Continuum Limits in Semi-Supervised Learning

Geometric and Topological Approaches to Data Analysis

Matthew Thorpe

Cantab Capital Institute for the Mathematics of Information
Department of Applied Mathematics and Theoretical Physics
University of Cambridge

13th June 2019

Semi-Supervised Learning

1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.

Semi-Supervised Learning

1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.

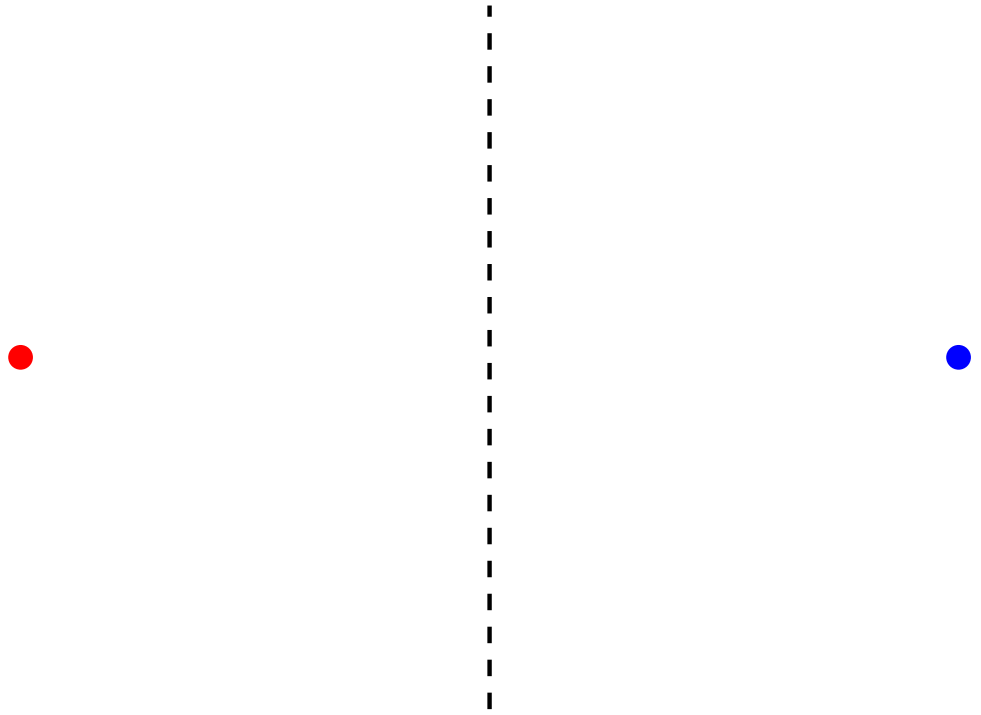
Semi-Supervised Learning

1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.

Example by Jeff Calder.

Semi-Supervised Learning

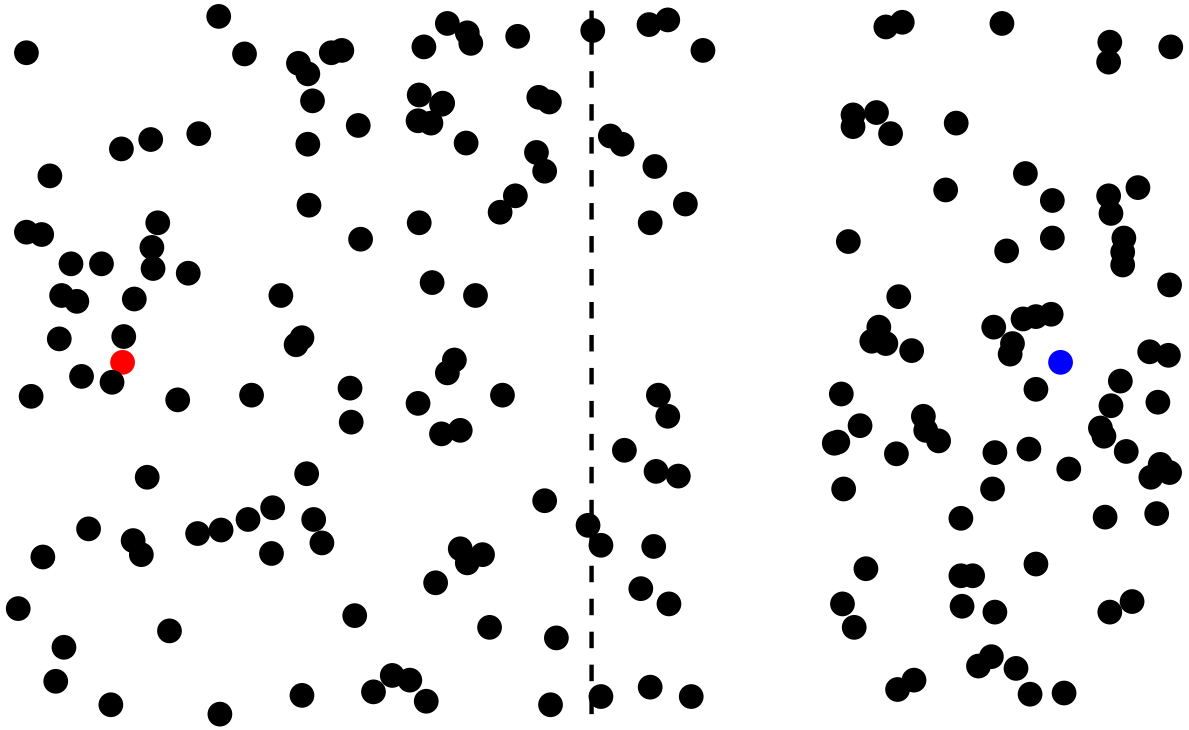
1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.



Example by Jeff Calder.

Semi-Supervised Learning

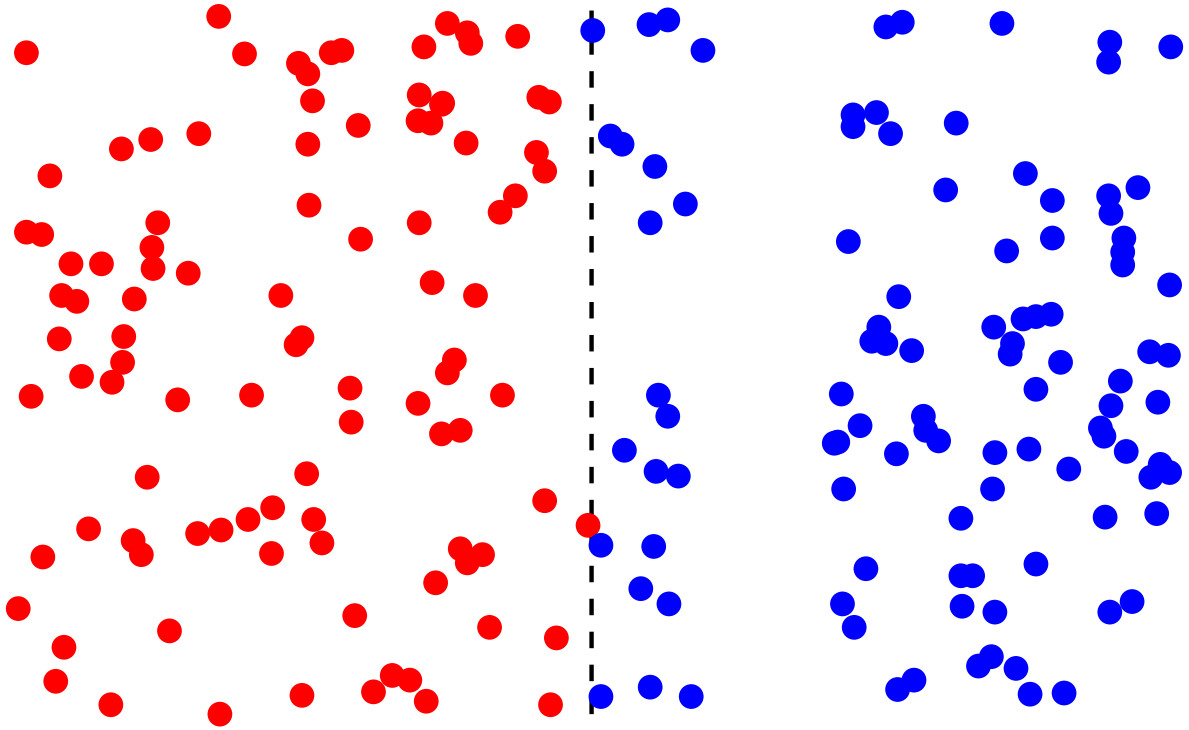
1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.



