博 士 学 位 論 文

The generalized second law based on information geometry and its extension to the q-geometry

（情報幾何学に基づく一般化された第二法則の再考とq-幾何学への拡張）

令和3年 3月

茨城大学大学院 理工学研究科
複雑系システム科学専攻

中村 文美

Abstract

Thermodynamics and statistical mechanics, as the name implies, are closely related to statistics and information theory. In the past, Maxwell's thought experiment of his demon and Landauer's principle showed the direct relationship between information manipulation and heat flow. With these foundations, energy and information were considered to be equal.

While many studies including thermodynamics of information are still being conducted, we have attempted to apply information geometry to thermodynamics. Information geometry provides a geometric understanding of optimization problems for the space of probabilities. It is recently attracting a lot of attention in the field of information theory such as artificial intelligence, machine learning, and neuroscience. Due to its potential to provide geometric understanding of statistical mechanics, it has increasingly become a focus of study.

In this paper, we apply information geometry to the generalized second law of the maximum work formulation, which is the most fundamental law of thermodynamics. In this application, we introduce a new concept of dimensional divergence scaled by a parameter such as temperature. We show that the geometric structure of the generalized second law is obtained by considering it as an optimization problem for this scaled divergence. Our main result is that scaling the divergence by physical dimensional quantities significantly changes the geometric structure of the space of probabilities. Scaled divergence breaks the invariance of Fisher's information matrix given for the original dimensionless divergence. It has, however, a dually flat structure that fits the maximum work formulation. We also extend this application to the systems obeying power laws in correlations. In this case, the equilibrium distribution of the system becomes a power distribution such as a Tsallis distribution. The generalized second law of the maximum q-work formulation for the Tsallis statistics

is also obtained by considering it as an optimization problem for a dimensional scaled q-divergence.

The second law of thermodynamics is originally restricted to transitions between equilibrium states. The generalized second law of the maximum work formulation is valid for transitions between nonequilibrium states. The maximum work extractable from a nonequilibrium initial state is described by the minimum value of Kullback-Leibler (KL) divergence between the nonequilibrium initial distribution of the system and the corresponding canonical distribution, multiplied by temperature. Thus, the maximum work formulation is reduced to an optimization problem of the scaled KL divergences. KL divergence plays a central role in information geometry because it satisfies invariance and is a kind of Bregman divergence. To solve the optimization problem in information geometry, Amari's generalized Pythagorean theorem is provided for a Bregman divergence.

It is well known that Amari's generalized Pythagorean theorem for three KL divergences gives the maximum entropy principle. For the maximum work formulation, a temperature parameter provides an energy dimension if we set Boltzmann's constant to 1. In applying information geometry to the generalized second law, we need the divergence scaled by the temperature. We obtain the maximum work formulation as a consequence of the generalized Pythagorean theorem. The scaled divergence has dimension, while typical divergences in information geometry are dimensionless. The metric of the scaled KL divergence breaks the invariance of Fisher's information matrix of the original KL divergence. In applying the generalized Pythagorean theorem, the temperature line and the isentropic surface are orthogonal for dimensional scaled KL divergences, while the inverse temperature line and the isoenergetic surface are orthogonal for the original KL divergences. Temperature and entropy are related to each other by the Legendre transformation, where the Bregman divergence for the free energy as a convex function leads to this dimensional divergence in the space of canonical distributions.

We also discuss the dual structure of the scaled KL divergence. An important property of information geometry is that divergence is not a mathematical distance in general. The mathematical distance of two points is symmetric for the exchange of the two points. The divergence of two distributions is generally not symmetric for the exchange of the two distributions. The divergence exchanged of the two distributions is the Legendre transformed divergence. This dual structure leads to the other important scaled divergence. We propose a divergence scaled by the inverse of the energy that is a dual parameter of the temperature. We expect this scaled divergence to be useful in considerations of heat engines in contact with finite high/low-temperature heat baths. These three divergences: the original dimensionless KL divergence, the KL divergence scaled by temperature, and the KL divergence scaled by the inverse of energy, have a symmetric structure based on the dual structure of information geometry. A list of these divergences is provided in this paper.

Finally, we extend our arguments based on statistical mechanics of exponential distributions, such as canonical distribution, to the statistical mechanics of power distributions. In Tsallis's q-statistics, we can extend our arguments using Amari-Ohara's normalized q-geometry. Dimensional normalized q-divergences give us the maximum q-work formulation and its geometric structures.

Acknowledgements

I am deeply grateful to my supervisor Professor Hiroshi H. Hasegawa for his guidance, encouragement, and support. He always listened to me and gave me suggestions whenever I had problems. In addition, he was always very caring about my health. I also thank Naomi Hasegawa for her advice on English writing.

I would like to express my gratitude to my collaborator, Dr. Dean J. Driebe at the Center for Complex Quantum Systems, University of Texas. I got a lot of insights and a deep understanding through discussions with him. I would also like to express my gratitude to Professor Shun-ichi Amari at RIKEN Brain Science Institute for his useful suggestions.

In my doctoral degree examination process, I would like to thank Professor Takehisa Hasegawa, Professor Hiraku Nishimori, Professor Tatsuaki Wada, and Professor Shinya Watanabe for their encouragement and comments. I also would like to thank all professors of the mathematics and informatics and mathematics courses, Hasegawa laboratory's members, and the administrative staff of Ibaraki University for their great support.

Finally, I am deeply thankful to my family. They have been a great support throughout my undergraduate and graduate studies.

This work was supported by a Grant-in-Aid for JSPS Research Fellow (No. 20J13492) from the Japan Society for the Promotion of Science (JSPS).

Contents

1 Introduction

1.1 Background

The chief concern of thermodynamics is the interconversion of heat and work. Whereas the first law of thermodynamics puts these two forms of energy on an equal footing, the second law expresses a dissymmetry between them. Work can be turned into heat without restriction, but heat can only be harnessed to do work with limited efficiency [1, 2].

We have recognized for some time that information processing can play an important role in a deeper understanding of the second law, an idea further exemplified through Maxwell's thought experiment [3–5]. By introducing a being whose sharpened faculties enabled it to perceive and manipulate individual molecules, Maxwell showed that appropriate operation based on information provides energy. Landauer [6], Bennett [7], and others [8–10], also pointed out that information processing itself may be accompanied by thermal effects, sometimes in subtle ways. For example, Landauer's principle dictates that erasing one bit requires at least $k\mathcal{T}\log 2$ of energy (k: Boltzmann's constant, $\mathcal{T}$: temperature). These considerations have led researchers to recognize information as a resource on par with heat and work in formulating the thermodynamics of information [11–15].

Thermodynamics and information theory are closely related, even in light of statistical mechanics. Statistical mechanics applies both probability theory and mechanics at the microscopic level to the study of thermodynamic behavior of many-body systems. In these cases, the state of the system is represented by a probability distribution. The study of statistical mechanics was initiated by Boltzmann at the end of the nineteenth century. With his contribution, Maxwell and Gibbs later established the study of statistical mechanics for the equilibrium state.

