Stein's method, logarithmic Sobolev and transport inequalities

Ivan Nourdin (Luxembourg University)

A talk in honor of Aline Bonami

The history of the Hypercontractivity Theorem

This part of my talk is taken from Ryan O'Donnel's blog, entitled "Analysis of Boolean Functions".

The history of the Hypercontractivity Theorem is complicated.

Its earlier roots are in the work of Paley (1932). [...]

In 1968, Bonami stated the following variant of the Hypercontractivity Theorem: If $f : \{-1,1\}^n \to \mathbb{R}$ is homogeneous of degree k then for all $q \geq 2$,

$$\|f\|_q \leq c_k \sqrt{q} \|f\|_2,$$

where the constant c_k may be taken to be 1 if q is an even integer.

She remarks that this theorem can be deduced from Paley's result but with a much worse (exponential) dependence on q.

The proof she gives is combinatorial and actually only treats the case $k = 2$ and q an even integer.

Independently in 1969, Schreiber considered multilinear polynomials f over a general orthonormal sequence $x_1, \ldots, x_n$ of centred real (or complex) random variables.

He showed that if f has degree at most k then for any integer $q \geq 4$ it holds that $\|f\|_q \leq C\|f\|_2$, where C depends only on k, q, and the q-norms of the x_i's.

Schreiber was interested mainly in the case that the x_i's are Gaussian; indeed, it is a generalization of his earlier work from 1967 specific to the Gaussian case.

In 1970, Bonami published her Ph.D. thesis **which contains the full Hypercontractivity Theorem**.

Her proof follows the standard template seen in essentially all proofs of hypercontractivity: first an elementary proof for the case $n = 1$ and then an induction to extend the general n.

Bonami's work was published in French, and it remained unknown to most English-language mathematicians for about a decade.

In the late '60s and early '70s, researchers in quantum field theory developed the theory of hypercontractivity for the Ornstein-Uhlenbeck operator P_t.

This is now recognized as essentially being a **special case** of hypercontractivity for bits, in light of the fact that $\frac{x_1+\ldots+x_n}{\sqrt{n}}$ tends to a Gaussian as $n \to \infty$ by the CLT.

We summarize here some of the works in this setting.

In 1966, Nelson showed that $\|P_{1/\sqrt{q-1}} f\|_q \leq C_q \|f\|_2$ for all $q \geq 2$.

Glimm gave in 1968 the alternative result that for each $q \geq 2$ there is a sufficiently small $t_q > 0$ such that $\|P_{t_q} f\|_q \leq \|f\|_2$.

Segal observed in 1970 that hypercontractive results can be proved by induction on the dimension n.

In 1973, Nelson gave the full Hypercontractivity Theorem in the Gaussian setting:

$$\|P_{\sqrt{(p-1)/(q-1)}} f\|_q \leq \|f\|_p$$

for all $1 \leq p < q \leq \infty$.

In 1975, Gross introduced the notion of Log-Sobolev inequalities and showed how to deduce hypercontractivity inequalities from them.

He established the Log-Sobolev inequality from 1-bit functions, used induction to obtain it for n-bit functions, and then used the CLT to transfer results to the Gaussian setting. This gave a new proof of Nelson's result and also independently established Bonami's full hypercontractivity Theorem.

Also, in 1975, Beckner published his Ph.D. thesis which proved a sharp form of the hypercontractivity inequality for purely complex t.

It is unfortunate that the influential paper of Kahn, Kalai and Linial (KKL 88) miscredited the hypercontractivity theorem to Beckner.

An example of application in statistics

(We use an idea introduced by Talagrand called the $L^1 - L^2$ bound.)

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector composed of independent $N(0, 1)$ random variables, and let $X_{(k)}$ denote the kth order statistic associated with $\mathbf{X}$, that is,

$$\{X_1, \ldots, X_n\} = \{X_{(1)}, \ldots, X_{(n)}\}$$

and

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)}.$$

It is easy to see that $X_{(k)} = f_k(\mathbf{X})$, where f_k satisfies

$$\partial_i f_k(\mathbf{x}) = \mathbf{1}_{\Delta_{i,k}}(\mathbf{x}),$$

with $\Delta_{i,k} = \{\mathbf{x} \in \mathbb{R}^n : x_i = x_{(k)}\}$.

