Estimation of High-Dimensional Densities

Joan Bruna, Stéphane Mallat,

École Normale Supérieure
High-Dimensional Density Estimation

• Estimation $\tilde{p}(x)$ of a probability density $p(x)$ for $x \in \mathbb{R}^d$ given $n$ realizations $\{x_i\}_{i \leq n}$ of a random vector $X$.

• $p(x)$ is the space $C^1(\mathbb{R}^d)$ of Lipschitz functions if then at best $\mathbb{E}(\|p - \tilde{p}\|_2^2) = O(n^{-\frac{2}{d+2}})$

• If $d > 10$ then $n$ must be huge: impossible.

Problem:

Find regularity properties which can break the curse of dimensionality.
Markov Hypothesis

- Markov hypothesis: local conditional dependence

\[
p\left(x(u) / x(u'), u' \neq u\right) = p\left(x(u) / x(u') , u' \in N_u\right)
\]

- Hammersely-Clifford theorem proves that

\[
\log p(x) = \beta_0 + \sum_{k=1}^{K} \phi_k(x(u), u \in C_k)
\]

separation over small cliques of neighbour variables of conditionally independent components.

- Problem: Markov hypothesis often not valid
Gibbs Distributions

Approximation of $p(x)$ conditioned on $K$ moments $\mathbb{E}_p(\phi_k(x))$ by $\tilde{p}$ which maximizes the entropy $H_{\tilde{p}} = -\int \tilde{p}(x) \log \tilde{p}(x) \, dx$

**Theorem**  [Canonical Gibbs] If $\tilde{p}(x)$ satisfies

$$\forall k \leq K, \quad \mathbb{E}_{\tilde{p}}(\phi_k(x)) = \int_{\mathbb{R}^N} \phi_m(x) \tilde{p}(x) \, dx = \mathbb{E}_p(\phi_k(x))$$

and maximizes $H_{\tilde{p}} = -\int \tilde{p}(x) \log \tilde{p}(x) \, dx$ then

$$\log \tilde{p}(x) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(x) \quad \text{(separation)}$$

**Problems:**

- How to choose the $\phi_k$ to approximate $p$?
- Can the $\phi_k(x)$ be quadratic if $\tilde{p}$ is Gaussian?
Key Ideas

We want \( \log p(x) \approx \log \tilde{p}(x) = \beta_0 + \sum_{k=1}^{K} \beta_k \phi_k(x) : \text{separation} \)

\( \Rightarrow \) the regularity of the \( \phi_k \) is defined by the regularity of \( p \)

- Regularity of \( p(x) \) defined by \textit{diffeomorphism groups} acting on \( x \)

- Separations are \textit{scale separations} (not Markov) \( \Rightarrow \) \textit{wavelets}

- \( H_{\tilde{p}} \geq H_p \) and if \( H_{\tilde{p}} = H_p \) then \( \tilde{p} = p \)
  - The \( \phi_k \) should minimize the maximum entropy \( H_{\tilde{p}} \)
  - Obtained with \textit{sparsity} and intersections of \( l^1 \) balls

- Approximate the \textit{canonical} \( \tilde{p} \) by a \textit{microcanonical} distribution

- Implemented by a \textit{deep convolutional network}
Lipschitz Regularity on a Group

- Group $G$ of operators acting on $x$ with a metric.

- An $f(x)$ is in $\mathbf{C}^1(G)$ of Lipschitz functions for the action of $G$

$$\forall (g, x) \in G \times \mathbb{R}^d, \quad |f(x) - f(g.x)| \leq C \text{dist}(g, Id)$$

The usual Lipschitz space is $\mathbf{C}^1(\mathbb{R}^d)$: $g.x = x - g$ for $g \in \mathbb{R}^d$.

$$\text{dist}(d, Id) = \|g\|$$

- Lipschitz continuity to spatial diffeomorphisms: deformations

Images $x(u) \in \mathbf{L}^2(\mathbb{R}^2)$ \quad $g.x(u) = x(g(u))$ for $g \in \text{Diff}(\mathbb{R}^2)$