Recently, there has been a gradual emergence of attempts to apply information geometry [16, 17], a branch of information theory, to thermodynamics. Information geometry is a differential geometric theory for the space of probability distributions. It has received a lot of attention in a variety of fields, such as machine learning, neuroscience, and optimal transport problem. It is written in terms of divergence such as the Kullback–Leibler (KL) divergence (also known as relative entropy in physics) in the Riemann geometry. Probability distributions are the points of a Riemannian manifold and the Fisher information metric provides the Riemannian metric. KL divergence is not a true metric for distance in mathematics as it is not symmetric when replaced by two probabilities. Even so, Amari et. al. showed the generalized Pythagorean theorem for a "right-angled triangle" of three probability densities and introduced "projection," an important concept in information geometry.

A unified understanding of nature based on geometry is the ultimate goal in Einstein's ideas. The general theory of relativity and gauge theory are representative examples. However, thermal and statistical mechanics, which deals with phenomena in a many-body system, has not been regarded as a target of this geometrization. Although Weinhold [18], Ruppeiner [19], Crooks [20, 21], and others [22–24] have discussed statistical mechanics from the differential geometrical view of point, it is still incompletely understood. They introduced the concept of "thermodynamic length", but they only took into account symmetrical distances. Most of their discussions can only be applied to near-equilibrium states or linear-response regimes. After the development of information geometry, which takes information as the object of geometry, the geometrical interpretation of thermal and statistical mechanics is now a possibility. A recent study of thermodynamics based on information geometry has been made by Ito et. al. [25, 26] in the context of stochastic thermodynamics [27–29]. They showed that in a Markov process, the positivity of entropy production is given by the optimized KL divergence between a discrete distribution of the system and the backward manifold. The optimization is done by the general-

ized Pythagorean theorem and the positivity is valid since KL divergence is always nonnegative. They also pointed out a relationship between information geometry and the Glansdorff-Prigogine criterion for stability [30].

In this thesis, we consider a Hamiltonian system. The total information, Gibbs-Shannon (GS) entropy for the whole system, is rigorously conserved in a Hamiltonian system. Change in entropy means loss of information. We can understand how entropy increases in a Markov process at the level of underlying Hamiltonian dynamics.

In a thermally isolated Hamiltonian system, we reconsider the maximum work formulation of the second law generalized to transitions between nonequilibrium states from the point of view of information geometry [31]. The original second law of thermodynamics is restricted to transitions between two equilibrium states. Independently of one another, Takara et. al. [32–35] and Esposito and Van den Broeck [36, 37] proposed the maximum work formulation generalized for a nonequilibrium initial state (nonequilibrium maximum work formulation). The nonequilibrium maximum work formulation is based on two fundamental properties: entropy conservation in Hamiltonian dynamics and the non-negativity of the KL divergence in information theory. It is a universal law and valid even for an integrable system. This fact means that it does not have anything to do with dynamical properties such as relaxation to an equilibrium state.

The generalized second law is stated in terms of the initial and final KL divergences between each nonequilibrium state and corresponding canonical states, as well as the Helmholtz free energy of the canonical states. The lower bound of work is given by the KL divergence between the initial state and corresponding canonical state. It is intuited as a change of "nonequilibriumness" since KL divergence gives the difference between the nonequilibrium probability density and the equilibrium canonical distribution. If the process is cyclic and the initial state is an equilibrium canonical state, KL divergence becomes zero and no more work can be derived from

the system. This is the original second law.

In this thesis, we discuss the geometrical structure behind this generalized second law based on information geometry, especially the generalized Pythagorean theorem [31] developed by Amari et. al. [16, 17]. It is well known that the generalized Pythagorean theorem for three KL divergences gives the maximum entropy principle. For the maximum work formulation, a temperature parameter provides an energy dimension if we set Boltzmann's constant to 1. In applying information geometry to the generalized second law, we need the divergence scaled by the temperature. We obtain the maximum work formulation as a consequence of the generalized Pythagorean theorem. The scaled divergence has dimension, while typical divergences in information geometry are dimensionless. The metric of the scaled KL divergence breaks the invariance of Fisher's information matrix of the original KL divergence. However, the scaled KL divergence has a dually flat structure that fits the maximum work formulation. In applying the generalized Pythagorean theorem, the temperature line and the isentropic surface are orthogonal for the dimensional scaled KL divergences, while the inverse temperature line and the isoenergetic surface are orthogonal for the original KL divergences. Temperature and entropy are related to each other by the Legendre transformation, where the Bregman divergence for free energy as a convex function leads to this dimensional divergence in the space of canonical distributions.

We also discuss the dual structure of the scaled KL divergence. An important property of information geometry is that divergence is not a mathematical distance in general. The mathematical distance of two points is symmetric for the exchange of the two points. The divergence of two distributions is generally not symmetric for the exchange of the two distributions. The divergence exchanged of the two distributions is the Legendre transformed divergence. This dual structure leads to the other important scaled divergence. We propose a divergence scaled by the inverse of the energy that is a dual parameter of the temperature. We expect

this scaled divergence to be useful in considerations of heat engines in contact with finite high/low-temperature heat baths. These three divergences: the original dimensionless KL divergence, the KL divergence scaled by temperature, and the KL divergence scaled by the inverse of energy, have a symmetric structure based on the dual structure of information geometry. A list of these divergences is provided in this paper.

Finally, we extend our argument for a system obeying power law such as Tsallis's statistics [38–41]. Systems obeying power law appear when there is a long-time correlation between particles and are found in many situations such as thermodynamics, network science, and economics. Tsallis used parameter q and introduced his q-exponential and q-logarithmic functions, q-entropy, q-density, etc. Furthermore, he pointed out the relationship between Tsallis statistics and information geometry [42, 43]. The q-divergence in Tsallis's statistics and the q-divergence in information geometry were related by $\alpha = 1 - 2q$. After that, Amari-Ohara introduced the Bregman divergence, which is defined as Tsallis's q-divergence divided by the conformal factor [44, 45]. Since Bregman divergence is a key concept in the structure of information geometry, the generalized Pythagorean theorem for this divergence also holds. We show that the maximum work formulation is extended to the maximum q-work formulation. Then we introduce Amari-Ohara's q-divergence scaled by temperature. By applying Amari-Ohara's generalized Pythagorean theorem for scaled Amari-Ohara's q-divergence, we obtain the maximum q-work formulation and its geometrical structure.

