CONSERVATIVE DIFFUSION AS ENTROPIC GRADIENT FLUX OF STEEPEST DESCENT

IOANNIS KARATZAS
Department of Mathematics
Columbia University, New York

Joint work with
Walter SCHACHERMAYER and Bertram TSCHIDERER
University of Vienna

van EENAM Lecture at the University of Michigan
Ann Arbor, September 2019

PROLOGUE

We provide a probabilistic and trajectorial interpretation, based on stochastic calculus, for the variational characterization of diffusion as **entropic gradient flux.**

This was first established in the seminal paper by Jordan-Kinderlehrer-Otto (1998). It was shown by those authors [1] that, for diffusions of the Langevin-Smoluchowski type

$$\mathrm{d}X(t) = -\nabla\Psi\big(X(t)\big)\,\mathrm{d}t + \mathrm{d}W(t)\,,$$

the associated Fokker-Planck probability density flow *follows, at each instant in time, the direction of steepest descent for the associated relative entropy (or free energy) functional.*

[1] Through an Euler discretization scheme on the Fokker-Planck probability density flow. This was then extended by Otto to more general settings via a set of tools that became known colloquially as "Otto Calculus".

The notion of steepest descent, or "gradient flux", makes sense only in the context of an appropriate metric. In this case, of the **quadratic Wasserstein distance** traveled in the *ambient space* – the space of probability measures with finite second moments.

We obtain here novel, stochastic-process versions of these features. These are valid along *(almost) every trajectory of the diffusive motion* in both the forward and, most transparently, the **backward**, [2] directions of time. We use for this a very direct perturbation analysis.

By **averaging** these trajectorial results with respect to the underlying measure on path space – i.e., by *just taking expectations* – we establish the maximality of the rate of entropy dissipation along the Fokker-Planck flow.

. And what is more, we measure *precisely* the deviation from this maximum that corresponds to any given perturbation. [3]

---

[2] As in Fontbona-Jourdain (2016).

[3] Our approach can be described pithily, as doing and extending "Otto calculus via Itô calculus"; or working at the very microscopic, particulate level.

AN UNEXPECTED BONUS

As a bonus of this perturbation analysis, the famous so-called **HWI inequality** of Otto and Villani (2000), relating

- relative entropy **(H)**,
- Wasserstein distance **(W)** and
- relative Fisher information **(I)**,

literally falls on our lap.

And with it, so do some basic inequalities of Functional Analysis:

- Talagrand,
- Log-Sobolev (P. Federbush for $\Psi(x) = |x|^2$), and
- Poincaré.

THE SETTING

We start with a "confinement potential well" $\Psi : \mathbb{R}^n \to [0, \infty)$.
Think quadratic,

$$\Psi(x) = |x|^2\,;$$

but our conditions allow "double well" potentials of the sort

$$\Psi(x) \;=\; \left(x^2 - a^2\right)^2, \qquad x \in \mathbb{R}\,,$$

or even $\Psi(x) \equiv 0$.

. We place Brownian particles in such a potential well.
They diffuse, but also "slide along the edges of the well",
according to the **Langevin-Smoluchowski equation**

$$\mathrm{d}X(t) = -\nabla\Psi\big(X(t)\big)\,\mathrm{d}t + \mathrm{d}W(t)\,,$$

where $W(\cdot)$ is standard Brownian motion in $\mathbb{R}^n$.

Because of the *sliding towards the bottom of the well,* this motion is "conservative": has an invariant distribution Q on $\mathcal{B}(\mathbb{R}^n)$, the so-called **Gibbs measure** with density

$$q(x) := e^{-2\Psi(x)}$$

relative to Lebesgue measure.

No need to assume here that this Q is finite
(so it can be normalized to a probability):

It is enough that Q be a $\sigma-$finite measure.

We posit now an initial distribution $P(0)$ of particles with density $p_0(\cdot)$ that admits finite second moment: $\int_{\mathbb{R}^n} |x|^2 p_0(x)\mathrm{d}x < \infty$.

. Under the **coercivity condition** on the confinement potential [4]

$$\Big\langle x, \nabla\Psi(x)\Big\rangle_{\mathbb{R}^n} + c\,|x|^2 \geq 0\,, \qquad \forall \;\; |x| > R\,,$$

for suitable positive real constants c, R, such finiteness propagates:

. With $P(t)$ the distribution of particles at time $t \in [0, \infty)$, the corresponding density $p(t, \cdot)$ also admits a finite second moment: $\int_{\mathbb{R}^n} |x|^2 p(t,x)\mathrm{d}x < \infty$.

This probability density function satisfies the **Fokker-Planck** (or **forward Kolmogorov**) equation

$$\partial_t p(t,x) \;=\; \frac{1}{2}\,\Delta p(t,x) + \mathrm{div}\Big(\nabla\Psi(x)\,p(t,x)\Big).$$

[4] This condition is not needed, when Ψ dominates a quadratic: $\Psi(x) \geq c\,|x|^2$ for some real $c > 0$.

WASSERSTEIN DISTANCE

The coercivity condition ensures that the resulting flow of probability measures

$$\Big(P(t)\Big)_{0\leq t<\infty}, \qquad P(t,A) = \int_A p(t,x)\,\mathrm{d}x$$

is a curve in the space

$$\mathcal{P}_2(\mathbb{R}^n)\ ,$$

the so-called **quadratic Wasserstein space** of probability measures with finite second moment.

