An Informal Introduction of Information Geometry and Its Applications in Non-equilibrium Thermodynamics

Group Meeting Lecture Notes

Lu Group @ UNC Chapel Hill
Nonequilibrium Thermodynamics in Chemistry

Jiming Zheng
jiming@unc.edu

October 23, 2023

Contents

1 Introduction[1]

This note aims to introduce information geometry without delving too deeply into differential geometry. Information geometry treats probability distributions as points on a manifold. Its metric tensor is given by the Fisher Information matrix. The most important structure in information geometry is the dually flat structure. Information geometry has wide applications in statistical inference, machine learning, optimization, open quantum systems and stochastic thermodynamics. It can describe the evolutionary dynamics of probability distributions, decomposition of the entropy production, and it can also be used in thermodynamic inference. These correspond to the Boltzmann, Clausius, and Gibbs' approach in classical thermodynamics.

Research on geometric thermodynamics has a long history[2]. Gibbs has already found the geometric structure under classical thermodynamics[3]. After that, some mathematicians found that the geometric structure of thermodynamics can be described by the contact geometry induced by Legendre transformation, which is an odd-dimensional counterpart of symplectic geometry. Many of the work in this field is done by Arnold and Zorich. About 44 years ago, George Ruppeiner find that the geometry of equilibrium thermodynamics can be described by Riemannian geometry, where the metric tensor is the Hessian matrix of thermodynamic entropy. In recent years, with the rise of stochastic thermodynamics, some scholars have begun to look for a geometric language that can describe it. Stochastic thermodynamics is based on stochastic process and stochastic calculus. So this geometric language turns out to be the geometric structure of probability distributions. As far as I know, there are at least two types of geometric structure for probability distributions. One is information geometry, the other one is Wasserstein geometry. Some scholars already shown that in equilibrium states, the information geometry just goes back to the Ruppeiner geometry. In recent years, information geometry has also shown great power in non-equilibrium thermodynamics.

This note is organized as follows. In section 2, we introduce the metric tensor on the manifold. In section 3, we introduce the dually flat structure of the manifold, which contains the dual connection in section 3.1, dual flatness in section 3.2 and dual coordinate systems in section 3.3. In section 4, we introduce general divergence functions on the manifold. The information geometric structure can also be introduced from divergence functions. In section 5, we introduce the generalized Pythagorean theorem on the manifold, which is induced by the flatness property. In section 6, we introduce the thermodynamic length, speed limit and gradient flow on the manifold, which is related to the Boltzmann's approach to thermodynamics. In section 7, we introduce the geometrical interpretation of entropy production and decompose it, which is related to the Clausius' approach to thermodynamics. In section 8, we introduce the exponential distribution and its relation to Maximum Entropy Principle, which is related to the Gibbs' approach to thermodynamics.

2 Fisher Information Metric

Let's consider probability distributions $p(\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{x}$ is in the sample space and $\boldsymbol{\theta} = \{\theta^1, \theta^2, \cdots, \theta^n\}$ is a vector of parameters. All the probability distributions lies on a manifold. The local coordinates on this manifold are these parameters. Now we want to figure out what is the geometric structure on this sort of manifold. Some mathematicians have already proven that the only metric tensor which is covariant under reparameterization and invariant under changing of variables is the Fisher Information metric tensor[4].

The elements of Fisher Information metric is

$$g_{\mu\nu} = \int \mathrm{d}\boldsymbol{x}\ p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_\mu \ln p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_\nu \ln p(\boldsymbol{x}, \boldsymbol{\theta}), \tag{1}$$

[1] As an informal introductory note, I skipped all the citations in the Introduction section.

[2] When I use the word "geometry" in this note, I'm talking about modern geometry, which is based on differential geometry.

[3] Classical thermodynamics here means equilibrium thermodynamics for classical systems.

[4] If these conditions are not necessary, we can also choose the Wasserstein metric as the metric tensor on the probability manifold.

where ∂_μ denotes $\frac{\partial}{\partial \theta_\mu}$. The logarithm of probability is the so-called likelihood function in statistics.