1.2 Outlines

In the next section, we start by recalling the maximum work formulation of the generalized second law. First, in subsection 2.1, we briefly review the normal second law for the transition between two equilibrium states. This thermodynamic

formulation of the second law has the advantage of being stated just in terms of energetic quantities: the work and the change in the Helmholtz free energy. It says that for a system in contact with a heat reservoir the work that can be extracted from the system is at most the decrease in its Helmholtz free energy. In subsection 2.2, we give a derivation of the generalized second law for a thermally isolated system and a system in contact with a heat reservoir. To consider transitions between nonequilibrium states we go to the finer level of description of the system provided by statistical mechanics. The nonequilibrium state is described by a distribution function in classical mechanics or a density matrix in quantum mechanics. In the last subsection, we consider the generalized second law for a system in contact with a heat reservoir.

In Sections 3, we reconsider the generalized second law based on information geometry. In the first subsection, we consider the case of a cyclic operation. Since there is no change in the Helmholtz free energy for a cyclic operation, the work is written only by the KL divergences scaled by temperature. "Information distance" measured by the KL divergence is dimensionless. We need energy dimensional "distance" to measure extractable work from a nonequilibrium state in the generalized second law. This energy dimensional "distance," hereafter referred to as "thermodynamic distance," is the KL divergence scaled by a temperature.

The work is bounded from below by the scaled KL divergence between a nonequilibrium initial state and a canonical state. The canonical state is parameterized by temperature. The maximum work is determined by minimizing the "thermodynamic distance." Amari's generalized Pythagorean theorem for the scaled KL divergences gives orthogonality between the one-parameter line of canonical states and the isentropic surface of state space. The maximum work is given by the canonical state that intersects the surface, with the temperature determined by the isentropic condition.

In subsection 3.2, we briefly explain the basic concepts of information geometry such as Bregman divergence and the proof of Amari's generalized Pythagorean

theorem. Bregman divergence is an important type of divergence which leads to the dually flat structure. We show that the scaled KL divergence is a new kind of Bregman divergence derived from using the free energy as a convex function. In subsection 3.3, we discuss the scaling of the KL divergence between more general exponential distributions. We find that the temperature scaling that appears in the generalized second law was a natural scaling for the canonical distribution. We also find that the scaling with the inverse of the energy appears naturally as well. These three divergences have a symmetrical relationship.

We extend our arguments to the case of a general non-cyclic operation in subsection 3.4. We also illustrate a concrete simple example of Amari's generalized Pythagorean theorem for the scaled KL divergences using the adiabatic expansion of the ideal gas in subsection 3.5. Finally, in subsection 3.6, we apply the geometrical interpretation of the generalized maximum work formulation to a simple two-level quantum system. The optimal cyclic operation to extract work from a nonequilibrium state is determined by minimalizing the scaled KL divergence between the final state and the final canonical state. The geometrical interpretation of the generalized maximum work formulation gives us a systematic method to figure out a protocol to realize the optimal operation.

In Appendix A, a geometrical interpretation of thermodynamics for Tsallis statistics is presented. The discussion is completely parallel to the generalized second law based on the KL divergence, although all functions have been extended by parameter q. Other details of the calculations that were not mentioned in the main text are also supplemented in the appendices.

This thesis relies on references [31, 35]. Several statements, equations, and figures are mentioned here without reference to these references.

2 The generalized second law

2.1 The second law of thermodynamics

We start with the maximum work formulation of the second law for transitions between equilibrium states. Consider a system in thermal contact with a heat reservoir at temperature $\mathcal{T}$ and able to perform work through the change of external parameters. This is the ordinary situation considered in elementary thermodynamics. The changing parameter performing work is, for example, a movable wall of a container [46] or a magnetic field interacting with the system of interest [47].

The first law of thermodynamics for our system of interest undergoing a process is

$$\Delta E = W + Q, \tag{2.1}$$

where $\Delta E \equiv E_T - E_0$ is the change in internal energy of the system from its initial state at time $t = 0$ to its final state at time $t = T$, W is the work done on the system and Q is the heat added to the system from the reservoir. Our interest is in work extraction from the system so we rewrite the first law as

$$W = \Delta E - Q. \tag{2.2}$$

Since a positive value for W means that energy is added to the system, extraction of energy in the form of work occurs when W is negative.

A convenient thermodynamic potential to use for a system in contact with a heat reservoir is the Helmholtz free energy, which is given by

$$F = E - \mathcal{T}S, \tag{2.3}$$

where S is the thermodynamic entropy of the system and $\mathcal{T}$ is the absolute tem-

perature. The change in the free energy during the process we are considering is

$$\Delta F = \Delta E - \mathcal{T}\Delta S. \tag{2.4}$$

Using this in our expression Eq. (2.2) for the work gives

$$W = \Delta F + \mathcal{T}\Delta S - Q. \tag{2.5}$$

Now, we bring in the second law of thermodynamics in the form of the Clausius inequality

$$\mathcal{T}\Delta S \geq Q, \tag{2.6}$$

which, when used in Eq. (2.5), gives us the work inequality

$$W \geq \Delta F. \tag{2.7}$$

This is the maximum work formulation of the second law. It tells us that the most work that can be extracted from the system in a transition between two equilibrium states is the decrease in the free energy of the system. Most work is obtained in a dissipationless process.

For the case of a cyclic transition, when the system returns to its initial configuration so that the free energy returns to its initial value, the work inequality becomes

$$W \geq 0. \tag{2.8}$$

This tells us that we can't get work out of a single heat reservoir by a process that returns the system to its original configuration, i.e., thermodynamical considerations prohibit the construction of a perpetual motion machine (of the second kind).

Perhaps the simplest application of the maximum work formulation of the second law is to an ideal gas, in contact with a heat reservoir at temperature $\mathcal{T}$, in a container

with a movable wall, such as a piston, that performs work. This is a standard example considered in all textbooks but we briefly recall it here for reference. Work is done by the system when the container expands and the most work is done in a dissipationless quasi-static expansion. The work, in this case, is given by

$$W_{qs} = -\int_{V_0}^{V_T} p dV = -Nk\mathcal{T}\int_{V_0}^{V_T}\frac{dV}{V} = -Nk\mathcal{T}\log\frac{V_T}{V_0}, \tag{2.9}$$

where V_t is the volume at time t, p is the pressure, N is the number of particles and k is the Boltzmann's constant. We also used the equation of state for the ideal gas: $pV = Nk\mathcal{T}$. Since the internal energy of the ideal gas doesn't change when its temperature remains the same, the work done is equal to the heat received from the bath. The heat flow is responsible for the entropy increase of the gas given by

$$\Delta S = \frac{Q}{\mathcal{T}} = -\frac{W_{qs}}{\mathcal{T}}. \tag{2.10}$$

So, the change in the Helmholtz free energy of the gas is precisely the work done by the gas found in Eq. (2.9) above,

$$\Delta F = \Delta E - \mathcal{T}\Delta S = W_{qs}. \tag{2.11}$$

Since the expansion is dissipationless the work inequality (2.7) is saturated.

