34 (2.42)**	75.13 (3.88)	86.94 (0.46)	**82.99 (0.81)**	90.65 (1.19)
	20	**82.95 (1.37)**	73.53 (3.31)	**79.16 (1.37)**	90.80 (0.34)	**85.73 (0.50)**	**93.55 (0.38)**
GRAND-l	1	52.53 (16.40)	50.06 (17.98)	62.11 (10.58)	59.15 (5.73)	48.67 (1.66)	81.25 (2.50)
	2	64.82 (11.16)	59.55 (10.89)	69.00 (7.55)	73.83 (5.58)	74.77 (1.85)	82.13 (3.27)
	5	76.07 (5.08)	68.37 (5.00)	**73.98 (5.08)**	85.29 (2.19)	80.72 (1.09)	88.27 (1.94)
	10	80.25 (3.40)	71.90 (7.66)	**76.33 (3.41)**	87.81 (1.36)	82.42 (1.10)	**90.98 (0.93)**
	20	82.86 (2.39)	73.02 (5.89)	78.76 (1.69)	91.03 (0.47)	84.54 (0.90)	93.53 (0.47)
GCN	1	47.72 (15.33)	48.94 (10.24)	58.61 (12.83)	**65.22 (2.25)**	49.46 (1.65)	82.94 (2.17)
	2	60.85 (14.01)	58.06 (9.76)	60.45 (16.20)	**83.61 (1.49)**	**76.90 (1.49)**	83.61 (0.71)
	5	73.86 (7.97)	67.24 (4.19)	68.69 (7.93)	86.66 (0.43)	82.47 (0.97)	**88.86 (1.56)**
	10	78.82 (5.38)	72.18 (3.47)	72.59 (3.19)	88.60 (0.50)	82.53 (0.74)	90.41 (0.35)
	20	82.07 (2.03)	**74.21 (2.90)**	76.89 (3.27)	91.09 (0.35)	82.94 (1.54)	91.95 (0.11)
GAT	1	47.86 (15.38)	50.31 (14.27)	58.84 (12.81)	51.13 (5.24)	37.14 (7.81)	73.58 (8.15)
	2	58.30 (13.55)	55.55 (9.19)	60.24 (14.44)	63.12 (6.09)	65.07 (8.86)	76.89 (4.89)
	5	71.04 (5.74)	67.37 (5.08)	68.54 (5.75)	71.65 (4.53)	71.43 (7.34)	83.01 (3.64)
	10	76.31 (4.87)	71.35 (4.92)	72.44 (3.50)	74.71 (3.35)	76.04 (0.35)	87.42 (2.38)
	20	79.92 (2.28)	73.22 (2.90)	75.55 (4.11)	79.95 (2.88)	80.05 (1.81)	89.38 (2.48)
GraphSage	1	43.04 (14.01)	48.81 (11.45)	55.53 (12.71)	61.35 (1.35)	27.65 (2.39)	45.36 (7.13)
	2	53.96 (12.18)	54.39 (11.37)	58.97 (12.65)	76.51 (1.31)	42.63 (4.29)	51.93 (4.21)
	5	68.14 (6.95)	64.79 (5.16)	66.07 (6.16)	**89.06 (0.69)**	64.83 (1.62)	78.26 (1.93)
	10	75.04 (5.03)	68.90 (5.08)	70.74 (3.11)	**89.68 (0.39)**	74.66 (1.29)	84.38 (1.75)
	20	80.04 (2.54)	72.02 (2.82)	74.55 (3.09)	**91.33 (0.36)**	79.98 (0.96)	91.29 (0.67)
MoNet	1	47.72 (15.53)	39.13 (11.37)	56.47 (4.67)	58.99 (5.17)	23.78 (7.57)	34.72 (8.18)
	2	60.85 (14.01)	48.52 (9.52)	61.03 (6.93)	76.57 (4.06)	38.19 (3.72)	43.03 (8.22)
	5	73.86 (7.97)	61.66 (6.61)	67.92 (2.50)	87.02 (1.67)	59.38 (4.73)	71.80 (5.02)
	10	78.82 (5.38)	68.08 (6.29)	71.24 (1.54)	88.76 (0.49)	68.66 (3.30)	78.66 (3.17)
	20	82.07 (2.03)	71.52 (4.11)	76.49 (1.75)	90.31 (0.41)	73.66 (2.87)	88.61 (1.18)

Table: Classification accuracy of different GNNs trained with different number of labelled data per class (#per class) on six benchmark graph node classification tasks. (Unit: %)

Thank you for listening!

In theory, there is no difference between theory and practice. But in practice, there is.

— Yogi Berra