We turn this space into a **metric space**, with the familiar distance [5]

$$W(\mu,\nu) := \left(\inf_{\substack{Y\sim\mu \\ Z\sim\nu}} E\left|Y - Z\right|^2 \right)^{1/2}.$$

This $\mathcal{P}_2(\mathbb{R}^n)$ is the ambient space, where the configuration (particle density function) of our system of particles will live.

[5] The quadratic cost of transporting μ to ν. The infimum is actually attained; more on this down the road. This metric space has very interesting geometric (non-Euclidean) features: Otto, then Lott, Villani and Sturm.

But why this distance? Why is it relevant in our context?

Because of its Benamou-Brenier (1997) minimum action, or *"minimum total kinetic energy"*, representation of the Wasserstein distance

$$W^2(\mu,\nu) = \inf_{(\rho,v)} \int_0^1 \int_{\mathbb{R}^n} \big|v(t,x)\big|^2 \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t.$$

Here the infimum is taken over all probability density $\rho(t,\cdot)$ and velocity $v(t,\cdot)$ (scalar and vector, respectively) fields, that satisfy the **transport equation of fluid mechanics**

$$\partial_t \rho(t,x) + \mathrm{div}\Big(v(t,x)\,\rho(t,x)\Big) = 0\,, \qquad 0 < t < 1$$

and the initial/terminal conditions
$\rho(0,\cdot)$ the density of μ,
$\rho(1,\cdot)$ the density of ν.

. Our Fokker-Planck equation

$$\partial_t p(t,x) \;=\; \frac{1}{2}\,\Delta p(t,x) + \mathrm{div}\Big(\nabla\Psi(x)\,p(t,x)\Big).$$

is *precisely* of this fluid transport equation form

$$\partial_t p(t,x) + \mathrm{div}\Big(v(t,x)\,p(t,x)\Big) = 0\,,$$

with velocity field which is a gradient:

$$v(t,x) \;=\; -\nabla\Psi(x) - \frac{1}{2}\,\frac{\nabla p(t,x)}{p(t,x)}\,.$$

This captures the effects of both drift, and diffusion (speed of the transport induced by the random motion with transitions $p(t,\cdot)$).

It is the *"rayonnement des probabilités"* of Bachelier (1900). We'll be seeing an awful lot of it throughout this talk.

RELATIVE ENTROPY

For every probability measure $P \in \mathcal{P}_2(\mathbb{R}^n)$ in this quadratic Wasserstein space, it is possible to define the **Shannon relative entropy** with respect to the invariant (Gibbs) measure Q, as

$$H(P|Q) := \int_{\mathbb{R}^n} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P \quad \in \quad (-\infty, \infty]$$

if $P \ll Q$; and as $H(P|Q) := \infty$ otherwise. [6]

. *IF this invariant Q is a probability measure,* then the relative entropy $H(P|Q)$ is well-defined and **non-negative**. And the density function q of Q minimizes then the free energy functional of the next slide.

[6] We are relying here on a construction of Ch. Léonard.

ENERGY AND ENTROPY

For any probability density function $\rho : \mathbb{R}^n \to [0, \infty)$, let us introduce the **confinement potential** energy and the Gibbs-Boltzmann internal energy or **entropy**

$$E(\rho) \;:=\; \int_{\mathbb{R}^n} \Psi(x)\,\rho(x)\,\mathrm{d}x\,, \qquad S(\rho) \;:=\; \int_{\mathbb{R}^n} \rho(x)\,\log \rho(x)\,\mathrm{d}x\,,$$

respectively, as well as the "free energy" (sum of potential and internal energies) [7]

$$F(\rho) \;:=\; E(\rho) + \frac{1}{2}\,S(\rho)\,.$$

[7] It is possible to include in this sum an "interaction potential" (generalized Curie-Weiss) component

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} \Phi(x-y)\,\rho(x)\,\rho(y)\,\mathrm{d}x\,\mathrm{d}y\,.$$

Results carry through, but not as crisply, or as sharply focused, as they are here.

We shall assume throughout, that our initial particle configuration has finite free energy:

$$\boxed{F\big(p_0(\cdot)\big) < \infty\,.}$$

. Then the free energy $t \mapsto F\big(p_t(\cdot)\big)$ **decreases** along the flow of probability density functions $\big(p_t(\cdot)\big)_{0 \le t < \infty}$, or equivalently along the flow of the corresponding measures

$$\Big(P(t)\Big)_{0 \le t < \infty} \subset \mathcal{P}_2(\mathbb{R}^n)\,.$$

. And is a constant multiple of the relative entropy with respect to the invariant (Gibbs) distribution: To wit, the function

$$t \longmapsto 2\,F\big(p_t(\cdot)\big) \;=\; H\big(P(t)\big|\,Q\big) \in \mathbb{R} \quad \text{decreases}$$

in accordance with the second law of thermodynamics. More about this decrease, and about its temporal and "ambient" rates, in a moment.

PROBABILISTIC SETTING AND NOTATION

Let us denote by $\mathbb{P}$ the probability measure induced by the diffusion $X(\cdot)$ governed by the equation

$$\mathrm{d}X(t) = -\nabla\Psi\big(X(t)\big)\,\mathrm{d}t + \mathrm{d}W(t)$$

on the space $C([0,\infty);\mathbb{R}^n)$ of continuous functions, and under which $W(\cdot)$ is Brownian motion.