Weak topology: $\text{dist}(g, Id) = \|\nabla g\|_\infty$

$$\Rightarrow \quad |f(x) - f(g.x)| \leq C \|\nabla g\|_\infty \quad \Rightarrow \text{translation invariance}$$
- Amplitude deformation of \( x(u) \in L^2(\mathbb{R}^2) \) with \( g \in \text{Diff}(\mathbb{R}) \)

\[
g.x(u) = g(x(u))
\]
• The action of $\text{Diff}(\mathbb{R}^3)$ on $x$ deforms the 3D measure

$$\bar{x}(u_1, u_2, u_3) = \delta(u_3 - x(u_1, u_2))$$
Amplitude-Space Deformations

- The action of $\text{Diff}(\mathbb{R}^3)$ on $x$ deforms the 3D measure
  \[
  \bar{x}(u_1, u_2, u_3) = \delta(u_3 - x(u_1, u_2))
  \]

- Image classification functions are typically in $C^1(\text{Diff}(\mathbb{R}^3))$
Lipschitz Approximations

- We want to approximate \( \log p \) in \( C^1(\text{Diff}(\mathbb{R}^3)) \) with

\[
\log \tilde{p}(x) = \sum_{k=0}^{K-1} \beta_k \phi_k(x) = \langle \Phi(x), \beta \rangle
\]

\( \log \tilde{p} \in C^1(\text{Diff}) \) if \( \Phi \) is in \( C^1(\text{Diff}(\mathbb{R}^3))^K \) with

\[
\| \Phi(x) - \Phi(g.x) \| \leq C \| \nabla g \|_{\infty}
\]

How can we build such \( \Phi \)?
Marginal Distributions

Cramer-Wold theorem

• A stationary density $p$ of $X$ is characterised by the $1D$ marginals of $X \star \psi_\alpha(u)$ for all $\psi_\alpha \in \mathbb{R}^d$

$\Rightarrow$ choose a ”large” family of $\{\psi_\alpha\}_\alpha$; Mumford, Zhu estimate the distribution of $X \star \psi_\alpha(u)$ with a histogram $\hat{p}$: maximum entropy conditioned to these histogram values

\textit{A bit too optimistic}...

spatial deformations

• To approximate $\log p$ is in $C^1(\text{Diff}(\mathbb{R}^2))$ we need that

$\forall \alpha \; , \; \|u \cdot \nabla \psi_\alpha(u)\|_1 \leq C$

dilated filters: \textit{scale separation}
Scale separation with Wavelets

- Wavelet filter $\psi(u)$: \[ \begin{array}{cc} \text{real parts} & \text{imaginary parts} \\ \end{array} \]

rotated and dilated: $\psi_{2^j, \theta}(u) = 2^{-j} \psi(2^{-j} r\theta u)$

$x \ast \psi_{2^j, \theta}(u) = \int x(v) \psi_{2^j, \theta}(u - v) \, dv$

- Wavelet transform: $W x = \left( d^{-1} \sum_u x(u) \ x \ast \psi_{2^j, \theta}(u) \right)_{j, \theta}$

Preserves norm: $\|W x\|^2 = \|x\|^2$. 

\[ x(u) \]

\[ \downarrow \]

\[ \text{average} \]

\[ \text{higher frequencies} \]
CHAPTER 2. TRANSLATION SCATTERING AND CONVOLUTIONAL NETWORKS

\[ J = 3 \]
\[ C = 6 \]
\[ Q = 1 \]

\[ J = 5 \]
\[ C = 8 \]
\[ Q = 1 \]

\[ J = 3 \]
\[ C = 4 \]
\[ Q = 2 \]

\[ \phi_J \]

\[ \{ \psi_{\theta,j} \} \]

\[ A(\omega) \]

**Figure 2.3**: Three Morlet wavelet families with different sets of parameters. For each set of parameters, we show, from left to right, the Gaussian window \( \phi_J \), the Morlet wavelets \( \psi_{\theta,j} \), and the associated Littlewood Paley sum \( A(\omega) \). When the number of scales \( J \) increases, so does the width of the low pass wavelet \( \phi_J \). When the number of orientations \( C \) increases or when the number of scales per octave \( Q \) decreases, the Morlet wavelets become more elongated in the direction perpendicular to the orientation, and hence have an increased angular sensitivity.
Chapter 2. Translation Scattering and Convolutional Networks

\( J = 3 \)
\( C = 6 \)
\( Q = 1 \)

\( J = 5 \)
\( C = 8 \)
\( Q = 1 \)

\( J = 3 \)
\( C = 4 \)
\( Q = 2 \)

\( \phi_J \)

\( \{ \psi_{\theta, j} \} \)

\( A(\omega) \)

Figure 2.3: Three Morlet wavelet families with different sets of parameters. For each set of parameters, we show, from left to right, the Gaussian window \( \phi_J \), the Morlet wavelets \( \psi_{\theta, j} \), and the associated Littlewood Paley sum \( A(\omega) \).

