Entropic Methods in Geometry, Imaging and Statistics Part 1: Statistics Theory

Keith Patarroyo*

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, Montréal, QC H3C 3J7

Wolfram Physics Project
Wolfram Research, Champaign, IL 61820-7237

Abstract

We present a new family of kernel density estimators that emerge from variational optimal transport of statistical divergences(entropies). The optimal transport framework allows us to unify three seemingly different density estimation techniques(kernel density estimation via difussion, cross-entropy method and a variational approach to maximum penalized likelihood estimation) into a joint formalism. The various optimal transport kernel estimators combine the advantages of each technique and provide a great flexibility on what properties we might want to give to the estimation. In this document we present the theoretical foundations of the framework and some elementary properties of the estimators.

Contents

*March 11 2021.

1 Introduction

The notion of entropy is ubiquitous now in many scientific disciplines, all the way from biology to mathematics. Introduced initially for the problem of information transmission, it is now one of the foundations of signal processing. In the context of geometry and image processing, the notion of entropy is usually only explored in compression or sampling, only recently entropic methods taking advantage of the geometry of images and shapes have been explored to exploit this additional structure to solve different tasks. In this document we discuss and propose entropic methods and their consequences in geometry, image processing and statistics.

We start by giving the foundations for each of these methods, where we stress on the fundamental intuitions behind the specific notion of entropy utilized by each of the numerical methods. Each of the entropies make sense in a particular context, and allow in different ways to quantify the concept of uncertanty of a probability distribution over an alphabet. In our case we explore different frameworks to solve several tasks relevant to statistics, geometry and image processing.

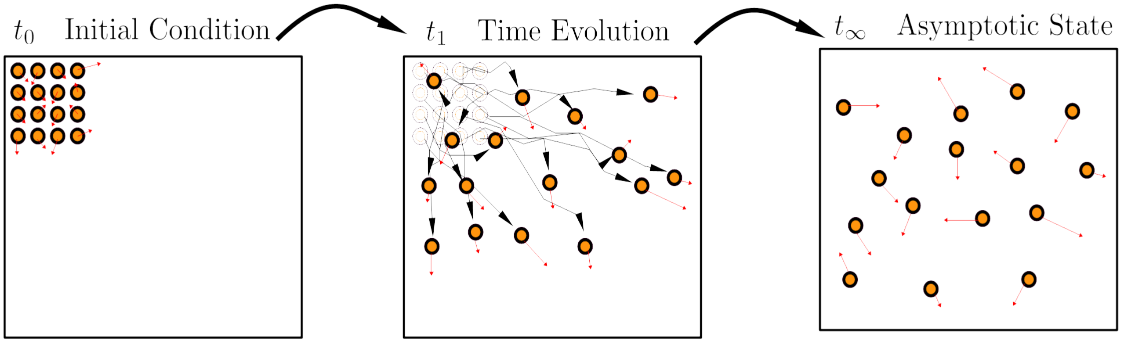

Figure 1: Experiment of gas of molecules diffussing. The molecules start at a corner of a box at an intial state t_0, then when released their properties evolve according to the dynamics of the system, as shown in time t_1. After a long time t_∞ it is reasonable to assume the velocities and position in the gas has been randomized.

2 The Phenomena of Difussion

It turns out that for many current methods involving entropy in image and geometry processing, the notion of diffusion is of fundamental importance. In some sense, this is not a surprise, since at its most basic the *physical* phenomena of heat difussion is an irreversible process, and we know from the second law of thermodynamics [67], **physical entropy** behaves irreversibly as it only increases towards the future.

More intuitively we can interpret this process in a stochastic sense following [67, 79] we can consider the experiment of setting up bunch of gas molecules in the corner of a box [39], then when we release them, they would spread out(diffuse) all over the box, Figure 1. In the following sections we explain this experiment using different complementary notions that aim to put the concept of physical entropy in a intuitive way in order to understand the motivations for the numerical methods. For a comparisson between these and classical thermodynamic interpretations see [67].

2.1 Information-Computation interpretation

We explore this experiment in two complementary ways, both related with entropy, first in the sense of classical information theory[68] and the second is to think as the evolution as a computation[79].

2.1.1 Information Interpretation

This interpretation consists in realizing that monitoring the exact state and interactions(fine-grain properties) of every molecule requires an enormous amount of work. In practice we cannot know the exact trajectory of all molecules, we have a limited resolution(experimental uncertanty) of position and velocity, hence we are constained to group particles properties in regions(coarse-graining). Effectively if we divide the effort of track the particles properties by first identifing if they lie in a region and then finding where exactly they are within the region, Figure 2, we still have complete information about the system.

Clearly we require much less effort to only do the coarse graining, hence if we ignore the fine-grain properties, we still have a good approximate idea of what is happening. This is the **key** idea, unless we have the capacity to always track the fine-grain properties of the system we have to be satisfied with a low resolution view of the world. The proccess of information loss by coarse-graining increases our amount of incertitude or entropy, hence every time we coarse-grain our entropy is bound to increase.

If we monitor the system only in this approximate way, at each time-step we will have a more and more distorted view of the real dynamics until the point where the approximate dynamics is so distorted that is better to describe the system in a global or averaged way(thermodynamic state), in some sense this is the reason why at after a long time of not knowing anything about the system the best guess is to consider the position and velocity of the molecules of the gas as random[1]. Since our amount incertitude or entropy is maximum, the best idea is to assume nothing about the system. From now on we would refer to this entropy as **Shannon entropy**. Interestingly, when we say that a *whole* room temperature is 21°C we are effectively ignoring the variation of temperature in the room and then claiming that the distribution of molecules in a room is Boltzman distributed and that its average energy or global temperature is 21°C.

[1] This does not mean that the molecules would never attain a "ordered" configuration at a certain point on time, in fact the Poincare Recurrence Theorem [39, 38, 1], states that a determinstic system would eventually arrive arbitrary close to its initial condition(it might take millions of years[71]). The randomness is **not** a property of the system, it is only the consequence of a very coarse modeling of the physical system generated by a uncompetent observer.

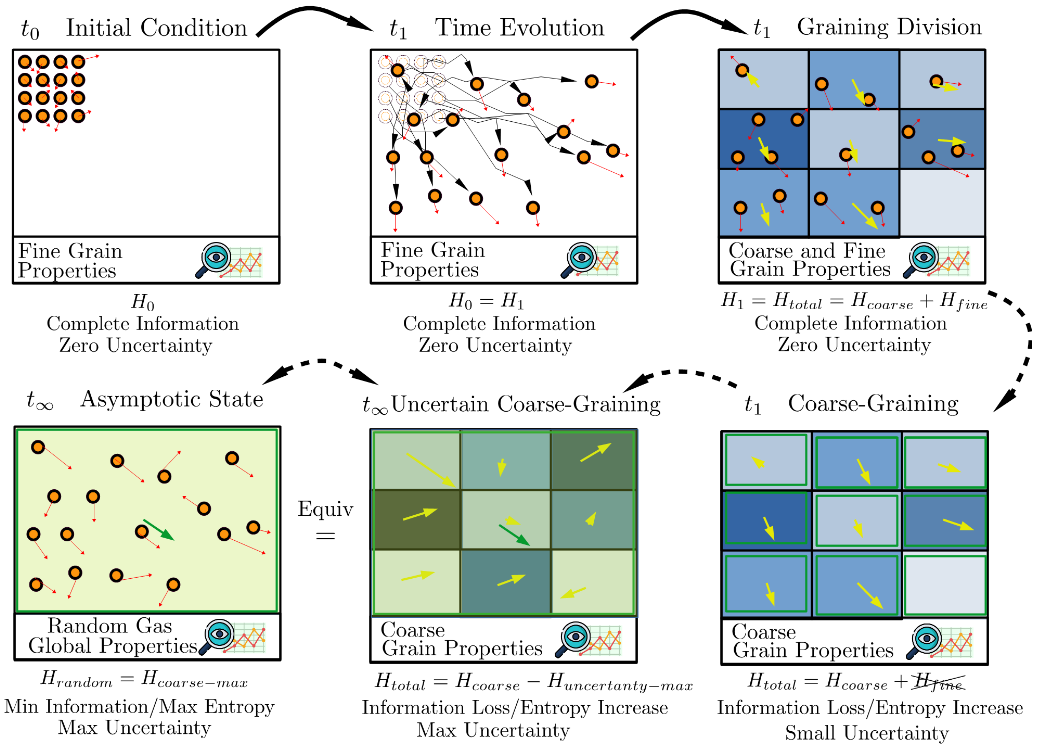
t_0 Initial Condition
Fine Grain Properties
H_0
Complete Information
Zero Uncertainty
t_1 Time Evolution
Fine Grain Properties
$H_0 = H_1$
Complete Information
Zero Uncertainty
t_1 Graining Division
Coarse and Fine Grain Properties
$H_1 = H_{total} = H_{coarse} + H_{fine}$
Complete Information
Zero Uncertainty
t_∞ Asymptotic State
Random Gas Global Properties
$H_{random} = H_{coarse-max}$
Min Information/Max Entropy
Max Uncertainty
Equiv =
t_∞Uncertain Coarse-Graining
Coarse Grain Properties
$H_{total} = H_{coarse} - H_{uncertanty-max}$
Information Loss/Entropy Increase
Max Uncertainty
t_1 Coarse-Graining
Coarse Grain Properties
$H_{total} = H_{coarse} + \cancel{H_{fine}}$
Information Loss/Entropy Increase
Small Uncertainty

Figure 2: Experiment of gas of molecules diffussing. The molecules start at a corner of a box, then when released we monitor their fine grain properties. By grouping nearby molecules in regions, we decide if we want to monitor the exact properties of the region. If we don't monitor them anymore, effectively we have lost track of the fine-grain information of each molecule. After a large time has passed t_∞ , the approximate dynamics are so distorted that our best guess is to assume the gas is essentially in a random state.

2.1.2 Computation Interpretation

The second interpretation consists in considering the evolution of positions and velocities of the molecules as a computation, more specifically as an encryption process(e.g. a hash function f that takes as input the positions and velocities at t_0 and outputs positions and velocities at t_1), we can think that the initial positions and velocities have been encrypted via an unknown procedure. After this procedure an attacker that wants to break the encryption has a restricted toolset starting from the output(velocities and positions at t_1), he can try by guessing randomly a decryption process that would lead him to the initial conditions, or if he knows the dynamical laws he can guess some initial conditions such that by evolving the physical system, they would lead him to the current conditions(essentially the physical representation of the verification notion NP problem [20]).

In some sense this decryption procedure amounts to a lossy decompression procedure, where we start at a state with very little information(lots of incertitude or Shannon entropy) about the position and velocity(starting as random) and then we arrive to a simplified or compressed representation, i.e. the initial condition plus the rules of the evolution of the physical system allow us to recover the full state of the system, in particular the positions and velocities at t_1. This analogy can be developed much further, and in fact it is highly relevant for many methods of random number generation by dynamical systems, the bedrock of many cryptographic algorithms[49]. In the following table we mention some of the simmilar properties of both frameworks,

Hash Functions	Dynamical Systems
Deterministic (Discrete)	Deterministic (Continious)
Quick to compute Hash Value from Message (Low Complexity)	Run the system Forward (Time Evolution)
Infeasible to find two different Messages with the same Hash Value (Collision)	Fine-Grain vs Coarse-Grain (Limited Resolution)
Small change to a Message should extensively change the Hash Value (Avalanche)	Chaos (Sensitivity on Initial conditions)
Infeasible to generate a Message given Hash Value (Intractability)	Second Law of Thermodynamics (Irreversibility)

Table 1: Comparison of Properties for Hash functions and Dynamical Systems

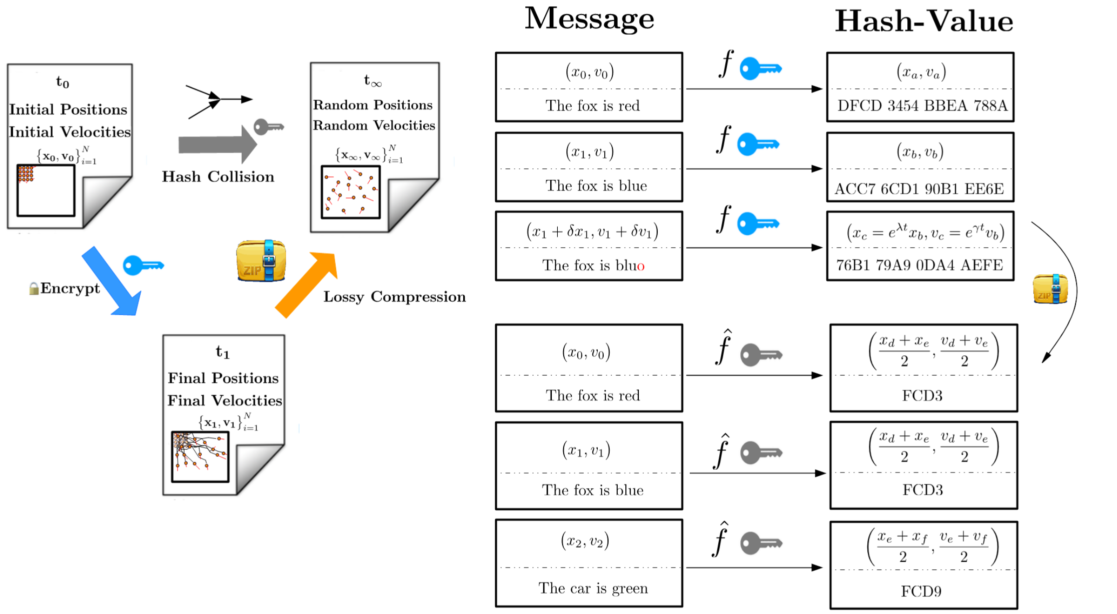

Figure 3: Experiment of gas of molecules diffussing. We only care about the state data at t_0, and then in time t_1. We can map the data at time t_1 into a compressed representation t_∞. This compressed representation is equivalent to a map with high degree of collisions from the initial data t_0. We give examples of the properties of these mapings with particles in only one dimension.

In both cases we have deterministic systems(one continious the other discrete), where it requires little effort to obtain the hash value or next time-step particle state given the message or the intial condition. A desirable Hash function will avoid collisions, i.e. different messages will have the different Hash values, since otherwise this would create errors when retriving the message, in the same way when we coarse-grain we are asigning different particle states a common value due to our limited resolution.

Remarkably other properties like the avalanche effect is analogous to the Chaotic notion of sensitivity on initial conditions, we want that a small change on the initial conditions or message to generate a big change in the final particle state or the hash value. However the most interesting property is the intractability of breaking the Hash, this is in some sense analogous to the lack of inherent increase of order within physical systems, otherwise known as the second law of thermodynamics[78]. Effectively the difficult part is to unscramble the eggs not to scramble them.

This two interpretations are results of two silent revolutions in physics, namely the introduction of the notions of information and computation in physics. The first started with Leo Szilard analysis of the Maxwell demon, arguing for a "memory resource" that compensate entropy, then after Shannon introduced formally Information theory [68] in 1948, the connections started to clarify, by the beggining of 1950 Brillouin [12] worked strongly to connect both fields. Among some remarkable work, we find Jayne's papers [42, 43] where he introduced the maximum entropy principle, the introduction of generalized entropies [72, 62, 23] and the foundation of the field of algorithmic information theory [16]. This field was very active during the following decades and perhaps lasted until the 1990's "resolution" of the Maxwell's Demon, a summary of this program can be found in [84].

The second silent revolution started maybe also in the 1940-50's with von Neumann introduction of cellular automation [77] and it gave birth in the next decades different areas of knowledge(Algoritmic Complexity [51], Non-Equilibrium Thermodynamics [28], Complexity Science [53], Quantum Computation [25], and much more) and it is still going today [83, 81]. In summary both programs claim a kind of duality between information-physics and computation-physics. There are many remarkable results by these two programs, specially about the notion of physical entropy, however they are out of scope of this document, interested readers can refer to [28, 78, 82, 80]. In this document we don't pursue directly the study of the nature of physical entropy, but we remark on those to stress on the fact that physical entropy has a natural information-computation aspect to it.

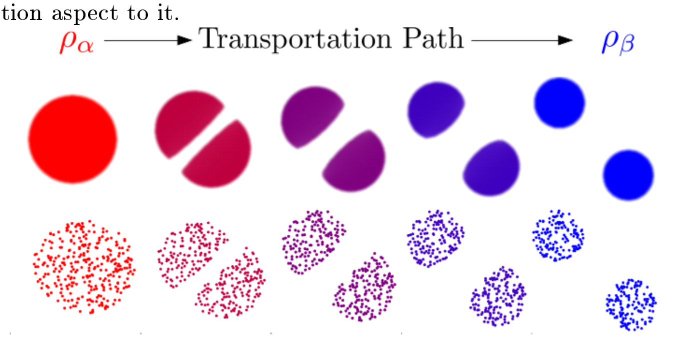

Figure 4: Optimal transport problem between densities ρ_α and ρ_β, where the red/blue-white color pallete represent the weight of the probability. In this formulation the problem states to find the minimal-path betweem the distributions equivalently as to minimize a total square cost distance of the transformation map between the original distributions. Figure modified from [60].

2.2 Optimal Transport Interpretation

Another interpretation perhaps from a not so silent revolution, comes from the field of optimal transport [60], where one asks: what is the optimal way to transport a probability density ρ_α into a different desired density function ρ_β in a way where the *cost* of doing so is the smallest ? In the case of our molecule gas, we can interpret both the initial condition and the randomized state as two probability densities, and we want to find a way to transport one to the other in a way that a desired *cost function* is minimized, Figure 4.