We introduce the likelihood ratio **function**

$$\ell(t,x) := \frac{p(t,x)}{q(x)} = p(t,x)\,e^{2\,\Psi(x)}\,, \qquad x \in \mathbb{R}^n$$

and the likelihood ratio **process**

$$\ell\big(t, X(t)\big) \;=\; \frac{\mathrm{d}P(t)}{\mathrm{d}Q}\big(X(t)\big)\,, \qquad 0 \le t < \infty\,.$$

TWO FUNDAMENTAL INFORMATION MEASURES

We have the representation of the **Shannon relative entropy**

$$H\big(P(t)\big|Q\big) \;=\; \mathbb{E}^{\mathbb{P}}\big[\;\log \ell\big(t, X(t)\big)\;\big]$$

$$= \int_{\mathbb{R}^n} \Big(\log p(t,x) + 2\Psi(x)\big)\Big) p(t,x)\,\mathrm{d}x\,,$$

and the definition of the **Fisher information**, or "kinetic energy"

$$I\big(P(t)\big|Q\big) \;:=\; \mathbb{E}^{\mathbb{P}}\Big[\,\big|\nabla \log \ell\big(t, X(t)\big)\big|^2\,\Big]$$

$$= \int_{\mathbb{R}^n} \Big|\nabla\big(\log p(t,x) + 2\Psi(x)\big)\Big|^2 p(t,x)\,\mathrm{d}x$$

$$= 4\int_{\mathbb{R}^n} \big|v(t,x)\big|^2\, p(t,x)\,\mathrm{d}x\,.$$

(Recall the "rayonnement des probabilités" velocity vector field v from the bottom of slide 11.)

BASIC IDENTITIES

We have the (HI, or **de Bruijn**), (WI) and (HWI) identities

$$\lim_{t \downarrow t_0} \frac{H\big(P(t)\big|Q\big) - H\big(P(t_0)\big|Q\big)}{t - t_0} = -\frac{1}{2}\, I\big(P(t_0)\big|Q\big)$$

$$\lim_{t \downarrow t_0} \frac{W\big(P(t), P(t_0)\big)}{t - t_0} = \frac{1}{2} \sqrt{I\big(P(t_0)\big|Q\big)}$$

$$\lim_{t \downarrow t_0} \frac{H\big(P(t)\big|Q\big) - H\big(P(t_0)\big|Q\big)}{W\big(P(t), P(t_0)\big)} = -\sqrt{I\big(P(t_0)\big|Q\big)}\,.$$

. The negative quantity in the first (HI, or "de Bruijn") identity, measures the **temporal** rate of decrease for the relative entropy along the curve $\big(P(t)\big)_{0 \le t < \infty}$. Thus, the (HI) identity contains the Second Law. We discuss it in detail below.

. The positive quantity in the second (WI) identity, measures the temporal velocity on the ambient space. Quite tricky to establish, as we do not have an integral representation for the numerator $W\big(P(t), P(t_0)\big)$ – only a variational representation.

. The negative quantity in the last (HWI) identity, measures the rate of decrease ("descent") of relative entropy along the curve

$$\Big(P(t)\Big)_{0 \le t < \infty}$$

in terms of the distance traveled in the ambient space $\mathcal{P}_2(\mathbb{R}^n)$.

. We shall see also that this descent is the "steepest possible" in a sense we are about to make very precise.

PERTURBATION ANALYSIS

Fix an arbitrary time $t_0 \in (0, \infty)$. Keep the same dynamics

$$\mathrm{d}X(t) = -\nabla\Psi\big(X(t)\big)\,\mathrm{d}t + \mathrm{d}W(t)\,, \qquad 0 < t < t_0$$

for the diffusive particles, on $[0, t_0]$; but from that time onward, perturb their drift by the gradient $\beta = \nabla B$ of a smooth, compactly supported potential:

$$\mathrm{d}X(t) = -\big[\nabla\Psi\big(X(t)\big) + \nabla B\big(X(t)\big)\big]\mathrm{d}t + \mathrm{d}W^{\beta}(t)\,, \quad t > t_0\,.$$

Denote by $\mathbb{P}^{\beta}$ the probability measure induced in this manner on the space $C([0, \infty); \mathbb{R}^n)$ of continuous functions, under which W^{β} is Brownian motion.

And denote by $P^{\beta}(t)$ the $\mathbb{P}^{\beta}-$distribution of the random variable $X(t)$; clearly, $P^{\beta}(t_0) \equiv P(t_0)$.

We introduce the random vectors

$$\mathfrak{a} := \nabla \log \ell\big(t_0, X(t_0)\big)\,,$$

$$\mathfrak{b} := \beta\big(X(t_0)\big)$$

in $L^2(\mathbb{P})$, and denote by

$$\Big\langle\Big\langle\, \mathfrak{a}, \mathfrak{b}\, \Big\rangle\Big\rangle_{L^2(\mathbb{P})} = \mathbb{E}^{\mathbb{P}}\Big(\big\langle \mathfrak{a}, \mathfrak{b} \big\rangle_{\mathbb{R}^n}\Big)\,, \qquad \big\|\, \mathfrak{a}\, \big\|_{L^2(\mathbb{P})}$$

the inner product and norm, respectively, in this Hilbert space $L^2(\mathbb{P})$.

