An Introduction to Logarithmic Sobolev Inequalities

Samworth Group, Virtual, 27 January 2025

Sam Power, University of Bristol

Some Motivating Questions

“What makes a class of probability measures 'nice' to work with?”

–SP, 26 January 2025

$$\mu = \mathcal{N}(0,1)$$

$$\mu = \mathcal{N}(0, \mathbf{Id})$$

$$\mu = \mathcal{N}(\mathsf{m}, \mathsf{C})$$

$$\mu = \exp(-V)$$

$$\mu = \exp(-V)$$

V **convex**

$$\mu = \exp(-V)$$

V *strongly* **convex**

What are we doing here?

- Move beyond Gaussian; *flexibility*
  - Don’t usually need { exact formulas, symmetry, … }
  - Do often need { concentration, smoothness, composition, … }
- What assumptions *contain Gaussianity* while *implying same conclusions?*
  - Implicit: what conclusions do we care about?

A Warning

- In a moment, I will show a large collection of such conditions
- Some will be familiar, some will not be
- We will really focus on <u>one</u>
- The point is to situate today’s discussion within a broader landscape

The Bogachev-Kolesnikov Hierarchy

The Hierarchy

Strong Log-Concavity	$\nabla^2 V \succeq m \cdot \mathbf{I}$
Gaussian Isoperimetry	$\mathcal{I}_\mu \geq m^{\frac{1}{2}} \cdot \mathcal{I}_\gamma$
Logarithmic Sobolev Inequality	$\mathtt{I}\left(\nu, \mu\right) \geq 2 \cdot m \cdot \mathsf{KL}\left(\nu, \mu\right)$
Transport-Entropy Inequality	$\mathsf{KL}\left(\nu, \mu\right) \geq \frac{m}{2} \cdot \mathcal{T}_2^2\left(\nu, \mu\right)$
Tail Bounds	$\mu\left(A^r\right) \geq 1 - \exp\left(-m \cdot r^2/2\right)$
Exponential Integrability	$\mu\left(\exp\left(\varepsilon \cdot \lVert x \rVert^2/2 \cdot m\right)\right) < \infty$

The Hierarchy

Strong Log-Concavity	pointwise geometry
Gaussian Isoperimetry	small-set geometry
Logarithmic Sobolev Inequality	moderate-scale geometry
Transport-Entropy Inequality	Lipschitz concentration
Tail Bounds	large-scale concentration
Exponential Integrability	large-scale integrability

Unpacking the Hierarchy

- These conditions go from strongest (at the top) to weakest (at the bottom)
- Each condition implies the next, *quantitatively*
- In general, the implications *cannot* be reversed
- However, under curvature lower bounds, they *can* be, *quantitatively*
  - see E. Milman (A, B, C, D, E, F, G), Ledoux

Today’s Main Character

Strong Log-Concavity	pointwise geometry
Gaussian Isoperimetry	small-set geometry
Logarithmic Sobolev Inequality	**moderate-scale geometry**
Transport-Entropy Inequality	Lipschitz concentration
Tail Bounds	large-scale concentration
Exponential Integrability	large-scale integrability

The Logarithmic Sobolev Inequality

The Information Theorist's LSI

- For all probability measures ν,

$$\mathtt{I}\left(\nu, \mu\right) \geq 2 \cdot m \cdot \mathsf{KL}\left(\nu, \mu\right)$$

with

$$\mathtt{I}\left(\nu, \mu\right) = \int \nu \cdot \left\| \nabla \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right\|_2^2, \qquad \mathsf{KL}\left(\nu, \mu\right) = \int \nu \cdot \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$$

The Functional Analyst's LSI

- For all functions f,

$$\mathbf{E}_\mu \left[\| \nabla f(X) \|^2 \right] \geq 2 \cdot m \cdot \mathsf{ent}_\mu \left(f^2 \right)$$

with

$$\mathsf{ent}_\mu \left(g \right) = \mathbf{E}_\mu \left[g \cdot \log g \right] - \mathbf{E}_\mu \left[g \right] \cdot \log \left(\mathbf{E}_\mu \left[g \right] \right)$$