Thus, f_k is Lipschitz with constant 1 and the classical Poincaré inequality yields

$$\mathsf{Var}(X_{(k)}) \leq 1.$$

One cannot deduce from this that $X_{(k)}$ concentrates as $n \to \infty$.

One has

$$\begin{aligned}
\mathsf{Var}(X_{(k)}) &= E[f_k(\mathbf{X})(f_k(\mathbf{X}) - E[f_k(\mathbf{X})])] \\
&= E[f_k(\mathbf{X})(P_0 f_k(\mathbf{X}) - P_\infty f_k(\mathbf{X}))] \\
&= -\int_0^\infty E[f_k(\mathbf{X})\frac{d}{dt}P_t f_k(\mathbf{X})]dt \\
&= \sum_{i=1}^n \int_0^\infty e^{-t} E[\partial_i f_k(\mathbf{X}) P_t \partial_i f_k(\mathbf{X})]dt \\
&\leq \sum_{i=1}^n \int_0^\infty e^{-t} \|\partial_i f_k(\mathbf{X})\|_{L^2} \|P_t \partial_i f_k(\mathbf{X})\|_{L^2} dt \\
&\leq \sum_{i=1}^n \int_0^\infty e^{-t} \|\partial_i f_k(\mathbf{X})\|_{L^2} \|\partial_i f_k(\mathbf{X})\|_{L^{1+e^{-2t}}} dt.
\end{aligned}$$

We have just seen that

$$\mathsf{Var}(X_{(k)}) \ \leq \ \sum_{i=1}^{n} \int_0^\infty e^{-t} \|\partial_i f_k(\mathbf{X})\|_{L^2} \|\partial_i f_k(\mathbf{X})\|_{L^{1+e^{-2t}}} dt.$$

But

$$\|\partial_i f_k(\mathbf{X})\|_{L^r} = P(\mathbf{X} \in \Delta_{i,k})^{1/r} = n^{-1/r},$$

leading to

$$\mathsf{Var}(X_{(k)}) \leq \int_0^\infty e^{-t} n^{\frac{1}{2} - \frac{1}{1+e^{-2t}}} dt \leq \frac{2}{\log n}.$$

The rest of this talk is mainly based on the material developed in the following two papers:

- I. Nourdin, G. Peccati and Y. Swan (2014): Entropy and the fourth moment phenomenon, *Journal of Functional Analysis* **266**, no. 5, 3170–3207.

- M. Ledoux, I. Nourdin and G. Peccati (2014): Stein's method, logarithmic Sobolev and transport inequalities. Submitted.

PART I: Stein's method

Let F be a given real random variable with, say, mean zero and variance one.

Let $N \sim N(0,1)$.

In many situations of interest, we may expect the law of F to be close of that of N, and one is interested in quantifying it. How to formalize this ?

Stein's lemma. Assume $h : \mathbb{R} \to [0,1]$ is continuous. Consider φ_h defined, for $x \in \mathbb{R}$, as

$$\begin{aligned} \varphi_h(x) &= e^{\frac{x^2}{2}} \int_{-\infty}^{x} (h(a) - E[h(N)]) e^{-\frac{a^2}{2}} da \\ &= -e^{\frac{x^2}{2}} \int_{x}^{\infty} (h(a) - E[h(N)]) e^{-\frac{a^2}{2}} da. \end{aligned}$$

Then φ_h is C^1, satisfies

$$\varphi_h'(x) - x\varphi_h(x) = h(x) - E[h(N)]$$

and is such that $\|\varphi_h\|_\infty \leq 2$.