It can be rewritten as an expectation

$$g_{\mu\nu} = \mathbb{E}[\partial_\mu \ln p(\boldsymbol{x}, \boldsymbol{\theta}) \partial_\nu \ln p(\boldsymbol{x}, \boldsymbol{\theta})]. \tag{2}$$

This expression is equivalent to the second order derivatives of the likelihood function

$$g_{\mu\nu} = -\mathbb{E}[\partial_\mu \partial_\nu \ln p(\boldsymbol{x}, \boldsymbol{\theta})] \tag{3}$$

The metric tensor tells us how to do the inner product in this space. The inner product of two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ can be defined as

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{\mu,\nu} g_{\mu\nu} \boldsymbol{A}^\mu \boldsymbol{B}^\nu. \tag{4}$$

Where A^μ and B^ν are components of vectors A and B respectively.

The metric tensor also tells us how to calculate the length in this space. The length is the so-called statistical length in statistics or thermodynamic length in stochastic thermodynamics. It is defined as

$$\mathcal{L} = \int_0^\tau \mathrm{d}t \, \sqrt{\sum_{\mu,\nu} g_{\mu\nu} \frac{\mathrm{d}\theta^\mu}{\mathrm{d}t} \frac{\mathrm{d}\theta^\nu}{\mathrm{d}t}}. \tag{5}$$

3 Dual Structures

The most important and interesting structure of information geometry is its dual structure.

3.1 Dual Affine Connections

The tangent vectors of different points are not in the same space. They lies in their own tangent space respectively. We need to figure out what is the relation between different tangent spaces. That's why we need to define connections. Roughly speaking, tangent vector is the ”velocity” along a curve, and connections tells us the ”acceleration” along the curve. It's the directional derivatives “∇” in calculus. The connection also tells us how to do the parallel transport along an arbitrary curve.

In Riemannian geometry, we usually take the connection that is compatible with the metric tensor. This compatibility means the inner product of any two vectors starting from a point is invariant under the parallel transport along any curve. It means

$$\langle \boldsymbol{A}(0), \boldsymbol{B}(0) \rangle = \langle \boldsymbol{A}(t), \boldsymbol{B}(t) \rangle, \quad \forall t \geq 0. \tag{6}$$

Connection is a map from the tangent space of a point to the tangent space of its nearby points. So basically there are lots of way to define the map. In this sense, we find a family of connections in information geometry, namely α-connections ($\nabla^\alpha, \alpha \in \mathbb{R}$). The invariant property under parallel transport is not true for a general connection in information geometry. But the good thing is, the inner product can keep invariant under the parallel transport of a connection and its dual connection. It means

$$\langle \boldsymbol{A}(0), \boldsymbol{B}(0) \rangle = \langle \boldsymbol{A}(t), \boldsymbol{B}^*(t) \rangle, \quad \forall t \geq 0. \tag{7}$$

The parallel transport is taking on the same curve with different connections. This pair of connections are called dual connections information geometry. A pair of dual connections can be expressed as (∇, ∇^*). The dual of a dual connection is the original connection

$$(\nabla^*)^* = \nabla. \tag{8}$$

Furthermore, mathematicians found that the dual connection of ∇^α is $\nabla^{-\alpha}$

$$(\nabla^\alpha)^* = \nabla^{-\alpha}, \quad (\nabla^\alpha)^{**} = (\nabla^{-\alpha})^* = \nabla^\alpha. \tag{9}$$

3.2 Dual Flatness

In Riemannian geometry, geodesics are defined by connections

$$\nabla_{\dot{\boldsymbol{\theta}}}\dot{\boldsymbol{\theta}} = 0. \tag{10}$$

This means the velocity vector $\dot{\theta}$ is moving along the curve parallel to itself. In other words, ∇-geodesics generalize the notion of "straight Euclidean" lines.

Flatness is a property of connections. If a connection is flat, then the geodesic lines compatible with the connection are straight lines under some coordinate systems. This means the coordinates of any points on the geodesic curve connecting point P and Q can be written as

$$\boldsymbol{\theta}(t) = t\boldsymbol{\theta}_P + (1-t)\boldsymbol{\theta}_Q. \tag{11}$$

Where $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_Q$ are the coordinates of points P and Q respectively.