Example by Jeff Calder.

Semi-Supervised Learning

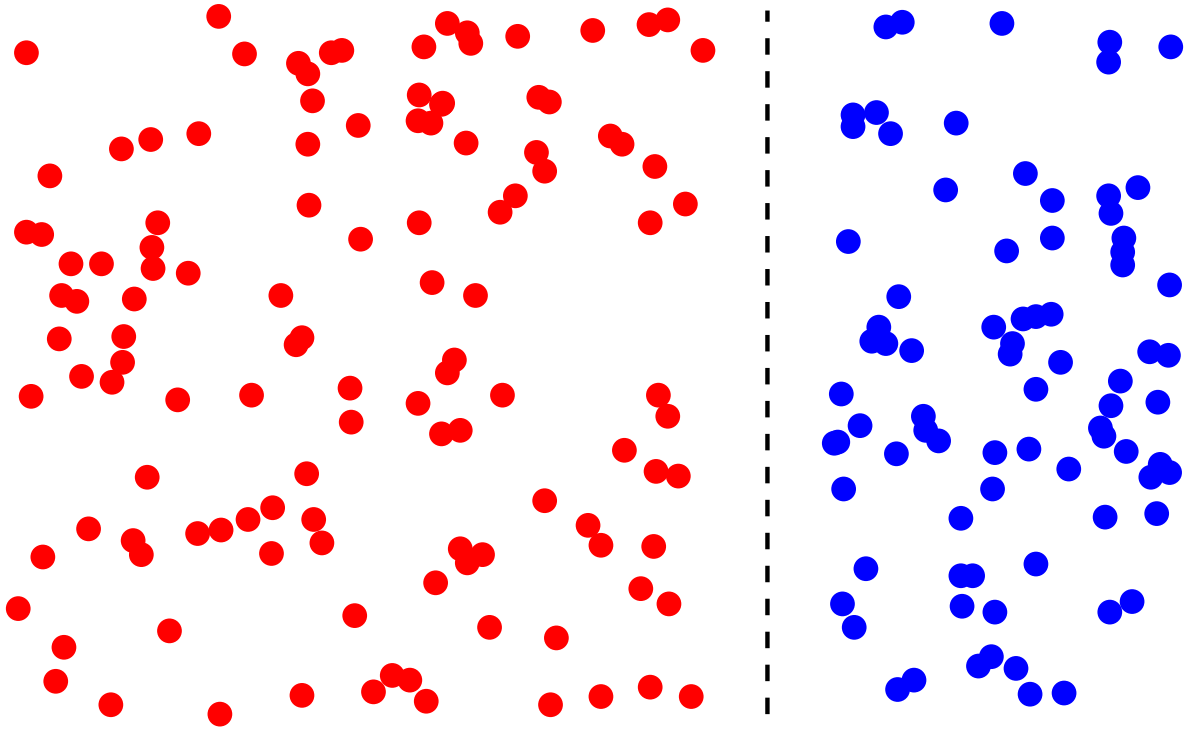
1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.



Example by Jeff Calder.

Semi-Supervised Learning

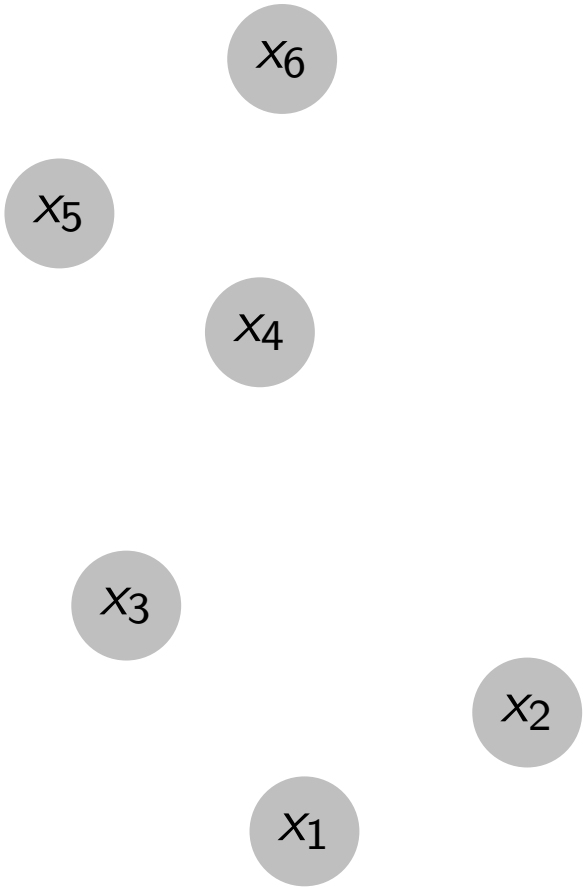
1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.



Example by Jeff Calder.

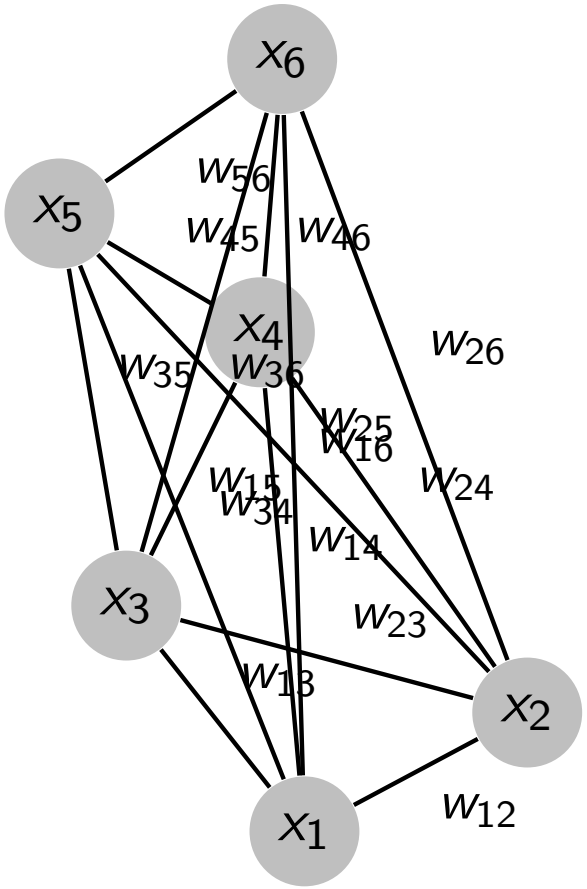
Capturing Geometry: Graphs

1. Let $X_n = \{x_i\}_{i=1}^n$, where $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(X)$, $X \subset \mathbb{R}^d$, be a point cloud that forms the nodes of a graph.