When the number of scales \( J \) increases, so does the width of the low-pass wavelet \( \phi_J \). When the number of orientations \( C \) increases or when the number of scales per octave \( Q \) decreases, the Morlet wavelets become more elongated in the direction perpendicular to their orientation, and hence have an increased angular sensitivity.
Wavelet transform

\[ x \ast \psi_{\lambda} \text{ (real part)} : \lambda = (2^j, \theta) \]
Wavelet Transform Marginals

Marginal distribution of wavelet coeffs $X \ast \psi_{j,\theta}(u)$

 histograms of real part
Laplacian: sparse
Gaussian

log of histograms

uniform phase distributions
histograms over $\mathbb{C}$
• If $X \star \psi_\lambda(u)$ has a Laplacian density $\alpha e^{-\beta |y|}$ then

$$\|X \star \psi_\lambda\|_1 = \sum_u |X \star \psi_\lambda(u)|$$

is a sufficient statistics of maximum entropy models.

• If $X \star \psi_\lambda(u)$ has a Gaussian density $\alpha e^{-\beta |y|^2}$ then

$$\|X \star \psi_\lambda\|_2^2 = \sum_u |X \star \psi_\lambda(u)|^2$$

is a sufficient statistics of maximum entropy models.
- Wavelet model

\[ \Phi(x) = \left\{ \sum_u x(u), \sum_u |x \ast \psi_{j, \theta}(u)|, \sum_u |x \ast \psi_{j, \theta}(u)|^2 \right\}_{(j, \theta)} \]

- Separates scales \( j \) and angles \( \theta \)
- Markovian along \( u \) over cliques of size \( \sim 2^j \) for each \( j, \theta \)

- **Canonical** max entropy distribution conditioned by \( \mathbb{E}_p(\Phi(x)) \)

\[ \log \tilde{p}(x) = \langle \Phi(x), \beta \rangle + \beta_0 \].

**Problem:** computing \( \beta \) is too expensive

\( \Rightarrow \) *microcanonical* approximation of \( \tilde{p} \)
Ergodic Microcanonical Model

Only \( n = 1 \) realisation \( x_1 \) of \( X \) is known

Microcanonical set: \( \Omega_{x_1} = \{ x : \| \Phi x - \Phi x_1 \| \leq \epsilon \} \)

Microcanonical model \( \bar{\rho} \): maximum entropy supported in \( \Omega_{x_1} \)

\( \Rightarrow \) uniform in \( \Omega_{x_1} \) if bounded set.

\[ \Omega_{x_1} \]

\[ \Phi(x_1) \]

Ergodicity: \( \text{Prob} \left( | \Phi X - \mathbb{E}(\Phi X) | < \epsilon \right) \xrightarrow{d \to \infty} 1 \Rightarrow \Phi x_1 \approx \mathbb{E}(\Phi X) \)

Gibbs conjecture: conditioning on \( \Phi x_1 \) or on \( \mathbb{E}(\Phi X) \) converges to the same Gibbs measure when \( d \) goes to \( \infty \).
Uniform Distribution on Balls

- Sphere in $\mathbb{R}^d$
  \[ \Phi x = d^{-1} \| x \|_2^2 = d^{-1} \sum_{k=1}^{d} |x(k)|^2 \]

- Simplex in $\mathbb{R}^d$
  \[ \Phi x = d^{-1} \| x \|_1 = d^{-1} \sum_{k=1}^{d} |x(k)| = \mu \]

Borel 1914
Diaconis, Freedman 1987

\[ \overline{X}(1), ..., \overline{X}(d) \xrightarrow{d \to \infty} \text{i.i.d Gaussian} \sim e^{-u^2/2\sigma^2} \]