BASIC IDENTITIES: THE PERTURBED CASE

Then we have the **"perturbed"** versions $(\mathrm{HI})^\beta$, $(\mathrm{WI})^\beta$ and $(\mathrm{HWI})^\beta$ of the previous identities:

$$\lim_{t \downarrow t_0} \frac{H\big(P^\beta(t)\big|Q\big) - H\big(P^\beta(t_0)\big|Q\big)}{t - t_0} = -\frac{1}{2} \Big\langle\Big\langle\, \mathfrak{a}, \mathfrak{a} + 2\mathfrak{b} \,\Big\rangle\Big\rangle_{L^2(\mathbb{P})}$$

$$\lim_{t \downarrow t_0} \frac{W\big(P^\beta(t), P^\beta(t_0)\big)}{t - t_0} = \frac{1}{2} \Big\| \mathfrak{a} + 2\mathfrak{b} \Big\|_{L^2(\mathbb{P})}$$

$$\lim_{t \downarrow t_0} \frac{H\big(P^\beta(t)\big|Q\big) - H\big(P^\beta(t_0)\big|Q\big)}{W\big(P^\beta(t), P^\beta(t_0)\big)} = -\Bigg\langle\Bigg\langle\, \mathfrak{a}\,, \frac{\mathfrak{a} + 2\mathfrak{b}}{\big|\mathfrak{a} + 2\mathfrak{b}\big|} \Bigg\rangle\Bigg\rangle_{L^2(\mathbb{P})} .$$

- We recall here from the previous slide the random vectors

$\mathfrak{a} := \nabla \log \ell\big(t_0, X(t_0)\big)$, $\qquad \mathfrak{b} := \beta\big(X(t_0)\big)$.

Now let us compare this last "perturbed" slope

$$S^\beta(t_0) \;:=\; \lim_{t \downarrow t_0} \frac{H\big(P^\beta(t)\big|Q\big) - H\big(P^\beta(t_0)\big|Q\big)}{W\big(P^\beta(t), P^\beta(t_0)\big)}$$

$$= -\left\langle\!\!\left\langle \mathfrak{a}\,, \frac{\mathfrak{a}+2\mathfrak{b}}{\big|\mathfrak{a}+2\mathfrak{b}\big|} \right\rangle\!\!\right\rangle_{L^2(\mathbb{P})} ,$$

with the "unperturbed" slope

$$S(t_0) := \lim_{t \downarrow t_0} \frac{H\big(P(t)\big|Q\big) - H\big(P(t_0)\big|Q\big)}{W\big(P(t), P(t_0)\big)} = -\big\|\,\mathfrak{a}\,\big\|_{L^2(\mathbb{P})}$$

from a few slides upstream. We recall again

$$\mathfrak{a} := \nabla \log \ell\big(t_0, X(t_0)\big)\,, \qquad \mathfrak{b} := \beta\big(X(t_0)\big).$$

Their difference in non-negative (Cauchy-Schwarz)

$$S^{\beta}(t_0) - S(t_0) \;=\; \| \,\mathfrak{a}\, \|_{L^2(\mathbb{P})} - \left\langle\!\!\left\langle \mathfrak{a}\,, \frac{\mathfrak{a}+2\mathfrak{b}}{|\mathfrak{a}+2\mathfrak{b}|} \right\rangle\!\!\right\rangle_{L^2(\mathbb{P})} \geq 0\,,$$

in fact strictly positive unless the random vectors $\mathfrak{a} = \nabla \log \ell\big(t_0, X(t_0)\big)$, $\mathfrak{b} = \beta\big(X(t_0)\big)$ are collinear.

. The **(negative)** slope along the original, unperturbed Fokker-Planck flow $\big(P(t)\big)_{t\geq 0}$, namely

$$S(t_0) := \lim_{t\downarrow t_0} \frac{H\big(P(t)\big|Q\big) - H\big(P(t_0)\big|Q\big)}{W\big(P(t), P(t_0)\big)} = -\,\| \,\mathfrak{a}\, \|_{L^2(\mathbb{P})}\,,$$

emerges in this manner as the slope of **steepest descent** among all such perturbations.

This is the “entropic gradient flux” property of the title.

THE JOYS AND BEAUTY OF PROBABILITY

How do we get to do all this? The short (and smug) answer is,

*"Well, we let the trajectories do all the work for us;
then we just go in, and take expectations."*

Or, as my colleague and collaborator Nicole El Karoui likes to say:

"Elles sont jolies, les Probas!"

But here with a twist: **It pays to look at trajectories BACKWARDS, in the reverse direction of time.**

TIME REVERSAL: A FIRST EYE-OPENER

. A very strong hint, as to why such an approach might pay off, comes from the fact that the likelihood ratio

$$\ell(t,x) := \frac{p(t,x)}{q(x)} = p(t,x)\, e^{2\,\Psi(x)}\,, \qquad x \in \mathbb{R}^n\,,$$

a quantity of great importance in everything that goes on here, satisfies not a forward (like the Fokker-Planck), but a **backward**, Kolmogorov equation:

$$\partial_t \ell(t,x) \;=\; \frac{1}{2}\,\Delta\ell(t,x) - \Big\langle \nabla\ell(t,x), \nabla\Psi(x) \Big\rangle_{\mathbb{R}^n}\,.$$

. Please note that the velocity vector field $v(t,\cdot)$ from slide 11, is cast now as a multiple of the gradient of the log-likelihood ratio:

$$-\frac{1}{2}\,\nabla \log \ell(t,x) \;=\; v(t,x)\,.$$

We fall back now on a long line that stretches back to Schrödinger (1931), Kolmogorov (1937) and passes, via Nagasawa (1963) and Nelson (1966), to Föllmer (1985, 86), Haussmann & Pardoux (1986), and finally Meyer (1994).