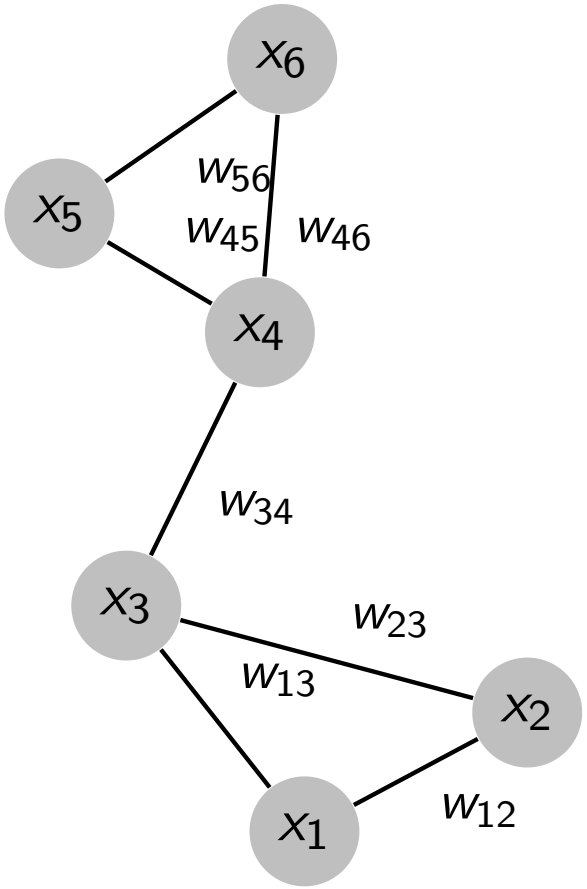
Capturing Geometry: Graphs

1. Let $X_n = \{x_i\}_{i=1}^n$, where $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(X)$, $X \subset \mathbb{R}^d$, be a point cloud that forms the nodes of a graph.
2. Place edges between all vertices with weights $w_{ij} = \eta_\epsilon(|x_i - x_j|)$, where $\eta_\varepsilon = \frac{1}{\varepsilon^d}\eta(\cdot/\varepsilon)$, e.g. $\eta(|x|) = \mathbb{I}_{|x|\leq 1}$.



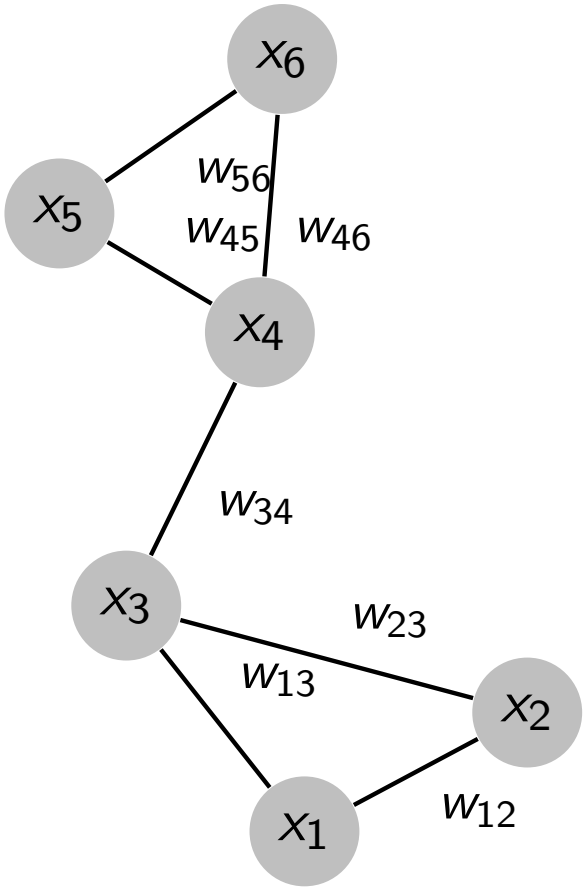
Capturing Geometry: Graphs

1. Let $X_n = \{x_i\}_{i=1}^n$, where $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(X)$, $X \subset \mathbb{R}^d$, be a point cloud that forms the nodes of a graph.
2. Place edges between all vertices with weights $w_{ij} = \eta_\epsilon(|x_i - x_j|)$, where $\eta_\varepsilon = \frac{1}{\varepsilon^d}\eta(\cdot/\varepsilon)$, e.g. $\eta(|x|) = \mathbb{I}_{|x|\leq 1}$.
3. Keep edges with weights $w_{ij} > 0$.



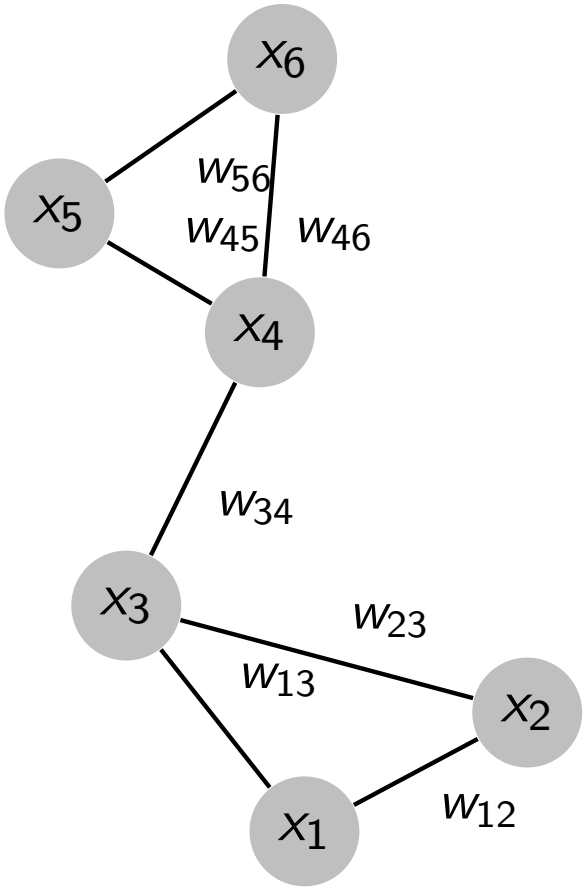
Capturing Geometry: Graphs

1. Let $X_n = \{x_i\}_{i=1}^n$, where $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(X)$, $X \subset \mathbb{R}^d$, be a point cloud that forms the nodes of a graph.
2. Place edges between all vertices with weights $w_{ij} = \eta_\epsilon(|x_i - x_j|)$, where $\eta_\varepsilon = \frac{1}{\varepsilon^d}\eta(\cdot/\varepsilon)$, e.g. $\eta(|x|) = \mathbb{I}_{|x|\leq 1}$.
3. Keep edges with weights $w_{ij} > 0$.
4. A lower bound is needed on $\varepsilon = \varepsilon_n$ in order for the graph to be connected.