More importantly for our work, the optimal transport problem is a way to enhance distances between probability measures in a way where the geometry of the space is taken into account(later we would see another way, section 4). The remarkable paper [45] showed that if at each timestep we minimize a special distance (the 2-Wasserstein distance) and maximize the Shannon entropy between the previous time-step density and the next one, this flow yields exactly the diffusion equation.

This result is remarkable since it shows that the heat equation is in some sense solving a regularized optimal transport problem, and in this way maximizing the Shannon entropy at each timestep. The notion of entropy maximization is a good intuition to have for the schemes we introduce, hence we describe it extensively in the following section.

2.3 Entropy and Relative Entropy

In this section we introduce the notion of entropy and relative entropy in a more mathematical rigorous way and analyse their usage and interpretation to understand the rest of the document. We also remark on the notion of maximum entropy and the relation of entropy and stochastic processes.

2.3.1 Entropy

First we introduce more formally the notion of Shannon Entropy, for a discrete random variable X defined in the countable set $\{x_{1,}, x_2, \ldots\}$ with $p(X = x_i) = p_i$, the Shannon entropy is defined as,

$$H(p) = -\sum_{i\geq 1} p_i \log p_i. \tag{1}$$

In the case of continious probability distributions, let's consider a space $\mathcal{X}$ and a measure $\alpha \in \mathcal{M}(\mathcal{X})$(the set of all normalized positive measures on $\mathcal{X}$), such that we can have a density $d\alpha(x) = \rho_\alpha(x)\mathrm{d}m(x)$, where $x \in \mathcal{X}$ and $\mathrm{d}m(x)$ is the volume element.

Then for a random variable X on $\mathcal{X}$ with distribution α such that $p(X \in A) = \alpha(A)$ the Shannon entropy with respect to the measure α is defined as,

$$H(\alpha) = -\int_{\mathcal{X}} \rho_\alpha \log(\rho_\alpha)\ \mathrm{d}m(x). \tag{2}$$

Specifically for $\mathcal{X} = \mathbb{R}^n$ we have $d\alpha(x) = \rho_\alpha(x)dx$ with respect to the Lebesgue measure $\mathcal{L}_{\mathbb{R}^n}$, then the Shannon entropy of a random variable X with distribution α yields,

$$H(\alpha) = -\int_{\mathbb{R}^n} \rho_\alpha \log(\rho_\alpha)\ \mathrm{d}x. \tag{3}$$

For a remarkable guide to the Shannon entropy and other alternative notions of entropy we refer to [30]. Now we are prepared state the principle of maximum entropy(we remove the reference to Shannon, but it refers to the principle of maximum Shannon entropy), it was initially developed in the context of statistical mechanics [42, 43], then popularized by its creator Jaynes[44] and now is used several scientific disciplines [19, 7, 58, 24].

Maximum Entropy Principle[19]: *Suppose we are seeking a probability density function subject to certain constraints (e.g., a given mean or variance), then to determine an estimate of this function, use the density satisfying those constraints that has entropy as large as possible.*

To get a further intuition, we follow [7] and use the enlightening words of Jaynes[35],

> *...the fact that a certain probability distribution maximizes entropy subject to certain constraints representing our incomplete information is the fundamental property which justifies the use of that distribution for inference; it agrees with everything that is known but carefully avoids assuming anything that is not known. It is a transcription into mathematics of an ancient principle of wisdom...*

This leads to an interesting discussion regarding the concept of entropy, that is, the notion of scale, in the previous section we defined discrete entropy with equation (1), however hidden in this definition is the notion that there is an absolute scale in the definition of entropy.

Theorem 1 ([21]). *Let X be a discrete random variable with alphabet $\mathcal{X}$, then $H(X) \leq \log|\mathcal{X}|$ where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality if and only X has a uniform distribution over $\mathcal{X}$.*

This theorem implies that a discrete entropy is always between $0 \leq H(X) \leq \log|\mathcal{X}|$, therefore a quantity of entropy makes sense by itself(because implicitly you are always normalizing to the maximum possible amount $\log|\mathcal{X}|$). This sounds redundant, however this is key to obtain results like, ML-ME(Maximum Likelihood-Maximum Entropy) Duality[7, 69]. Interestingly enough this characterestic is transfered naturally to the differential notion of entropy (3) with a few of caveats. The first caveat is that the maximum amount of entropy depends on the space, $\mathcal{X}$. We see that for two different spaces $\mathcal{X}_1 = [a,b]$ with no constraints and $\mathcal{X}_2 = \mathbb{R}$ with variance σ^2, the maximum entropy distributions are different.

Theorem 2 ([21]). *For a continuous probability density function p on $\mathcal{X}_1 = [a,b]$, with no other constraints,*

$$H(p) \leq \ln(b-a),$$

with equality if and only if p is the uniform distribution over this range.

Theorem 3 ([19]). *For a continuous probability density function p on $\mathbb{R}$ with variance σ^2 ,*

$$H(p) \leq \frac{1}{2}(1+\log(2\pi\sigma^2)),$$

with equality if and only if p is Gaussian with variance σ^2, i.e., for some μ we have $p(x) = (1/\sqrt{2\pi}\sigma)e^{-(1/2)((x-\mu)/\sigma)^2}$.

There are two more caveats that sadly make the differential version of entropy in some sense much less expressive that its differential counterpart. Number one has to be with the fact that the differential notion of entropy can be negative, this is a really hard conceptual roadblock compared to the discrete version. Ignoring how to interpret this result, we can see that this is the maifestation that a continious probability distribution can be in some sense more concentrated than the uniform $\mathcal{U}(0,1)$.

The next roadblock is perhaps even more problematic for the conceptual interpretation, note that we are taking the logarithm of probabilites or densities. This is not a problem for discrete probabilities, since they are naturally ratios of quantities(ratios of measures[8, 36]), thereby physically adimensional[74]. However a density function ρ_α is not necessary dimensionless, the key fact is that the product $\rho_\alpha(x)dx$ has the same units as the measure α since $d\alpha = \rho_\alpha(x)dx$ and $\alpha = \int d\alpha$. It is **not** always the case that dx has the same units as $d\alpha$, this can be easily seen if we work with non-uniform density functions. A direct consequence of this fact is that the differential entropy is **not** invariant under non-rigid Euclidean motions(scalings), in other words our quantitiy of information **do** depends on what meter stick do we use to make our measurement and this is not what we experience in the real world.

These roadblocks is what does not allow us to make the differential entropy itself a measure of information. This is a very non-straight forward feature, in fact the contemporary matematician L.C. Evans [27] mistakenly defined the notion of entropy for parabolic PDE(partial differential equations) using differential entropy. This lead him astray, even having to define two different entropies and ignoring the units of the phenomena in the process, ultimately he struggled in the conceptual understanding of the entropy for parabloic PDEs. The correct approach in both information theory and PDE's is to consider the idea of relative entropy.

2.3.2 Relative Entropy

Given two discrete random variables X and Y defined in the countable sets $\{x_{1,}, x_2, \dots\}$ and $\{y_{1,}, y_2, \dots\}$ respectively with $p(X = x_i) = p_i$ and $q(Y = y_i) = q_i$, their relative entropy, or Kuller-Leibler(KL) divergence or more precisely the information of p relative to q, is

$$\mathrm{KL}(p|q) = \sum_{i\geq 1} p_i \ln\left(\frac{p_i}{q_i}\right) \tag{4}$$

In the case of continious probability distributions, consider two measures α and β with $d\alpha(x) = \rho_\alpha(x)\mathrm{d}m(x)$ and $d\beta(x) = \rho_\beta(x)\mathrm{d}m(x)$. Also the random variables X and Y on $\mathcal{X}$ with distribution α and β such that $p(X \in A) = \alpha(A)$ and $p(Y \in B) = \beta(B)$ then the relative entropy or Kuller-Leibler(KL) divergence between the measures α and β is defined as:

$$\mathrm{KL}(\alpha|\beta) = \int_{\mathcal{X}} \rho_\alpha \log\left(\frac{\rho_\alpha}{\rho_\beta}\right)\ \mathrm{d}m(x) = \mathbb{E}_{x\sim\alpha}\left[\log\left(\frac{\rho_\alpha}{\rho_\beta}\right)\right], \tag{5}$$

specifically for $\mathcal{X} = \mathbb{R}^n$ we have $d\alpha(x) = \rho_\alpha(x)dx$ and $d\beta(x) = \rho_\beta(x)dx$ with respect to the Lebesgue measure $\mathcal{L}_{\mathbb{R}^n}$, then the KL divergence between the measures α and β is

$$\mathrm{KL}(\alpha|\beta) = \int_{\mathbb{R}^n} \rho_\alpha \log\left(\frac{\rho_\alpha}{\rho_\beta}\right)\ \mathrm{d}x = \mathbb{E}_{x\sim\alpha}\left[\log\left(\frac{\rho_\alpha}{\rho_\beta}\right)\right]. \tag{6}$$

Even though the previous definition of KL divergence is the standard form, it would be useful to consider the following form for future generalizations,

$$\mathrm{KL}(\alpha|\beta) = 1 + \mathbb{E}_{x\sim\alpha}\left[\log\left(\frac{\rho_\alpha(x)}{\rho_\beta(x)}\right)\right] - \mathbb{E}_{x\sim\beta}\left[\frac{\rho_\alpha(x)}{\rho_\beta(x)}\right]. \tag{7}$$

Relative entropy solves all the previous mentioned problems, first $\mathrm{KL}(\alpha|\beta) \geq 0$ always with equality if, and only if, $\rho_\alpha(x) = \rho_\beta(x)$. Second it is invariant to scalings or more generally parameter transformations, $\mathrm{KL}(\mathbf{A}\alpha|\mathbf{A}\beta) = \mathrm{KL}(\alpha|\beta)$[21] . However it does have a small tradeoff, while the discrete and differential entropy are able to measure a property from a specific distribution, the relative entropy impose us to always compare two distributions. It is a relative measure, not an absolute one, in some sense this seems like a disadvantage, however as we mentioned before whenever we used the discrete or diferential entropy we are really always comparing with the "maximum" entropy distribution in order to make sense of our measurement, so in some sense this measure is also relative.

Next we compute explicitly the relative entropy between two gaussians, this is a key example to unveil what is the underlying geometry that the relative entropy induces.

Example 4. In $\mathbb{R}$, let's consider $\alpha = \mathcal{N}\left(\mu_\alpha, \sigma_\alpha^2\right)$ and $\beta = \mathcal{N}\left(\mu_\beta, \sigma_\beta^2\right)$, then one has,

$$\begin{aligned}
\mathrm{KL}(\alpha|\beta) &= \frac{1}{\sigma_\alpha\sqrt{2\pi}}\int_{\mathbb{R}} e^{\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2}\left(\log\left(\frac{1}{\sigma_\alpha\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2}\right) - \log\left(\frac{1}{\sigma_\beta\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2}\right)\right)\,\mathrm{d}x \\
&= \frac{1}{2}\left(2\log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right) + \frac{\left(\mu_\alpha-\mu_\beta\right)^2}{\sigma_\beta^2} + \frac{\sigma_\alpha^2}{\sigma_\beta^2} - 1\right)
\end{aligned}$$

To see the derivation of this formula, please see Appendix A. In order to understand this result, one can look at the infinitesimal geometry of KL[60, 50] , which is obtained by performing a Taylor expansion at order 2,

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{N}\left(\mu+\delta_\mu, (\sigma+\delta_\sigma)^2\right)|\mathcal{N}\left(\mu,\sigma^2\right)\right) &= \frac{1}{2}\left(2\log\left(\frac{\sigma}{\sigma+\delta_\sigma}\right) + \frac{(\mu-\mu+\delta_\mu)^2}{\sigma^2} + \frac{(\sigma+\delta_\sigma)^2}{\sigma^2} - 1\right) \\
&= \frac{\delta_\sigma}{\sigma} + \frac{\delta_\sigma^2+\delta_\mu^2}{2\sigma^2} - \log\left(1+\frac{\delta_\sigma}{\sigma}\right) \\
&= \frac{\delta_\sigma}{\sigma} + \frac{\delta_\sigma^2+\delta_\mu^2}{2\sigma^2} - \left(\frac{\delta_\sigma}{\sigma} - \frac{\delta_\sigma^2}{2\sigma^2} + \cdots\right) \\
&= \frac{\delta_\sigma^2+\frac{1}{2}\delta_\mu^2}{\sigma^2} + O(\delta_\sigma^3, \delta_\mu^3)
\end{aligned}$$

The key importance of this formula lies on the object called **Fisher-Rao Information metric**. For this case we have that this is the metric of the hyperbolic Poincare Half Plane, an interesting behavior is the geodesics in this space. These are half circles centered along the $\sigma = 0$ line and have an exponential speed, i.e. they only reach the limit $\sigma = 0$ after an infinite time. The computation of these geodesics is described in Apendix B.

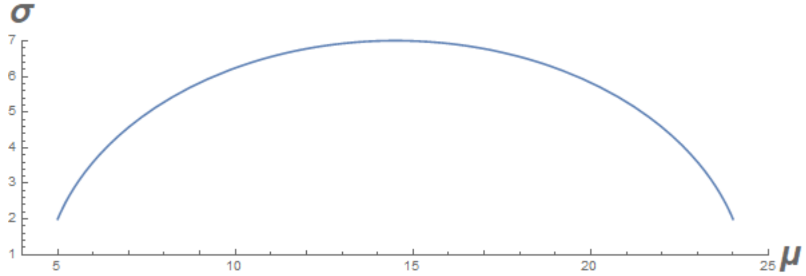

Figure 5: Geodesics of the Fisher-Rao metric between ($\mu_1 = 5, \sigma_1 = 2$)and ($\mu_2 = 24, \sigma_2 = 2$) .

The important idea of these geodesics is that interpolating alongside the geodesics of this curved space is a natural generalization of linear interpolation in euclidean space, since both are families of isometric interpolations. Therefore an interpolation scheme that naturally comes from the KL divergence would be as in Figure 6. Of course this idea can be used with more metrics, a natural comparison is linear interpolation in Euclidean space. So one migth ask, which "distance" between **functions** induces the euclidean metric in (μ, σ)-space. This turns out to be the Wasserstein-2 metric, widely used in optimal transport. In fact this is one of the reasons why is it argued that the Wasserstein metric is the "best" geometric loss functions[50], we'll expand much more on this metric in the next sections of the document.

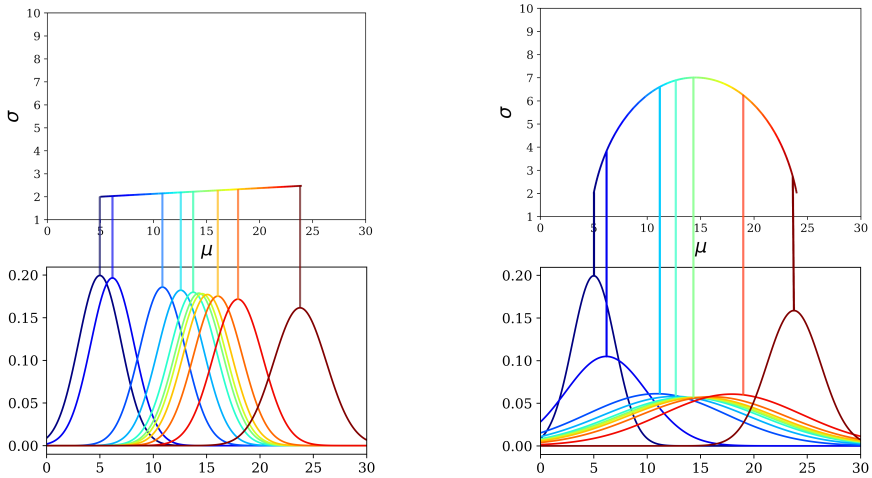

Figure 6: KL Geodesic interpolation and Euclidean interpolation between $(\mu_1 = 5, \sigma_1 = 2)$and $(\mu_2 = 24, \sigma_2 = 2)$.

2.3.3 Entropy and Stochastich Processes

If we focus again in the Maximum Entropy Principle, it was introduced in the context of statistical inference, however it has direct relation with the processes of diffusion, for starters the maximum entropy density distribution in $\mathbb{R}$ with given mean μ and variance σ^2 is a Gaussian $\mathcal{N}(\mu, \sigma^2)$ as we saw in Theorem 3, and as its well known [63] this is the fundamental solution of the Heat equation in $\mathbb{R}$. Furthermore it was shown in [46] that in some sense the one-dimensional Focker-Planck-Kolmogorov equation(a general form of the Difussion equation) is a consequence of the maximum entropy principle, more specifically

Proposition 5. *[[46]] Let there be given a scalar valued random process $x_t(\omega)$, and assume that all the information which we have about $x_t(\omega)$, is summarized in the knowledge of the first moment and of the second moment of the transition probability, that is to say,*

$$\begin{aligned}\int_{-\infty}^{\infty} zq(z, \Delta t/x, t)\ \mathrm{d}z &= \gamma(x, t)\Delta t,\\ \int_{-\infty}^{\infty} z^2 q(z, \Delta t/x, t)\ \mathrm{d}z &= \nu(x, t)\Delta t, \quad \beta > 0,\end{aligned}$$

where $q(z, \Delta t/x, t)$ is the probability density,

$$p(x', t + \Delta t/x, t) \coloneqq q(z, \Delta t/x, t), \quad z \coloneqq x' - x.$$

Then, according to the maximum entropy principle, the probability density $p(x, t)$ is given by F-P-K equation(Focker-Planck-Kolmogorov equation),

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[\gamma(x,t)p(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\nu(x,t)p(x,t)\right]. \tag{8}$$

This roughly says that a density evolving under the F-P-K equation has a transition probability that wants at each time-step to maximize its Shannon entropy, in some sense this gives an understanding on how the notion Shannon entropy enters in the time evolution of an irreversible difussion process. Another connection between shannon entropy and difussion related with optimal transport would be explored later in section 3.1.