Thus, we fix $T \in (t_0, \infty)$, and look at the time-reversed process $X(t-s)$, $0 \le s \le T$ and at the filtration it generates:

$$\mathcal{G}(T-s) := \sigma\big(X(T-u), 0 \le u \le s\big)\,, \qquad 0 \le s \le T.$$

• *How does this look like?*
Is it (strongly) Markovian?
Is it a semimartingale?
Is it a diffusion?

It stands to reason that, if you want to go backwards from where you are, you'd better know how you got there in the first place.

This is the Ariadne-and-Theseus truism. Here, it means we have to know the transition probabilities.

. Turns out, Ariadne's thread is given here by *the gradient of the log-transition-probability-density function* — which acts as an additive component to the potential. And when the transition probability density is smooth, the time-reversed process is a diffusion (in particular, strongly Markovian), in fact again with *gradient drift:*

$$\mathrm{d}X(T-s) = \nabla\Big(\log p(T-s,\cdot) + \Psi\Big)\big(X(T-s)\big)\mathrm{d}s + \mathrm{d}\overline{W}(T-s)$$

$$= \nabla\Big(\log \ell\big(T-s, X(T-s)\big) - \Psi\big(X(T-s)\big)\Big)\mathrm{d}s + \mathrm{d}\overline{W}(T-s).$$

Here $\left(\overline{W}(T-s)\right)_{0\leq s\leq T}$ is a $\mathbb{P}-$Brownian motion of the time-reversed filtration.

. Please note that the drift vector field in this equation, is given as

$$-2v(T-s,\cdot)-\nabla\Psi(\cdot)$$

in terms of the velocity vector field in the fluid-mechanical picture of slides 11, 12.

Here is a basic result, the bedrock of the analysis we carry out here.

THEOREM: Backwards de Bruijn Martingale.
Fix a time-window $[0, T]$ *of finite length, and consider the cumulative Fisher information process*

$$F(T-s) := \frac{1}{2}\int_0^s \left|\nabla \log \ell(T-u, X(T-u))\right|^2 \mathrm{d}u\,, \quad 0 \le s \le T$$

accumulated from the right. Then we have $\mathbb{E}^{\mathbb{P}}(F(0)) < \infty\,,$ *and the process*

$$M(T-s) := \log\left(\frac{\ell(T-s, X(T-s))}{\ell(T, X(T))}\right) - F(T-s)$$

$$= \int_0^s \left\langle \nabla \log \ell(T-u, X(T-u)), \mathrm{d}\overline{W}(T-u)\right\rangle_{\mathbb{R}^n}$$

for $0 \le s \le T$*, is a square-integrable* $\mathbb{P}-$*martingale of the backwards filtration – with quadratic variation* $[M, M] = 2F\,.$

TAKING STOCK

The gist of this result, and of the one that follows, is that the trade-off between

- the decay of relative entropy
- and the Wasserstein transportation cost,

both of which are characterized in terms of the cumulative Fisher information process, are valid not only in expectation but also **along (almost) every trajectory** of the diffusive motion, provided time is run in the reverse direction.

Working at the level of individual particles, rather than their ensembles.

The **de Bruijn** [8] **identity** for the dissipation of relative entropy

$$H(P(t)|Q) - H(P(t_0)|Q) = \mathbb{E}^{\mathbb{P}}\left[\log\left(\frac{\ell(t, X(t))}{\ell(t_0, X(t_0))}\right)\right]$$

$$= -\frac{1}{2}\int_{t_0}^{t} I(P(\theta)|Q)\,\mathrm{d}\theta$$

with $t \geq t_0$, follows from this right away.

. Just by taking expectations, as we were saying – and using the fact that martingales have constant expectations:

$$\mathbb{E}^{\mathbb{P}}\left[M(T)\right] = \mathbb{E}^{\mathbb{P}}\left[M(0)\right].$$

[8] The seminal paper of Stam (1959) from the early days of Information Theory, establishes this "blue" identity for $\Psi(x) = |x|^2/4$, i.e., when the Gibbs measure Q is standard Gaussian. Stam credits his teacher, the analyst, number theorist, combinatorialist and logician Nicolaus DE BRUIJN.

TIME REVERSAL: THE PERTURBED CASE

In a completely analogous manner, we carry out this analysis also in the "perturbed" case.

Turns out, the dynamics are (Ariadne's thread, plus perturbation)

$$\mathrm{d}X(T-s) = \nabla\big(\log p^{\beta} + \Psi + B\big)(T-s, X(T-s))\mathrm{d}s + \mathrm{d}\overline{W}^{\beta}(T-s)$$

$$= \nabla\big(\log \ell^{\beta} - \Psi + B\big)(T-s, X(T-s))\mathrm{d}s + \mathrm{d}\overline{W}^{\beta}(T-s)$$

with $\big(\overline{W}^{\beta}(T-s)\big)_{0 \le s \le T}$ a $\mathbb{P}^{\beta}-$Brownian motion of the time-reversed filtration.