Capturing Geometry: Graphs

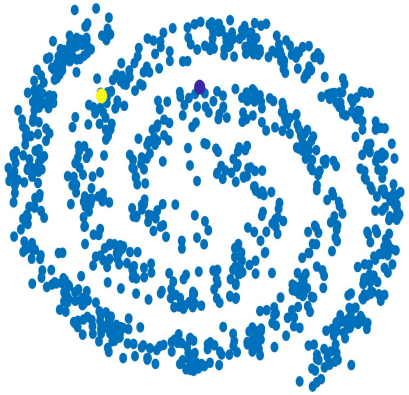
1. Let $X_n = \{x_i\}_{i=1}^n$, where $x_i \overset{\text{iid}}{\sim} \mu \in \mathcal{P}(X)$, $X \subset \mathbb{R}^d$, be a point cloud that forms the nodes of a graph.
2. Place edges between all vertices with weights $w_{ij} = \eta_\epsilon(|x_i - x_j|)$, where $\eta_\varepsilon = \frac{1}{\varepsilon^d}\eta(\cdot/\varepsilon)$, e.g. $\eta(|x|) = \mathbb{I}_{|x|\leq 1}$.
3. Keep edges with weights $w_{ij} > 0$.
4. A lower bound is needed on $\varepsilon = \varepsilon_n$ in order for the graph to be connected.



Theorem (Penrose, 03)

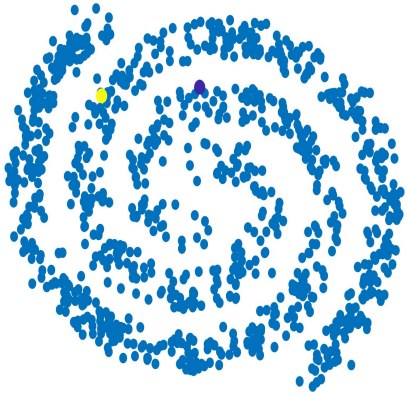
If $R(n)$ is the connectivity radius of the graph then $R(n) \sim \sqrt[d]{\frac{\log n}{n}}$.

Sensitivity to ε

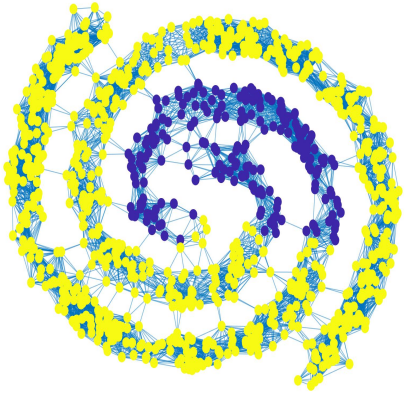


(a) Data.

Sensitivity to ε

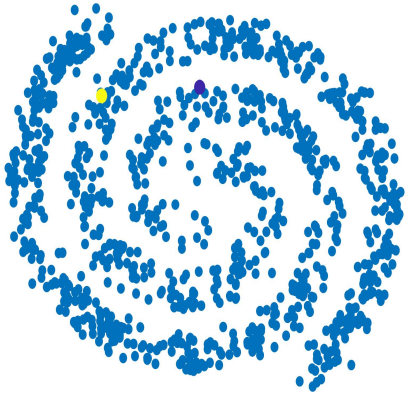


(a) Data.

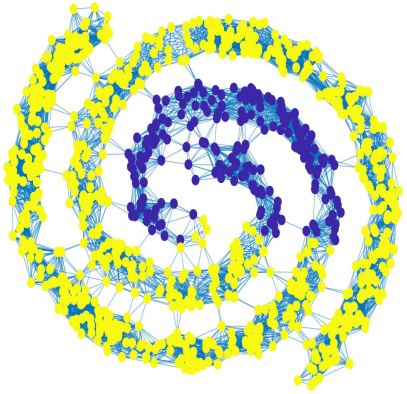


(b) Large ε.

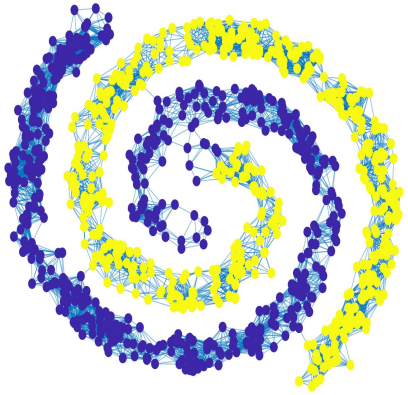
Sensitivity to ε



(a) Data.



(b) Large ε.



(c) Small ε.

p-Laplacian Regularisation

1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.

p-Laplacian Regularisation

1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.
3. p-**Laplacian regularisation:** (Zhu, Ghahramani and Lafferty 03, and Zhou and Schölkopf 05)

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

p-Laplacian Regularisation

1. **Problem:** Given data $X_n = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and a subset of labels $\{y_i\}_{i \in Z_n} \subset \mathbb{R}$, where $Z_n \subseteq \{1, \ldots, n\}$, find the 'best' $f_n : X_n \to \mathbb{R}$ such that $f_n(x_i) = y_i$ for all $i \in Z_n$.
2. **'Classical' approach:** minimise $\|\nabla f_n\|_{L^2}^2$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.
3. p-**Laplacian regularisation:** (Zhu, Ghahramani and Lafferty 03, and Zhou and Schölkopf 05)

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

4. **Variational Problem:** Minimise $\mathcal{E}_n^{(p)}$ over $f_n : X_n \to \mathbb{R}$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$.

p-Laplacian Regularisation: Formal Derivation of Limit

The following formal calculation gives intuition as to what we should expect. We assume $x_i \overset{\text{iid}}{\sim} \mu$ and μ has density ρ.

p-Laplacian Regularisation: Formal Derivation of Limit

The following formal calculation gives intuition as to what we should expect. We assume $x_i \overset{\text{iid}}{\sim} \mu$ and μ has density ρ.

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

p-Laplacian Regularisation: Formal Derivation of Limit

The following formal calculation gives intuition as to what we should expect. We assume $x_i \overset{\text{iid}}{\sim} \mu$ and μ has density ρ.

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

p-Laplacian Regularisation: Formal Derivation of Limit

The following formal calculation gives intuition as to what we should expect. We assume $x_i \overset{\text{iid}}{\sim} \mu$ and μ has density ρ.

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

p-Laplacian Regularisation: Formal Derivation of Limit

The following formal calculation gives intuition as to what we should expect. We assume $x_i \overset{\text{iid}}{\sim} \mu$ and μ has density ρ.

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

p-Laplacian Regularisation: Formal Derivation of Limit

The following formal calculation gives intuition as to what we should expect. We assume $x_i \overset{\text{iid}}{\sim} \mu$ and μ has density ρ.

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

where

$$
\sigma_\eta = \int_{\mathbb{R}^d} \eta(|z|) |z_1|^p \, \mathrm{d}z.
$$

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No!

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

- Problems occur when spikes have low energy.

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

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

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

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

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

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

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

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

Passing Constraints to the Limit

If we fix the number of training data points, say $|Z_n| = N$, we see that $p > d$ is necessary for constraints to respected (i.e. constraints for the finite data problem pass to the limit), is it sufficient?

No! Why not?

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

This elementary 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.

Discrete-To-Continuum Topologies

- Before we can say more we need a topology in which we can say $f_n \to f$ where $f_n : X_n \to \mathbb{R}$ and $f : X \to \mathbb{R}$.

Discrete-To-Continuum Topologies

- Before we can say more we need a topology in which we can say $f_n \to f$ where $f_n : X_n \to \mathbb{R}$ and $f : X \to \mathbb{R}$.
- The idea is to define a space which contains both discrete and continuum functions; we do this by treating the coupling (f_n, μ_n) and (f, μ) where $f_n \in L^p(\mu_n)$ and $f \in L^p(\mu)$.