A bound on the total variation distance. One has

$$\begin{aligned}
d_{TV}(F,N) &:= \sup_{A\in\mathcal{B}(\mathbb{R})}\Big|P[F\in A]-P[N\in A]\Big| \\
&\le \sup_{h:\mathbb{R}\to[0,1]}\Big|E[h(A)]-E[h(N)]\Big| \\
&= \sup_{h:\mathbb{R}\to[0,1]\in C^0}\Big|E[h(A)]-E[h(N)]\Big| \quad \text{(Lusin)} \\
&\le \sup_{\varphi\in C^1:\,\|\varphi\|_\infty\le 2}\Big|E[\varphi'(F)]-E[F\varphi(F)]\Big| \quad \text{(Stein)}.
\end{aligned}$$

Now the question is: how to relate $E[\varphi'(F)]$ and $E[F\varphi(F)]$?

Definition. We say that $\tau_F : \mathbb{R} \to \mathbb{R}$ is a *Stein factor* for F if

$$E[F\varphi(F)] = E[\tau_F(F)\varphi'(F)]$$

for all test function φ. The *Stein discrepancy* is

$$S(F|N) = \sqrt{E[(1 - \tau_F(F))^2]} = \sqrt{\mathrm{Var}(\tau_F(F))}.$$

Theorem. One has $d_{TV}(F, N) \leq 2\, S(F|N)$.

Examples: 1. If $F \sim N(0,1)$ then $\tau_F(x) = 1$ is a Stein factor for F. We then have $S(F|N) = 0$.

2. If $F_n = \frac{1}{\sqrt{n}}(X_1 + \ldots + X_n)$ with X_i iid, centered and unit variance, then $\tau_{F_n} = \frac{1}{n}\sum_{i=1}^n E[\tau_X(X_i)|F_n]$ is a Stein factor for F_n, so that

$$S(F_n|N)^2 \leq \frac{1}{n}\mathrm{Var}(\tau_X(X)).$$

3. If $F = I_p(f)$ is a multiple integral of order p, then $\tau_F = E[\langle DF, -DL^{-1}F\rangle|F]$ is a Stein factor for F, implying in turn that

$$S(F|N)^2 \leq \frac{p-1}{3p}(E[F^4] - 3).$$

Multivariate extension. Let F be a centered random vector of $\mathbb{R}^d$. Let $N \sim N_d(0, \mathsf{Id})$.

Definition. 1) A measurable matrix-valued map on $\mathbb{R}^d$

$$x \mapsto \tau_F(x) = \Big\{\tau_F^{ij}(x) : i, j = 1, \ldots, d\Big\}$$

is said to be a *Stein matrix* for F if $\tau_F^{ij}(F) \in \mathsf{L}^1(\Omega)$ for every i, j and, for every smooth $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$E[F \cdot \nabla\varphi(F)] = E[\langle \tau_F(F), \mathsf{Hess}(\varphi)(F)\rangle_{\mathsf{HS}}],$$

with $\langle A, B\rangle_{\mathsf{HS}} = \sum_{i,j=1}^d a_{ij} b_{ij}$.

2) The *Stein discrepancy* is $\mathsf{S}\Big(F \,|\, N\Big)$, with

$$\mathsf{S}^2\Big(F \,|\, N\Big) = E\|\tau_F(F) - \mathsf{Id}\|_{\mathsf{HS}}^2.$$

PART II: Logarithmic Sobolev inequality

Let F be any random vector of $\mathbb{R}^d$ whose law, noted ν, is absolutely continuous (wrt Lebesgue).

Let $N \sim N_d(0, \mathrm{Id})$ and denote its law by γ, that is, $d\gamma(x) = (2\pi)^{-d/2} e^{-|x|^2/2} dx$ on $\mathbb{R}^d$.

Let $h = \frac{d\nu}{d\gamma}$.