Diaconis, Freedman 1987

\[ \overline{X}(1), ..., \overline{X}(d) \xrightarrow{d \to \infty} \text{i.i.d Exponential} \sim e^{-\lambda|u|} \]
• Intersection of a Sphere and a Simplex in $\mathbb{R}^d$

$$\Phi x = (\|x\|_1, \|x\|_2^2)$$

$\Omega_x$

**Chatterjee 2015**

• If $d$ goes to $\infty$ then $\bar{X}(1), ..., \bar{X}(d)$ converges to:

- a canonical Gibbs: $e^{-\alpha|x| - \beta|x|^2}$ if $r = \|x\|_2 / \|x\|_1 < 2$

- Gaussian if $r = \sqrt{\pi/2}$

- Laplacian if $r = \sqrt{2}$

a singular sparse distribution if $r > 2$
Theorem \((H. \text{ Georgii})\)

If \(\Phi x = \sum_u U_x(u)\) where \(U_x\) has a bounded range for \(u \in \mathbb{Z}^d\)

If the macro canonical distribution exists and converges to a unique Gibbs measure when \(d\) goes to \(\infty\)

then the microcanonical model converges to the same measure for a weak topology.

Proof: large deviation principle
Microcanonical Sampling

Joan Bruna

- Sample max entropy $\overline{X}$ in $\Omega_{x_1}$: $\|\Phi \overline{X} - \Phi x_1\| \leq \epsilon$

Algorithm:
Initialized with $X_0$ Gaussian white noise
Iteratively reduce $\|\Phi X_n - \Phi x_1\|^2$ with gradient descent

- Proof of convergence to a stationary process $X_\infty$
  The algorithm defines a transport of measure.

Math problems:
- No proof on maximum entropy
- Entropy lower bounds depend upon the Jacobian of $\Phi$...
Ising at Critical Temperature

\[ x(u) \in \{0, 1\} \quad p(x) = Z^{-1} \exp \left( \frac{1}{T} \sum_{(u,u') \in C_I} x(u) x(u') \right) \]

\[ \Phi(x) = \left\{ d^{-1} \sum_u x(u) , \| x \ast \psi_\lambda \|_1 , \| x \ast \psi_\lambda \|_2^{2} \right\}_\lambda \]

\[ T = T_{\text{critic}} + \epsilon \]

Realization \( x_1 \) of \( X \)

Microcanonical \( X_\infty \)
\[ \Phi(x) = \left\{ d^{-1} \sum_{u} x(u), \| x \ast \psi_{\lambda} \|_1, \| x \ast \psi_{\lambda} \|_2^2 \right\}_\lambda \]

Realization \( x_1 \) of \( X \) \hspace{1cm} \text{Microcanonical} \ X_\infty
Wavelet Model

\[ |x \ast \psi_{j,\theta}(u)| = w(u, j, \theta) \]

\[ \Phi(x) = \left\{ \sum_u x(u), \sum_u |x \ast \psi_{j,\theta}(u)|, \sum_u |x \ast \psi_{j,\theta}(u)|^2 \right\}_{(j,\theta)} \]

"Conditional independence" may be violated along \( u, \theta, j \).
Higher Order Wavelet Coefficients

Loss of information:

\[ \| x \ast \psi_{\lambda_1} \|_1 = \sum_u |x \ast \psi_{\lambda_1}(u)| \]

eliminates all variations of \( |x \ast \psi_{\lambda_1}(u)| \) along \( u \)

Lipschitz to diffeomorphisms:

recover them as wavelet coefficients of \( |x \ast \psi_{\lambda_1}(u)| \)

\[ |W_2| |x \ast \psi_{\lambda_1}| = \left( \frac{\sum_u |x \ast \psi_{\lambda_1}(u)|}{\|x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2}(u)|} \right)_{\lambda_2} \]
Wavelet Scattering Network

\[ \Phi = |W_{\log d/2}| \cdots |W_2| |W_1| \]

\[ \Phi x = \left\{ \sum_{u} x(u) \right\} \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \cdots \ast \psi_{\lambda_m} \|_1 \right\}_{\lambda_k} \]
Scattering Properties

\[ \Phi x = \left( \sum_u x(u) \right) \begin{pmatrix} \|x \ast \psi_{\lambda_1}\|_1 \\ \|x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2}\|_1 \\ \|x \ast \psi_{\lambda_2} \ast \psi_{\lambda_2} \ast \psi_{\lambda_3}\|_1 \\ \vdots \end{pmatrix} = \ldots |W_3| |W_2| |W_1| x \]

\[ \|W_k x\| = \|x\| \Rightarrow \|W_k x - W_k x'\| \leq \|x - x'\| \]