Here the "perturbed likelihood ratio" is

$$\ell^{\beta}(t,x) := \frac{p^{\beta}(t,x)}{q(x)} = p^{\beta}(t,x)\, e^{2\,\Psi(x)}\,, \qquad x \in \mathbb{R}^n\,.$$

THEOREM: "Perturbed" Backwards de Bruijn Martingale.
Fix a time-window $[0, T]$ *of finite length* $0 < t_0 < T < \infty$, *and consider the Fisher information process for the perturbed dynamics, accumulated from the right*

$$F^\beta(T-s) \;:=\; \int_0^s \Big[\, \frac{1}{2}\big|\nabla \log \ell^\beta(T-u, X(T-u))\big|^2 + \Big(\big\langle \beta, 2\,\nabla\Psi \big\rangle_{\mathbb{R}^n} - div(\beta)\Big)\big(X(T-u)\big)\Big]\, \mathrm{d}u\,,$$

for $0 \le s \le T - t_0$. *Then* $\mathbb{E}^{\mathbb{P}^\beta}\big(F^\beta(t_0)\big) < \infty$, *and the process*

$$M^\beta(T-s) \;:=\; \log\left(\frac{\ell^\beta(T-s, X(T-s))}{\ell^\beta(T, X(T))}\right) - F^\beta(T-s)$$

$$= \int_0^s \Big\langle \nabla \log \ell^\beta(T-u, X(T-u)), \mathrm{d}\overline{W}^\beta(T-u) \Big\rangle_{\mathbb{R}^n}$$

for $0 \le s \le T - t_0$, *is a square-integrable* $\mathbb{P}^\beta-$*martingale of the backwards filtration, with* $[M^\beta, M^\beta] = 2F^\beta$.

Once again, the integral version **"perturbed" (HI), or de Bruijn, identity**

$$H(P^\beta(t)|Q) - H(P^\beta(t_0)|Q) = \mathbb{E}^{\mathbb{P}^\beta}\left[\log\left(\frac{\ell^\beta(t,X(t))}{\ell^\beta(t_0,X(t_0))}\right)\right]$$

$$= \int_{t_0}^{t} \mathbb{E}^{\mathbb{P}^\beta}\Big(-\frac{1}{2}\left|\nabla \log \ell^\beta(\theta, X(\theta))\right|^2 +$$

$$+ \Big(\operatorname{div}(\beta) - \langle \beta, 2\nabla\Psi\rangle_{\mathbb{R}^n}\Big)(X(\theta))\Big)\mathrm{d}\theta$$

for $t \geq t_0$ follows from this right away, just by taking expectations.

. We obtain now the differential version of the perturbed (HI) identity, the blue identity on slide 21, simply dividing by $t - t_0$, letting $t \downarrow t_0$, and integrating by parts.

TECHNICAL WORK

Lots of technical details are hidden here under the rug.
The indicated derivatives exist only outside a (countable, at most) set of exceptional points.

Quite delicate analysis is necessary, in order to show that the exceptional set N, for the temporal dissipation of relative entropy, is *the same for the perturbed case, as for the unperturbed.*

And even greater delicacy is needed, in showing that the above set N is also the exceptional set for the temporal growth of the Wasserstein distance along the Fokker-Planck flows — *both unperturbed, and perturbed.*

CONDITIONAL TRAJECTORIAL RATES OF RELATIVE ENTROPY DISSIPATION, "à la De BRUIJN"

With $0 < t_0 < T - s < T$, we have the $\mathbb{L}^1(\mathbb{P})-$limits:

$$\lim_{s \uparrow T - t_0} \frac{1}{T - t_0 - s} \mathbb{E}^{\mathbb{P}} \left[\log \left(\frac{\ell(T-s, X(T-s))}{\ell(t_0, X(t_0))} \right) \Big| \mathcal{G}(T-s) \right]$$

$$= -\frac{1}{2} \left| \nabla \log \ell \left(t_0, X(t_0) \right) \right|^2 ,$$

$$\lim_{s \uparrow T - t_0} \frac{1}{T - t_0 - s} \log \left(\frac{\ell^{\beta}(T-s, X(T-s))}{\ell\,(T-s, X(T-s))} \right)$$

$$= \left(\operatorname{div}(\beta) + \left\langle \beta, \nabla \log p(t_0, \cdot) \right\rangle_{\mathbb{R}^n} \right) \left(X(t_0) \right) ,$$

$$\lim_{s\uparrow T-t_0} \frac{1}{T-t_0-s} \mathbb{E}^{\mathbb{P}}\left[\log\left(\frac{\ell^\beta(T-s, X(T-s))}{\ell^\beta(t_0, X(t_0))}\right) \,\Big|\, \mathcal{G}(T-s)\right]$$

$$= \left(\mathrm{div}(\beta) + \Big\langle \beta, \nabla \log p(t_0,\cdot)\Big\rangle_{\mathbb{R}^n}\right)(X(t_0)) - \frac{1}{2}\left|\nabla \log \ell(t_0, X(t_0))\right|^2,$$

$$\lim_{s\uparrow T-t_0} \frac{1}{T-t_0-s} \mathbb{E}^{\mathbb{P}^\beta}\left[\log\left(\frac{\ell^\beta(T-s, X(T-s))}{\ell^\beta(t_0, X(t_0))}\right) \,\Big|\, \mathcal{G}(T-s)\right]$$

$$= \left(\mathrm{div}(\beta) - \Big\langle \beta, \nabla \log p(t_0,\cdot)\Big\rangle_{\mathbb{R}^n}\right)(X(t_0)) - \frac{1}{2}\left|\nabla \log \ell(t_0, X(t_0))\right|^2.$$

Hard to tell as yet, what additional information might be gleaned from these very precise, trajectorial descriptions of relative entropy dissipation. Perhaps CLT and LD-type results as well.

A SECOND EYE-OPENER, IN A SPECIAL CASE

Suppose $q(x) = e^{-2\Psi(x)}$ is a probability density function.