Next we talk more specifically about the general difussion equation and some fundamental properties that would be relevant later in the document.

2.4 General Difussion Equation

Starting from the FPK equation (8), we can consider that $\gamma(x)$ and $\nu(x)$ only depend on x only and make the change, $\gamma(x) = \frac{a'(x)}{2b(x)}$ and $\nu(x) = \frac{a(x)}{b(x)}$, hence (8) can be written as,

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\frac{d}{dx}\left(a(x)\frac{d}{dx}\left(\frac{p(x,t)}{b(x)}\right)\right), \tag{9}$$

which is going to be the general form of the difussion equation we would refer in the rest of the document. Clearly if $a(x) = b(x) = 1$, we are in the case of the classical heat equation.

For more general spaces $\mathcal{X}$, we can write the difussion equation in a fairly similar way,

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2}\nabla_{\mathcal{X}} \cdot \left(a(x)\nabla_{\mathcal{X}}\left(\frac{p(x,t)}{b(x)}\right)\right), \tag{10}$$

where the operator $\nabla_{\mathcal{X}}$ refers to the nabla operator associated to the space $\mathcal{X}$.

As it was mentioned in Proposition (5), the difussion equation is deeply related with stochastic procesess, we briefly expand on this in the next section.

2.4.1 Itô Process and FPK Equation

From the theory of stochastic processes [9, 45, 31] it is well known that equation (8) is the so-called Kolmogorov Forward equation from the Itô process$(X_t, t > 0)$ given by stochastic partial differential equation,

$$dX_t = \gamma(X_t)dt + \sqrt{\nu(X_t)}dB_t, \qquad X(0) = X^0 \tag{11}$$

where $(B_t, t > 0)$ is a standard Brownian motion or Wienner process, $\gamma(x)$ is the drift coefficient, $\nu(x)$ is the difussion coefficient and X^0 has a density $p(x, 0)$.

Equation (11) is a formalization of the notion of Brownian motion as described by Langevein[70], where we model the motion of a particle undergoing difussion in the potential field Ψ, where $\gamma = -\nabla\Psi$ and $\sqrt{\nu(X_t)}dB_t$ represents a random force due to molecular collisions(basically the experiment we described in Figure (1)). With that X_t then represents the position of the particle, and ν is proportional to the square of the temperature. In this model the solution $p(x,t)$ of the FPK equation represents the probability density at time t for finding the particle at position x.

Next we emphasize on a different manner to get the difussion equation that is commonly used in differential geometry and that would be useful later in the document.

2.4.2 Gradient Flow of General Dirichlet Energy

We start by defining the generalized Dirichlet energy functional,

$$E[u] = \frac{1}{4}\int_{\mathcal{X}} |\nabla_{\mathcal{X}} u(x)|^2 \ \mathrm{d}m(x), \tag{12}$$

where as before $\mathcal{X}$ is an arbitrary space, $x \in \mathcal{X}$ and $\mathrm{d}m(x)$ is the volume element. Next we would introduce three definitions,

Definition 6 (L^2 Space)**.** *Let Ω be a Lebesgue measurable set in $\mathcal{X}$. We denote by $L^2(\Omega)$ the set of functions $f : \Omega \to \mathbb{R}$ (or $\mathbb{C}$) such that $|f|^2$ is Lebesgue integrable in Ω. Identifying two functions f and g when they are equal almost everywhere in, $L^2(\Omega)$ becomes a Banach space when equipped with the norm (integral norm of order 2),*

$$\|f(x)\|_{L^2(\Omega)} = \left(\int_\Omega |f(x)|^2 \ \mathrm{d}m(x)\right)^{1/2}.$$

Taking $u \in L^2(\Omega)$ and $v \in L^2(\Omega)$, then $L^2(\Omega)$ is a Hilbert space with respect to the inner product,

$$(u(x), v(x))_{L^2(\Omega)} = \int_\Omega u(x)v(x) \ \mathrm{d}m(x)$$

Definition 7 (Gradient with respect to the L^2 structure)**.** *Let $L^2(\Omega)$ be the Hilbert space of square-integrable functions on a Lebesgue measurable set Ω in a space $\mathcal{X}$, and $F : L^2(\Omega) \to \mathbb{R}$ be a smooth functional. If g is regular for F, we call $\frac{\delta F}{\delta g}(g) = \nabla_{L^2(\Omega)}F$, if it exists, any square-integrable function such that*

$$\frac{d}{d\varepsilon}F[g+\varepsilon h]\Big|_{\varepsilon=0} = \int_\Omega h\frac{\delta F}{\delta g}(g) \ \mathrm{d}m(x)$$

for every perturbation $h = g - \tilde{g}$ with $\tilde{g} \in L^2(\Omega)$.

This gradient is denoted as having the L^2 structure, since it naturally defines the gradient in the $L^2(\Omega)$ inner product,

$$\left\langle \nabla_{L^2(\Omega)}F(g), h\right\rangle = \int_\Omega h\frac{\delta F}{\delta g}(g) \ \mathrm{d}m(x) = DF[g]\cdot h$$

Having this we can pose the following definition [18],

Definition 8 (Gradient Flow in Linear Space)**.** *Let $\mathcal{P}$ be a linear space, and $F : \mathcal{P} \to \mathbb{R}$ is smooth. The Gradient flow (or steepest descent curve) of F is a smooth curve $p : \mathbb{R} \to \mathcal{P}$ such that,*

$$p'(t) = -\nabla F(p(t)).$$

For $\mathcal{P} = L^2(\Omega)$ we can find the Gradient flow for the Dirichlet energy functional using the gradient with respect to the L^2 structure under *usual* boundary conditions,

$$p'(t) = -\nabla E(p(t)) = \frac{1}{2}\nabla_{\mathcal{X}}\cdot\left(\nabla_{\mathcal{X}}\left(p(x,t)\right)\right), \tag{13}$$

which is precisely the general Difussion equation (10) for $a(x) = b(x) = 1$.

The Dirichlet energy (12) usually plays a role in order to prove the uniqueness of the solution of the heat equation under *usual* boundary conditions and also to show that the difussion equation is effectively a dissipative process[63].

Next we show a couple discretizations of this equation since they would be useful for the future developements, first we can propose an **explicit** time discretization of the form,

$$p(t^{\ell+1}) = p(t^{\ell}) - \tau \nabla E(p(t^{\ell})),$$

for a small step size τ. Alternatively we can use an **implicit** discretization,

$$p(t^{\ell+1}) = p(t^{\ell}) - \tau \nabla E(p(t^{\ell+1})),$$

This discretization has the advantage of being stable, furthermore we can reformulate this scheme into the following **variational** discretization scheme, also called proximal-point algorithm[4],

$$p(t^{\ell+1}) = \underset{t}{\operatorname{argmin}} \frac{1}{2} \left\| p(t) - p(t^{\ell}) \right\|_{L^2(\Omega)} + \tau E(p(t)), \tag{14}$$

where we casted problem of time-steping as a optimization problem. This at first might seem like an overkill, but this formulation turns out to be useful for both non-diferentiable energies E [60], large time-steppings [32] and it generalizes naturally to more general spaces. The latter scheme will be discussed later in the document.

3 Difussion Estimator

Let's now turn our attention to the notion of entropy in the context of Partial Differential Equations(PDE), in this context the notion of entropy has usually been concieved as a way to quantify the irreversablity in dissipative processes [27]. However this is not the full picture, the more general notion of relative entropy happens to take an important role to describe important transport phenomena. In the past 25 years this has been explored in the community of Kinetic Theory [76] and Optimal Transport [60, 75, 64] since they want to describe the transport between two distributions and as an special case if the target distribution(convergence to equilibrium) is an uniform distribution, then we can think of the process as a dissipative process. Nowadays these ideas have spreaded further and they are well known in many communities [22, 47].

Let's take a look at two fundamental examples that reveal the fundamental ideas of entropy in PDEs:

Example 9. Suppose we want to model the experiment in Figure 1, one way we could idealize this is by taking a particle density profile inside a box and let it spread under a difussion process. We can formalize this by saying we have a profile $f(\mathbf{x})$ located inside a box $\Omega \subset \mathbb{R}^n$ and we let evolve this distribution under the heat equation

$$\frac{\partial p(\mathbf{x},t)}{\partial t} - \Delta p(\mathbf{x},t) = 0 \qquad \mathbf{x} \in \Omega \subset \mathbb{R}^n,\ t > 0, \tag{15}$$

$$p(\mathbf{x},0) = f(\mathbf{x}) \qquad \mathbf{x} \in \Omega \subset \mathbb{R}^n, \tag{16}$$

where we impose no flux on the boundaries

$$\frac{\partial p(\mathbf{x},t)}{\partial \boldsymbol{\nu}} = 0, \qquad \forall \mathbf{x} \in \partial\Omega. \tag{17}$$

where $\boldsymbol{\nu}$ denotes the outward normal unit vector to $\partial\Omega$.

In general this problem is solved numerically [9], however we can get an understanding of the behavior looking at a couple of special cases, first lets take the case where $\Omega = \mathbb{R}^n$, also lets consider the profile as a single particle $\delta(\mathbf{x}-\mathbf{s})$ located at $\mathbf{s} \in \mathbb{R}^n$ and we consider the global cauchy conditions $\lim_{|\mathbf{x}|\to\infty} p(\mathbf{x},t) = 0$. In this manner it is well known [63], that the so called fundamental solution of the heat equation is

$$p(\mathbf{x},t) = \frac{1}{(4\pi t)^{n/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)},\ t > 0$$

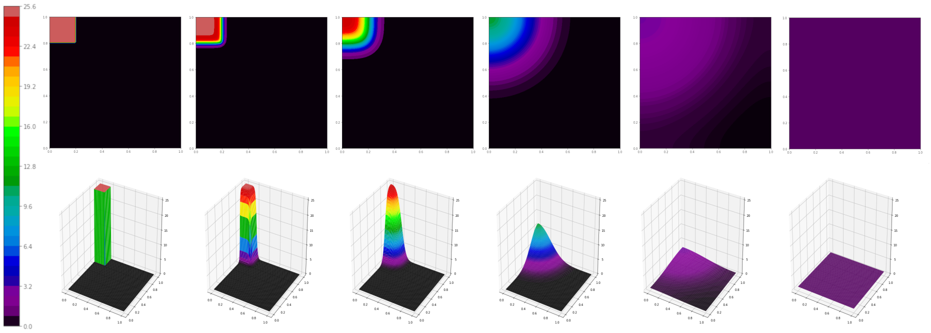

Figure 7: Experiment of gas of molecules diffussing. The molecules start at a corner of a box, then when released they spread through the box. The probability to find a molecule at a particular point in the box gets more uniform over time, until its equaly probable everywhere.

In this way we see that the particle starts at a well defined position and then it spreads all over space, very much like figure 1. Now in the general case of a profile $f(\mathbf{x})$ and a box $\Omega \subset \mathbb{R}^n$ we cannot solve the problem analyticaly, however there are still many things we can say. Perhaps the most important thing is that $p(\mathbf{x},\infty) = p_\infty(\mathbf{x}) = 1/\text{Area}(\Omega)$ is the equilibrium solution of (15). This is the case since, $p_\infty(\mathbf{x})$ satisfies the asymptotic equation $\Delta p(\mathbf{x}) = 0$ and the boundary conditions (17). Hence this is exactly the realization of the experiment of Figure 1, we start with a well defined initial condition $f(\mathbf{x})$ and little by little we lose information until we regard the profile as random $p_\infty(\mathbf{x}) = 1/\text{Area}(\Omega)$.

However this is not the full story, what if instead of a uniform random distribution on Ω, we suspect for some reason(maybe some inhomogeneity in the box) that the final random profile is a n-simplex distribution(triangular distribution in n dimensions). Is there a way we can adapt our difussion process to take this into account ? The answer relies on the notion of **relative entropy.**

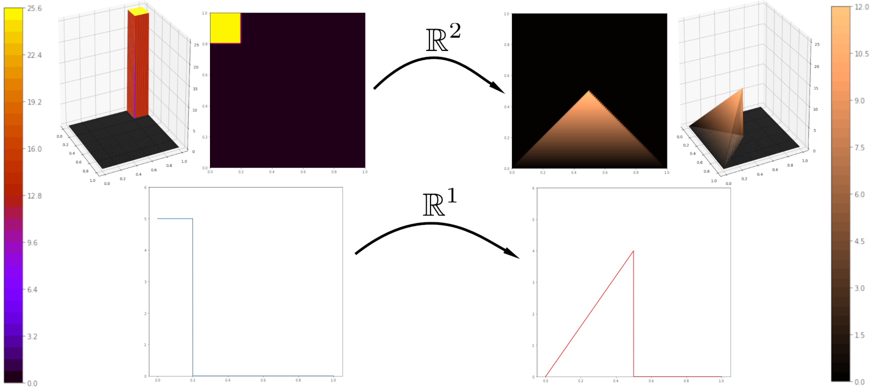

Figure 8: Experiment of gas of molecules transporting from a density to another in a one and two dimensional box. What is necessary for having a n-simplex distribution as an equilibrium distribution ?

Let's first take a look at how differential entropy changes over time,

$$\frac{d}{dt}H(p) = \frac{d}{dt}\left(-\int_\Omega p\log(p)\ \mathrm{d}\mathbf{x}\right) = \int_\Omega -\frac{dp}{dt}\left(\log(p)+1\right)\ \mathrm{d}\mathbf{x} = -\int_\Omega \Delta p(\mathbf{x},t)\left(\log(p)+1\right)\ \mathrm{d}\mathbf{x}$$

Now applying green first identity twice and the boundary condition [63],

$$\frac{d}{dt}H(p) = -\cancel{\int_{\partial\Omega}\frac{\partial p(\mathbf{x},t)}{\partial\boldsymbol{\nu}}\left(\log(p)+1\right)}\ \mathrm{d}\sigma + \int_\Omega \nabla p(\mathbf{x},t)\cdot\nabla\left(\log(p)\right)\ \mathrm{d}\mathbf{x} = \int_\Omega \frac{\|\nabla p(\mathbf{x},t)\|^2}{p}\ \mathrm{d}\mathbf{x} \geq 0.$$

Furthermore if Ω is convex then $\frac{d^2}{dt^2}H(p) \leq 0$ [27]. All of this sounds wonderful, the diferential entropy behaves just like we expect thermodynamic entropy to behave, it increases over time and it attains a maximum at the stationary solution. However we know that the differential entropy is a quantity with a lot of difficulties. How is it that it behaves so neatly in this example ? we can explain this by looking at the relative entropy of the density at a time t with respect to the equilibrium solution,

$$KL(p|p_\infty) = \int_\Omega p\log\left(\frac{p}{p_\infty}\right)\ \mathrm{d}\mathbf{x} = \int_\Omega p\log\left(\text{Volume}(\Omega)p\right)\ \mathrm{d}\mathbf{x} = -H(p) + \log\left(\text{Volume}(\Omega)\right)$$

Therefore for this example the diferential entropy is really the realtive entropy minus a constant. Therefore the diferential entropy is really measuring how far is the profile at a time t with respect to the equilibrium solution. This also explains why the related quantity

$$\begin{aligned} S(p) = \int_\Omega \log(p) \ \mathrm{d}\mathbf{x} &= \mathrm{Volume}(\Omega) \int_\Omega \frac{1}{\mathrm{Volume}(\Omega)} \log(\mathrm{Volume}(\Omega)p) \ \mathrm{d}\mathbf{x} - \int_\Omega \log(\mathrm{Volume}(\Omega)) \ \mathrm{d}\mathbf{x} \\ &= -\mathrm{Volume}(\Omega) \left(KL(p_\infty|p) + \log(\mathrm{Volume}(\Omega)) \right) \end{aligned}$$

is also sometimes associated with entropy[27]. Furthermore it explains why $\frac{d}{dt}S(p) \geq 0$ even though $\frac{d^2}{dt^2}S(p) \not\leq 0$, both are properties of the KL divergence($KL(g|f)$ is a distance between probability distributions such that $KL(g|f) = 0$ iff $g = f$ means that $p \to p_\infty$ as $t \to \infty$ and $KL(g|f)$ is only convex in g explains the difference of convexity between $S(p)$ and $H(p)$).

3.1 Wasserstein Flow

With this in place it is not surprising that the notion of relative entropy relates deeply with the general-difussion equation. To understand this further we'll first introduce some notions from optimal transport.

The problem of optimal transport[60, 75] consist in figuring out what is the minimum total cost of transporting in space a measure α to another measure β (where a "transportation" involves "disassembling" the measure α, transporting it, and "reassembling it" in the form β). More rigorously let $\alpha \in \mathcal{M}(\mathcal{X})$ and $\beta \in \mathcal{M}(\mathcal{Y})$ be two arbitrary normalized positive measures, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a cost function, then we define the joint probability distribution(transportation map)

$$\Pi(\alpha, \beta) \overset{\text{def.}}{=} \{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : \pi(A \times \mathcal{Y}) = \alpha(A) \text{ and } \pi(\mathcal{X} \times B) = \beta(B)\},$$

for sets $A \subset \mathcal{X}$ and $B \subset \mathcal{Y}$.