2.2 Generalized second law for a thermally isolated system

For the generalization of the second law, we need to go beyond the purely thermodynamical description in the previous subsection to a statistical mechanics description. We first consider a thermally isolated Hamiltonian system and focus on the period of the operation starting at $t = 0$ and finishing at $t = T$. The state is described using the probability density, $\rho(x, t)$ (or ρ_t in abbreviated form), at time t and

phase-space point x, which is either specified or is obtained by dynamical evolution from a previously specified state. The dynamics of the system is governed by its Hamiltonian, $H(x, \kappa(t))$ (abbreviated as H_t), that has an explicit time dependence due to the external parameters, $\kappa(t)$, under our control. Hereafter we abbreviate a function of time $A(t)$ as A_t.

The time evolution of the state is written using the time evolution operator U as

$$\rho_T = U^T \rho_0. \tag{2.12}$$

where U^T is defined as

$$U^T = \hat{T} \exp\left(-\mathrm{i} \int_0^T L(\kappa(t)) dt\right), \tag{2.13}$$

where $\hat{T}$ is the time-ordered product and L is the Liouvillian for the Hamiltonian.

We employ the bracket notation to describe the expectation value of an observable $A(x)$ as,

$$\langle A|\rho\rangle = \int_\Gamma A^*(x)\rho(x)dx, \tag{2.14}$$

where Γ is the phase space of the Hamiltonian system and A^* is the (transposed) complex conjugate of A. We employ the bracket notation for a quantum system as $\langle A|\rho\rangle = \mathrm{Tr}[A\rho]$. This quantum version will be used later when dealing with concrete examples of a simple two-level quantum system.

Thus the expectation value of the Hamiltonian of the system at time t is given by $\langle H_t|\rho_t\rangle$. The work done on this thermally isolated system is given just by the change in the internal energy of the system, i.e.,

$$W = E_T - E_0 \tag{2.15}$$

where the internal energy of the system at time t is

$$E_t = \langle H_t | \rho_t \rangle. \tag{2.16}$$

The canonical distribution of the system with respect to the Hamiltonian and with a parameter α will play an important role:

$$\rho_{\mathrm{can}}(\alpha) = \frac{e^{-\alpha H}}{Z(\alpha)}. \tag{2.17}$$

Here, the partition function is

$$Z(\alpha) = \left\langle 1 | e^{-\alpha H} \right\rangle. \tag{2.18}$$

We are going to rewrite by using its relation to the Helmholtz free energy

$$F(\alpha) = -\alpha^{-1} \log Z(\alpha), \tag{2.19}$$

so that the canonical distribution may be expressed simply as the exponential function

$$\rho_{\mathrm{can}}(\alpha) = e^{\alpha\{F(\alpha) - H\}}. \tag{2.20}$$

We also will employ the Gibbs–Shannon (GS) entropy of the system as

$$S[\rho] = -\langle \log \rho | \rho \rangle \tag{2.21}$$

where we set the Boltzmann constant $k = 1$ to make the thermodynamic entropy compatible with the dimensionless entropy in information theory.

The cross entropy appearing in information theory is defined as

$$S_C[\rho_A, \rho_B] = -\langle \log \rho_B | \rho_A \rangle. \tag{2.22}$$

Specifically, the cross entropy between a probability density and a canonical distribution is given as,

$$S_C[\rho, \rho_{\text{can}}(\alpha)] = -\alpha F(\alpha) + \alpha E \tag{2.23}$$

where we used the property of a normalized distribution, $< 1|\rho >= 1$.

Of chief importance for the generalized second law is the KL divergence between two states, ρ_A and ρ_B, given by

$$D_{KL}[\rho_A \parallel \rho_B] \equiv \left\langle \log \frac{\rho_A}{\rho_B} \middle| \rho_A \right\rangle = \langle \log \rho_A | \rho_A \rangle - \langle \log \rho_B | \rho_A \rangle \tag{2.24}$$

$$= -S[\rho_A] + S_C[\rho_A, \rho_B]. \tag{2.25}$$

The KL divergence measures the "distance" between the two states. Its most important property in the derivation of the generalized second law will be its non-negativity:

$$D_{KL}[\rho_A \parallel \rho_B] \geq 0. \tag{2.26}$$

The KL divergence is zero only when $\rho_A = \rho_B$. See Appendix B for proofs of these facts. In Appendix B, we also show the non-negativity of KL divergence in the quantum case.

The basic quantitative relationship we need to obtain the generalized second law emerges when we take the KL divergence of an arbitrary state, ρ, with respect to a canonical state, $\rho_{\text{can}}(\alpha)$. This gives

$$D_{KL}[\rho \parallel \rho_{\text{can}}(\alpha)] = \langle \log \rho | \rho \rangle - \langle \log \rho_{\text{can}}(\alpha) | \rho \rangle \tag{2.27}$$

$$= -S[\rho] - \langle \alpha(F(\alpha) - H) | \rho \rangle. \tag{2.28}$$

Now, ρ is a normalized distribution and $F(\alpha)$ does not depend on the phase space

variables so it has nothing to integrate. Thus,

$$D_{KL}[\rho \parallel \rho_{\text{can}}(\alpha)] = -S[\rho] - \alpha F(\alpha) + \alpha E. \tag{2.29}$$

The KL divergence of an arbitrary state with a canonical state gives us a sum of terms involving the entropy of the arbitrary state, the free energy of the canonical state and the internal energy of the arbitrary state.

Using the above relation, the internal energy of the state at time t is written as

$$E_t = F_t(\alpha) + \alpha^{-1} S[\rho_t] + \alpha^{-1} D_{KL}[\rho_t \parallel \rho_{\text{can},t}(\alpha)]. \tag{2.30}$$

Since the work is given by the change in the internal energy as in Eq. (2.15), we obtain

$$W = \Delta F(\alpha) + \alpha^{-1} D_{KL}[\rho_T \parallel \rho_{\text{can},T}(\alpha)] - \alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\text{can},0}(\alpha)] \tag{2.31}$$

where we used the conservation of the GS entropy in a thermally isolated Hamiltonian system, i.e., $S[\rho_T] = S[\rho_0]$; and $\Delta F(\alpha) \equiv F_T(\alpha) - F_0(\alpha)$ is the change in the Helmholtz free energy of the system. Since there is no heat bath and we do not have an *a priori* temperature, we call the parameter corresponding to an inverse temperature as α.

The work equality, Eq. (2.31), is just a statement of energy accounting when we know the initial and final state of the system as well as the initial and final Hamiltonians. Let us suppose that a nonequilibrium initial state is just given and associated Hamiltonian with external parameters are under our control. Until we choose a schedule for the parameter control of the Hamiltonian the final state is not determined. Due to the non-negativity of the KL divergence term involving the unknown final state, Eq. (2.31) gives us under the condition of a known initial state

the following work inequality:

$$W \geq \Delta F(\alpha) - \alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\text{can},0}(\alpha)] \equiv \mathcal{W}_{LB}(\alpha) \tag{2.32}$$

where we denote the lower bound for the work as $\mathcal{W}_{LB}(\alpha)$.