Discrete-To-Continuum Topologies

- Before we can say more we need a topology in which we can say $f_n \to f$ where $f_n : X_n \to \mathbb{R}$ and $f : X \to \mathbb{R}$.
- The idea is to define a space which contains both discrete and continuum functions; we do this by treating the coupling (f_n, μ_n) and (f, μ) where $f_n \in L^p(\mu_n)$ and $f \in L^p(\mu)$.
- Let
$$TL^p := \{(f, \mu) \,:\, f \in L^p(\mu), \mu \in \mathcal{P}(X)\}\,.$$

Discrete-To-Continuum Topologies

- Before we can say more we need a topology in which we can say $f_n \to f$ where $f_n : X_n \to \mathbb{R}$ and $f : X \to \mathbb{R}$.
- The idea is to define a space which contains both discrete and continuum functions; we do this by treating the coupling (f_n, μ_n) and (f, μ) where $f_n \in L^p(\mu_n)$ and $f \in L^p(\mu)$.
- Let
$$TL^p := \{(f, \mu) \,:\, f \in L^p(\mu), \mu \in \mathcal{P}(X)\}\,.$$
- The TL^p distance is the defined for $(f, \mu), (g, \nu) \in TL^p$ by
$$d^p_{TL^p}((f, \mu), (g, \nu)) = \inf_{\{T : T_{\#}\mu = \nu\}} \int_X |x - T(x)|^p + |f(x) - g(T(x))|^p \, \mathrm{d}\mu(x)$$

Discrete-To-Continuum Topologies

- Before we can say more we need a topology in which we can say $f_n \to f$ where $f_n : X_n \to \mathbb{R}$ and $f : X \to \mathbb{R}$.
- The idea is to define a space which contains both discrete and continuum functions; we do this by treating the coupling (f_n, μ_n) and (f, μ) where $f_n \in L^p(\mu_n)$ and $f \in L^p(\mu)$.
- Let
$$TL^p := \{(f, \mu) \,:\, f \in L^p(\mu), \mu \in \mathcal{P}(X)\}\,.$$
- The TL^p distance is the defined for $(f, \mu), (g, \nu) \in TL^p$ by
$$d^p_{TL^p}((f, \mu), (g, \nu)) = \inf_{\{T : T_{\#}\mu = \nu\}} \int_X |x - T(x)|^p + |f(x) - g(T(x))|^p \, \mathrm{d}\mu(x)$$
- Compare this to the Wasserstein distance
$$d^p_W(\mu, \nu) = \inf_{\{T : T_{\#}\mu = \nu\}} \int_X |x - T(x)|^p \, \mathrm{d}\mu(x).$$

Discrete-To-Continuum Topologies

- Before we can say more we need a topology in which we can say $f_n \to f$ where $f_n : X_n \to \mathbb{R}$ and $f : X \to \mathbb{R}$.
- The idea is to define a space which contains both discrete and continuum functions; we do this by treating the coupling

Proposition (García Trillos and Slepčev 16)

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

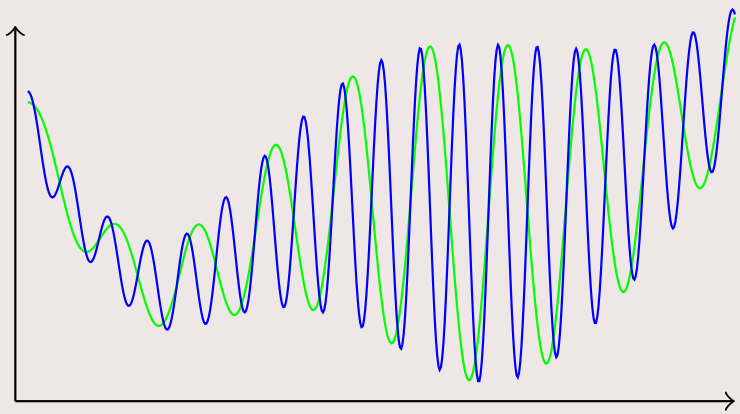
$$\|f_n \circ T_n - f\|_{L^p(\mu)} \to 0.$$

$$d_{TL^p}((f,\mu),(g,\nu)) = \inf_{\{T : T_{\#}\mu = \nu\}} \int_X |x - T(x)|^p + |f(x) - g(T(x))|^p \, \mathrm{d}\mu(x)$$

- Compare this to the Wasserstein distance

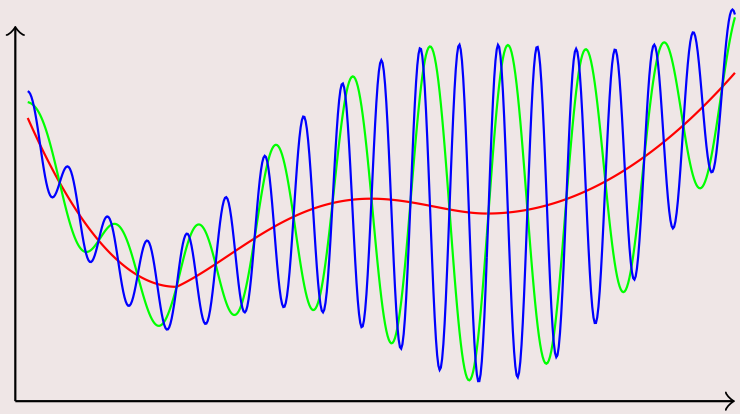
$$d_W^p(\mu, \nu) = \inf_{\{T : T_{\#}\mu = \nu\}} \int_X |x - T(x)|^p \, \mathrm{d}\mu(x).$$

Γ-Convergence



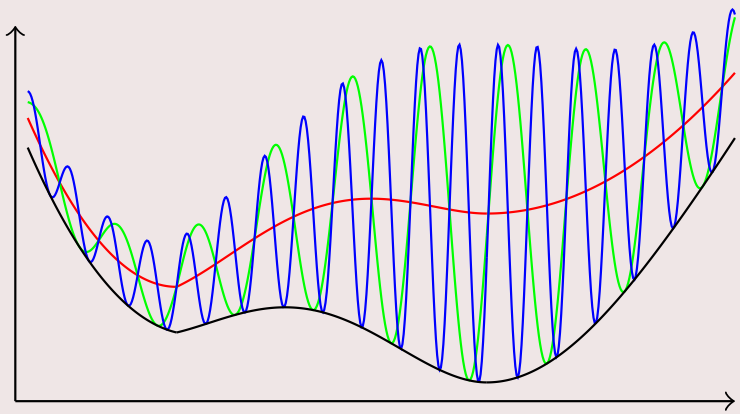
Green - $\mathcal{F}_n$, Blue - $\mathcal{F}_m$ for $m > n$

Γ-Convergence



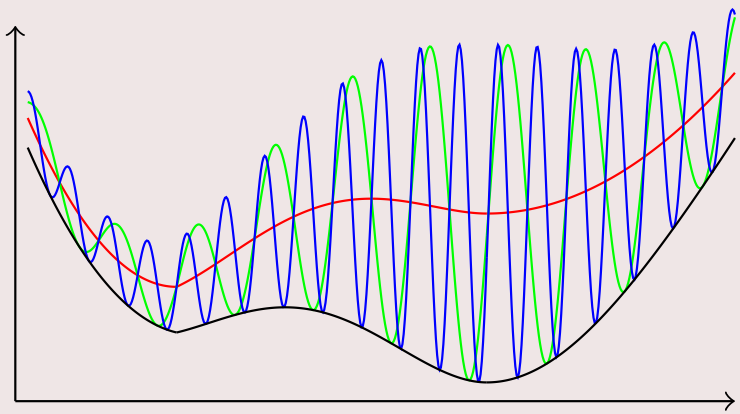
Green - $\mathcal{F}_n$, Blue - $\mathcal{F}_m$ for $m > n$, Red - weak limit