The optimal transportation problem from α to β seeks a coupling $\pi_{\alpha\beta} \in \Pi(\alpha, \beta)$ with minimal cost, computed as the integral of squared cotst c^2 against $\pi_{\alpha\beta}$. Formally, the 2-Wasserstein distance between α and β is thus defined as,

$$\mathcal{W}_2(\alpha, \beta) \overset{\text{def}}{=} \left[\inf_{\pi_{\alpha\beta} \in \Pi(\alpha,\beta)} \iint_{\mathcal{X} \times \mathcal{Y}} c(x,y)^2 \ \mathrm{d}\pi(x,y) \right]^{1/2}.$$

If we consider $\mathcal{X}$ and $\mathcal{Y}$ to be a Riemannian manifold M we have $\mathrm{d}\pi(x,y) = \pi_{\alpha\beta}(x,y) \ \mathrm{d}m(x)\mathrm{d}m(y)$ and $c = d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ to be the geodesic distance function, so $d(x,y)$ is the shortest distance from x to y along M, then the 2-Wasserstein distance between α and β is thus defined as,

$$\mathcal{W}_2(\alpha, \beta) \overset{\text{def}}{=} \left[\inf_{\pi \in \Pi(\alpha,\beta)} \iint_{\mathcal{X} \times \mathcal{Y}} \pi_{\alpha\beta}(x,y) d(x,y)^2 \ \mathrm{d}m(x)\mathrm{d}m(y) \right]^{1/2}.$$

Next recall our discussion of measures in the beginning of the document, now consider functionals of the form

$$\alpha = \rho \cdot \lambda \to \int_{\mathcal{X}} f(\rho(x)) \ \mathrm{d}\lambda(x),$$

where λ is a given positive measure on $\mathcal{X}$. The object ρ is in general the Radon-Nikodym derivative or density of α with respect to λ. Next we'll introduce the Wasserstein gradient, but its useful first to introduce another concept first, note that in section 2.4.2 we introduced the gradient with respect to the L^2 structure, now we introduce a simmilar notion but for measures[64],

Definition 10 (First variation)**.** *Consider a space $\mathcal{X}$, a measure $\alpha \in \mathcal{M}(\mathcal{X})$ (if α is an absolute continious measure then we will denote it by its density ρ with respect to λ), and a functional $F : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$. If ρ is regular for F, then we call $\frac{\delta F}{\delta \rho}(\rho) = \nabla_{L^2} F$, if it exists, any measurable function such that*

$$\frac{d}{d\varepsilon} F[\alpha + \varepsilon\chi]\Big|_{\varepsilon=0} = \int_{\mathcal{X}} \frac{\delta F}{\delta \rho}(\rho)\ \mathrm{d}\chi$$

for every perturbation $\chi = \alpha - \tilde{\alpha}$ with $\tilde{\alpha} \in \mathcal{M}(\mathcal{X})$.

In our case we will consider $d\lambda(x) = \mathrm{d}m(x)$ as the volume element of $\mathcal{X}$, hence $\rho = \rho_\alpha$, and the functional $F : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ is defined by

$$F(\alpha) = \int_{\mathcal{X}} f(\rho_\alpha)\ \mathrm{d}m(x),$$

For this functional the first variation can be shown to be $\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha) = \dot{f}(\rho_\alpha)$. With this on hand we can introduce the Wasserstein gradient,

Definition 11 (Gradient with respect to the W^2 structure)**.** *Consider a space $\mathcal{X}$, a measure $\alpha \in \mathcal{M}(\mathcal{X})$ (if α is an absolute continious measure then we will denote it by its density ρ_α), and a functional $F : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$. If ρ_α is regular for F, we call $\nabla_{W^2} F$,*

$$\nabla_{W^2} F = -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha) \right) \right)$$

This gradient is denoted as having the W^2 structure, since it naturally defines the gradient in the geometric tangent cone $\mathrm{Tan}_{\rho_\alpha}\mathcal{M}(\mathcal{X})$ inner product [75, 22],

$$\langle \nabla_{W^2} F(\rho_\alpha), -\nabla_{\mathcal{X}} \cdot (\xi\rho_\alpha) \rangle = \int_{\mathcal{X}} \rho_\alpha \left\langle \nabla_{\mathcal{X}} \frac{\delta F}{\delta g}(\rho_\alpha), \xi \right\rangle\ \mathrm{d}m(x) = DF[g] \cdot [-\nabla_{\mathcal{X}} \cdot (\xi\rho_\alpha)],$$

for $-\nabla_{\mathcal{X}} \cdot (\xi\rho_\alpha)$ a "tangent" vector, $-\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha) \right) \right) = \frac{\partial \rho_\alpha}{\partial t}$ and $-\nabla_{\mathcal{X}} \cdot (\xi\rho_\alpha) = \frac{\partial \rho_\alpha}{\partial t}$. This inner product is deeply related with the hydrodynamic interpretation of the Optimal Transport problem, investigated by Otto, Benamou, Bernier and others [75, 54, 5].

Now using Definition 8 For $\mathcal{P} = \mathcal{M}(\mathcal{X})$ we can find the Gradient flow for a functional $F : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ using the gradient with respect to the W^2 structure,

$$\rho_\alpha'(t) = -\nabla_{W^2} F(\alpha) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha) \right) \right), \tag{18}$$

where we assume Neumman boundary conditions $\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha) \right) \Big|_{\partial\mathcal{X}} = 0.$ in the boundary,

This is the generalization of the Difussion equation. As an special case take $F(\alpha) = -H(\alpha) = \int_{\mathcal{X}} \rho_\alpha \log \rho_\alpha\ \mathrm{d}m(x)$, the functional as the differential entropy, first note $\dot{f}(\rho_\alpha) = \log \rho_\alpha + 1$, then

$$\rho_\alpha'(t) = -\nabla_{W^2} F(\alpha) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \frac{\nabla_{\mathcal{X}} \rho_\alpha}{\rho_\alpha} \right) = \nabla_{\mathcal{X}}^2 \rho_\alpha,$$

which is precisely the heat equation. Henceforth the density that solves the heat equation evolves in the direction of steepest ascent of the differential entropy with respect to the Wasserstein-2 metric structure. Which is a precise mathematical statement of the second law of thermodynamics in this phenomena, i.e. the process evolves always increasing the entropy.

So we just described a way to obtain the heat equation by the means of a 2-Wasserstein gradient flow, although this is quite impresive, what new insight does it bring to the table ? Something quite remarkable is the fact that we are taking the gradient of a zeroth-order functional $F(\alpha) = \int_{\mathcal{X}} \rho_\alpha \log \rho_\alpha \; \mathrm{d}m(x)$, whereas with a L^2 gradient flow (13) we are taking the gradient of a first-order functional $E(\alpha) = \frac{1}{4}\int_{\mathcal{X}} |\nabla_{\mathcal{X}}\rho_\alpha(x)|^2 \; \mathrm{d}m(x)$, in fact we cannot generate a second order L^2 gradient flow with a zeroth order functional, this would have important consequences in the proximal-point algorithm of each of the gradient formulations as we'll see next.

3.1.1 Relation with Density Estimation and Thermodynamic Quantities

Just like we discussed in section 2.4.2, a possible discretization for equation (18) is given by an anlaogous form [45] of the proximal-point algorithm, where we solve the following optimization problem,

$$\rho_\alpha(t^{\ell+1}) = \underset{t}{\operatorname{argmin}}\frac{1}{2}\left\|\rho_\alpha(t) - \rho_\alpha(t^{\ell})\right\|_{W^2(\Omega)} + \tau F(p(t)), \tag{19}$$

this is the so-called JKO-scheme, a scheme that directly connects the solution of a regularized optimal transport problem with the solution of the solution of a second order partial differential equation. An striking difference between this discretization and the variational scheme (14) is that for $F(\alpha) = -H(\alpha) = \int_{\mathcal{X}} \rho_\alpha \log \rho_\alpha \; \mathrm{d}m(x)$, the functional that yields heat equation we don't impose a first-order differentiability constraint on the density, contrary to both (13) with the first-order functional $E(\alpha) = \frac{1}{4}\int_{\mathcal{X}} |\nabla_{\mathcal{X}}\rho_\alpha(x)|^2 \; \mathrm{d}m(x)$, and other Heat equation discretization schemes, Section 2.4.2,

This scheme has sparked the interest of many researchers, from theory [75, 64] to applications [60, 13, 59]. In particular for us it was used for density estimation in [13]. In this paper they introduce a family of optimal transport methods essentially inspired on the method of **Maximum Penalized Likelihood Estimation**, where they try to estimate a density by solving a regularized optimization problem, however there was no bandwidth selection algorithm and no comparison with other competing methods.

In this document we aim to show that we can extend the variational approach of [13]by merging features from competing methods such as the kernel density estimation via difussion [9] and the cross entropy method[10], thus generating an unified framework of density estimation. Under this formalism we claim that the three methods are best seen under a unified perspective based on the gradient flow interpretation.

Next we'll consider some interesting quantities that emerge naturally from this formalism, first note that,

$$-\frac{d}{dt}F = -\frac{d}{dt}\int_{\mathcal{X}} f(\rho_\alpha)\;\mathrm{d}m(x) = -\int_{\mathcal{X}} \dot{f}(\rho_\alpha)\frac{\partial \rho_\alpha}{\partial t}\;\mathrm{d}m(x) = -\int_{\mathcal{X}} \frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha)\nabla_{\mathcal{X}}\cdot\left(\rho_\alpha \nabla_{\mathcal{X}}\left(\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha)\right)\right)\;\mathrm{d}m(x)$$

Then using Green first identity and the boundary conditions,

$$-\frac{d}{dt}F = \int_{\mathcal{X}} \rho_\alpha \left|\nabla_{\mathcal{X}}\left(\frac{\delta F}{\delta \rho_\alpha}(\rho_\alpha)\right)\right|^2 \;\mathrm{d}m(x) = I(\alpha)$$

This is the so called Fisher Information or Entropy Production, in particular for $F(\alpha) = -H(\alpha)$, we have, $I(\alpha) = \int_{\mathcal{X}} \rho_\alpha \left\|\nabla \log \rho_\alpha\right\|^2 \; \mathrm{d}m(x)$, which is the usual form of the Fisher Information [21]. Next let's define one more interesting quantity, first the Fisher Information Metric $\|\alpha(t_0)\|^2$,

$$\|\alpha(t_0)\|^2 = \frac{d^2}{dt^2} KL(\rho_\alpha(t)|\rho_\alpha(t_0))\Big|_{t_0}$$

Now looking at the infinitesimal geometry of KL[60, 50] , which is obtained by performing a Taylor expansion at order 2 we can get some insight,

$$\begin{aligned}
KL(\alpha(t+\Delta t)|\alpha(t)) &= \int_{\mathcal{X}} \rho_\alpha(t+\Delta t)\left(\log\rho_\alpha(t+\Delta t) - \log\rho_\alpha(t)\right)\ \mathrm{d}m(x) \\
&= \int_{\mathcal{X}} \left(\Delta t\frac{\partial\rho_\alpha}{\partial t} + \rho_\alpha(t)\right)\frac{\partial}{\partial t}\left(\log\rho_\alpha(t)\right)\Delta t\ \mathrm{d}m(x) \\
&= \int_{\mathcal{X}} \left(\Delta t\frac{\partial\rho_\alpha}{\partial t} + \rho_\alpha(t)\right)\frac{\Delta t}{\rho_\alpha(t)}\frac{\partial\rho_\alpha}{\partial t}\ \mathrm{d}m(x) \\
&= \int_{\mathcal{X}} \left(\frac{\partial\rho_\alpha}{\partial t}\Delta t + \frac{1}{\rho_\alpha(t)}\left(\frac{\partial\rho_\alpha}{\partial t}\right)^2\Delta t^2\right)\ \mathrm{d}m(x) \\
&= \cancel{\Delta t\frac{d}{dt}\left(\int_{\mathcal{X}}\rho_\alpha dm\right)} + \Delta t^2\int_{\mathcal{X}}\frac{1}{\rho_\alpha(t)}\left(\frac{\partial\rho_\alpha}{\partial t}\right)^2\ \mathrm{d}m(x)
\end{aligned}$$

Hence the formula for the Fisher Information metric $\|\alpha(t)\|^2$,

$$\|\alpha(t)\|^2 = \int_{\mathcal{X}}\frac{1}{\rho_\alpha(t)}\left(\frac{\partial\rho_\alpha}{\partial t}\right)^2\ \mathrm{d}m(x) = \int_{\mathcal{X}}\frac{1}{\rho_\alpha(t)}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\frac{\delta F}{\delta\rho_\alpha}(\rho_\alpha)\right)\right)\right|^2\ \mathrm{d}m(x)$$

in particular for $F(\alpha) = -H(\alpha)$, we have, $\|\alpha(t)\|^2 = \int_{\mathcal{X}}\frac{1}{\rho_\alpha}\|\Delta\rho_\alpha\|^2\ \mathrm{d}m(x)$, which is a second order functional. Now we'll describe the fundamental example from this formalism.

3.1.2 Focker-Planck Equation

So far he have seen how several quantites of information theory and statistical mechanics emerge in the context of the difussion equation. Now we will see how relative entropy appears in the picture and clarify several things, consider the following functional(free energy)

$$F(\alpha) = E(\alpha) - \beta^{-1}H(\alpha) = \int_{\mathcal{X}}\rho_\alpha\Psi\ \mathrm{d}m(x) + \beta^{-1}\int_{\mathcal{X}}\rho_\alpha\log\rho_\alpha\ \mathrm{d}m(x), \qquad \rho_\mu = \frac{1}{Z}e^{-\beta\Psi},\ \text{with } Z = \int_{\mathcal{X}} e^{-\beta\Psi}\ \mathrm{d}m(x)$$

where ρ_μ is the density of a Gibbs distribution and β can be regarded as the inverse temperature of the ensemble. It turns out that this functional is precisely the KL divergence between a configuration at t and the μ density,

$$F(\alpha) = \beta^{-1}KL(\alpha|\mu) - \beta^{-1}\log(Z).$$

The relation of relative entropy and free energy holds universaly in continious and discrete dynamical processes, interestingly it has been explored recently trying to unite biology and information [61, 3, 2].

Returning to our topic, we can use the full free energy $F(\alpha) = \beta^{-1}KL(\alpha|\mu) - \beta^{-1}\log(Z)$ to compute the Wasserstein gradient flow , note that $\dot{f}(\rho_\alpha) = \Psi + \beta^{-1}\log\rho_\alpha + 1$, then

$$\rho_\alpha'(t) = -\nabla_{W^2}F(\alpha) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha\left(\nabla_{\mathcal{X}}\Psi + \nabla_{\mathcal{X}}\left(\beta^{-1}\log\rho_\alpha\right)\right)\right) = \nabla_{\mathcal{X}} \cdot (\rho_\alpha\nabla_{\mathcal{X}}\Psi) + \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}}\left(\beta^{-1}\rho_\alpha\right)\right)$$

Now taking $2\beta^{-1} = \nu$ as the difussion coefficient and $\gamma = -\nabla\Psi$ as the drift coefficient inspired from the Itô Process of equation (11), we obtain the F-P-K equation(Focker-Planck-Kolmogorov equation (8)),

$$\rho_\alpha'(t) = -\nabla_{\mathcal{X}} \cdot (\rho_\alpha\gamma) + \frac{1}{2}\nabla_{\mathcal{X}} \cdot (\nabla_{\mathcal{X}}(\rho_\alpha\nu)), \tag{20}$$

Henceforth the density that solves the F-P-K equation evolves in the direction of steepest descent of the relative entropy with respect to the Wasserstein-2 metric. Also alternatively taking $a(x) = 2\beta^{-1}\rho_\mu$ and $b(x) = \rho_\mu$we recover (10)

$$\frac{\partial\rho_\alpha}{\partial t} = \frac{1}{2}\nabla_{\mathcal{X}} \cdot \left(a(x)\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{b(x)}\right)\right),$$

From this we see that the stationary solution is $b(x) = \rho_\mu$, therefore we'll denote this distribution as $\mu = \alpha_\infty$. This alongside $\frac{d}{dt}F(\alpha) \leq 0$ means that the distribution obtaines its minimum free energy at the stationary solution.

This form is precisely the **Kernel Density Estimation via Diffusion** from [9] with particular choice of $a(x)$ and $b(x)$. The method of density estimation via difussion is very widely used because of its robustness and its bandwith selection algorithm. As a consequence it has been used in many applications, from crime [34] to clustering[52], and is now the state of the art in the field of density estimation. Our previous derivation shows that the formalism of Wasserstein flow might offer a extended view to this powerful density estimator.