First, we note that the inequality (2.32) is true for any value of α. Second, $\mathcal{W}_{LB}(\alpha)$ is a concave function of α^{-1}, which means that it has a global maximum as a function of α as discussed in Appendix C. Thus, there exists a best value for α; namely, the value for which $\mathcal{W}_{LB}(\alpha)$ is a maximum. This is so because the maximum of $\mathcal{W}_{LB}(\alpha)$ is a value for the work that may be achieved. A lesser value for $\mathcal{W}_{LB}(\alpha)$ (i.e., a more negative value when considering work extraction) while giving a true inequality would not be an achievable value for the work since this value would not satisfy the inequality of $W \geq \max.[\mathcal{W}_{LB}(\alpha)]$, which must be true.

To find the maximum of $\mathcal{W}_{LB}(\alpha)$ we make its dependence on α explicit by writing the KL divergence, as in Eq. (2.29), in terms of the entropy, free energy and internal energy as

$$\mathcal{W}_{LB}(\alpha) = F_T(\alpha) + \alpha^{-1} S[\rho_0] - E_0. \tag{2.33}$$

Now, taking the derivative with respect to α, and using the fact that the derivative of the free energy with respect to temperature is minus the entropy gives us

$$\frac{\partial \mathcal{W}_{LB}(\alpha)}{\partial \alpha} = \frac{1}{\alpha^2} S[\rho_{\text{can},T}(\alpha)] - \frac{1}{\alpha^2} S[\rho_0]. \tag{2.34}$$

The value of $\alpha = \tilde{\beta}$ where this equals zero we call the effective (inverse) temperature. This gives us the isentropic condition of

$$S[\rho_{\text{can},T}(\tilde{\beta})] = S[\rho_0]. \tag{2.35}$$

So, the effective temperature is determined as the temperature of the canonical

distribution with respect to the final Hamiltonian that has the same entropy as the initial nonequilibrium state. This condition is entirely reasonable as the dynamics preserves the entropy and the final distribution being canonical means that the maximum energy permissible has been extracted. The isentropic condition is also naturally derived from Amari's generalized Pythagorean theorem in information geometry, which will be discussed in the next section.

The generalized second law for a nonequilibrium initial state in a thermally isolated system is thus,

$$W \geq \Delta F(\tilde{\beta}) - \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\text{can},0}(\tilde{\beta})], \tag{2.36}$$

where the effective temperature, $\tilde{\beta}^{-1}$, is determined by the isentropic condition, Eq. (2.35).

If the Hamiltonian is changed back to its original form, as in a cyclic process, then $\Delta F = 0$ and the work on the system from the initial nonequilibrium state is just

$$W \geq -\tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\text{can},0}(\tilde{\beta})]. \tag{2.37}$$

In both cases the work extracted from the system is $-W$.

The KL divergence in the right-hand-side of Eq. (2.37) measures the "information distance" corresponding to the available informational resource. The effective temperature converts the available informational resource to the extractable work. Considering the Kelvin principle in the context of a thermally isolated system leads us to regard a canonical state as an equilibrium state for such a system as well [48].

At the end of this subsection, we comment on the physical depiction of work equality, Eq. (2.31) and consider the generalized second law to formulate a protocol for work extraction. To extract maximum work as in Eq. (2.36) or Eq. (2.37), it is best that the final state of the system is in equilibrium. The relative entropy of

a nonequilibrium initial state with the canonical state corresponding to the initial Hamiltonian of the system quantifies the maximum extractable work from such a nonequilibrium state as seen in Eq. (2.37). We know in general that dissipation is associated with nonequilibrium states so in order to get the maximum extractable work we should immediately stop the evolution of the nonequilibrium state and make it an equilibrium state for a new Hamiltonian. This new Hamiltonian should be reachable from the initial Hamiltonian through the time-dependent external parameters that are under our control. Once the state is stabilized and made an equilibrium state a process (in general quasi-static) is performed by changing the external parameters and putting the system into its final state while work is extracted. In general work extraction may occur in either or both of the stabilization and restoration processes. With such an operation, entropy is conserved as in Eq. (2.35) and maximum work is achieved as in Eq. (2.36) or Eq. (2.37).

2.3 Generalized second law for a system in contact with heat reservoir

Based on the generalized second law for thermally isolated systems, we derive the generalized second law for the case where the system is in contact with a heat reservoir with inverse temperature β. First, we consider the whole system as a thermally isolated system and divide it into two subsystems. The isothermal system in contact with a heat reservoir is then realized by taking the thermodynamic limit for one of the two subsystems.

Consider a situation where a finite system of coupled with a finite heat reservoir. We represent functions related to the system of interest and heat reservoir with superscript (S) and (R), respectively. They are interacting and exchanging energy. The state is described using the joint probability, $\rho_t(X) = \rho_t(x, y)$ at time t where the phase-space point is $X = (x, y)$, x is a point with respect to the system and y is

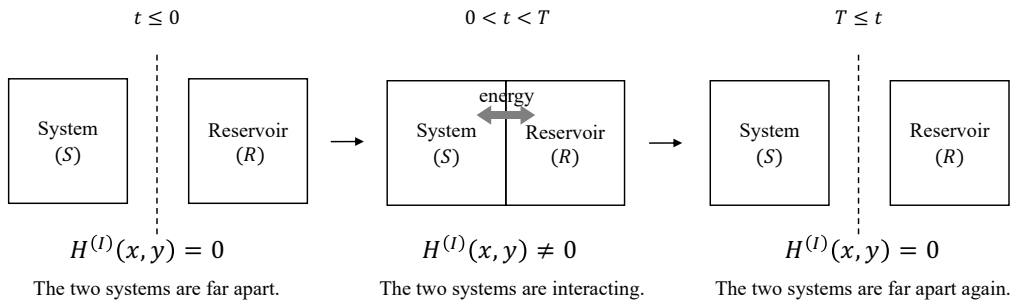

Figure 1: Image of a situation where a system and a heat reservoir interact each other. The interaction occurs during $0 < t < T$, and the rest of the time they are far apart.

a point with respect to the heat reservoir. The total Hamiltonian consists of three parts:

$$\begin{aligned} H_t(X) &= H(X, \kappa(t)) \\ &= H^{(S)}(x, \kappa_S(t)) + H^{(R)}(y, \kappa_R(t)) + H^{(I)}(x, y) \\ &= H_t^{(S)}(x) + H_t^{(R)}(y) + H^{(I)}(x, y) \end{aligned} \tag{2.38}$$

where $H_t^{(S)}(x)\left(= H^{(S)}(x, \kappa_S(t))\right)$ and $H_t^{(R)}(y)\left(= H^{(R)}(y, \kappa_R(t))\right)$ are the Hamiltonian for the system and reservoir respectively, and $H^{(I)}(x, y)$ is the interaction Hamiltonian between them. We assume that the interaction occurs during $0 < t < T$. So the interaction Hamiltonian is

$$H^{(I)}(x, y) \begin{cases} = 0 & (t \leq 0, \ \ T \leq t) \\ \neq 0 & (0 < t < T) \end{cases} \tag{2.39}$$

as described in Figure 1.