Imagine starting the Langevin-Smoluchowski diffusion

$$\mathrm{d}X(t) = -\nabla\Psi\big(X(t)\big)\,\mathrm{d}t + \mathrm{d}W(t)$$

with $X(0)$ — thus also $X(t)$ for all $t > 0$ — having this invariant density. Denote by $\mathbb{Q}$ the probability measure induced on the space $C\big([0,\infty);\mathbb{R}^n\big)$ by the continuous process $X(\cdot)$.

. Then the time-reversed likelihood ratio process

$$\ell\big(T-s, X(T-s)\big), \quad 0 \le s \le T \qquad \text{is a } \mathbb{Q} - \textbf{martingale}$$

of the backwards filtration, for any given $T \in (0,\infty)$; Fontbona and Jourdain (2016). VERY SIMPLE PROOF IS GIVEN HERE.

And the decrease of the relative entropy

$$H\big(P(T)\big|Q\big) \;=\; \mathbb{E}^{\mathbb{P}}\big[\log \ell\big(T, X(T)\big)\big]$$

$$=\; \mathbb{E}^{\mathbb{Q}}\big[\ell\big(T, X(T)\big)\,\log \ell\big(T, X(T)\big)\big]$$

as a function of $T \in (0, \infty)$, is now a direct consequence of the convexity of the function

$$f(x) \;=\; x \log x$$

and of the Jensen inequality – **always in the special case when Q is a probability measure.** [9]

[9] If there is a simpler proof of the second law, I have not seen it. It is also possible, though with quite a bit of extra work, to make sense of these properties also in the general, $\sigma-$finite setting (Bertram's Master's Thesis in Vienna).

THE HWI INEQUALITY, UNDER CONVEXITY

Let us suppose now that $Q \in \mathcal{P}_2(\mathbb{R}^n)$, and that the potential satisfies the convexity condition

$$D^2\Psi(x) \geq \kappa\,\mathrm{Id}, \quad \forall \ x \in \mathbb{R}^n$$

for some real number κ. We claim that, in this case, the perturbed (HI) and (HWI) identities of slide 21 contain, in seminal form, the celebrated Otto-Villani (2000) **HWI inequality:** [10]

$$H(P_0|Q) - H(P_1|Q) \leq W(P_0, P_1)\sqrt{I(P_0|Q)} - (\kappa/2)\, W^2(P_0, P_1)\ .$$

for any two probability measures $P_0 \in \mathcal{P}_2(\mathbb{R}^n)$, $P_1 \in \mathcal{P}_2(\mathbb{R}^n)$,

[10] Here, though, for a $\sigma-$finite measure Q, that need not be a probability.

. When $\kappa > 0$ in the convexity requirement

$$D^2\Psi(x) \geq \kappa \, \mathrm{Id}, \quad \forall \ x \in \mathbb{R}^n,$$

this inequality

$$\boxed{H(P_0|Q) - H(P_1|Q) \leq W(P_0, P_1)\sqrt{I(P_0|Q)} - (\kappa/2)\, W^2(P_0, P_1)}$$

leads, *very directly,* to some celebrated inequalities of Functional Analysis for $\mathcal{P}_2(\mathbb{R}^n)$:
the **Talagrand** (1996) inequality

$$W^2(P, Q) \leq \frac{2}{\kappa} H(P|Q),$$

the **Log-Sobolev** inequality of **Federbush** (1969) (Gross (1975))

$$H(P|Q) \leq \frac{1}{2\,\kappa} I(P|Q),$$

and also the **Poincaré** inequality.

. It gives also the exponential dissipation of relative entropy for our original Fokker-Planck flow:

$$H\big(P(t)|Q\big) \le H\big(P(0)|Q\big)\, e^{-\kappa t}, \qquad 0 \le t < \infty .$$

OK, let's try to explain how this is done.

• Pick and fix two probability measures P_0, P_1 in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^n)$, with smooth, compactly supported densities to make life simple (though ultimately not necessary).

Transport P_0 to P_1 by means of a **constant-speed geodesic** $\big(P_t\big)_{0\le t\le 1}$, as follows: [11]

[11] Ask our car's GPS to find the shortest route, between where we are and where we want to go (geodesic). Then put the car on cruise control, to get there along this geodesic and at the "right" constant speed (which?). We imagine also that the landscape of the Wasserstein space has hills and valleys: that is, an associated notion of *curvature*. More on this soon.

For some convex $G : \mathbb{R}^n \to \mathbb{R}$, we have $P_1 = (\nabla G)_\# P_0$, and

$$W^2(P_0, P_1) = \inf_{\substack{X \sim P_0 \\ Y \sim P_1}} E|X - Y|^2$$

$$= \int_{\mathbb{R}^n} |x - \nabla G(x)|^2 P_0(\mathrm{d}x) = \|\gamma\|^2_{L^2(P_0)}$$

with $\gamma(x) := \nabla G(x) - x$ by Brenier (1991). [12] Furthermore,

$$W(P_0, P_t) = t \cdot \|\gamma\|_{L^2(P_0)}, \qquad 0 \le t \le 1$$

holds for the "constant-speed geodesic"

$$P_t := \left(T_t^\gamma\right)_\# P_0 \, ; \qquad T_t^\gamma(x) := t \cdot \nabla G(x) + (1 - t) \cdot x \, , \quad x \in \mathbb{R}^n .$$

Gee, we have found out what the "right" speed is!

---

[12] "There is a convex mapping, whose gradient pushes P_0 forward to P_1 in a manner that attains the infimum in the definition of the Wasserstein distance".