Γ-Convergence



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

Γ-Convergence



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

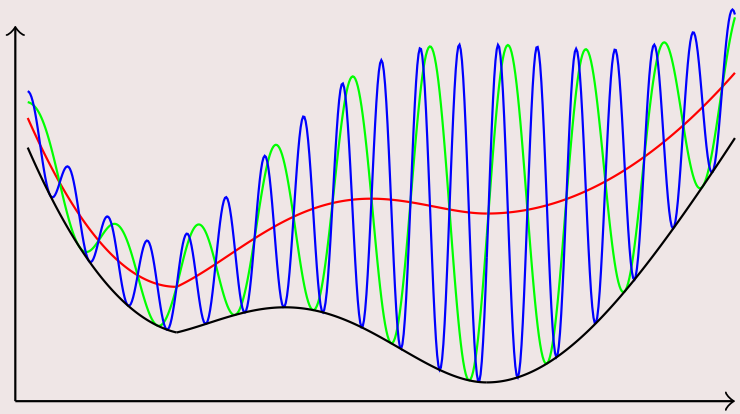
Definition

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

(i) $\forall\, f_n \to f$,
$\mathcal{F}_\infty(f) \leq \liminf_{n\to\infty} \mathcal{F}_n(f_n)$;

(ii) $\exists\, f_n \to f$,
$\mathcal{F}_\infty(f) \geq \limsup_{n\to\infty} \mathcal{F}_n(f_n)$.

Γ-Convergence



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

Definition

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

(i) $\forall f_n \to f$,
$\mathcal{F}_\infty(f) \leq \liminf_{n\to\infty} \mathcal{F}_n(f_n)$;

(ii) $\exists f_n \to f$,
$\mathcal{F}_\infty(f) \geq \limsup_{n\to\infty} \mathcal{F}_n(f_n)$.

Theorem

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

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

Furthermore any cluster point of f_n minimizes $\mathcal{F}_\infty$.

p-Laplacian Regularisation: Results

Theorem (Slepčev and T. 17)

Let $p > 1$ and assume ε_n satisfies a lower bound. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints where $Z_n = \{1, \ldots, N\}$ for N fixed. Then, almost surely, f_n converges in TL^p along subsequences to some $f \in W^{1,p}(X)$. Furthermore,

p-Laplacian Regularisation: Results

Theorem (Slepčev and T. 17)

Let $p > 1$ and assume ε_n satisfies a lower bound. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints where $Z_n = \{1, \ldots, N\}$ for N fixed. Then, almost surely, f_n converges in TL^p along subsequences to some $f \in W^{1,p}(X)$. Furthermore,

(i) **(well-posed regime)** *if $n\varepsilon_n^p \to 0$ as $n \to \infty$ then*

p-Laplacian Regularisation: Results

Theorem (Slepčev and T. 17)

Let $p > 1$ and assume ε_n satisfies a lower bound. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints where $Z_n = \{1, \ldots, N\}$ for N fixed. Then, almost surely, f_n converges in TL^p along subsequences to some $f \in W^{1,p}(X)$. Furthermore,

(i) **(well-posed regime)** *if $n\varepsilon_n^p \to 0$ as $n \to \infty$ then*

(a) *the whole sequence f_n converges to f locally uniformly, meaning that for any X' with $\overline{X'} \subset X$*

$$\lim_{n\to\infty} \max_{\{k \le n \,:\, x_k \in X'\}} |f(x_k) - f_n(x_k)| = 0,$$

p-Laplacian Regularisation: Results

Theorem (Slepčev and T. 17)

Let $p > 1$ and assume ε_n satisfies a lower bound. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints where $Z_n = \{1, \ldots, N\}$ for N fixed. Then, almost surely, f_n converges in TL^p along subsequences to some $f \in W^{1,p}(X)$. Furthermore,

(i) **(well-posed regime)** *if $n\varepsilon_n^p \to 0$ as $n \to \infty$ then*

(a) *the whole sequence f_n converges to f locally uniformly, meaning that for any X' with $\overline{X'} \subset X$*

$$\lim_{n\to\infty} \max_{\{k \le n \,:\, x_k \in X'\}} |f(x_k) - f_n(x_k)| = 0,$$

(b) *f is a minimizer of $\mathcal{E}_\infty^{(p)}$ subject to the constraints;*

p-Laplacian Regularisation: Results

Theorem (Slepčev and T. 17)

Let $p > 1$ and assume ε_n satisfies a lower bound. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints where $Z_n = \{1, \ldots, N\}$ for N fixed. Then, almost surely, f_n converges in TL^p along subsequences to some $f \in W^{1,p}(X)$. Furthermore,

(i) **(well-posed regime)** *if $n\varepsilon_n^p \to 0$ as $n \to \infty$ then*

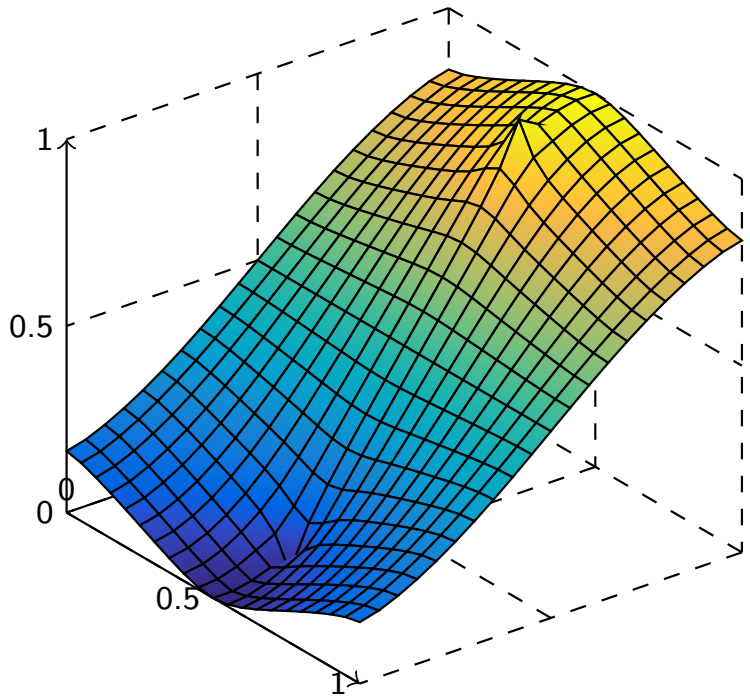
(a) *the whole sequence f_n converges to f locally uniformly, meaning that for any X' with $\overline{X'} \subset X$*

$$\lim_{n\to\infty} \max_{\{k \le n \,:\, x_k \in X'\}} |f(x_k) - f_n(x_k)| = 0,$$

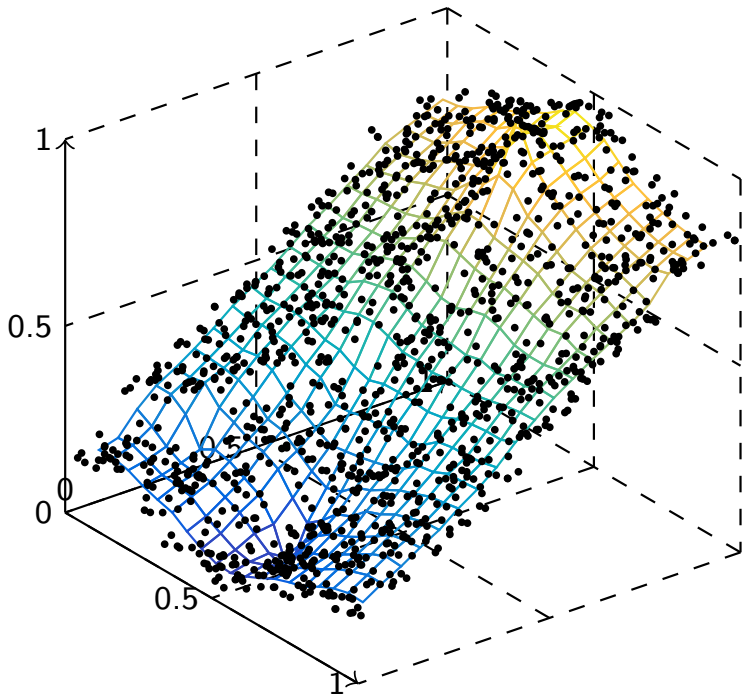
(b) *f is a minimizer of $\mathcal{E}_\infty^{(p)}$ subject to the constraints;*