The Functional Analyst's LSI, Bis

- For all functions f,

$$\mathbf{E}_{\mu}\left[\frac{\| \nabla f(X) \|^2}{f(X)}\right] \geq \frac{m}{2} \cdot \mathrm{ent}_{\mu}(f)$$

with

$$\mathrm{ent}_{\mu}(g) = \mathbf{E}_{\mu}[g \cdot \log g] - \mathbf{E}_{\mu}[g] \cdot \log\left(\mathbf{E}_{\mu}[g]\right)$$

Simple Implications of the LSI

- Take $f = \exp\left(\lambda \cdot g\right)$ with g Lipschitz; observe sub-Gaussian MGF
  - ‘Herbst’ argument implies that tails are sub-Gaussian
  - Gaussian indeed satisfies LSI, so no better implication can hold
  - Converse false in general

Simple Implications of the LSI

- Introduce OLD $\left(\mu\right)$:

$$\mathrm{d}X_t = \nabla \log \mu \left(X_t\right) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}W_t$$

- ‘Overdamped Langevin Diffusion’; converges towards μ
- Rich font of insights for ‘geometry of μ’; see Bakry-Gentil-Ledoux

Simple Implications of the LSI

- With $\nu_t = \mathsf{Law}\left(X_t; X_0 \sim \nu\right)$, there holds *de Bruijn's identity*:

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\nu_t, \mu\right) = \mathtt{I}\left(\nu_t, \mu\right),$$

  and so Grönwall's inequality yields that

$$\mathsf{KL}\left(\nu_t, \mu\right) \leq \exp\left(-2 \cdot m \cdot t\right) \cdot \mathsf{KL}\left(\nu_0, \mu\right)$$

- 'Reverse-Grönwall' argument shows equivalence (to entropy contraction)

Simple Implications of the LSI

- What intuition do we have so far?
  - ‘Gaussian in the tails’ from MGF bound
  - ‘nothing too nasty’ in the bulk, from OLD mixing; ‘well-connected’
  - little control on *pointwise* behaviour
    - this affords quite a lot of flexibility

Verifying the LSI

- When does the LSI hold?
  - Easiest case: when we have m*-strong log-concavity*
    - For me: strong log-concavity implies everything I could ever want
- How do we move beyond this?
  - '*transfer principles*'

Transfer Principles

- I will present a sort of ‘tasting menu’ to give a sense of what is possible
  - Proofs are often (but not always) easy
  - Some general references: Chafaï-Lehec, Chen, Chewi, van Handel
- Before trying to prove an LSI, first check that it doesn’t ‘obviously’ fail …

Transfer Principles (1)

- Tensorisation
- If $\mu_i \in \mathtt{LSI}\left(m_i\right)$, then $\bigotimes_i \mu_i \in \mathtt{LSI}\left(\min_i m_i\right)$
    - ‘never worse than your worst marginal’
    - *dimension-free* estimate

Transfer Principles (2)

- Bounded Change-of-Measure
- If $\mu \in \mathtt{LSI}\,(m)$, and $0 < \alpha \leq \dfrac{\mathrm{d}\mu'}{\mathrm{d}\mu} \leq \beta < \infty$, then $\mu' \in \mathtt{LSI}\left(m \cdot \dfrac{\alpha}{\beta}\right)$
    - 'stability under log-bounded perturbation'
    - in applications, $\dfrac{\beta}{\alpha}$ may be $\exp\left(\Omega\,(d)\right)$

Transfer Principles (3)

- Lipschitz Transport
  - If $\mu \in \mathtt{LSI}(m)$, and $\| T \|_{\mathrm{Lip}} \leq L$, then $T_{\#}\mu \in \mathtt{LSI}\left(m \cdot L^{-2}\right)$
    - ‘stability under Lipschitz pushforward’
    - natural to consider $T = T^{\mathrm{OT}}$; fruitful to consider $T = T^{\mathrm{HF}}$
      - see e.g. Kim-Milman, Mikulincer-Shenfeld, Brigati-Pedrotti