In the language of differential geometry, flat means the curvature vanishes under some proper coordinate systems. A flat space just looks like the Euclidean space. As a special case, Euclidean space is a flat space.

In information geometry, it turns out that if the manifold is flat with respect to a connection ∇^{α}, then it is also flat with respect to the dual connection $\nabla^{-\alpha}$. There are only two flat connections in the α-family. They are ∇^1 and ∇^{-1}.

3.3 Dual Coordinate Systems

Two coordinate systems $\{\theta^\mu\}$ and $\{\tilde{\theta}_\nu\}$ are said to be dual to each other if their coordinate basis $\hat{e}_\mu$ and $\hat{\tilde{e}}^{\tilde{\nu}}$ satisfy

$$\langle \hat{e}_\mu, \hat{\tilde{e}}^\nu \rangle = \delta^\mu_\nu. \tag{12}$$

The two basis are related by Jacobi matrix

$$\hat{e}_\mu = \frac{\partial \tilde{\theta}_\nu}{\partial \theta^\mu}\hat{\tilde{e}}^\nu, \quad \hat{\tilde{e}}^\nu = \frac{\partial \theta^\mu}{\partial \tilde{\theta}_\nu}\hat{e}_\mu. \tag{13}$$

When $\{\theta^\mu\}$ and $\{\tilde{\theta}_\mu\}$ are dual coordinate systems, there exist a pair of potential functions $\Theta(\boldsymbol{\theta})$ and $\tilde{\Theta}(\tilde{\boldsymbol{\theta}})$, such that

$$\theta^\mu = \tilde{\partial}^\mu \tilde{\Theta}(\tilde{\boldsymbol{\theta}}), \quad \tilde{\theta}_\mu = \partial_\mu \Theta(\boldsymbol{\theta}) \tag{14}$$

The elements of Fisher Information metric tensor $g_{\mu\nu}$ and its inverse $\tilde{g}^{\mu\nu}$ is given by the second order derivatives of (convex) potential functions respectively

$$g_{\mu\nu} = \partial_\mu \partial_\nu \Theta(\boldsymbol{\theta}), \quad \tilde{g}^{\mu\nu} = \tilde{\partial}^\mu \tilde{\partial}^\nu \tilde{\Theta}(\tilde{\boldsymbol{\theta}}). \tag{15}$$

Moreover,

$$\Theta(\boldsymbol{\theta}) + \tilde{\Theta}(\tilde{\boldsymbol{\theta}}) = \theta^\mu \tilde{\theta}_\mu. \tag{16}$$

This equation means $\{\theta^\mu\}$ and $\{\tilde{\theta}_\mu\}$ are Legendre dual. Actually, the potential functions need not to be strictly convex. In that case, the Legendre dual is reduced to Legendre-Fenchel dual.

Conversely, when a potential function $\Theta(\boldsymbol{\theta})$ exists such that $g_{\mu\nu} = \partial_\mu \partial_\nu \Theta(\boldsymbol{\theta})$, then equation 14 gives the dual coordinate systems.

4 Divergence

We can define divergence function between point P and point Q as

$$\mathscr{D}[P\|Q] = \Theta(\boldsymbol{\theta}_P) + \tilde{\Theta}(\tilde{\boldsymbol{\theta}}_Q) - \sum_\mu \theta^\mu_P \tilde{\theta}_{Q,\mu}. \tag{17}$$

This type of divergence is called Fenchel-Young divergence. It is induced by a pair of dual convex functions. The KL-divergence is a special case of Fenchel-Young divergence, where the $\tilde{\Theta}$ becomes the negative Shannon entropy.

When $P = Q$, the divergence function goes back to equation 16, otherwise it is always greater than zero.