(ii) **(ill-posed regime)** *if $n\varepsilon_n^p \to \infty$ as $n \to \infty$ then f is constant.*

Numerical Comparisons

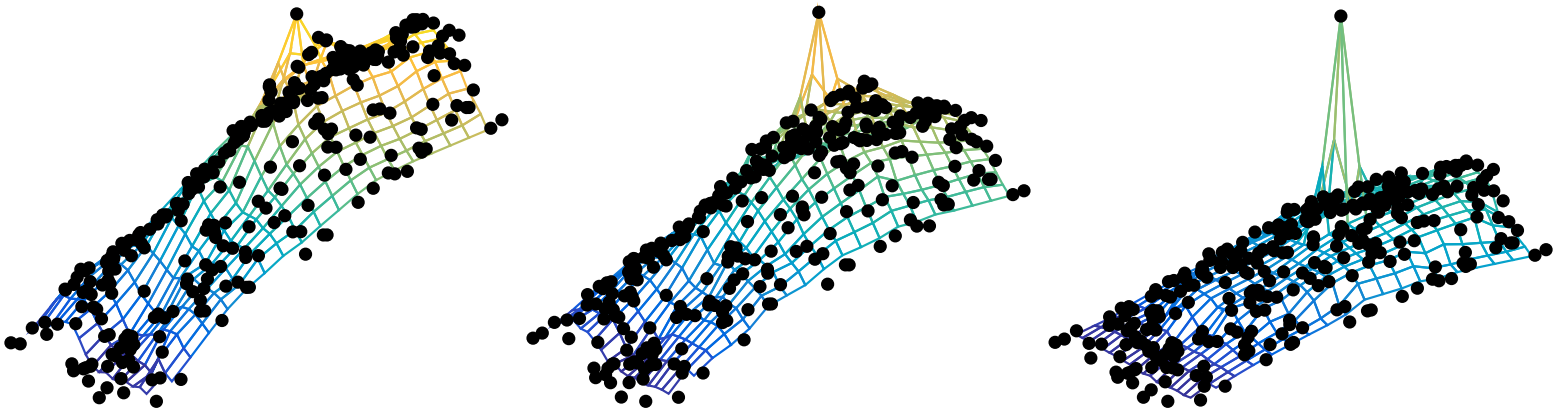


(a) $p = 4$ continuum limit minimiser.



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

Development of Spikes ($p = 4$)



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

Intuition on the Proof for Finite Training Size

1. **Step 1:** We show $\mathcal{E}_n^{(p)}(f_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{f}_n)$ where $\tilde{f}_n = f_n \circ T_n$ and J is a mollifier, $T_n : X \to X_n$.

Intuition on the Proof for Finite Training Size

1. **Step 1:** We show $\mathcal{E}_n^{(p)}(f_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{f}_n)$ where $\tilde{f}_n = f_n \circ T_n$ and J is a mollifier, $T_n : X \to X_n$.
2. **Step 2:** We show $\mathrm{osc}_{\varepsilon_n}^{(n)}(f_n) \leq C \sqrt[p]{n\varepsilon_n^p \mathcal{E}_n^{(p)}(f_n)}$ where

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

Intuition on the Proof for Finite Training Size

1. **Step 1:** We show $\mathcal{E}_n^{(p)}(f_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{f}_n)$ where $\tilde{f}_n = f_n \circ T_n$ and J is a mollifier, $T_n : X \to X_n$.
2. **Step 2:** We show $\mathrm{osc}_{\varepsilon_n}^{(n)}(f_n) \leq C \sqrt[p]{n\varepsilon_n^p \mathcal{E}_n^{(p)}(f_n)}$ where

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

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

$$\lim_{n \to \infty} \max_{\{k \leq n \,:\, x_k \in X'\}} |f(x_k) - f_n(x_k)| = 0.$$

Intuition on the Proof for Finite Training Size

1. **Step 1:** We show $\mathcal{E}_n^{(p)}(f_n) \approx \mathcal{E}_\infty^{(p)}(J_{\varepsilon_n} * \tilde{f}_n)$ where $\tilde{f}_n = f_n \circ T_n$ and J is a mollifier, $T_n : X \to X_n$.
2. **Step 2:** We show $\mathrm{osc}_{\varepsilon_n}^{(n)}(f_n) \leq C \sqrt[p]{n\varepsilon_n^p \mathcal{E}_n^{(p)}(f_n)}$ where

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

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

$$\lim_{n\to\infty} \max_{\{k \leq n \,:\, x_k \in X'\}} |f(x_k) - f_n(x_k)| = 0.$$

4. **Step 4:** Γ-convergence of $\mathcal{E}_n^{(p)}$ to $\mathcal{E}_\infty^{(p)}$ plus a TL^p compactness result is now enough to get convergence of constrained minimizers.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).
4. **This talks approach:** We take $|Z_n| \to \infty$.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).
4. **This talks approach:** We take $|Z_n| \to \infty$.
5. **Model:** Recall $\{(x_i, y_i)\}_{i \in Z_n}$ is the training data.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).
4. **This talks approach:** We take $|Z_n| \to \infty$.
5. **Model:** Recall $\{(x_i, y_i)\}_{i \in Z_n}$ is the training data.
   1. Assume $\mathbb{P}(i \in Z_n) = \beta_n$.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).
4. **This talks approach:** We take $|Z_n| \to \infty$.
5. **Model:** Recall $\{(x_i, y_i)\}_{i \in Z_n}$ is the training data.
   1. Assume $\mathbb{P}(i \in Z_n) = \beta_n$.
   2. If $i \in Z_n$ we set $y_i = g^\dagger(x_i)$.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).
4. **This talks approach:** We take $|Z_n| \to \infty$.
5. **Model:** Recall $\{(x_i, y_i)\}_{i \in Z_n}$ is the training data.
   1. Assume $\mathbb{P}(i \in Z_n) = \beta_n$.
   2. If $i \in Z_n$ we set $y_i = g^\dagger(x_i)$.
6. **Well-posed case:** Minimisers of $\mathcal{E}_n^{(p)}$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$ converge to $g^\dagger$.

Minimal Number of Labels

1. The reason we had asymptotic ill-posedness whenever $n\varepsilon_n^p \to \infty$ or $p \leq d$ is because we worked in $W^{1,p}$ with a fixed the number of constraints.
2. **Question:** How can we get a well-posed result when $p = 2$?
3. **One approach:** Increase the weight around labelled data (Calder and Slepčev 18).
4. **This talks approach:** We take $|Z_n| \to \infty$.
5. **Model:** Recall $\{(x_i, y_i)\}_{i \in Z_n}$ is the training data.
   1. Assume $\mathbb{P}(i \in Z_n) = \beta_n$.
   2. If $i \in Z_n$ we set $y_i = g^\dagger(x_i)$.
6. **Well-posed case:** Minimisers of $\mathcal{E}_n^{(p)}$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$ converge to $g^\dagger$.
7. **Ill-posed case:** Minimisers of $\mathcal{E}_n^{(p)}$ subject to $f_n(x_i) = y_i$ for all $i \in Z_n$ converge to constants.