Transfer Principles (4)

- Convolution
  - If $\mu_i \in \mathsf{LSI}\left(m_i\right)$, then $\star_i \, \mu_i \in \mathsf{LSI}\left(\sum_i m_i\right)$
    - ‘stability under additive convolution’
    - useful when one of the μ_i is Gaussian
    - see e.g. Chafaï

Transfer Principles (5)

- 'Overlapping' Mixtures
  - If $\mu_z \in \mathtt{LSI}(m)$ for all z, and $\chi^2\left(\mu_z, \mu_{\bar{z}}\right) \leq \bar{\chi}^2$ for all $z, \bar{z}$, then

$$\alpha\mu := \int \alpha\left(\mathrm{d}z\right) \cdot \mu_z \implies \alpha\mu \in \mathtt{LSI}(m')$$

with m' depending only on $m, \bar{\chi}$; from CCNW

- 'stability under overlapping mixture'
- assumptions can be weakened, but qualitatively reasonable

Transfer Principles (6)

- Some other possibilities
  - ‘dependent tensorisation’ (Blower-Bolley, Ledoux)
  - ‘multi-scale’ mixtures (Bauerschmidt-Bodineau-Dagallier)
  - log-Lipschitz perturbations (Brigati-Pedrotti, Fathi-Mikulincer-Shenfeld)
  - …

Working with the LSI

- Hopefully, you have a sense that there are many tools for verifying LSI
    - For any specific instance, different tools could be relevant
- Moreover, you should have a sense that …
    - we can easily become ‘far from Gaussian’ in a pointwise sense,
    - while being ‘Gaussian enough’ in a coarse sense

Working with the LSI

- Why is it worth moving *beyond strong log-concavity*?
    - We can now tolerate various ‘impurities’ in μ (fine vs. coarse features)
        - For ‘intrinsically coarse’ tasks like sampling, we may expect that uniform curvature is more than needed
    - We seem to retain many useful features of the original setting
        - Of course, we must lose some things too (e.g. conditionals)

The LSI for Statistics

- Imposing the LSI as a learning constraint (e.g. density estimation)
  - Interesting, though maybe too implicit for easy use (MLE?)
- Imposing the LSI on { data / noise / … }
  - Natural, painless; likely to be ‘as good as Gaussian’ for the statistician
- Using the LSI to replace KL by I
  - From MLE to Score Matching, at the population level (e.g. Koehler++)

An Analogy with Optimisation

The ‘LSI’ for Optimisation

- For minimising a function by gradient flow, …
  - Strong Convexity yields that $\|x_t - x_t'\| \lesssim \exp(-m \cdot t)$
  - Polyak-Łojasiewicz yields that $f\left(x_t\right) - \inf f \lesssim \exp(-m \cdot t)$
- Among (some) optimisation experts: PŁI is ‘just for proofs’
  - Not clear that it opens up the function class in interesting ways

The LSI for Optimisation

- For sampling a distribution by OLD, …
  - Strong Log-Concavity yields that $\|x_t - x_t'\| \lesssim \exp\left(-m \cdot t\right)$
  - Log-Sobolev yields that KL $\left(\nu_t, \mu\right) \lesssim \exp\left(-m \cdot t\right)$
- Among sampling experts: strong convexity is rarely realistic
  - Working with LSI genuinely opens up the class of interesting targets

The LSI Elsewhere

- LSI can be { strengthened, weakened } to treat other μ
- LSI can be adapted to treat other ‘geometries’
- LSI can be used to study μ arising in { statistical mechanics, ML, … }
- Ubiquity: many people need a way to robustly characterise ‘*niceness*’

Some Conclusions

- The LSI is one well-developed way to reason about ‘nice’ μ
- In trading off Assumptions and Consequences, it strikes a nice balance
- Moreover, the tools for verifying these Assumptions are quite usable