Because of the duality between Θ and $\tilde{\Theta}$, the first order derivatives of $\mathscr{D}[P\|Q]$ vanish. The second order derivative gives the Fisher Information metric at point P

$$\frac{\partial}{\partial\theta_P^\mu}\frac{\partial}{\partial\theta_P^\nu}\mathscr{D}[P\|Q] = g_{\mu\nu}(P). \tag{18}$$

In fact, these properties imply that information geometric structure can also be built from a (Bregman) divergence function. If we define the divergence function at first, then the metric tensor is given by its second order derivatives and the dual connections is given by its asymmetric multi-derivatives. Therefore, information geometry can also be regarded as a geometric structure of divergence function. In this case, the duality is given by the Legendre-Fenchel duality of the generation function of a Bregman function.

5 Generalized Pythagorean Theorem

Any flat structure has generalized Pythagorean theorem, so does information geometry (when $\alpha = \pm 1$). Given three points P, Q and R, if the geodesic connecting P and Q are orthogonal to the dual geodesic connecting Q and R, then the following Pythagorean theorem holds

$$\mathscr{D}[P\|R] = \mathscr{D}[P, Q] + \mathscr{D}[Q\|R]. \tag{19}$$

If the angle φ is greater than (less than) $90°$, the equal sign becomes $>$ ($<$).

In the flat space, geodesics can be written as

$$\gamma_{PQ} : t \mapsto \boldsymbol{\theta}_P + (\boldsymbol{\theta}_Q - \boldsymbol{\theta}_P)t \tag{20}$$

$$\gamma_{QR} : t \mapsto \tilde{\boldsymbol{\theta}}_Q + (\tilde{\boldsymbol{\theta}}_R - \tilde{\boldsymbol{\theta}}_Q)t \tag{21}$$

Therefore, the right hand side of equation 19 is

$$\begin{aligned}
&\mathscr{D}[P\|Q] + \mathscr{D}[Q\|R] \\
&= \Theta(\boldsymbol{\theta}_P) + \tilde{\Theta}(\boldsymbol{\theta}_Q) - \sum_\mu \theta_P^\mu \tilde{\theta}_{Q,\mu} + \Theta(\boldsymbol{\theta}_Q) + \tilde{\Theta}(\boldsymbol{\theta}_R) - \sum_\mu \theta_Q^\mu \tilde{\theta}_{R,\mu} && (22)\\
&= \mathscr{D}[P\|R] + \sum_\mu (\theta_Q^\mu \tilde{\theta}_{Q,\mu} - \theta_P^\mu \tilde{\theta}_{Q,\mu} - \theta_Q^\mu \tilde{\theta}_{R,\mu} + \theta_P^\mu \tilde{\theta}_{R,\mu}) && (23)\\
&= \mathscr{D}[P\|R] - \sum_\mu (\theta_Q^\mu - \theta_P^\mu)(\tilde{\theta}_{R,\mu} - \tilde{\theta}_{Q,\mu}) && (24)\\
&= \mathscr{D}[P\|R] + \|\dot{\gamma}_{PQ}\|\|\dot{\gamma}_{QR}\| \cos(\varphi) && (25)
\end{aligned}$$

Therefore the above statements on the equality and inequalities holds.

In a short summary, the structure of information geometry is $(\mathcal{M}, g, \nabla, \nabla^*)$. Every point on the manifold is a probability distribution. The metric tensor on the manifold is Fisher Information metric. The key point of information geometry is its dually flat structure. The dual structure is essentially defined by Legendre(-Fenchel) dual. The length is calculated by an integration on the metric tensor. The dissimilarity or "distence" is described by the divergence function.

6 Dynamical Properties on Probability Manifold

Dynamical properties can be obtained from the time evolution of probability distributions on the manifold. These dynamical properties are strongly related to the Boltzmann's approach of the second law.

6.1 Thermodynamic Length and Speed Limit

If we take parameters $\boldsymbol{\theta}$ to be time t, then the length of a evolution of a probability distribution is the thermodynamic length[1]

$$\mathcal{L}(\tau) = \int_0^\tau \mathrm{d}t \sqrt{\int \mathrm{d}\boldsymbol{x}\ p(\boldsymbol{x},t) \left[\frac{\partial \ln p(\boldsymbol{x},t)}{\partial t}\right]^2}. \tag{26}$$

This means the evolution speed of the probability distribution can be characterized by an intrinsic speed v_I, where

$$v_I = \sqrt{\int \mathrm{d}\boldsymbol{x}\ p(\boldsymbol{x},t) \left[\frac{\partial \ln p(\boldsymbol{x},t)}{\partial t}\right]^2} \tag{27}$$

is the square root of Fisher Information.