Next lets define a couple more quantites, the mean energy and energy production, first recall from the the definition of free energy in Thermodynamics $F = \langle E\rangle - TS$, hence our "thermodynamic" Mean Energy in this case is:

$$E(\alpha) = \int_{\mathcal{X}} \rho_\alpha\Psi \ \mathrm{d}m(x) = -\beta^{-1}\int_{\mathcal{X}} \rho_\alpha\log\rho_\infty \ \mathrm{d}m(x) - \beta^{-1}\log(Z)$$

Now the Mean Energy Production can be computed using,

$$\begin{aligned}
\frac{d}{dt}E(\alpha) &= -\beta^{-1}\int_{\mathcal{X}} \frac{\partial\rho_\alpha}{\partial t}\log\rho_\infty \ \mathrm{d}m(x) \\
&= -\beta^{-1}\int_{\mathcal{X}} \frac{1}{2}\nabla_{\mathcal{X}} \cdot \left(2\beta^{-1}\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\log\rho_\infty \ \mathrm{d}m(x) \\
&= -\beta^{-2}\int_{\mathcal{X}} \rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\nabla_{\mathcal{X}}(\log\rho_\infty) \ \mathrm{d}m(x) \\
&= -\beta^{-2}\int_{\mathcal{X}} \nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\nabla_{\mathcal{X}}(\rho_\infty) \ \mathrm{d}m(x)
\end{aligned}$$

Clearly only if $\nabla_{\mathcal{X}}(\rho_\infty) = 0$, where the equilibrium distribution is constant, then the mean energy is conserved. Next we'll consider some special functionals and their flows.

3.2 Flow of Statistical divergences

Performing a Wasserstein gradient flow with the KL-Divergence provide us with the very powerful KDE, we can extend this result considering a larger family of divergences that provide simmilar properties to the gradient flow. In general a divergence D satisfies $D(\alpha, \beta) \geq 0$, $D(\alpha, \beta) = 0$ if and only if $\alpha = \beta$ and sometimes a convexity property, but it does not need to be symmetric or satisfy the triangular inequality. This functions are used extensively as loss functions for problems of inference[60], moreover we can think of them intuitively in the following coginitive approach: A formalization of how much did I learned from β given that I knew α beforehand. Clearly this is asymmetric and not necessarily linearly cummulative[17].

Therefore it is not unreasonable to consider a bigger family of divergences D in the context of statistical inference since the KL divergence is in some sense as special as any other divergence measure[23]. It is just one more way to operationalize the notion of divergence with certain linearity properties.

3.2.1 Csiszár divergences (f-divergences)

Inspired in the fact that the free energy $F(\alpha) = \beta^{-1} KL(\alpha|\mu) - \beta^{-1} \log(Z)$ contained a KL-Divergence consider a more general functional of f-divergences or Csiszár divergences[23],

$$F(\alpha) = \mathcal{D}(\rho_\alpha \to \rho_\beta) = \int_{\mathcal{X}} \rho_\beta \varphi\left(\frac{\rho_\alpha}{\rho_\beta}\right) \, \mathrm{d}m(x),$$

where $\varphi : \mathbb{R}^+ \to \mathbb{R}$ is a twice continuous diferentiable function with $\varphi(1) = 0$ and $\varphi''(x) \geq 0$ for all $\mathbb{R}^+$.

With this we compute the Wasserstein gradient flow , note that $\dot{f}(\rho_\alpha) = \varphi'\left(\frac{\rho_\alpha}{\rho_\beta}\right)$, then

$$\rho'_\alpha(t) = -\nabla_{W^2} F(\alpha) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\varphi'\left(\frac{\rho_\alpha}{\rho_\beta}\right)\right)\right)$$

It is pretty useful to compute now the equilibrium distribution in the following way [47, 15, 57] since it will be useful later,

$$0 = \rho_\infty \left(\nabla_{\mathcal{X}} \left(\varphi'\left(\frac{\rho_\infty}{\rho_\beta}\right)\right)\right)$$

for $\rho_\infty > 0$,

$$\varphi'\left(\frac{\rho_\infty}{\rho_\beta}\right) = cte$$

Since $\varphi''(x) \geq 0$ for all $\mathbb{R}^+$, the function $\varphi'(x)$ has a unique inverse in the domain $\mathbb{R}^+$, hence,

$$\frac{\rho_\infty}{\rho_\beta} = \varphi'^{-1}(cte) \Longrightarrow \rho_\infty = C \cdot \rho_\beta$$

Since both ρ_β and ρ_∞ are normalized densities, in all cases ρ_β is the stationary distribution !!!! Therefore we'll denote this distribution as $\rho_\infty = \rho_\beta$ from now on.

- Taking as special case $\varphi(x) = \beta^{-1}(x \log(x) - \log(Z)\, x)$, we obtain the case of KL-Divergence

$$F(\alpha) = \int_{\mathcal{X}} \rho_\infty \beta^{-1} \left(\frac{\rho_\alpha}{\rho_\infty} \log\left(\frac{\rho_\alpha}{\rho_\infty}\right) - \frac{\rho_\alpha}{\rho_\infty}\right) dm = \beta^{-1}\left(KL(\alpha|\alpha_\infty) - \log(Z)\right),$$

Then the Wasserstein gradient flow is,

$$\begin{aligned}
\rho_\alpha'(t) &= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\beta^{-1} \log\left(\frac{\rho_\alpha}{\rho_\infty}\right) - \beta^{-1} \frac{\rho_\infty}{\rho_\alpha} \left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right) \\
&= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\beta^{-1} \log\left(\rho_\alpha\right) - \beta^{-1} \log\left(\rho_\infty\right) - \beta^{-1}\right)\right) \\
&= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\beta^{-1} \log\left(\rho_\alpha\right)\right) - \rho_\alpha \beta^{-1} \nabla_{\mathcal{X}} \left(\log\left(\rho_\infty\right)\right)\right) \\
&= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \Psi + \rho_\alpha \beta^{-1} \nabla_{\mathcal{X}} \rho_\alpha\right)
\end{aligned}$$

,where $\Psi = -\beta^{-1} \log \rho_\infty$. An **important** difference of the KL-Divergence compared to the more general case $\mathcal{D}(\rho_\alpha \to \rho_\beta)$, is that the resulting Wasserstein gradient flow is a linear diferential operator $L = \frac{1}{2} \nabla_{\mathcal{X}} \cdot \left(2\beta^{-1} \rho_\infty \nabla_{\mathcal{X}} \left(\frac{\cdot}{\rho_\infty}\right)\right)$ on ρ_α, this is not a trivial property and seems to be a coincidence than a feature of the Divergences. Moreover all of the following Wasserstein flows are going to produce **non-linear Difussion Operators**[59].

- Let's consider another example, that yields the Pearson's χ^2 CE distance, hence $\varphi(x) = \frac{1}{2}\left(x^2 - 1\right)$, then $\dot{f}(\rho_\alpha) = \frac{\rho_\alpha}{\rho_\infty}$, hence the Wasserstein gradient flow,

$$\rho_\alpha'(t) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)$$

- As a generalization of the previous distance consider the ℓ-parametrized family $\varphi(x) = \frac{x^\ell - x}{\ell(\ell-1)}$,then $\dot{f}(\rho_\alpha) = \frac{1}{\ell-1}\left(\frac{\rho_\alpha}{\rho_\infty}\right)^{\ell-1}$, hence the Wasserstein gradient flow,

$$\begin{aligned}
\rho_\alpha'(t) &= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{1}{\ell-1} \left(\frac{\rho_\alpha}{\rho_\infty}\right)^{\ell-1}\right)\right) \\
&= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha^\ell \rho_\infty^{1-\ell} \nabla_{\mathcal{X}} \left(\rho_\alpha\right) + \frac{1}{\ell-1} \rho_\alpha^\ell \nabla_{\mathcal{X}} \left(\rho_\infty^{1-\ell}\right)\right) \\
&= \nabla_{\mathcal{X}} \cdot \left(\frac{1}{\ell \rho_\infty^{\ell-1}} \nabla_{\mathcal{X}} \left(\rho_\alpha^\ell\right) + \frac{1}{\ell-1} \rho_\alpha^\ell \nabla_{\mathcal{X}} \left(\rho_\infty^{1-\ell}\right)\right)
\end{aligned}$$

where the parameter ℓ includes the Hellinger distance for $\ell = 1/2$, Pearson's χ^2 discrepancy measure for $\ell = 2$, Neymann's χ^2 measure for $\ell = -1$, the Kullback-Leibler distance in the limit as $\ell \to 1$, and Burg CE distance as $\ell \to 0$.

- Also consider the zeroth order Total Variation $\varphi(x) = \frac{1}{2}|x - 1|$, then $\dot{f}(\rho_\alpha) = \frac{\left(\frac{\rho_\alpha}{\rho_\infty}\right)}{2\left|\frac{\rho_\alpha}{\rho_\infty}\right|}$, hence the Wasserstein gradient flow,

$$\rho_\alpha'(t) = \frac{1}{2} \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{\left(\frac{\rho_\alpha}{\rho_\infty}\right)}{\left|\frac{\rho_\alpha}{\rho_\infty}\right|}\right)\right)$$

- Probability Simplex for Divergences Experiments:f-divergenceSimplex-Peyre-Twitter.

3.2.2 Cross Entropy and Maximum Entropy Method

Note for the following generalized functional,

$$F(\alpha) = \int_{\mathcal{X}} \rho_\beta \varphi\left(\frac{\rho_\alpha}{\rho_\beta}\right) dm + \sum_{i=1}^{n} \int_{\mathcal{X}} \rho_\alpha \lambda_i K_i(x)\ \mathrm{d}m(x),$$

where each $K_i : \mathcal{X} \to \mathbb{R}$ is an absolutely continuous function and $\kappa_i = \int_{\mathcal{X}} \rho_\alpha K_i(x)\ \mathrm{d}m(x)$ are the generalized moments of the density ρ_α.

The Wasserstein gradient flow yields easily,

$$\rho_\alpha'(t) = -\nabla_{W^2} F(\alpha) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\varphi'\left(\frac{\rho_\alpha}{\rho_\beta}\right)\right) + \rho_\alpha \sum_{i=1}^{n} \nabla_{\mathcal{X}} \lambda_i K_i(x)\right)$$

The equilibrium solution satisfies,

$$\varphi'\left(\frac{\rho_\infty}{\rho_\beta}\right) + \sum_{i=1}^{n} \lambda_i K_i(x) = cte$$

Next, set $cte = -\lambda_0 = \sum_{i=1}^{n} -\lambda_0 K_0(x)$ with $K_0(x) = 1$, now

$$\rho_\infty = -\rho_\beta {\varphi'}^{-1}\left(\sum_{i=0}^{n} \lambda_i K_i(x)\right),$$

which is precisely the functional solution of the Euler-Lagrange Equation of the Dual problem of the Cross Entropy(CE) postulate[10]. This implies that we are able to naturally adapt the solution of the **Cross-Entropy Method** with the stationary solution derived from the Wasserstein flow**.** This method of inference is of great power because it attempts to estimate a density given the known information about the density(the moments κ_i) and nothing more. As a result it is very minimal, hence it can be used in a great amount of circumstances, from Natural language[7] to analysis of popularity of programming languages[24].

Our current analysis via optimal transport in some sense allowed us to formalized the notion that every time we do a density estimation via gradient flow we are also doing a minimal information inference, starting from the initial density(empirical density) into a guess that represents the information known about the data prior to the density estimation. This surprising result is also something that follows from the analysis of [46].

In the particular case again of the KL divergence with the function $\varphi(x) = \beta^{-1}(x \log(x) - x)$, the equilibrium solution satisfies,

$$\log\left(\frac{\rho_\infty}{\rho_\beta}\right) + \sum_{i=1}^{n} \lambda_i K_i(x) = cte,$$

setting $\rho_\beta = e^{-\lambda_0 K_0(x)}$, we have

$$\log(\rho_\infty) = cte - \sum_{i=0}^{n} \lambda_i K_i(x) \Longrightarrow \rho_\infty = \frac{1}{Z} e^{\sum_{i=0}^{n} -\lambda_i K_i(x)}, \qquad Z = \int_{\mathcal{X}} e^{\sum_{i=0}^{n} -\lambda_i K_i(x)}\ \mathrm{d}m(x),$$

which is precisely the Gibbs Distribution, or in general the solution of the Dual problem of the Maximum entropy method[19, 24]. This method is the original version of the cross-entropy method, first pioneered in the physics of statistical mechanics[42, 43], and it has had great impact in the scientific community[24]. Furthermore the result that the stationary solution of a linear Focker-Planck Equation solves a version of the Maximum entropy method was already known in the physics community[29] for the past 20 years, however they had not conected it with optimal transport. Nevertheless we can say that many of the ideas discussed in this document have been rediscovered plenty of times in several communities.

3.2.3 Bregman Divergences

We can also consider the family of Bregman divergences[11], for smooth strictly "convex" functional $\psi : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$,

$$\mathbf{B}_\psi\left(\rho_\alpha \middle| \rho_\beta\right) = \psi\left(\rho_\alpha\right) - \psi\left(\rho_\beta\right) - \left\langle \frac{\delta\psi}{\delta\rho_\alpha}\left(\rho_\alpha\right), \rho_\alpha - \rho_\beta \right\rangle_{L^2(\mathcal{X})}$$

With this we compute the Wasserstein gradient flow , note that $\dot{f}(\rho_\alpha) = \frac{\delta\psi}{\delta\rho_\alpha}\left(\rho_\alpha\right) - \frac{\delta\psi}{\delta\rho_\beta}\left(\rho_\beta\right)$, then

$$\rho_\alpha'(t) = -\nabla_{W^2} F(\alpha) = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}}\left(\frac{\delta\psi}{\delta\rho_\alpha}\left(\rho_\alpha\right) - \frac{\delta\psi}{\delta\rho_\beta}\left(\rho_\beta\right)\right)\right)$$

where clearly if $\rho_\alpha = \rho_\beta$, then it satisfies the stationary equation, Therefore we'll denote this distribution as $\rho_\infty = \rho_\beta$ from now on.

- As a couple of examples, lets look at, $\psi(\alpha) = \beta^{-1} H(\alpha)$, that induces the KL-Divergence, then

$$\begin{aligned} \rho_\alpha'(t) &= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}}\left(\beta^{-1}\log\left(\rho_\alpha\right) - \beta^{-1}\log\left(\rho_\infty\right)\right)\right) \\ &= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}}\Psi + \rho_\alpha \beta^{-1}\nabla_{\mathcal{X}}\rho_\alpha\right) \end{aligned}$$

,where $\Psi = -\beta^{-1}\log\rho_\infty$.

- Also then consider, $\psi(\alpha) = \frac{1}{m-1}\left\|\rho_\alpha\right\|_{L^m(\mathcal{X})}$, that induces the L^m-Distance, then

$$\begin{aligned} \rho_\alpha'(t) &= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}}\left(\frac{m}{m-1}\rho_\alpha^{m-1} - \frac{m}{m-1}\rho_\infty^{m-1}\right)\right) \\ &= \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}}\left(\rho_\alpha^m\right)\right) - \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha m \rho_\infty^{m-2}\nabla_{\mathcal{X}}\left(\rho_\infty\right)\right) \end{aligned}$$

which is a porous media type equation [73, 54] that evolves towards ρ_∞.

- Finally consider, $\psi(\alpha) = -\int_{\mathcal{X}} \log\rho_\alpha \, \mathrm{d}m(x)$, that induces the Itakura–Saito distance, then

$$\begin{aligned} \rho_\alpha'(t) &= \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}}\left(-\frac{1}{\rho_\alpha} + \frac{1}{\rho_\infty}\right)\right) \\ &= \nabla_{\mathcal{X}} \cdot \left(\frac{1}{\rho_\alpha}\nabla_{\mathcal{X}}\left(\rho_\alpha\right) + \rho_\alpha \nabla_{\mathcal{X}}\left(\frac{1}{\rho_\infty}\right)\right) \end{aligned}$$

- Other examples: Bregman-Peyre-Twitter.