Large Data Limits for $|Z_n| \to \infty$

Theorem (Calder, Slepčev and T., 2019 (unpublished))

Ill-Posed Regime. *Let $p > 1$ and assume ε_n satisfies the same lower bound as before. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints. Assume $\beta_n \ll \varepsilon_n^p$ and $\frac{n\varepsilon_n^p}{\log n} \gg 1$. Then, almost surely, $\{f_n\}_{n\in\mathbb{N}}$ is precompact and any convergent subsequence converges to a constant.*

Large Data Limits for $|Z_n| \to \infty$

Theorem (Calder, Slepčev and T., 2019 (unpublished))

Ill-Posed Regime. *Let $p > 1$ and assume ε_n satisfies the same lower bound as before. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints. Assume $\beta_n \ll \varepsilon_n^p$ and $\frac{n\varepsilon_n^p}{\log n} \gg 1$. Then, almost surely, $\{f_n\}_{n\in\mathbb{N}}$ is precompact and any convergent subsequence converges to a constant.*

Theorem (Calder, Slepčev and T., 2019 (unpublished))

Well-Posed Regime. *Let $p = 2$ and assume ε_n satisfies the same lower bound as before. Let f_n be a sequence of minimizers of $\mathcal{E}_n^{(p)}$ satisfying the constraints. Assume $\beta_n \gg \varepsilon_n^p$ and $\beta_n\varepsilon_n^d \gg \frac{\log n}{n}$. Then, almost surely, f_n converges to $g^\dagger$ uniformly, i.e.*

$$\max_{i=1,\dots,n} |f_n(x_i) - g^\dagger(x_i)| \to 0 \quad \text{as } n \to \infty.$$

Random Walks on Graphs

Our proof for the well-posed regime when $p = 2$ makes explicit use of the random walk interpretation of minimisers.

Random Walks on Graphs

Our proof for the well-posed regime when $p = 2$ makes explicit use of the random walk interpretation of minimisers.

1. Let $G_n = (X_n, W)$ be the graph with edge weights $W = (w_{ij})$.

Random Walks on Graphs

Our proof for the well-posed regime when $p = 2$ makes explicit use of the random walk interpretation of minimisers.

1. Let $G_n = (X_n, W)$ be the graph with edge weights $W = (w_{ij})$.
2. Let Z_t^x be the random walk on X_n starting from $Z_0^x = x \in X_n$ and transitioning with probability

$$\mathbb{P}(Z_{t+1}^x = x_k \mid Z_t^x = x_\ell) = \frac{w_{k\ell}}{d_\ell}$$

where $d_\ell = \sum_{k=1}^n w_{k\ell}$.

Random Walks on Graphs

Our proof for the well-posed regime when $p = 2$ makes explicit use of the random walk interpretation of minimisers.

1. Let $G_n = (X_n, W)$ be the graph with edge weights $W = (w_{ij})$.
2. Let Z_t^x be the random walk on X_n starting from $Z_0^x = x \in X_n$ and transitioning with probability

$$\mathbb{P}(Z_{t+1}^x = x_k \mid Z_t^x = x_\ell) = \frac{w_{k\ell}}{d_\ell}$$

   where $d_\ell = \sum_{k=1}^n w_{k\ell}$.
3. Define the stopping time

$$S(x) = \min\left\{t \in \mathbb{N} : Z_s^x \in \{x_i\}_{i \in Z_n}\right\}.$$

Random Walks on Graphs

Our proof for the well-posed regime when $p = 2$ makes explicit use of the random walk interpretation of minimisers.

1. Let $G_n = (X_n, W)$ be the graph with edge weights $W = (w_{ij})$.
2. Let Z_t^x be the random walk on X_n starting from $Z_0^x = x \in X_n$ and transitioning with probability

$$\mathbb{P}(Z_{t+1}^x = x_k \mid Z_t^x = x_\ell) = \frac{w_{k\ell}}{d_\ell}$$

   where $d_\ell = \sum_{k=1}^n w_{k\ell}$.
3. Define the stopping time

$$S(x) = \min\left\{t \in \mathbb{N} : Z_s^x \in \{x_i\}_{i \in Z_n}\right\}.$$

Proposition

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

Intuition on the Proof when $p = 2$

1. **Step 1:** We show $\mathbb{P}[S(x_i) > k | G_n] < e^{-Ck\beta_n}$ for $i \notin Z_n$ with probability at least $1 - 4ne^{-\tilde{C}n\beta_n\varepsilon_n^d}$.

Intuition on the Proof when $p = 2$

1. **Step 1:** We show $\mathbb{P}[S(x_i) > k | G_n] < e^{-Ck\beta_n}$ for $i \notin Z_n$ with probability at least $1 - 4ne^{-\tilde{C}n\beta_n\varepsilon_n^d}$.
2. **Step 2:** We show, for $M > 0$ with high probability,

$$\mathbb{P}\left[|Z_k^{x_i} - x_i| \leq cM\varepsilon_n\sqrt{k} | G_n\right] \geq 1 - 2e^{-CM^2}.$$

Intuition on the Proof when $p = 2$

1. **Step 1:** We show $\mathbb{P}[S(x_i) > k | G_n] < e^{-Ck\beta_n}$ for $i \notin Z_n$ with probability at least $1 - 4ne^{-\tilde{C}n\beta_n\varepsilon_n^d}$.
2. **Step 2:** We show, for $M > 0$ with high probability,

$$\mathbb{P}\left[|Z_k^{x_i} - x_i| \leq cM\varepsilon_n\sqrt{k} | G_n\right] \geq 1 - 2e^{-CM^2}.$$

3. **Step 3:** Combining the two previous results we show, for $\alpha > 0$ with high probability,

$$\mathbb{P}\left[|Z_{S(x_i)}^{x_i} - x_i| \leq \alpha | G_n\right] \geq 1 - 3e^{-\frac{C\alpha\sqrt{\beta_n}}{\varepsilon_n}}.$$

Intuition on the Proof when $p = 2$

1. **Step 1:** We show $\mathbb{P}[S(x_i) > k | G_n] < e^{-Ck\beta_n}$ for $i \notin Z_n$ with probability at least $1 - 4ne^{-\tilde{C}n\beta_n\varepsilon_n^d}$.
2. **Step 2:** We show, for $M > 0$ with high probability,
$$\mathbb{P}\left[|Z_k^{x_i} - x_i| \leq cM\varepsilon_n\sqrt{k}|G_n\right] \geq 1 - 2e^{-CM^2}.$$
3. **Step 3:** Combining the two previous results we show, for $\alpha > 0$ with high probability,
$$\mathbb{P}\left[|Z_{S(x_i)}^{x_i} - x_i| \leq \alpha|G_n\right] \geq 1 - 3e^{-\frac{C\alpha\sqrt{\beta_n}}{\varepsilon_n}}.$$
4. **Step 4:** If $g^\dagger$ is Lipschitz then we can show, with high probability, for all $\xi > 0$ and by choosing α optimally
$$\|f_n - g^\dagger\|_{L^\infty(X_n)} \leq \alpha + C_1 e^{-\frac{C_2\alpha\sqrt{\beta_n}}{\varepsilon_n}} \approx O\left(\frac{\varepsilon_n}{\sqrt{\beta_n}}\right).$$

Thank you for listening!

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

— Yogi Berra