The evolution speed of any observable R can be defined as

$$v_R = \frac{|\mathrm{d}\langle R\rangle/\mathrm{d}t|}{\sqrt{\mathrm{Var}[R]}}. \tag{28}$$

By using Cramér-Rao inequality, we can get the Thermodynamic Speed Limit[4]

$$v_R \leq v_I. \tag{29}$$

It means the evolution speed of any observable R is upper bounded by the intrinsic speed v_I. This inequality holds both for steady states and non-stationary states. This deficiency comes from the loss of information.

6.2 Gradient Flow

Information monotonicity is a wild known property of the probability evolution controlled by master equation or Fokker-Planck equation. The monitonicity means the KL-divergence between the present distribution and the steady state distribution is always decreasing

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}_{\mathrm{KL}}[p(\boldsymbol{x},t)\|\pi(\boldsymbol{x})] \leq 0, \tag{30}$$

where $\pi(\boldsymbol{x})$ denotes the steady state distribution.

Furthermore, because the master equation and the Fokker-Planck equation are continuity equations for discrete and continuous systems respectively, the probability evolution controlled by them is the gradient flow on the probability manifold. For example, the Fokker-Planck equation can be rewritten as[2]

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t} = \mathbf{D}\left[\frac{\partial}{\partial p(\boldsymbol{x},t)}\mathscr{D}_{\mathrm{KL}}[p(\boldsymbol{x},t)\|\pi(\boldsymbol{x})]\right], \tag{31}$$

where $\mathbf{D}$ is the weighted Laplacian operator defined as

$$\mathbf{D}[*] = \nabla \cdot (\mu T p(\boldsymbol{x},t)\nabla[*]). \tag{32}$$

The gradient flow representation means the probability always evolves along the steepest direction to the steady state distribution on the probability manifold[5].

7 Pythagorean Theorem and Entropy Production Decomposition

Pythagorean theorem gives us a method to decompose the divergence into two parts. In thermodynamics, this theorem tells us how to decompose the entropy production (or entropy production rate). These methods of decomposition are non-equilibrium correspondences of the Clausius representation of the second law.

[5] In machine learning, this is also called the natural gradient.

7.1 Entropy Production Rate as a Projection

When we need to consider the entropy production, we cannot just consider the probability distribution on the sample space. Instead, we need to consider path probability distributions, i.e. trajectory probability distributions. That is because the entropy production is defined on the forward and backward trajectories.

The dimension of continuous time trajectories space is uncountable infinity. It's geometry remains unclear. However, we can consider the joint probability during a short time, i.e. $P(\boldsymbol{x}_{\tau+\mathrm{d}t}, \boldsymbol{x}_\tau)$.[6] The sample space of the joint probability is the Cartesian product of the original space[7]

$$\Omega'(\tau+\mathrm{d}t, \tau) = \Omega(\tau+\mathrm{d}t) \times \Omega(\tau) = \Omega^2. \tag{33}$$

Let's take the overdamped Langevin system as an example. The transition rate is given by

$$\mathbb{T}(\boldsymbol{x}_{\tau+\mathrm{d}t} \mid \boldsymbol{x}_\tau) = \frac{1}{(4\pi\mu T\mathrm{d}t)^{\frac{3}{2}}} \exp\left[-\frac{\|\boldsymbol{x}_{\tau+\mathrm{d}t} - \boldsymbol{x}_\tau - \mu\boldsymbol{F}_\tau(\boldsymbol{x}_\tau)\mathrm{d}t\|^2}{4\mu T\mathrm{d}t}\right]. \tag{34}$$

The short-time forward path probability is

$$P(\boldsymbol{x}_{\tau+\mathrm{d}t}, \boldsymbol{x}_\tau) = \mathbb{T}(\boldsymbol{x}_{\tau+\mathrm{d}t} \mid \boldsymbol{x}_\tau) p(\boldsymbol{x}_\tau). \tag{35}$$