Definitions. 1) The *relative entropy* of F with respect to N is

$$\mathsf{H}\big(F \,|\, N\big) = \int_{\mathbb{R}^d} h \log h \, d\gamma \, (= \mathsf{Ent}_\gamma(h)\,).$$

2) The *Fisher information* of F with respect to N is

$$\mathsf{I}\big(F \,|\, N\big) = \int_{\mathbb{R}^d} \frac{|\nabla h|^2}{h} \, d\gamma \, (= \mathsf{I}_\gamma(h)\,).$$

The classical **logarithmic Sobolev inequality** with respect to the standard Gaussian measure γ indicates that

$$\mathsf{H}\Big(F \,|\, N\Big) \;\leq\; \frac{1}{2}\,\mathsf{I}\Big(F \,|\, N\Big).$$

The HSI inequality (Ledoux, Nourdin, Peccati). One has

$$\mathsf{H}\Big(F\,|\,N\Big) \;\le\; \frac{1}{2}\,\mathsf{S}^2\Big(F\,|\,N\Big)\,\log\left(1+\frac{\mathsf{I}(F\,|\,N)}{\mathsf{S}^2(F\,|\,N)}\right).$$

Since $\log(1+x) \le x$, our inequality is a new improved form of the logarithmic Sobolev inequality.

PART III: Transport inequalities

Wasserstein quadratic distance. Given two probability measures ν and μ on the Borel sets of $\mathbb{R}^d$ whose marginals are square integrable, we define the *quadratic Wasserstein distance* between ν and μ as the quantity

$$\mathsf{W}_2(\nu, \mu) = \inf_{\pi} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\pi(x, y) \right)^{1/2}$$

where the infimum runs over all probability measures π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ν and μ.

Let F be any random vector of $\mathbb{R}^d$ whose law, noted ν, is absolutely continuous (wrt Lebesgue).

Let $N \sim N_d(0, \mathsf{Id})$ and denote its law by γ, that is, $d\gamma(x) = (2\pi)^{-d/2} e^{-|x|^2/2} dx$ on $\mathbb{R}^d$.

Talagrand inequality: One has the quadratic transportation cost inequality:

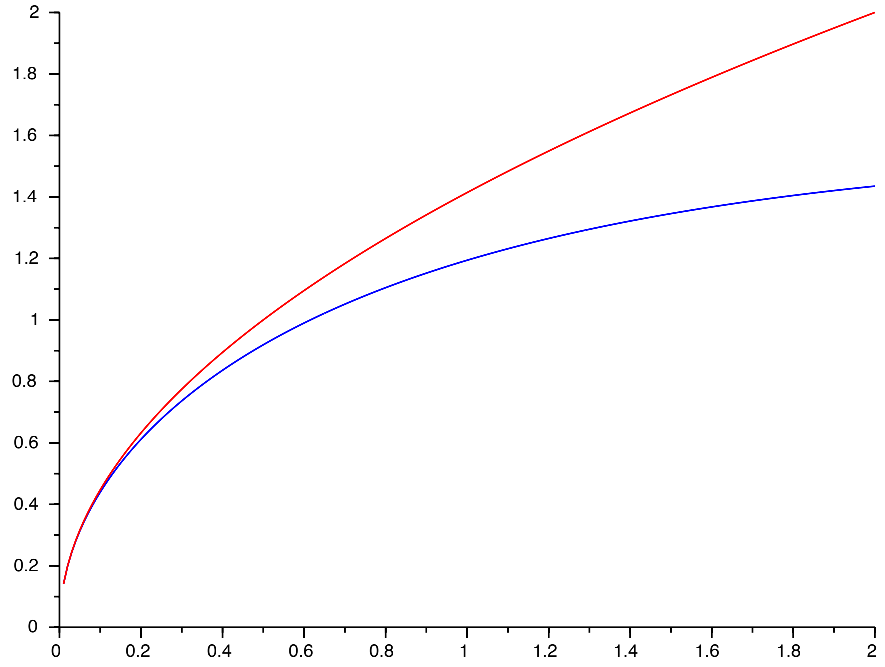
$$\mathsf{W}_2(F, N)^2 \leq 2\,\mathsf{H}\Big(F \,|\, N\Big).$$

WSH inequality (Ledoux, Nourdin, Peccati):

$$\mathsf{W}_2(F, N) \leq \mathsf{S}\Big(F \,|\, N\Big) \arccos\Big(e^{-\frac{\mathsf{H}(F|N)}{\mathsf{S}^2(F|N)}}\Big).$$

WSH implies Talagrand. This is because

$$\mathrm{arccos}(e^{-r}) \leq \sqrt{2r}, \quad r \geq 0.$$



Can we also recover the celebrated HWI inequality?