3.2.4 Renyi Divergences and Other Generalizations

We can introduce the Renyi divergence as,

$$\mathrm{KL}_q^{re}(\alpha|\beta) = \frac{1}{1-q}\log\left(\int_{\mathcal{X}} \rho_\alpha \left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1} \mathrm{d}m(x)\right),$$

hence,

$$\frac{\delta \mathrm{KL}_q(\alpha|\beta)}{\delta\rho_\alpha}(\rho_\alpha) = \frac{q}{1-q}\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1} e^{(q-1)\mathrm{KL}_q^{re}(\alpha|\beta)} = \frac{q}{1-q}\left(\int_{\mathcal{X}}\rho_\beta\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q} \mathrm{d}m(x)\right)^{-1}\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1}$$

therefore the Wasserstein gradient flow,

$$\rho'_\alpha(t) \quad = \quad \frac{1}{2}\nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\frac{q}{1-q}\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1} e^{(q-1)\mathrm{KL}_q^{re}(\alpha|\beta)}\right)\right)$$

Next consider the Tsallis divergence, it is related to the Renyi divergence through the following one-to-one function. For $\gamma \in (0,1) \cup (1,\infty)$and $z > 0$, let

$$\varphi_\gamma(z) = \exp(z - \gamma z),$$

then

$$\mathrm{KL}_q^{ts}(\alpha|\beta) = \frac{1-\varphi_q\left(\mathrm{KL}_q^{re}(\alpha|\beta)\right)}{1-q} = \frac{1}{1-q}\left(1 - \int_{\mathcal{X}}\rho_\alpha\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1} \mathrm{d}m(x)\right),$$

hence,

$$\frac{\delta \mathrm{KL}_q^{ts}(\alpha|\beta)}{\delta\rho_\alpha}(\rho_\alpha) = \frac{-q}{1-q}\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1},$$

therefore the Wasserstein gradient flow,

$$\rho'_\alpha(t) \quad = \quad \frac{1}{2}\nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\frac{-q}{1-q}\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1}\right)\right)$$

Moreover the Tsallis divergence is very related to the ℓ-parametrized family $\varphi_\ell(x) = \frac{x^\ell - x}{\ell(\ell-1)}$ divergence that we introdice in the previous section with $q - 1 = \ell$,

$$\mathcal{D}_{\varphi_{q-1}}(\rho_\alpha \to \rho_\beta) = \frac{1}{(q-1)(q-2)}\left(\int_{\mathcal{X}}\rho_\alpha\left(\frac{\rho_\alpha}{\rho_\beta}\right)^{q-1} \mathrm{d}m(x) - 1\right) = \frac{1}{(q-2)}\mathrm{KL}_q^{ts}(\alpha|\beta),$$

What is the difference between the standard KL and this generalizations, we can get some intuition by analyzing different properties that the discrete versions of this divergences satisfy. First let's define the entropies associated to these divergences,

$$\mathrm{KL}_q^{re}(\alpha|\mathcal{L}_{\mathcal{X}}) = H_q^{re}(\alpha) = \frac{1}{1-q}\log\left(\int_{\mathcal{X}} (\rho_\alpha)^q \ \mathrm{d}m(x)\right),$$

$$\mathrm{KL}_q^{ts}(\alpha|\mathcal{L}_{\mathcal{X}}) = H_q^{ts}(\alpha) = \frac{1}{1-q}\left(1-\int_{\mathcal{X}} (\rho_\alpha)^q \ \mathrm{d}m(x)\right),$$

where $\mathcal{L}_{\mathcal{X}}$ is the uniform measure over $\mathcal{X}$. These entropies also have their discrete version, for a discrete random variable X defined in the countable set $\{x_{1,}, x_2, \ldots\}$ with $p(X = x_i) = p_i$, the entropies are,

$$H_q^{re}(p) = \frac{1}{1-q}\log\left(\sum_{i\geq 1}(p_i)^q\right), \tag{21}$$

$$H_q^{ts}(p) = \frac{1}{1-q}\left(1-\sum_{i\geq 1}(p_i)^q\right), \tag{22}$$

Now if we compare the different properties that each of the entropies (1) ,(21) ,(22) have, we can list five basic properties:

1. **Continuity**: The entropy measure $H(p_1, \ldots, p_n)$ is a continuous function of all the probabilities p_k, which means that a small change in probability distribution will only result in a small change in the entropy.
2. **Symmetry**: $H(p_1, \ldots, p_n)$ is permutationally symmetric; i.e., the position change of any two or more p_k in $H(p_1, \ldots, p_n)$ will not change the entropy value. Actually, the permutation of any p_k in the distribution will not change the uncertainty or disorder of the distribution and thus should not affect the entropy.
3. **Maximality**: $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ is a monotonic increasing function of n. For an equiprobable distribution, when the number of choices n increases, the uncertainty or disorder increases, and so does the entropy measure.
4. **Recursivity or Coarse Graining**: If an entropy measure satisfies (23) or (24), then it has the recursivity property. It means that the entropy of n outcomes can be expressed in terms of the entropy of $n-1$ outcomes plus the weighted entropy of the combined 2 outcomes.

$$H_n(p_1, p_2, \ldots, p_n) = H_{n-1}(p_1+p_2, p_3, \ldots, p_n) + (p_1+p_2)\, H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right), \tag{23}$$

$$H_n(p_1, p_2, \ldots, p_n) = H_{n-1}(p_1+p_2, p_3, \ldots, p_n) + (p_1+p_2)^q\, H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right), \tag{24}$$

where q is the parameter in Renyi's entropy or Tsallis entropy.

5. **Additivity**: If $p = (p_1, \dots p_n)$ and $r = (r_1, \dots, r_n)$ are two independent probability distribution, and the joint probability distribution is denoted by $p \bullet r$, then the property
$$H(p \bullet r) = H(p) + H(r),$$
$$H(p \bullet r) = H(p) + H(r) + (1-q)H(p)H(r),$$
where q is the parameter in Renyi's entropy or Tsallis entropy.

We can list what properties satisfiy each entropy:

Properties	(1)	(2)	(3)	(4)	(5)
Shannon's	yes	yes	yes	yes(only Recursivity)	yes(only Additivity)
Renyi's	yes	yes	yes	none	yes(only Additivity)
Tsallis's	yes	yes	yes	yes(only q-Recursivity)	yes(only q-Additivity)

Table 2: Comparison of Properties from the Three Entropies

So we see here that in some sense Shannon's entropy is quite special, since it satisfies all the "desired" properties for a information measure. Even though the Renyi entropy satisfies the unmodified version of additivity it does not satisfy recursivity and the Tsallis entropy satsfies a modified vesion of recursivity and additivity, which in some sense makes it more robust than the shannon entropy. As a consequence as we saw in the previous section the generalization of this flow yields a porous medium equation, which for $\ell = 1$ yields the linear heat equation.

Next we consider the flow of some higher order divergences and a few of their properties.

3.2.5 Higher order Divergences

In the previous section we introduced a series of zeroth order divergences commonly used in statistics and information theory. We can also consider higher order divergences and their Wasserstein flows, hence we'll consider functionals of the form
$$\alpha = \rho \cdot \lambda \to \int_{\mathcal{X}} f(\rho(x), \nabla_{\mathcal{X}}\rho(x), \nabla_{\mathcal{X}} \cdot \nabla_{\mathcal{X}}\rho(x), \cdots) d\lambda(x),$$

Again this is very useful to discretize 4-th or ($2n$+2-th) higher order PDE's with a variational scheme (19) requiring only first-order differentiability or (n-th order differentiability) instead of a higher differentiability requirement (14), Section 2.4.2,.

- First consider a modified Dirichlet energy like we introduced in the begining of the document,
$$E_D(\alpha) = \frac{1}{2}\int_{\mathcal{X}} \rho_\infty \left| \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right|^2 \, \mathrm{d}m(x)$$

Computing the Wasserstein gradient flow,
$$\rho'_\alpha(t) \quad = \quad -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{1}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right) \right)$$

which is a type of thin film equation [66]. For this equation you have to impose two boundary conditions,
$$\frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) = 0, \qquad \rho_\alpha \frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{1}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right) = 0$$

- We can also consider the Fisher Information of the KL with $\beta = 1$,

$$I(\alpha) = \int_{\mathcal{X}} \rho_\alpha \left| \nabla_{\mathcal{X}} \left(\log \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|^2 \, \mathrm{d}m(x) = \int_{\mathcal{X}} \frac{1}{\rho_\alpha} \left| \rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right|^2 \, \mathrm{d}m(x)$$

Computing the Wasserstein gradient flow,

$$\begin{aligned} \rho'_\alpha(t) &= -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{4}{\sqrt{\rho_\alpha}} \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\sqrt{\rho_\alpha} \right) \right) - \frac{4}{\sqrt{\rho_\infty}} \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\sqrt{\rho_\infty} \right) \right) \right) \right) \\ &= -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\left(\frac{\rho_\infty}{\rho_\alpha} \right)^2 \left(\nabla_{\mathcal{X}} \frac{\rho_\alpha}{\rho_\infty} \right)^2 - \left(\frac{2}{\rho_\alpha} \right) \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right) \right) \end{aligned}$$

which is the Derrida-Lebowitz-Speer-Spohn equation [33, 48]. For this equation you have to impose two boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}} \left(\rho_\alpha \right) = 0, \qquad \rho_\alpha \frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{4}{\sqrt{\rho_\alpha}} \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\sqrt{\rho_\alpha} \right) \right) - \frac{4}{\sqrt{\rho_\infty}} \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\sqrt{\rho_\infty} \right) \right) \right) = 0$$

Surprisingly this equation behaves very simmilarly to the general difussion equation, with simmilar bounds, temporal and stationary solutions.

- This functional might be generalized into a family of first order entropies, with $s > 1$,

$$I^s(\alpha) = \int_{\mathcal{X}} \rho_\alpha \left| \nabla_{\mathcal{X}} \left(\log \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|^s \, \mathrm{d}m(x) = \int_{\mathcal{X}} \frac{1}{\rho_\alpha^{s-1}} \left| \rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right|^s \, \mathrm{d}m(x)$$

Computing the Wasserstein gradient flow, Computing the Wasserstein gradient flow,

$$\rho'_\alpha(t) = -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\left(\frac{\rho_\infty}{\rho_\alpha} \right)^s \left(\nabla_{\mathcal{X}} \frac{\rho_\alpha}{\rho_\infty} \right)^s - \left(\frac{s}{\rho_\alpha} \right) \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \left(\frac{\rho_\infty}{\rho_\alpha} \right)^{s-1} \left(\nabla_{\mathcal{X}} \frac{\rho_\alpha}{\rho_\infty} \right)^{s-1} \right) \right) \right)$$

For this equation you have to impose two boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}} \left(\rho_\alpha \right) = 0, \qquad \rho_\alpha \frac{\partial}{\partial \boldsymbol{\nu}} \left(\left(\frac{\rho_\infty}{\rho_\alpha} \right)^s \left(\nabla_{\mathcal{X}} \frac{\rho_\alpha}{\rho_\infty} \right)^s - \left(\frac{s}{\rho_\alpha} \right) \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \left(\frac{\rho_\infty}{\rho_\alpha} \right)^{s-1} \left(\nabla_{\mathcal{X}} \frac{\rho_\alpha}{\rho_\infty} \right)^{s-1} \right) \right) = 0$$

- As an special case of this functional we have $s = 1$, called the Total Variation(TV)[65],

$$TV(\alpha) = I^1(\alpha) = \int_{\mathcal{X}} \rho_\alpha \left| \nabla_{\mathcal{X}} \left(\log \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right| \, \mathrm{d}m(x) = \int_{\mathcal{X}} \left| \rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right| \, \mathrm{d}m(x)$$

Computing the Wasserstein gradient flow,

$$\rho'_\alpha(t) = -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\tfrac{1}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \frac{\nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right)}{\left| \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right|} \right) \right) \right) \quad ,$$

is an equation recently studied in [14, 26, 6].For this equation you have to impose two boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)=0, \qquad \rho_\alpha \frac{\partial}{\partial \boldsymbol{\nu}}\left(\frac{1}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty \frac{\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)}{\left|\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right|}\right)\right)=0$$

- Simmilarly we can propose second-order entropies first considering the Biharmonic energy [41],

$$E_B(\alpha)=\frac{1}{2}\int_{\mathcal{X}}\frac{1}{\rho_\infty}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right|^2\,\mathrm{d}m(x)$$

Computing the Wasserstein gradient flow,

$$\rho_\alpha'(t) \quad = \quad \nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\frac{1}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{1}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right)\right)\right)\right)$$

which is a sixth-order equation. For this equation you have to impose three boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)=0, \qquad \frac{\partial}{\partial \boldsymbol{\nu}}\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)=0, \qquad \rho_\alpha\nabla_{\mathcal{X}}\left(\frac{1}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{1}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right)\right)\right)=0$$

- Also consider the Fisher Information metric of the KL with $\beta=1$,

$$\|\alpha(t)\|^2=\int_\Omega\frac{1}{\rho_\alpha}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right|^2\,\mathrm{d}m(x)=\int_\Omega\frac{1}{\rho_\alpha}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\log\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right)\right|^2\,\mathrm{d}m(x),$$

Computing the Wasserstein gradient flow,

$$\rho_\alpha'(t) \quad = \quad \nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\frac{2}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{1}{\rho_\alpha}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right)\right)-\frac{1}{\rho_\alpha^2}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right|^2\right)\right)$$

For this equation you have to impose three boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)=0, \qquad \frac{\partial}{\partial \boldsymbol{\nu}}\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)=0,$$

$$\rho_\alpha\frac{\partial}{\partial \boldsymbol{\nu}}\left(\frac{2}{\rho_\infty}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{1}{\rho_\alpha}\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right)\right)-\frac{1}{\rho_\alpha^2}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right|^2\right)=0$$

- This functional might be generalized into a family of second order entropies,

$$\|\alpha(t)\|_s^2=\int_{\mathcal{X}}\frac{1}{\rho_\alpha^{s-1}}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\infty\nabla_{\mathcal{X}}\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right|^s\,\mathrm{d}m(x)=\int_\Omega\frac{1}{\rho_\alpha^{s-1}}\left|\nabla_{\mathcal{X}}\cdot\left(\rho_\alpha\nabla_{\mathcal{X}}\left(\log\left(\frac{\rho_\alpha}{\rho_\infty}\right)\right)\right)\right|^s\,\mathrm{d}m(x)$$

Computing the Wasserstein gradient flow,

$$\rho_\alpha'(t) \quad = \quad \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{s}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{1}{\rho_\alpha^{s-1}} \left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|^{s-1} \right) \right) - \frac{1}{\rho_\alpha^s} \left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|^s \right) \right)$$

For this equation you have to impose three boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) = 0, \qquad \frac{\partial}{\partial \boldsymbol{\nu}} \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) = 0,$$

$$\rho_\alpha \frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{s}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{1}{\rho_\alpha^{s-1}} \left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|^{s-1} \right) \right) - \frac{1}{\rho_\alpha^s} \left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|^s \right) = 0$$

- As an special case of this functional we have $s = 1$, called the second-order Total Variation(TV),

$$TV_2(\alpha) = \|\alpha(t)\|_1^2 = \int_{\mathcal{X}} \left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right| \, \mathrm{d}m(x)$$

Computing the Wasserstein gradient flow,

$$\rho_\alpha'(t) \quad = \nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\tfrac{1}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right)}{\left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|} \right) \right) \right) \right) \quad ,$$

For this equation you have to impose three boundary conditions,

$$\frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) = 0, \qquad \frac{\partial}{\partial \boldsymbol{\nu}} \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) = 0, \qquad \rho_\alpha \frac{\partial}{\partial \boldsymbol{\nu}} \left(\frac{1}{\rho_\infty} \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right)}{\left| \nabla_{\mathcal{X}} \cdot \left(\rho_\infty \nabla_{\mathcal{X}} \left(\frac{\rho_\alpha}{\rho_\infty} \right) \right) \right|} \right) \right) \right) = 0$$

3.2.6 Fundamental Solutions

In the next section we'll consider the global cauchy problem of some of the equations given in the previous section with $\rho_\infty = 1$,

1. (KL-Divergence) Heat Equation:
$$\frac{\partial \rho_\alpha}{\partial t} = \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\rho_\alpha \right) \right),$$
2. (L^m-Distance) Porous Media Equation with $m = 2$,
$$\frac{\partial \rho_\alpha}{\partial t} = \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\rho_\alpha^2 \right) \right),$$

3. (Dirichlet energy) Thin Film Equation,

$$\rho_\alpha'(t) \quad = \quad -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\rho_\alpha\right)\right)\right)\right)$$

4. (Fisher Information) Derrida-Lebowitz-Speer-Spohn(DLSS) equation,

$$\rho_\alpha'(t) \quad = \quad -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\frac{4}{\sqrt{\rho_\alpha}} \nabla_{\mathcal{X}} \cdot \left(\nabla_{\mathcal{X}} \left(\sqrt{\rho_\alpha}\right)\right)\right)\right)$$

5. (Total Variation) Wasserstein Flow,

$$\rho_\alpha'(t) \quad = -\nabla_{\mathcal{X}} \cdot \left(\rho_\alpha \nabla_{\mathcal{X}} \left(\nabla_{\mathcal{X}} \cdot \left(\frac{\nabla_{\mathcal{X}}(\rho_\alpha)}{|\nabla_{\mathcal{X}}(\rho_\alpha)|}\right)\right)\right) \quad ,$$

We'll consider the inital conditions as,

$$\rho_\alpha(x, 0) = \delta(x - s) \qquad x \in \Omega \subset \mathbb{R}^n,$$

and the boundary condition is the global cauchy condition

$$\lim_{|x| \to \infty} \rho_\alpha(x, t) = 0.$$

Let's simplify and work with $\mathbb{R}^1$, to obtain the fundamental solution of this equations we use the results of [13], where similarity arguments were used to obtain the solutions. This method is in some sense a generalazation of the invariant transformation for the heat equation[63]. If we normalize to the conditions given above, we obtain,

1. (KL-Divergence) Heat Equation:

$$\rho_\alpha(x, t) = \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|x-s|^2}{4t}\right)}, \; t > 0$$

2. (L^m-Distance) Porous Media Equation with $m = 2$,

$$\rho_\alpha(x, t) = \begin{cases} \frac{-(x-s)^2}{12t} + \frac{1}{4}\left(\frac{3}{t}\right)^{1/3} & x \in \left[-(9t)^{1/3} + s, (9t)^{1/3} + s\right] \\ 0 & \text{otherwise} \end{cases}$$

3. (Dirichlet energy) Thin Film Equation,

$$\rho_\alpha(x, t) = \begin{cases} \frac{(x-s)^4}{120t} - \frac{\left(\frac{1}{t^{3/2}}\right)^{2/5}(x-s)^2}{4 \cdot 15^{1/5} \cdot 2^{2/5}} + \frac{t\left(\frac{1}{t^{3/2}}\right)^{4/5} 15^{3/5}}{8 \cdot 2^{4/5}} & x \in \left[-\frac{\sqrt{t}\left(\frac{1}{t^{3/2}}\right)^{1/5} 15^{2/5}}{2^{1/5}} + s, \frac{\sqrt{t}\left(\frac{1}{t^{3/2}}\right)^{1/5} 15^{2/5}}{2^{1/5}} + s\right] \\ 0 & \text{otherwise} \end{cases}$$

4. (Fisher Information) Derrida-Lebowitz-Speer-Spohn(DLSS) equation,

$$\rho_\alpha(x, t) = \frac{1}{\left(2\pi\sqrt{2t}\right)^{1/2}} e^{\left(\frac{-|x-s|^2}{2\sqrt{2t}}\right)}, \; t > 0$$