The short-time backward path probability is given in a similar way

$$P^\dagger(\boldsymbol{x}_{\tau+\mathrm{d}t}, \boldsymbol{x}_\tau) = \mathbb{T}(\boldsymbol{x}_\tau \mid \boldsymbol{x}_{\tau+\mathrm{d}t}) p(\boldsymbol{x}_{\tau+\mathrm{d}t}). \tag{36}$$

We can define a backward manifold as

$$\mathcal{M}_\mathrm{B}(P) = \left\{P \,\middle|\, P(\boldsymbol{x}_{\tau+\mathrm{d}t}, \boldsymbol{x}_\tau) = \mathbb{T}(\boldsymbol{x}_\tau \mid \boldsymbol{x}_{\tau+\mathrm{d}t}) \int \mathrm{d}\boldsymbol{x}_\tau\ P(\boldsymbol{x}_{\tau+\mathrm{d}t}, \boldsymbol{x}_\tau)\right\}. \tag{37}$$

It's easy to check that all the short-time backward path probabilities lie on the backward manifold. The entropy production is the KL-divergence between the forward path probability and the backward path probability. In the sense of projection theorem, it can be regarded as a projection from the forward path probability onto the backward manifold. For any $Q \in \mathcal{M}_\mathrm{B}(P)$, the entropy production rate $sigma_\tau$ is given by the minimum projection[3]

$$\sigma_\tau = \lim_{\mathrm{d}t\to 0} \inf_{Q\in\mathcal{M}_\mathrm{B}(P)} \frac{1}{\mathrm{d}t}\mathscr{D}[P\|Q]. \tag{38}$$

7.2 Entropy Production Rate Decomposition

Because the entropy production is a KL-divergence in this space. If we choose appropriate intermediate points to build the "right trangle", then the entropy production can be decomposed into two parts, which are the so-called housekeeping part and excess park. These methods of decomposition are related to the Hatano-Sasa decomposition and Maes-Netočnỳ decomposition[2, 3].

$$\sigma = \sigma_\mathrm{hk}^\mathrm{HS} + \sigma_\mathrm{ex}^\mathrm{HS}, \tag{39}$$

$$\sigma = \sigma_\mathrm{hk}^\mathrm{MN} + \sigma_\mathrm{ex}^\mathrm{MN}. \tag{40}$$

In Hatano-Sasa decomposition, the intermediate point is the instantaneous steady state distribution. Its housekeeping part represents the effect of non-conservative force, and its excess part represents the effect of conservative force. In Maes-Netočnỳ decomposition, the intermediate point is give by the Wasserstein distance, which represents the minimum entropy consumption to change the probability distribution.

These formalisms have something in common. The excess part is zero only if the system is on equilibrium states. The housekeeping part is non-zero if the system is on non-stationary states.

[6] Here we use P to represent joint probability distributions, and p to represent the marginal probability distributions.

[7] It is a sort of discretization.

8 Exponential Family and Maximum Entropy Principle

There are two special families of probability distributions. One is the exponential family, another one is the mixture family. Any probability simplex belongs to both of the two families. And these two families give the dual structure on any probability simplex. We will see that the exponential family plays an important role in Maximum Entropy Principle. It implies that the information geometry is strongly related to the Gibbs' approach of the second law in thermodynamics.

8.1 Geometric Structures of Exponential Family

If a probability distribution can be written in the following form

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left[\sum_{\mu} \theta^{\mu} h_{\mu}(\boldsymbol{x}) + k(\boldsymbol{x}) - \varphi(\boldsymbol{\theta})\right], \tag{41}$$

then it belongs to the exponential family. The functions $h_\mu(\boldsymbol{x})$ are linearly independent, the function $k(\boldsymbol{x})$ is an arbitrary function and the function $\varphi(\boldsymbol{\theta})$ is the Log-normalizer. The function $k(\boldsymbol{x})$ allows us to change the measure from $\mathrm{d}\boldsymbol{x}$ to $\mathrm{d}\mu(\boldsymbol{x}) = \exp\{k(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x}$. It means the exponential family defines a family upon an arbitrary measure. The function $\varphi(\boldsymbol{\theta})$ is a convex function and it is the generating function of the exponential family.