Otto and Villani:

$$\mathsf{H}(F|N) \leq \mathsf{W}_2(F,N)\sqrt{\mathsf{I}(F|N)} - \frac{1}{2}\mathsf{W}_2(F,N)^2.$$

(It implies the log-Sobolev inequality since $xy - \frac{1}{2}y^2 \leq \frac{1}{2}x^2$.)

Can we consider further distributions?

On the basis of the Gaussian example, we are able to address the issue of HSI inequalities for distributions on $\mathbb{R}^d$, $d \geq 1$, that are not necessarily Gaussian.

Our basic ingredient is a semigroup approach *à la* Bakry-Emery. As such, one can deal with the family of invariant measures of second order differential operators.

These include gamma and beta distributions, as well as families of log-concave measures as illustrations.

But we will not give the details here!

REST OF THE TALK:
some elements of proof

Denote by τ_F a Stein matrix associated with F, having distribution $d\nu = h d\gamma$.

For every $t > 0$, set $d\nu_t = P_t h \, d\gamma$, and write $v_t = \log P_t h$, with $(P_t)_{t \geq 0}$ the Ornstein-Uhlenbeck semigroup associated with $d\gamma(x) = (2\pi)^{-d/2} e^{-|x|^2/2} dx$ on $\mathbb{R}^d$.

We will make intensive use of the following key inequalities.

* **Integrated de Bruijn's formula**

$$\mathsf{H}\Big(F \,|\, N\Big) \,=\, \mathsf{Ent}_\gamma(h) \,=\, \int_0^\infty \mathsf{I}_\gamma(P_t h) dt.$$

* **Exponential decay of Fisher information**

$$\mathsf{I}_\gamma(P_t h) \,\leq\, e^{-2t}\, \mathsf{I}_\gamma(h) \,=\, e^{-2t}\, \mathsf{I}\Big(F \,|\, N\Big).$$

* **Linking I and** S (crucial!)

$$\mathsf{I}_\gamma(P_t h) = \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\Big[\big(\tau_F(x) - \mathsf{Id}\big)y\cdot\nabla v_t\big(e^{-t}x + \sqrt{1-e^{-2t}}\,y\big)\Big]d\nu(x)d\gamma(y).$$

As a consequence,

$$\mathsf{I}_\gamma(P_t h) \ = \ \int_{\mathbb{R}^d}|\nabla v_t|^2 d\nu \le \frac{e^{-4t}}{1-e^{-2t}}\,\mathsf{S}^2\Big(F\,|\,N\Big).$$

* **Exponential decay of Stein discrepancy**

$$\mathsf{S}\Big(\nu_t\,|\,\gamma\Big) \le e^{-2t}\,\mathsf{S}\Big(F\,|\,N\Big)\,(\le \mathsf{S}\Big(F\,|\,N\Big)).$$

Proof of the HSI inequality. The idea is to bound the Fisher information $\mathsf{I}_\gamma(P_t h)$ differently for t around 0 and away from 0.

We can write, for every $u > 0$,

$$\begin{aligned}
\mathsf{H}\Big(F\,|\,N\Big) &= \int_0^u \mathsf{I}_\gamma(P_t h)dt + \int_u^\infty \mathsf{I}_\gamma(P_t h)dt \\
&\leq \mathsf{I}\Big(F\,|\,N\Big)\int_0^u e^{-2t}dt + \mathsf{S}^2\Big(F\,|\,N\Big)\int_u^\infty \frac{e^{-4t}}{1-e^{-2t}}\,dt \\
&\leq \frac{1}{2}\,\mathsf{I}\Big(F\,|\,N\Big)(1-e^{-2u}) + \frac{1}{2}\,\mathsf{S}^2\Big(F\,|\,N\Big)\Big(-e^{-2u} - \log(1-e^{-2u})\Big).
\end{aligned}$$

Optimizing in u (set $1 - e^{-2u} = r \in (0,1)$) leads to the desired inequality

$$\mathsf{H}\Big(F\,|\,N\Big) \leq \frac{1}{2}\,\mathsf{S}^2\Big(F\,|\,N\Big)\log\left(1 + \frac{\mathsf{I}(F\,|\,N)}{\mathsf{S}^2(F\,|\,N)}\right).$$