5. (Total Variation) Wasserstein Flow,

$$\rho_\alpha(x,t) = \begin{cases} 1/(2\sqrt[6]{9t}) & x \in \left[-\sqrt[6]{9t}+s,\, \sqrt[6]{9t}+s\right] \\ 0 & \text{otherwise} \end{cases}$$

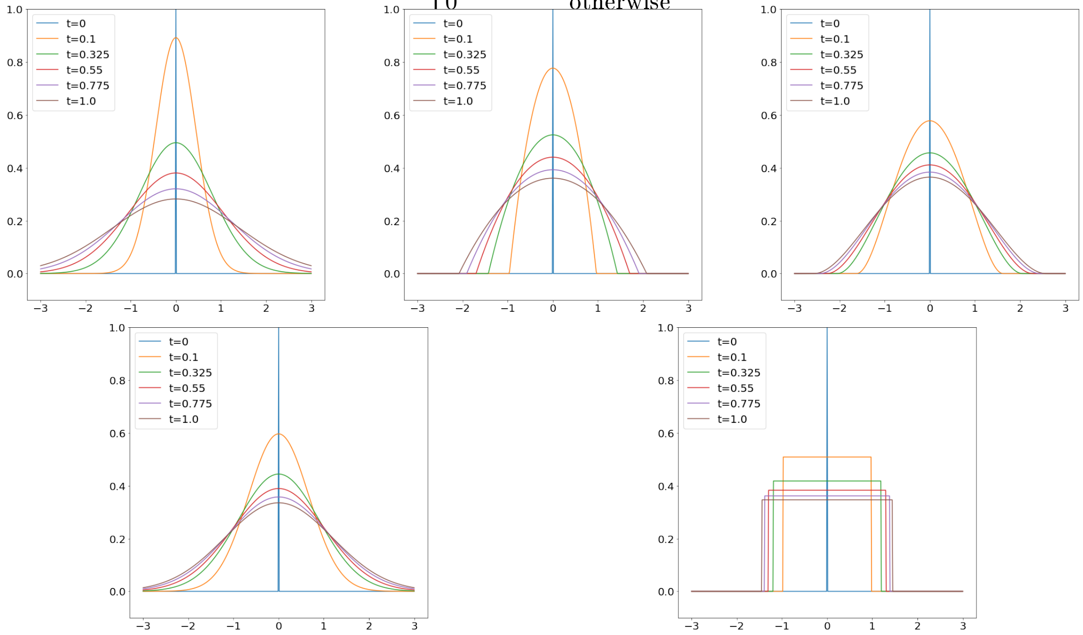

Figure 9: Fundamental Solutions at six different times t of Equations 1-5 with $s = 0$.

3.2.7 Elementary Examples

- Simplest example Gaussian in $\mathbb{R}$, compute Mean Energy, Entropy, Free Energy, Fisher information(generalized and normal), Mean Energy production, Fisher Information metric if possible to compute for fundamental solution(cauchy) and then plot them for numerical example(neumman). **Appendix C**.
- Transport of Two Gaussians in $\mathbb{R}$, compute Mean Energy, Entropy, Free Energy, Fisher information(generalized and normal), Mean Energy production, Fisher Information metric, and plot them for analytical(cauchy) and numerical example(neumman). Then introduce the transport for $\mathbb{R}$ and $\mathbb{R}^2$ from one to two gaussians as an example. **Appendix D**.
- Most "complex" distribution[61, 3], using the KDE examples and the Energy, Entropy, Free Energy, Fisher information(generalized and normal), Energy production, **Fisher Information metric**. Fisher-invariance-Nlab, Tweet Peyre, DivergenceSimon KL Applications.

3.3 Entropic Inequalities

- Exponential decay results for statisitcal divergences(PDE literature[47],Stochastic Processes[57],Easy[56]). High order entropies decay results ??
- Mention that with this method you can control the moments of the equilibrium distribution, but not for all the process[55]. What happens with high order entropies ??

4 Conclusions

References

[1] Arnold, V. I., Vogtmann, K., and Weinstein, A. *Mathematical Methods of Classical Mechanics.* Graduate Texts in Mathematics. Springer, New York, NY, 2013.

[2] BAEZ, J., and STAY, M. Algorithmic thermodynamics. *Mathematical Structures in Computer Science 22*, 5 (2012), 771â787.

[3] Baez, J. C., and Pollard, B. S. Relative entropy in biological systems. *Entropy 18*, 2 (2016).

[4] Bauschke, H. H., and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer New York, 2011.

[5] Benamou, J.-D., and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik 84*, 3 (Jan. 2000), 375–393.

[6] Benning, M., Calatroni, L., Düring, B., and Schönlieb, C.-B. A primal-dual approach for a total variation wasserstein flow. In *Geometric Science of Information* (Berlin, Heidelberg, 2013), F. Nielsen and F. Barbaresco, Eds., Springer Berlin Heidelberg, pp. 413–421.

[7] Berger, A. L., Della Pietra, V. J., and Della Pietra, S. A. A maximum entropy approach to natural language processing. *Computational linguistics - Association for Computational Linguistics 22*, 1 (1996), 39–71.

[8] Blanco Castaneda, L., Arunachalam, V., and Dharmaraja, S. *Introduction to Probability and Stochastic Processes with Applications*, 1. aufl. ed. Wiley, Hoboken, NJ, 2012.

[9] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. Kernel density estimation via diffusion. *The Annals of Statistics 38*, 5 (Oct. 2010), 2916–2957.

[10] Botev, Z. I., and Kroese, D. P. The generalized cross entropy method, with applications to probability density estimation. *Methodology and Computing in Applied Probability 13*, 1 (May 2009), 1–27.

[11] Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics 7*, 3 (Jan. 1967), 200–217.

[12] Brillouin, L. *Science and information theory.*, 2d ed. ed. Academic Press, New York, 1962.

[13] Burger, M., Franek, M., and Schonlieb, C.-B. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress* (Mar. 2012).

[14] Carlier, G., and Poon, C. On the total variation wasserstein gradient flow and the TV-JKO scheme. *ESAIM: Control, Optimisation and Calculus of Variations 25* (2019), 42.

[15] Carrillo, J. A., Jüngel, A., Markowich, P. A., Toscani, G., and Unterreiter, A. Entropy dissipation methods for degenerate ParabolicProblems and generalized sobolev inequalities. *Monatshefte fur Mathematik 133*, 1 (May 2001), 1–82.

[16] Chaitin, G. J. *Algorithmic information theory.* Cambridge tracts in theoretical computer science ; 1. 1987.

[17] Chang, K. K., and DeDeo, S. Divergence and the complexity of difference in text and culture. *Journal of Cultural Analytics* (Oct. 2020).

[18] Chang, L. Gradient flows. http://ml.cs.tsinghua.edu.cn/ changliu/static/Gradient-Flow.pdf, 2017. Talk at Tsinghua University.[Online; accessed 28-Dic-2020].

[19] Conrad, K. Probability distributions and maximum entropy. http://www.math.uconn.edu/kconrad/blurbs/entropy.pdf, 2005. [Online; accessed 28-Dic-2020].

[20] Cormen, T. H., Leiserson, C. E., Rivest, R., and Stein, C. *Introduction to algorithms*, third edition ed. 2009.

[21] Cover, T. M. *Elements of information theory*, 2nd ed. ed. Wiley-Interscience, Hoboken, N.J., 2006.

[22] Craig, K. Gradient flow in the wasserstein metric, 2017. NIPS Optimal Transport & Machine Learning[Online; accessed 28-Dic-2020].

[23] Csiszár, I. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica 2*, 1-4 (Mar. 1972), 191–213.

[24] DeDeo, S. Maximum entropy methods. https://www.complexityexplorer.org/, 2015. Current-2020 Complexity Explorer: Santa Fe Institute. License: Creative Commons BY-NC-SA.

[25] Deutsch, D. Quantum theory, the church-turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences 400*, 1818 (1985), 97–117.

[26] Düring, B., and Schönlieb, C.-B. A high-contrast fourth-order PDE from imaging: numerical solution by ADI splitting, 2012.

[27] Evans, L. Lecture notes for a graduate course "entropy and partial differential equations". https://math.berkeley.edu/ evans/entropy.and.PDE.pdf. [Online; accessed 28-Dic-2020].

[28] Ford, I. *Statistical physics an entropic approach.* Wiley, Chichester, 2013.

[29] Frank, T. D. *Nonlinear Fokker-Planck equations : fundamentals and applications.* Springer complexity. Springer, Berlin ; New York, 2005.

[30] Frigg, R., and Werndl, C. Entropy - a guide for the perplexed, 2011. In *Probabilities in Physics*, Oxford University Press.

[31] Gardiner, C. W., and Gardiner, C. W. *Stochastic methods: a handbook for the natural and social sciences*, 4th ed ed. Springer series in synergetics. Springer, Berlin, 2009.

[32] Gast, T. F., Schroeder, C., Stomakhin, A., Jiang, C., and Teran, J. M. Optimization integrator for large time steps. *IEEE Transactions on Visualization and Computer Graphics 21*, 10 (Oct. 2015), 1103–1115.

[33] Gianazza, U., Savaré, G., and Toscani, G. The wasserstein gradient flow of the fisher information and the quantum drift-diffusion equation. *Archive for Rational Mechanics and Analysis 194*, 1 (Oct. 2008), 133–220.

[34] Gomez, F., Torres, A., Galvis, J., Camargo, J., and Martinez, O. Hotspot mapping for perception of security. In *2016 IEEE International Smart Cities Conference (ISC2)* (2016), pp. 1–6.

[35] Grandy, W. T., and Schick, L. H., Eds. *Maximum Entropy and Bayesian Methods.* Springer Netherlands, 1991.

[36] Grinspun, E. Session details: Discrete differential geometry: an applied introduction. In *ACM SIGGRAPH 2006 Courses* (2006), SIGGRAPH '06, ACM.

[37] Halder, A., and Georgiou, T. T. Gradient flows in uncertainty propagation and filtering of linear gaussian systems. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* (2017), IEEE, pp. 3081–3088.

[38] Hand, L. N. *Analytical mechanics.* Cambridge University Press, Cambridge ; New York, 1998.

[39] Haran, B., and Numberphile. The LONGEST time - Numberphile ft. Tony Paddilla. https://youtu.be/1GCf29FPM4k, 2012. [Online; accessed 28-Dic-2020].

[40] Hartle, J. B. J. B. *Gravity : an introduction to Einstein's general relativity.* Addison-Wesley, San Francisco, 2003. Mathematica Programs.Christoffel Symbols and Geodesic Equations. http://web.physics.ucsb.edu/ gravitybook/mathematica.html.

[41] Jacobson, A., Baran, I., Popović, J., and Sorkine, O. Bounded biharmonic weights for real-time deformation. *ACM Transactions on Graphics 30*, 4 (July 2011), 1–8.

[42] Jaynes, E. T. Information theory and statistical mechanics. *Physical Review 106*, 4 (May 1957), 620–630.

[43] Jaynes, E. T. Information theory and statistical mechanics. II. *Physical Review 108*, 2 (Oct. 1957), 171–190.

[44] Jaynes, E. T. E. T. *Probability theory the logic of science.* Cambridge University Press, Cambridge, UK ; New York, NY, 2003.

[45] Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis 29*, 1 (Jan. 1998), 1–17.

[46] Jumarie, G. Solution of the fokker plank-kolmogorov equation by using the maximum entropy principle. *Journal of Information and Optimization Sciences 6*, 1 (Jan. 1985), 1–16.

[47] Jungel, A. *Entropy Methods for Diffusive Partial Differential Equations.* SpringerBriefs in Mathematics. Springer International Publishing AG, Cham, 2016.

[48] Jüngel, A., and Matthes, D. The derrida–lebowitz–speer–spohn equation: Existence, NonUniqueness, and decay rates of the solutions. *SIAM Journal on Mathematical Analysis 39*, 6 (Jan. 2008), 1996–2015.

[49] Knuth, D. E. *The Art of Computer Programming, Volume 2 (3rd Ed.): Seminumerical Algorithms.* Addison-Wesley Longman Publishing Co., Inc., USA, 1997.

[50] Levien, R. L. *Geometric data analysis, beyond convolutions.* PhD thesis, ENS Paris-Saclay, France, 2020.

[51] Li, M., and Vitaanyi, P. *An introduction to Kolmogorov complexity and its applications*, fourth edition. ed. Texts in computer science. 2019.

[52] Mehmood, R., Zhang, G., Bie, R., Dawood, H., and Ahmad, H. Clustering by fast search and find of density peaks via heat diffusion. *Neurocomputing 208* (2016), 210 – 217. SI: BridgingSemantic.

[53] Mitchell, M. *Complexity : a guided tour.* Oxford University Press, Oxford [England] ; New York, 2009.

[54] Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations 26*, 1-2 (Jan. 2001), 101–174.

[55] Patarroyo, K. Y. Mean conservation for density estimation via diusion using the finite element method, 2017.

[56] Pavliotis, G. The focker planck equation, 2015. Imperial College[Online; accessed 28-Dic-2020].

[57] Pavliotis, G. A. *Stochastic processes and applications : diffusion processes, the Fokker-Planck and Langevin equations*, 1st ed. 2014. ed. Texts in applied mathematics ; Volume 60. 2014.

[58] Penfield, P., and Lloyd, S. 6-050j information and entropy. https://ocw.mit.edu/, 2008. Massachusetts Institute of Technology: MIT OpenCourseWare. License: Creative Commons BY-NC-SA.

[59] Peyré, G. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences 8*, 4 (Jan. 2015), 2323–2351.

[60] Peyré, G., and Cuturi, M. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning 11*, 5-6 (2019), 355–607.

[61] Pollard, B. S. Open markov processes: A compositional perspective on non-equilibrium steady states in biology. *Entropy 18*, 4 (2016).

[62] Renyi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (Berkeley, Calif., 1961), University of California Press, pp. 547–561.

[63] Salsa, S. *Partial Differential Equations in Action: From Modelling to Theory.* Universitext. Springer Milan, Milano.

[64] Santambrogio, F. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, vol. 87 of *Progress in nonlinear differential equations and their applications.* Springer International Publishing AG, Cham, 2015.

[65] Sapiro, G. *Geometric Partial Differential Equations and Image Analysis.* Cambridge University Press, USA, 2006.

[66] Seis, C. The thin-film equation close to self-similarity. *Anal. PDE 11*, 5 (2018), 1303–1342.

[67] Sethna, J. P. *Statistical mechanics entropy, order parameters, and complexity.* Oxford master series in statistical, computational, and theoretical physics. Oxford University Press, Oxford ; New York, 2006.

[68] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal 27*, 3 (July 1948), 379–423.

[69] Shashua, A. Introduction to machine learning: Class notes 67577, 2009.

[70] Sommerfeld, A., Bopp, F., Meixner, J., and Kestin, J. *Lectures on Theoretical Physics: Thermodynamics and Statistical Mechanics.* Elsevier Science & Technology, Saint Louis, 1964.

[71] Susskind, L. Statistical Mechanics Lecture 8-Entropy, reversibility, and magnetism. http://theoreticalminimum.com/courses/statistical-mechanics/2013/spring/lecture-8, 2013. [Online; accessed 28-Dic-2020].

[72] Tsallis, C. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics 52*, 1-2 (July 1988), 479–487.

[73] Vazquez, J. L. *The Porous Medium Equation: Mathematical Theory.* Oxford Mathematical Monographs. Oxford University Press, Oxford, 2006.

[74] Verdu, S. See the section differential entropy 4 in relative entropy video lecture, 2009. NIPS[Online; accessed 28-Dic-2020].

[75] Villani, C. *Topics in optimal transportation.* Graduate studies in mathematics ; v. 58. American Mathematical Society, Providence, RI, 2003.

[76] Villani, C. *Entropy Production and Convergence to Equilibrium.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 1–70.

[77] Von Neumann, J. *Theory of self-reproducing automata.* University of Illinois Press, Urbana, 1966.

[78] Wolfram, S. *A new kind of science.* Wolfram Media, Champaign, IL, 2002.

[79] Wolfram, S. Buzzword Convergence: Making Sense of Quantum Neural Blockchain AI. https://writings.stephenwolfram.com/2018/04/buzzword-convergence-making-sense-of-quantum-neural-blockchain-ai/, 2018. [Online; accessed 28-Dic-2020].

[80] Wolfram, S. *A project to find the fundamental theory of physics.* Wolfram Media, Champaign, IL, 2020.

[81] Wolfram, S., et al. Wolfram Research Inc. Official Website of the Wolfram Physics Project. https://www.wolframphysics.org/, 2020. [Online; accessed 28-Dic-2020].

[82] Wolpert, D. H. The stochastic thermodynamics of computation. *Journal of Physics A: Mathematical and Theoretical 52*, 19 (Apr. 2019), 193001.

[83] Wolpert, D. H., Kempes, C., Stadler, P. F., and Grochow, J. A., Eds. *The energetics of computing in life and machines.* No. book 4 in Seminar. SFI Press, Santa Fe, 2019.

[84] Zurek, W. H. *Complexity, entropy, and the physics of information: the proceedings of the 1988 Workshop on Complexity, Entropy, and the Physics of Information held May-June, 1989, in Santa Fe, New Mexico*, vol. 8 of *Santa Fe Institute studies in the sciences of complexity.* CRC Press, 2018.