The two systems are far apart and independent at time $t \leq 0$, so the initial joint probability of the whole system is written as a product of two marginal distributions:

$$\rho_0(x, y) = \rho_0^{(S)}(x)\rho_0^{(R)}(y). \tag{2.40}$$

The initial and final canonical distributions are also written in a simple product since their Hamiltonians are additive,

$$\begin{aligned}
\rho_{\mathrm{can},0/T}(\alpha) &= \exp\left\{\alpha\left(F_{0/T}(\alpha) - H_{0/T}\right)\right\} \\
&= \exp\left\{\alpha\left(F_{0/T}^{(S)}(\alpha) + F_{0/T}^{(R)}(\alpha) - H_{0/T}^{(S)} - H_{0/T}^{(R)}\right)\right\} \\
&= \exp\left\{\alpha\left(F_{0/T}^{(S)}(\alpha) - H_{0/T}^{(S)}\right)\right\} \exp\left\{\alpha\left(F_{0/T}^{(R)}(\alpha) - H_{0/T}^{(R)}\right)\right\} \\
&= \rho_{\mathrm{can},0/T}^{(S)}(\alpha)\rho_{\mathrm{can},0/T}^{(R)}(\alpha),
\end{aligned} \tag{2.41}$$

where $0/T$ means 0 or T and we used the additivity of free energy $F_{0/T}(\alpha) = F_{0/T}^{(S)}(\alpha)+F_{0/T}^{(R)}(\alpha)$ which is also easily derived from the additivity of the Hamiltonian. The final state, however, are dependent. We only decompose it as

$$\rho_T(x, y) = \rho_T^{(S)}(x)\rho_T^{(R)}(y|x). \tag{2.42}$$

Since the whole system is assumed to be a thermally isolated system, we apply the generalized maximum work formulation (2.36) for it under the isentropic condition (2.35). Independency of Eqs. (2.40) and (2.41) gives us

$$\begin{aligned}
W &\geq \Delta F(\tilde{\beta}) - \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] \\
&= \Delta F^{(S)}(\tilde{\beta}) - \tilde{\beta}^{-1} D_{KL}[\rho_0^{(S)} \parallel \rho_{\mathrm{can},0}^{(S)}(\tilde{\beta})] + \Delta F^{(R)}(\tilde{\beta}) - \tilde{\beta}^{-1} D_{KL}[\rho_0^{(R)} \parallel \rho_{\mathrm{can},0}^{(R)}(\tilde{\beta})]
\end{aligned} \tag{2.43}$$

and the isentropic condition is

$$S[\rho^{(S)}_{\mathrm{can},T}(\tilde{\beta})] + S[\rho^{(R)}_{\mathrm{can},T}(\tilde{\beta})] = S[\rho^{(S)}_0] + S[\rho^{(R)}_0]. \tag{2.44}$$

We assume that the heat reservoir is initially canonical distribution with a temperature β, i.e.,

$$\rho^{(R)}_0 = \rho^{(R)}_{\mathrm{can}}(\beta). \tag{2.45}$$

Eq. (2.43) is transformed using Eqs. (2.44) and (2.45) as follows,

$$W \geq \Delta F^{(S)}(\beta) + \Delta F^{(R)}(\beta) - \beta^{-1} D_{KL}[\rho^{(S)}_0 \parallel \rho^{(S)}_{\mathrm{can},0}(\beta)] + \beta^{-1} D_{KL}[\rho_{\mathrm{can},T}(\tilde{\beta}) \parallel \rho_{\mathrm{can},T}(\beta)]. \tag{2.46}$$

We put

$$W^{(S)}_O = -\Delta F^{(S)}(\beta) + \beta^{-1} D_{KL}[\rho^{(S)}_0 \parallel \rho^{(S)}_{\mathrm{can},0}(\beta)] \tag{2.47}$$

which is a quantity that only depends on the system, and assume that no operations are performed on the reservoir (i.e. there is no change in the Hamiltonian of the reservoir) and that $\Delta F^{(R)}(\beta) = 0$. Then we have

$$W \geq -W^{(S)}_O + \beta^{-1} D_{KL}[\rho_{\mathrm{can},T}(\tilde{\beta}) \parallel \rho_{\mathrm{can},T}(\beta)]. \tag{2.48}$$

Finally, we consider the thermodynamic limit where the heat reservoir is infinitely large. The effective temperature becomes the temperature of reservoir, $\tilde{\beta} = \beta$ in this case. It is justified by the isentropic condition (2.44) and the fact that entropy is a quantitative quantity and that entropy is a monotonic function of temperature. Thus, the KL divergence in Eq. (2.48) vanishes and we obtain the maximum work formulation for the isothermal process as

$$W \geq \Delta F^{(S)}(\beta) - \beta^{-1} D_{KL}[\rho^{(S)}_0 \parallel \rho^{(S)}_{\mathrm{can},0}(\beta)]. \tag{2.49}$$

The difference from Eq. (2.36) in the thermally isolated system is that the effective temperature has been changed to the temperature of the heat reservoir.

The right-hand-side of Eq. (2.49) is the maximum work that can be extracted when we are given only the nonequilibrium initial state of the system. It is again generally obtained by instantaneous stabilization, which stops the relaxation of the system and eliminates the temperature difference with the heat reservoir, and by quasi-static operations, which conserve entropy. If the initial distribution of the system is a canonical distribution with the same temperature as the reservoir, $\rho_0^{(S)} = \rho_{\mathrm{can},0}^{(S)}(\beta)$, the extractable work is the difference in free energy of the system, reproducing Kelvin's principle.

In this section, we have shown the derivation of the generalized second law extended to transitions between nonequilibrium states. It was derived from two basic properties: the conservation of GS entropy in Hamiltonian dynamics and the non-negativity of KL divergence. It is important to note that the maximum work that can be extracted from the system is determined by the "nonequilibriumness" measured by the KL divergence. In the next section, we will consider the geometrical structure of this generalized second law on the basis of the geometry of the KL-divergence given by information geometry.

3 Reconsideration of the generalized second law based on information geometry

3.1 The case of cyclic processes

As shown in Eq. (2.31), the extractable work for a thermally isolated system is given as the difference between two KL divergences scaled by a temperature α^{-1} (and the change in the Helmholtz free energy). This work equality suggests that there is an information-geometric foundation of the generalized second law. In this subsection, we consider a cyclic operation because this case has a clearer geometric structure than the case of a non-cyclic operation, which we will consider in subsection 3.4.