$\varphi(\boldsymbol{\theta})$ can be rewritten as

$$\varphi(\boldsymbol{\theta}) = \ln \int \mathrm{d}\mu(\boldsymbol{x}) \ \exp\left\{\sum_{\mu} \theta^{\mu} h_{\mu}(\boldsymbol{x})\right\}. \tag{42}$$

The corresponding dual coordinate system of the exponential family is

$$\eta_\mu \equiv \tilde{\theta}_\mu = \frac{\partial}{\partial \theta^\mu} \varphi(\boldsymbol{\theta}) = \int \mathrm{d}\mu(\boldsymbol{x}) \ h_\mu(\boldsymbol{x}) p(\boldsymbol{x}, \theta) = \mathbb{E}[h_\mu(\boldsymbol{x})]. \tag{43}$$

It is called expectation coordinate system[8].

8.2 Dual Coordinate Systems in Maximum Entropy Principle

Maximum Entropy Principle maximize the entropy function under some constrains. We can define a function $\mathcal{C}$ as follows

$$\begin{aligned} \mathcal{C} = & - \int \mathrm{d}\mu(\boldsymbol{x}) \ p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \\ & - \sum_{\mu} \theta^{\mu} \left(\int \mathrm{d}\mu(\boldsymbol{x}) \ h_\mu(\boldsymbol{x}) - \mathbb{E}[h_\mu(\boldsymbol{x})] \right) \\ & - \alpha \left(\int \mathrm{d}\mu(\boldsymbol{x}) \ p(\boldsymbol{x}) - 1 \right). \end{aligned} \tag{44}$$

Maximum Entropy Principle asks us to maximize the function $\mathcal{C}$. It turns out that the probability distributions that satisfy the Maximum Entropy principle belong to the exponential family. The coordinates and the dual coordinates are the Lagrange multipliers $\{\theta^\mu\}$ and expectation constrains $\{\mathbb{E}[h_\mu(\boldsymbol{x})]\}$ respectively. They are Legendre dual of each other[5].

Maximum Caliber Principle is slightly different[9] from Maximum Entropy Principle. The probability distributions on states change to the probability distributions on trajectories. It is difficult to deal with (uncountable) infinite dimensional space where the path probability lies in, so the relation between information geometry and Maximum Caliber Principle still remains partially unclear[10].

[8] Here φ plays the role of Θ, $\boldsymbol{\eta}$ plays the role of $\tilde{\boldsymbol{\theta}}$

[9] But it's a big difference in mathematics, because dealing with infinite dimensional spaces requires functional analysis.

[10] at least for me.

The relations between information geometry and Maximum Entropy Principle shows that the Gibbs' approach of the second law also has a geometric interpretation. It implies that information geometry may shed light on the relation between Maximum Caliber Principle and other thermodynamic representations. This coincidence also implies that we can use the dual structure to infer the properties of a thermodynamic system from its dual system[6].

References

[1] Gavin E. Crooks. Measuring thermodynamic length. *Physical review letters*, 99(10):100602, Sep 2007.

[2] Andreas Dechant, Shin-Ichi Sasa, and Sosuke Ito. Geometric decomposition of entropy production into excess, housekeeping, and coupling parts. *Phys Rev E*, 106(2-1):024125, August 2022.

[3] Sosuke Ito. Geometric thermodynamics for the fokker–planck equation: stochastic thermodynamic links between information geometry and optimal transport. *Information Geometry*, Mar 2023.

[4] Sosuke Ito and Andreas Dechant. Stochastic time-evolution, information geometry and the cramer-rao bound. *Physical Review X*, 10(2):021056, Jun 2020.

[5] Frank Nielsen. The many faces of information geometry. *Notices of the American Mathematical Society. American Mathematical Society*, 69(01):1, Jan 2022.

[6] Naruo Ohga and Sosuke Ito. Inferring nonequilibrium thermodynamics from tilted equilibrium using information-geometric legendre transform. Dec 2022.