**Lemma:** If \( g \in \text{Diff}(\mathbb{R}^2) \) then

\[ \|[W_k, g]\| = \|W_k g - gW_k\| \leq C \|\nabla g\|_{\infty} \]

**Theorem:** For appropriate wavelets, a scattering is

- contractive \( \|\Phi x - \Phi y\| \leq \|x - y\| \) : in \( C^1(L^2(\mathbb{R}^2)) \)

- preserves norms \( \|\Phi x\| = \|x\| \)

- Lipschitz on diffeomorphisms \( \|\Phi x - \Phi(g.x)\| \leq C \|\nabla g\|_{\infty} \)
Energy conservation

\[ \| x \| = \| \Phi x \| \Rightarrow \| x \ast \psi_{\lambda_1} \|_2^2 = \| \Phi (x \ast \psi_{\lambda_1}) \|_2^2 \]

\[ \| x \ast \psi_{\lambda_1} \|_2^2 = \sum_{m=2}^{\infty} \sum_{\lambda_2, \ldots, \lambda_m} \| x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \ldots \ast \psi_{\lambda_m} \|_1^2 \]

All \( L^2 \) norms are derived from \( L^1 \) norms.

Non-negligible \( L^1 \) norms appear at order 1 and 2:

\[ \Phi(x) = \left\{ \sum_u x(u), \| x \ast \psi_{\lambda_1} \|_1, \| | x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \|_1 \right\}_{\lambda_1, \lambda_2} \]

If \( x \in \mathbb{R}^d \) then \( \Phi x \in \mathbb{R}^{O(\log^2 d)} \)
Texture Reconstructions

Texture of $d$ pixels

Ising-critical
Turbulence 2D

Gaussian process model with $d$ second order moments

Reconstructions from $\|X \ast \psi_{\lambda_1}\|_1$ and $\|\|X \ast \psi_{\lambda_1} \ast \psi_{\lambda_2}\|_1$

$O(\log^2 d)$ scattering coefficients
Microcanonical Reconstructions

\[ \Phi(x) = \left\{ \sum_{u} x(u), \| x \ast \psi_{\lambda_1} \|_1, \| x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \|_1 \right\}_{\lambda_1, \lambda_2} \]

Realization \( x_1 \) of \( X \)

Microcanonical \( X_\infty \)

order 1

order 2

Must further reduce entropy
- Scattering model of too high entropy

\[ |x \ast \psi_{\theta, j}(u)| = w(u, \theta, j) \]

- not sparse at intermediate scales \(2^j\) but not Gaussian

- joint dependance in \((u, \theta)\) \(\Rightarrow\) wavelet transforms in \((u, \theta)\)

- dependence on amplitude values?
3D Scattering for Amplitude

\[ \bar{x}(u_1, u_2, u_3) = \delta(u_3 - x(u_1, u_2)) \]

We want \( \Phi \) in \( \mathbb{C}^1(\text{Diff}(\mathbb{R}^3)) \)

3D wavelets: \( \psi_{\lambda}(u_1, u_2, u_3) = 2^{-2j} \psi(2^{-j} r_\theta(u_1, u_2)) \) \( 2^{-\ell} \psi(2^{-\ell} u_3) \)

Joint dependance on amplitude and spatial geometry

Wavelet coefficients are much more sparse at intermediate scales

\[ \Phi \bar{x} = \left( \begin{array}{c} \sum_u \bar{x}(u) \\ \| \bar{x} \ast \psi_{\lambda_1} \|_1 \\ \| x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \|_1 \\ \vdots \end{array} \right) \]

\( \lambda_1, \lambda_2, \ldots \)
3D Scattering Models

Preliminary results

Realization $x_1$ of $X$  2D Scat on $x$  3D Scat on $\bar{x}$
Conclusions

• Regularity in high dimension as regularity to action of diffeomorphisms on different groups

• Long range dependence: variable separation through scales

• Entropy reduction with sparsity: $L_1$ geometry