Extra

- Chen, X., & Yang, Y. (2020-a). Diffusion K-means clustering on manifolds: Provable exact recovery via semidefinite relaxationsApplied and Computational Harmonic Analysis. arXiv:1903.04416
- Chen, X. (2020-b). Continuity equations, PDEs, probabilities, and gradient flows. `http://www.mit.edu/~xiaohui/continuity_equation.pdf`

Appendix A

In $\mathbb{R}$, let's consider $\alpha = \mathcal{N}\left(\mu_\alpha, \sigma_\alpha^2\right)$ and $\beta = \mathcal{N}\left(\mu_\beta, \sigma_\beta^2\right)$, then one has,

$$\begin{aligned}
\mathrm{KL}(\alpha|\beta) &= 1 + \mathbb{E}_{x\sim\alpha}\left[\log\left(\frac{\rho_\alpha(x)}{\rho_\beta(x)}\right)\right] - \mathbb{E}_{x\sim\beta}\left[\frac{\rho_\alpha(x)}{\rho_\beta(x)}\right] \\
&= \frac{1}{\sigma_\alpha\sqrt{2\pi}}\int_{\mathbb{R}} e^{\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2}\left(\log\left(\frac{1}{\sigma_\alpha\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2}\right) - \log\left(\frac{1}{\sigma_\beta\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2}\right)\right)dx \\
&= \frac{1}{\sigma_\alpha\sqrt{2\pi}}\int_{\mathbb{R}} e^{\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2}\left(\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2 + \frac{1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2 + \log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right)\right)dx \\
&= \frac{e^{\frac{-\mu_\alpha^2}{2\sigma_\alpha^2}}}{\sigma_\alpha\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2\sigma_\alpha^2}x^2+\frac{\mu_\alpha}{\sigma_\alpha^2}x}\left(\left(\frac{1}{2\sigma_\beta^2}-\frac{1}{2\sigma_\alpha^2}\right)x^2 - \left(\frac{\mu_\beta}{\sigma_\beta^2}-\frac{\mu_\alpha}{\sigma_\alpha^2}\right)x\right)dx \\
&\quad + \frac{1}{\sigma_\alpha\sqrt{2\pi}}\int_{\mathbb{R}} e^{\frac{-1}{2}\left(\frac{x-\mu_\alpha}{\sigma_\alpha}\right)^2}\left(\frac{\mu_\beta^2}{2\sigma_\beta^2}-\frac{\mu_\alpha^2}{2\sigma_\alpha^2}+\log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right)\right)dx
\end{aligned}$$

Using the formulas

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx = 1, \int_{-\infty}^{\infty} xe^{-ax^2+bx}dx = \frac{\sqrt{\pi}b}{2a^{3/2}}e^{\frac{b^2}{4a}}, \int_{-\infty}^{\infty} x^2e^{-(ax^2+bx)}dx = \frac{\sqrt{\pi}\left(2a+b^2\right)}{4a^{5/2}}e^{\frac{b^2}{4a}}$$

$$\begin{aligned}
\mathrm{KL}(\alpha|\beta) &= \frac{e^{-\frac{\mu_\alpha^2}{2\sigma_\alpha^2}}}{\sigma_\alpha\sqrt{2\pi}}\left[\left(\frac{1}{2\sigma_\beta^2}-\frac{1}{2\sigma_\alpha^2}\right)\sqrt{2\pi}\left(2\frac{1}{2\sigma_\alpha^2}+\frac{\mu_\alpha^2}{\sigma_\alpha^4}\right)\sigma_\alpha^5 e^{\left(\frac{\mu_\alpha}{\sigma_\alpha^2}\right)^2\frac{2\sigma_\alpha^2}{4}} - \left(\frac{\mu_\beta}{\sigma_\beta^2}-\frac{\mu_\alpha}{\sigma_\alpha^2}\right)\frac{\sqrt{2\pi}\mu_\alpha\sigma_\alpha^3}{\sigma_\alpha^2}e^{\left(\frac{\mu_\alpha}{\sigma_\alpha^2}\right)^2\frac{2\sigma_\alpha^2}{4}}\right] \\
&\quad + \left(\frac{\mu_\beta^2}{2\sigma_\beta^2}-\frac{\mu_\alpha^2}{2\sigma_\alpha^2}+\log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right)\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\alpha|\beta) &= \left(\frac{1}{2\sigma_\beta^2}-\frac{1}{2\sigma_\alpha^2}\right)\left(2\frac{1}{2\sigma_\alpha^2}+\frac{\mu_\alpha^2}{\sigma_\alpha^4}\right)\sigma_\alpha^4 - \left(\frac{\mu_\beta}{\sigma_\beta^2}-\frac{\mu_\alpha}{\sigma_\alpha^2}\right)\mu_\alpha + \left(\frac{\mu_\beta^2}{2\sigma_\beta^2}-\frac{\mu_\alpha^2}{2\sigma_\alpha^2}+\log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right)\right) \\
&= \left(\frac{1}{2\sigma_\beta^2}-\frac{1}{2\sigma_\alpha^2}\right)\left(\sigma_\alpha^2+\mu_\alpha^2\right) - \frac{\mu_\beta\mu_\alpha}{\sigma_\beta^2} + \frac{\mu_\alpha^2}{\sigma_\alpha^2} + \frac{\mu_\beta^2}{2\sigma_\beta^2} - \frac{\mu_\alpha^2}{2\sigma_\alpha^2} + \log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right) \\
&= \frac{\sigma_\alpha^2}{2\sigma_\beta^2} - \frac{1}{2} + \frac{\mu_\alpha^2}{2\sigma_\beta^2} - \frac{\mu_\alpha^2}{2\sigma_\alpha^2} - \frac{\mu_\beta\mu_\alpha}{\sigma_\beta^2} + \frac{\mu_\alpha^2}{2\sigma_\alpha^2} + \frac{\mu_\beta^2}{2\sigma_\beta^2} + \log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right) \\
&= \frac{1}{2}\left(2\log\left(\frac{\sigma_\beta}{\sigma_\alpha}\right) + \frac{\left(\mu_\alpha-\mu_\beta\right)^2}{\sigma_\beta^2} + \frac{\sigma_\alpha^2}{\sigma_\beta^2} - 1\right)
\end{aligned}$$

Appendix B

To compute the geodesics of a metric $g_{\mu\nu}dx_\nu dx\mu$, we compute the Christoffel symobols according to the formula,

$$\Gamma^\lambda_{\mu v} = \frac{1}{2}g^{\lambda\sigma}\left(\partial_\mu g_{\sigma v} + \partial_v g_{\sigma\mu} - \partial_\sigma g_{\mu v}\right),$$

where $g^{\lambda\sigma}$ is the inverse of the metric matrix. Then using this expression the geodesic equation,

$$\frac{d^2x^\alpha}{d\tau^2} = -\Gamma^\alpha_{\beta\gamma}\frac{dx^\beta}{d\tau}\frac{dx^\gamma}{d\tau}$$

For our the KL induced-metric in the coordinates (μ, σ)

$$g_{\mu\nu} = \left(\begin{array}{cc} \frac{1}{2\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{array}\right)$$

The non-zero Christoffel symbols are

$$\Gamma^{2,1}_1 = \frac{-1}{\sigma}, \qquad \Gamma^{1,1}_2 = \frac{1}{2\sigma}, \qquad \Gamma^{2,2}_2 = \frac{-1}{\sigma}.$$

Hence the Geodesic equations are,

$$\frac{d^2\mu}{d\tau^2} = \frac{2}{\sigma}\frac{d\mu}{d\tau}\frac{d\sigma}{d\tau}, \qquad \frac{d^2\sigma}{d\tau^2} = \frac{-(\mu^2 - 2\sigma^2)}{2\sigma}.$$

Solving these coupled equations [40] gives the Figure 5.

Appendix C

As the simplest example,

$$\rho_\alpha(\mathbf{x},t) = \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)},\ t > 0$$

$$\frac{\partial \rho_\alpha}{\partial t} = \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)} \left(\frac{-1}{2t} + \frac{|\mathbf{x}-\mathbf{s}|^2}{4t^2}\right)$$

Now,

$$\begin{aligned}
\frac{1}{\rho_\alpha(t)}\left(\frac{\partial \rho_\alpha}{\partial t}\right)^2 &= \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)} \left(\frac{-1}{2t} + \frac{|\mathbf{x}-\mathbf{s}|^2}{4t^2}\right)^2 \\
&= \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)} \left(\frac{1}{4t^2} + \frac{|\mathbf{x}-\mathbf{s}|^2}{4t^3} + \frac{|\mathbf{x}-\mathbf{s}|^4}{16t^4}\right)
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\alpha(t)\|^2 &= \int_{\mathbb{R}} \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)} \left(\frac{1}{4t^2} - \frac{|\mathbf{x}-\mathbf{s}|^2}{4t^3} + \frac{|\mathbf{x}-\mathbf{s}|^4}{16t^4}\right) dx \\
&= \frac{1}{4t^2} - \frac{1}{4t^3}\int_{\mathbb{R}} |\mathbf{x}-\mathbf{s}|^2 \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)} dx + \frac{1}{16t^4}\int_{\mathbb{R}} |\mathbf{x}-\mathbf{s}|^4 \frac{1}{(4\pi t)^{1/2}} e^{\left(\frac{-|\mathbf{x}-\mathbf{s}|^2}{4t}\right)} dx \\
&= \frac{1}{4t^2} - \frac{1}{4t^3}2t + \frac{1}{16t^4}12t^2 \\
&= \frac{1}{2t^2}
\end{aligned}$$

Hence the norm of a vector

$$\|\alpha(t)\| = \frac{1}{t\sqrt{2}}$$

The time where the hyperbola changes abrutly is

$$t = 2^{-\frac{1}{4}} \approx 0.84$$

Hence the "optimal" gaussian is

$$2t = \sigma^2 \Longrightarrow \sigma = 2^{\frac{3}{8}} \approx 1.3$$

More calculations...

Appendix D

Lets take the difussion equation with $\beta^{-1} = 1$ and $a(x) = \rho_\infty(x)$ with $\frac{d}{dx}\left(\frac{p(x,t)}{a(x)}\right) = 0$ on the boundary

$$\begin{aligned}
\frac{\partial \rho_\alpha(x,t)}{\partial t} &= \frac{d}{dx}\left(a(x)\frac{d}{dx}\left(\frac{\rho_\alpha(x,t)}{a(x)}\right)\right) \\
&= \frac{d}{dx}\left(a(x)\left(-\frac{\rho_\alpha(x,t)}{a^2(x)}\frac{da(x)}{dx} + \frac{1}{a(x)}\frac{d\rho_\alpha(x,t)}{dx}\right)\right) \\
&= \frac{d}{dx}\left(-\frac{\rho_\alpha(x,t)}{a(x)}\frac{da(x)}{dx} + \frac{d\rho_\alpha(x,t)}{dx}\right)
\end{aligned}$$

Now let's take $a(x) = \frac{1}{\sigma_\beta\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2}$, hence,

$$\begin{aligned}
\frac{\partial \rho_\alpha(x,t)}{\partial t} &= \frac{d}{dx}\left(-\rho_\alpha(x,t)e^{\frac{1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2}e^{\frac{-1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2}\frac{-2}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\frac{1}{\sigma_\beta} + \frac{d\rho_\alpha(x,t)}{dx}\right) \\
&= \frac{1}{\sigma_\beta}\frac{d}{dx}\left(\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\right)\rho_\alpha(x,t) + \frac{1}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\frac{d\rho_\alpha(x,t)}{dx} + \frac{d^2\rho_\alpha(x,t)}{dx^2} \\
&= \frac{1}{\sigma_\beta^2}\rho_\alpha(x,t) + \frac{1}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\frac{d\rho_\alpha(x,t)}{dx} + \frac{d^2\rho_\alpha(x,t)}{dx^2}
\end{aligned}$$

with $\frac{d}{dx}\left(\rho_\alpha(x,t)e^{\frac{1}{2}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)^2}\right) = 0$ on the boundary. Now using the ansatz [37]

$$\rho_\alpha(x,t) = \frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2},$$

hence we find,

$$\begin{aligned}
\frac{\partial \rho_\alpha(x,t)}{\partial t} &= \frac{-1}{\sigma(t)}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2}\sigma'(t) + \frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2}\frac{-2}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)\left(\frac{-\mu'(t)\sigma(t)-\sigma'(t)(x-\mu(t))}{\sigma(t)^2}\right) \\
&= \frac{1}{\sigma_\beta^2}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2} + \frac{1}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\frac{-2}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)\frac{1}{\sigma(t)}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2} \\
&\quad + \frac{d}{dx}\left(\frac{-2}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)\frac{1}{\sigma(t)}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2}\right) \\
&= \frac{1}{\sigma_\beta^2}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2} + \frac{1}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\frac{-2}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)\frac{1}{\sigma(t)}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2} \\
&\quad + \left(\frac{-1}{\sigma(t)}\right)\frac{1}{\sigma(t)}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2} + \left(\frac{x-\mu(t)}{\sigma(t)}\right)^2\frac{1}{\sigma(t)}\frac{1}{\sigma(t)\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)^2}
\end{aligned}$$

Doing some algebra,

$$\frac{-\sigma'(t)}{\sigma(t)} + \frac{-2}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)\left(\frac{-\mu'(t)\sigma(t)-\sigma'(t)(x-\mu(t))}{\sigma(t)^2}\right) = \frac{1}{\sigma_\beta^2} + \frac{1}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right)\frac{-2}{2}\left(\frac{x-\mu(t)}{\sigma(t)}\right)\frac{1}{\sigma(t)} + \left(\frac{-1}{\sigma(t)}\right)\frac{1}{\sigma(t)} + \left(\frac{x-\mu(t)}{\sigma(t)}\right)$$

Multiplying by $\sigma(t)^3$,

$$\begin{aligned}
-\sigma'(t)\sigma(t)^3 - \sigma(t)\left(x-\mu(t)\right)\left(-\mu'(t)\sigma(t) - \sigma'(t)(x-\mu(t))\right) &= \frac{\sigma(t)^4}{\sigma_\beta^2} - \frac{\left(x-\mu(t)\right)\sigma(t)^2}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right) - \sigma(t)^2 + (x-\mu(t))^2 \\
-\sigma'(t)\sigma(t)^3 + \mu'(t)\sigma(t)^2\left(x-\mu(t)\right) + \sigma(t)\sigma'(t)\left(x-\mu(t)\right)^2 &= \frac{\sigma(t)^4}{\sigma_\beta^2} - \frac{\left(x-\mu(t)\right)\sigma(t)^2}{\sigma_\beta}\left(\frac{x-\mu_\beta}{\sigma_\beta}\right) - \sigma(t)^2 + (x-\mu(t))^2
\end{aligned}$$

Gathering the terms for each power of x

$$\begin{aligned}
0 &= \left(-1 + \frac{\sigma(t)^2}{\sigma_\beta^2} + \sigma'(t)\sigma(t)\right)x^2 + \left(\mu'(t)\sigma(t)^2 - 2\mu(t)\sigma(t)\sigma'(t) + 2\mu(t) - \frac{\sigma(t)^2}{\sigma_\beta^2}(\mu(t)+\mu_\beta)\right)x. \\
&+ \left(-\frac{\sigma(t)^4}{\sigma_\beta^2} + \sigma(t)^2 - \sigma'(t)\sigma(t)^3 - \mu(t)\mu'(t)\sigma(t)^2 + \sigma(t)\sigma'(t)\mu(t)^2 - \mu(t)^2 + \frac{\sigma(t)^2\mu(t)\mu_\beta}{\sigma_\beta^2}\right)
\end{aligned}$$

Now we obtain,

$$\begin{aligned}
-1 + \frac{\sigma(t)^2}{\sigma_\beta^2} + \sigma'(t)\sigma(t) &= 0 \\
\mu'(t)\sigma(t)^2 - 2\mu(t)\sigma(t)\sigma'(t) + 2\mu(t) - \frac{\sigma(t)^2}{\sigma_\beta^2}(\mu(t)+\mu_\beta) &= 0 \\
-\mu(t)\mu'(t)\sigma(t)^2 + \sigma(t)\sigma'(t)\mu(t)^2 - \mu(t)^2 + \frac{\sigma(t)^2\mu(t)\mu_\beta}{\sigma_\beta^2} &= 0
\end{aligned}$$

Now we see that the second line can be reduced to,

$$\begin{aligned}
\mu'(t)\sigma(t)^2 + 2\mu(t)\left(-1 + \frac{\sigma(t)^2}{\sigma_\beta^2}\right) + 2\mu(t) - \frac{\sigma(t)^2}{\sigma_\beta^2}(\mu(t)+\mu_\beta) &= 0 \\
\mu'(t)\sigma(t)^2 - \frac{\sigma(t)^2}{\sigma_\beta^2}(-\mu(t)+\mu_\beta) &= 0
\end{aligned}$$

The last line also is reduced to an identity,

$$-\mu(t)\left(\frac{\sigma(t)^2}{\sigma_\beta^2}(-\mu(t)+\mu_\beta)\right) + \left(1 - \frac{\sigma(t)^2}{\sigma_\beta^2}\right)\mu(t)^2 - \mu(t)^2 + \frac{\sigma(t)^2\mu(t)\mu_\beta}{\sigma_\beta^2} = 0$$

Hence the two coupled equations are

$$\begin{aligned}
-1 + \frac{\sigma(t)^2}{\sigma_\beta^2} + \sigma'(t)\sigma(t) &= 0 \\
\mu'(t) - \frac{1}{\sigma_\beta^2}(-\mu(t)+\mu_\beta) &= 0
\end{aligned}$$