Information geometry considers the geometric structure in parameter space of families of probability distributions [16, 17]. A parameter (vector) θ specifies the distribution $\rho(\theta)$ and the notion of "distance" between two distributions is provided by a divergence function $D[\rho(\theta_A) \parallel \rho(\theta_B)]$, which satisfies the following three conditions:

1. Non-negativity: $D[\rho(\theta_A) \parallel \rho(\theta_B)] \geq 0$.
2. Uniqueness: $D[\rho(\theta_A) \parallel \rho(\theta_B)] = 0$, if and only if $\theta_A = \theta_B$.
3. Positive-definiteness of the metric tensor g_{ij}:

$$D[\rho(\theta) \parallel \rho(\theta + d\theta)] = \sum (g_{ij}/2) d\theta_i d\theta_j .$$

The geometric structure is determined by the metric tensor appearing in the above local divergence. The KL divergence satisfies the above conditions and its metric

tensor is the well-known Fisher information matrix,

$$g_{ij} = \langle \frac{\partial \log \rho(\theta)}{\partial \theta_i} \frac{\partial \log \rho(\theta)}{\partial \theta_j} | \rho(\theta) \rangle. \tag{3.1}$$

Now we define the key concept, "thermodynamic distance." The "thermodynamic distance" is the KL divergence between an arbitrary probability distribution and a canonical distribution scaled by the temperature of the canonical distribution, $\alpha^{-1} D_{KL}[\rho \parallel \rho_{\mathrm{can}}(\alpha)]$. A divergence scaled by a positive parameter satisfies the above three conditions so that the scaled divergence is also a divergence. This scaled KL divergence has an energy dimension and measures the extractable work in the generalized second law as a "thermodynamic distance." The metric tensor of the scaled KL divergence is the Fisher information matrix scaled by the temperature. The different metric tensor gives us a completely different geometric structure as will be shown in the next subsection.

The maximum extractable work for a cyclic process is determined by the following inequality obtained from Eq. (2.32),

$$W \geq -\alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] \tag{3.2}$$

where we used $\Delta F(\alpha) = 0$ and the non-negativity of the KL divergence involving the final distribution. In order to find the lower bound of W in Eq. (3.2) (i.e., the upper bound for the extractable work), we have to find the value of α that makes the right-hand-side of Eq. (3.2) maximum. In the previous section we derived the condition of the effective temperature by differentiating the right-hand-side of Eq. (3.2). Now we use the information-geometric structure to derive the same condition.

The minimum (or shortest) distance from a point to a plane is obtained by the Pythagorean theorem in elementary geometry. Similarly, the minimum divergence from a probability distribution to a curved surface is obtained by the generalized

Pythagorean theorem in information geometry. Hereafter, for the sake of brevity, we will write it as GPT. The GPT is based on three divergences. Suppose that a point P is a probability distribution and $\mathcal{S}$ is a curved surface which does not include point P in the space of probability distribution. When there exists a point Q within $\mathcal{S}$ such that the geodesic line from P to Q is orthogonal to $\mathcal{S}$ as illustrated in Figure 2, the GPT holds as

$$D[P \parallel R] = D[P \parallel Q] + D[Q \parallel R] \quad \text{for } {}^{\forall}R \in \mathcal{S}. \tag{3.3}$$

From the non-negativity of the divergence, we obtain

$$D[P \parallel R] \geq D[P \parallel Q] \quad \text{for } {}^{\forall}R \in \mathcal{S} \tag{3.4}$$

which means that the divergence between P and Q is minimum [16, 17].

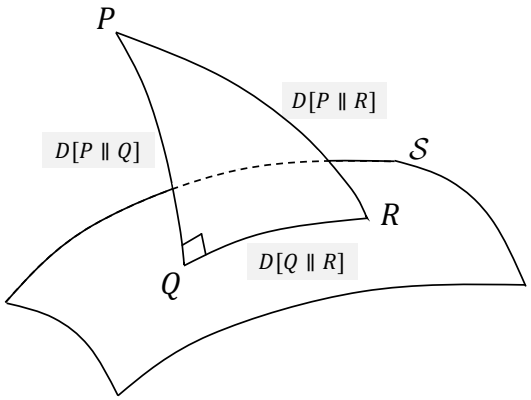

Figure 2: The image of the GPT.

The GPT based on three bare KL divergences has been well studied in information

geometry. The validity condition of the GPT based on three bare KL divergences is the isoenergetic condition. It determines the minimum KL divergence from a probability distribution to a set of canonical distributions (one-parameter geodesic line of canonical distribution). The principle of maximum entropy is derived from the GPT based on three bare KL divergences. We show the details in Appendix D.

Since the (information) entropy is dimensionless and the principle of maximum entropy is based on dimensionless KL divergences, we expect that the maximum work formulation is based on energy-dimensional divergences. Our energy-dimensional divergence is equal to (the minus of) the right-hand-side of Eq. (3.2) for the pair of the initial probability distribution and the initial canonical distribution. The change of scaled KL divergences is the (dissipative) work in Eq. (2.31) for a cyclic process. The minimization of this scaled KL divergence means the maximization of the work in Eq. (3.2).

The GPT based on scaled KL divergences holds as the following theorem:

Theorem 1. *If the isentropic condition is satisfied, i.e.,*

$$S[\rho_{\mathrm{can},0}(\tilde{\beta})] = S[\rho_0], \tag{3.5}$$

then the GPT holds,

$$\alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] = \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] + \alpha^{-1} D_{KL}[\rho_{\mathrm{can},0}(\tilde{\beta}) \parallel \rho_{\mathrm{can},0}(\alpha)]. \tag{3.6}$$

Proof. The derivation is straightforward,

$$\begin{aligned}
&\alpha^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] - \tilde{\beta}^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] - \alpha^{-1}D_{KL}[\rho_{\mathrm{can},0}(\tilde{\beta}) \parallel \rho_{\mathrm{can},0}(\alpha)] \\
&= -\alpha^{-1}S[\rho_0] - F_0(\alpha) + \langle H_0|\rho_0\rangle + \tilde{\beta}^{-1}S[\rho_0] + F_0(\tilde{\beta}) - \langle H_0|\rho_0\rangle \\
&\qquad + \alpha^{-1}S[\rho_{\mathrm{can},0}(\tilde{\beta})] + F_0(\alpha) - \langle H_0|\rho_{\mathrm{can},0}(\tilde{\beta})\rangle \\
&= -\alpha^{-1}S[\rho_0] + \tilde{\beta}^{-1}S[\rho_0] + F_0(\tilde{\beta}) + \alpha^{-1}S[\rho_{\mathrm{can},0}(\tilde{\beta})] - \langle H_0|\rho_{\mathrm{can},0}(\tilde{\beta})\rangle. \qquad (3.7)
\end{aligned}$$