Now, denote the density function of this probability measure P_t by $p_t(\cdot)$, and define the likelihood ratio

$$\ell_t(x) := \frac{p_t(x)}{q(x)} = p_t(x)\, e^{2\Psi(x)} \,, \qquad x \in \mathbb{R}^n \,.$$

Then, by complete analogy with the perturbed (HI) identity

$$\lim_{t \downarrow t_0} \frac{H\big(P^\beta(t)\big|Q\big) - H\big(P(t_0)\big|Q\big)}{t - t_0} = \Big\langle\Big\langle\, \mathfrak{a}, \mathfrak{c} \,\Big\rangle\Big\rangle_{L^2(\mathbb{P})}$$

already discussed on slide 21, with

$$\mathfrak{a} = \nabla \log \ell\big(t_0, X(t_0)\big) \,, \qquad \mathfrak{b} = \beta\big(X(t_0)\big), \qquad \mathfrak{c} = -\frac{1}{2}\, \mathfrak{a} - \mathfrak{b}$$

and identifying $P(0) = P_0$, $t_0 = 0$, $\ell(0, \cdot) = \ell_0$, we obtain the limiting relation

$$\lim_{t \downarrow 0} \frac{H(P_t|Q) - H(P_0|Q)}{t} = \Big\langle\Big\langle \nabla \log \ell_0 \, , \gamma \Big\rangle\Big\rangle_{L^2(P_0)}$$

with

$$\gamma(x) = \nabla G(x) - x = \nabla\Big(G(x) - \frac{x^2}{2}\Big) .$$

. We have then this "blue slope", along the **straight** flow, or constant-speed geodesic, $\big(P_t\big)_{0 \le t \le 1}$.

. As well as the "green slope "

$$\lim_{t \downarrow t_0} \frac{H(P^\beta(t)|Q) - H(P(t_0)|Q)}{t - t_0} = \Big\langle\Big\langle \mathfrak{a}, \mathfrak{c} \Big\rangle\Big\rangle_{L^2(\mathbb{P})} = \Big\langle\Big\langle \nabla \log \ell_0 \, , \mathfrak{c} \Big\rangle\Big\rangle_{L^2(\mathbb{P})}$$

along the **curved** Fokker-Planck flow $\big(P^\beta(t)\big)_{t \ge 0}$, where

$$\mathfrak{a} = \nabla \log \ell_0\big(X(0)\big) , \qquad \mathfrak{b} = \beta\big(X(0)\big), \qquad \mathfrak{c} = -\frac{1}{2}\,\mathfrak{a} - \mathfrak{b} .$$

- These two slopes are EXACTLY THE SAME, if we manage $\mathfrak{c} = \gamma$; that is, if we select the gradient vector field β via

$$\gamma = -\frac{1}{2} \nabla \log \ell_0 - \beta .$$

That is, we view this time around β more as an element of "control", or of "steering", than as a perturbation ... .

Note that this β is a gradient:

$$2 \beta(x) = \nabla \Big(x^2 - \log \ell_0(x) - 2\, G(x) \Big) .$$

Now write for the function $h(t) := H\big(P_t \big| Q\big)$ the Taylor expansion

$$h(1) - h(0) \,=\, h'(0+) + \int_0^1 (1-t)\, h''(t)\,\mathrm{d}t\,.$$

. From the above computation and Cauchy-Schwarz, we get

$$h'(0+) \,=\, \Big\langle\Big\langle \nabla \log \ell_0\,, \gamma \Big\rangle\Big\rangle_{L^2(P_0)} \,\geq\, -\big\| \nabla \log \ell_0 \big\|_{L^2(P_0)} \cdot \big\| \gamma \big\|_{L^2(P_0)}$$

$$= -\sqrt{I(P_0|Q)} \cdot W(P_0, P_1).$$

. Whereas, McCann's (1994) "displacement convexity" [13] gives

$$h''(t) \,\geq\, \kappa \cdot W^2\big(P_0, P_1\big)\,, \qquad 0 \leq t \leq 1$$

and the HWI inequality follows:

$$H(P_1|Q) - H(P_0|Q) \geq -W(P_0, P_1)\sqrt{I(P_0|Q)} + (\kappa/2)\, W^2(P_0, P_1)\,.$$

---

[13] A link between this notion, and the geometry of $\mathcal{P}_2(\mathbb{R}^n)$: $h(\cdot)$ is convex if, and only if, this space has nonnegative "Ricci Curvature" (no valleys).

EPILOGUE

There is a lot more to say, of course.
Hope I have said nothing but the truth.

Something resembling the *whole truth* can be found in a paper under the title "Trajectorial Otto Calculus" on *https://arxiv.org/abs/1811.08686* .

We are exploring the applicability of this trajectorial methodology to more general, and to different, settings. Some of these seem to need a lot of work.

SOME REFERENCES

Jordan, R., Kinderlehrer, D. & Otto, F. (1998) The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.* **29**, 1-17.

Fontbona, J. & Jourdain, B. (2016) A trajectorial interpretation of the dissipations of entropy and Fisher information for stochastic differential equations. *Annals of Probability* **44**, 131-170.

Föllmer, H. (1985) An entropy approach to the time reversal of diffusion processes. *Lecture Notes in Control & Information Sciences* **69**, 156-163.

Haussmann, U.G. & Pardoux, E. (1986) Time reversal of diffusions. *Annals of Probability* **14**, 1188-1205.

Villani, C. (2003) *Topics in Optimal Transportation.* Graduate Studies in Mathematics **58**. AMS Publications, Providence, RI.

Ambrosio, L., & Gigli, N. & Savaré, G. (2008) *Gradient Flows.* Lectures in Mathematics, ETH Zürich.