Proof of the WSH inequality. For any $t \geq 0$, recall $d\nu_t = P_t h d\gamma$ (in particular, $\nu_0 = \nu$ and $\nu_t \to \gamma$ as $t \to \infty$). The HSI inequality applied to ν_t yields that

$$\mathsf{H}\big(\nu_t \,|\, \gamma\big) \,\leq\, \frac{1}{2}\,\mathsf{S}^2\big(\nu_t \,|\, \gamma\big)\,\log\left(1 + \frac{\mathsf{I}(\nu_t \,|\, \gamma)}{\mathsf{S}^2(\nu_t \,|\, \gamma)}\right).$$

Now, $\mathsf{S}^2(\nu_t \,|\, \gamma) \leq \mathsf{S}^2(\nu \,|\, \gamma)$ (due to the exponential decay of the Stein discrepancy) and $r \mapsto r \log\big(1 + \frac{s}{r}\big)$ is increasing for any fixed s. It follows that

$$\mathsf{H}\big(\nu_t \,|\, \gamma\big) \,\leq\, \frac{1}{2}\,\mathsf{S}^2\big(\nu \,|\, \gamma\big)\,\log\left(1 + \frac{\mathsf{I}(\nu_t \,|\, \gamma)}{\mathsf{S}^2(\nu \,|\, \gamma)}\right).$$

By exponentiating both sides, this inequality is equivalent to:

$$\sqrt{\mathsf{I}\Big(\nu_t \,|\, \gamma\Big)} \;\le\; \frac{\mathsf{I}(\nu_t \,|\, \gamma)}{\mathsf{S}(\nu \,|\, \gamma)\sqrt{e^{\frac{2\mathsf{H}(\nu_t|\gamma)}{\mathsf{S}^2(\nu|\gamma)}} - 1}}\,.$$

But Otto and Villani showed in their pathbreaking paper that

$$\frac{d^+}{dt}\mathsf{W}_2(\nu, \nu_t) \;\le\; \sqrt{\mathsf{I}\Big(\nu_t \,|\, \gamma\Big)}.$$

Moreover, one has the de Bruijn identity:

$$\mathsf{I}\Big(\nu_t \,|\, \gamma\Big) = -\frac{d}{dt}\mathsf{H}(\nu_t \,|\, \gamma).$$

Plugging everything together leads to

$$\begin{aligned}\frac{d^+}{dt}\mathsf{W}_2(\nu,\nu_t) \;\le\; \sqrt{\mathsf{I}\big(\nu_t\,|\,\gamma\big)} \;&\le\; -\frac{\frac{d}{dt}\mathsf{H}(\nu_t\,|\,\gamma)}{\mathsf{S}(\nu\,|\,\gamma)\sqrt{e^{\frac{2\mathsf{H}(\nu_t|\gamma)}{\mathsf{S}^2(\nu|\gamma)}}-1}} \\ &=\; -\frac{d}{dt}\bigg(\mathsf{S}\big(\nu\,|\,\gamma\big)\,\mathsf{arccos}\Big(e^{-\frac{H(\nu_t|\gamma)}{\mathsf{S}^2(\nu|\gamma)}}\Big)\bigg).\end{aligned}$$

In other words,

$$\frac{d^+}{dt}\bigg(\mathsf{W}_2(\nu,\nu_t)+\mathsf{S}\big(\nu\,|\,\gamma\big)\,\mathsf{arccos}\Big(e^{-\frac{H(\nu_t|\gamma)}{S^2(\nu|\gamma)}}\Big)\bigg) \;\le\; 0,$$

and we get the WSH inequality, namely

$$\mathsf{W}_2(F,N) \;\le\; \mathsf{S}\big(F\,|\,N\big)\,\mathsf{arccos}\Big(e^{-\frac{\mathsf{H}(F|N)}{\mathsf{S}^2(F|N)}}\Big),$$

after integrating between $t=0$ and $t=\infty$.