Substituting the following relation into Eq. (3.7),

$$F_0(\tilde{\beta}) - \langle H_0|\rho_{\mathrm{can},0}(\tilde{\beta})\rangle = -\tilde{\beta}^{-1}S[\rho_{\mathrm{can},0}(\tilde{\beta})], \qquad (3.8)$$

we obtain

$$\begin{aligned}
&\alpha^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] - \tilde{\beta}^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] - \alpha^{-1}D_{KL}[\rho_{\mathrm{can},0}(\tilde{\beta}) \parallel \rho_{\mathrm{can},0}(\alpha)] \\
&= -\alpha^{-1}S[\rho_0] + \tilde{\beta}^{-1}S[\rho_0] + \alpha^{-1}S[\rho_{\mathrm{can},0}(\tilde{\beta})] - \tilde{\beta}^{-1}S[\rho_{\mathrm{can},0}(\tilde{\beta})] \\
&= \left(\tilde{\beta}^{-1} - \alpha^{-1}\right)\left(S[\rho_0] - S[\rho_{\mathrm{can},0}(\tilde{\beta})]\right) \\
&= 0, \qquad (3.9)
\end{aligned}$$

where we used the isentropic condition (3.5) in the last line. □

From Eq. (3.6) and the non-negativity of the scaled KL divergence, we immediately obtain the following inequality,

$$-\alpha^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] \leq -\tilde{\beta}^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]. \qquad (3.10)$$

Since Eq. (3.2) is valid for any α, the right-hand-side of Eq. (3.10) provides the greatest lower bound of W consistent with the isentropic condition (3.5); hence, its negative gives the maximum extractable work from the system.

The geometric image of the GPT based on scaled KL divergences is illustrated

in Figure 3. The set of canonical distributions is drawn as a line parametrized by the temperature. The (geodesic) line connecting ρ_0 and $\rho_{\mathrm{can},0}(\tilde{\beta})$ on the isentropic surface is orthogonal to the parametric line of canonical distributions in terms of the scaled KL divergence. The difference in structure between using dimensionless KL divergences and scaled KL divergences, in particular the change of metric and the dually flat structure with Bregman divergence in information geometry, are discussed in detail in subsection 3.2 and 3.3.

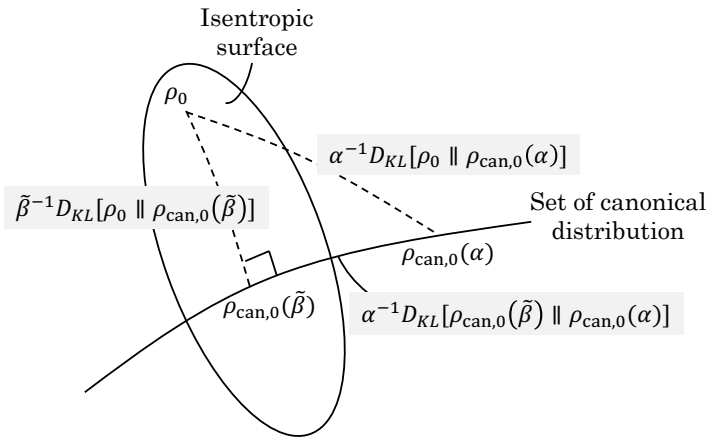

Figure 3: The image of the GPT based on the scaled KL divergences.

3.2 Geometric structure of the scaled KL divergence

In this subsection, we show the new geometric structure that scaled KL divergence leads to. First, we briefly explain the Bregman divergence and the dually flat structure in information geometry. The KL divergence is a Bregman divergence with $-\alpha F(\alpha)$ as a convex function and its Riemannian metric is the Fisher information matrix. On the other hand, we show that the scaled KL divergence is the Bregman divergence with $-F(\alpha)$ as a convex function. The Riemannian metric is also scaled

by temperature and the way of taking the coordinates in the parameter space has changed.

The space of interest in information geometry is a statistical manifold constructed from probability distributions. It is assumed that a point in n-dimensional manifold is represented using n-dimensional coordinates. One example of such a manifold is a family of one-dimensional Gaussian distributions. Since the Gaussian distribution is uniquely determined by the mean μ and the variance σ^2 as

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{3.11}$$

the family of Gaussian distributions can be regarded as a two-dimensional manifold if we take (μ, σ^2) as the coordinate system.

In information geometry, the exponential family plays an important role. The exponential family consists of exponential distributions in the form of

$$\rho(x;\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}\cdot\mathcal{H}(x) - \psi(\boldsymbol{\theta})\right\} \tag{3.12}$$

where

$$\boldsymbol{\theta} = (\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(n-1)}), \tag{3.13}$$

$$\mathcal{H}(x) = (\mathcal{H}^{(0)}(x), \mathcal{H}^{(1)}(x), \cdots, \mathcal{H}^{(n-1)}(x)), \tag{3.14}$$

and $\psi(\boldsymbol{\theta})$ is the potential function determined by the normalization condition as

$$\psi(\boldsymbol{\theta}) = \log\left\langle 1\right| \exp\left(\boldsymbol{\theta}\cdot\mathcal{H}\right)\rangle\,. \tag{3.15}$$

The parameter that identifies the distribution is $\boldsymbol{\theta}$. Assuming that $-\mathcal{H}^{(0)}$ is the Hamiltonian of the system, H, then the parameter $\boldsymbol{\theta}_{\text{can}} = (\alpha, 0, \cdots, 0)$ identifies the

canonical distributions

$$\rho(x;\boldsymbol{\theta}_{\mathrm{can}}) = \exp\{\alpha(F(\alpha) - H(x))\} = \rho_{\mathrm{can}}(x;\alpha) \tag{3.16}$$

where the potential function and the free energy are connected by the following relationship: $\psi(\boldsymbol{\theta}_{\mathrm{can}}) = -\alpha F(\alpha)$.

We then introduce a Bregman divergence between two exponential distributions. It is derived naturally from the convex function and we use $\psi(\boldsymbol{\theta})$ as the convex function. The convexity of $\psi(\boldsymbol{\theta})$ is guaranteed by the fact that its Hessian is a positive definite matrix because a simple calculation shows that

$$\frac{\partial^2}{\partial\theta^{(i)}\partial\theta^{(j)}}\psi(\boldsymbol{\theta}) = \left\langle \left(\mathcal{H}^{(i)} - \langle\mathcal{H}^{(i)}|\rho(\boldsymbol{\theta})\rangle\right)\left(\mathcal{H}^{(j)} - \langle\mathcal{H}^{(j)}|\rho(\boldsymbol{\theta})\rangle\right)|\rho(\boldsymbol{\theta})\right\rangle \tag{3.17}$$

which means that the Hessian is a variance-covariance matrix and it is positive definite in general. The Bregman divergence is defined as follows:

$$D[\boldsymbol{\theta}_A \parallel \boldsymbol{\theta}_B] = \psi(\boldsymbol{\theta}_A) - \psi(\boldsymbol{\theta}_B) - \nabla\psi(\boldsymbol{\theta}_B)\cdot(\boldsymbol{\theta}_A - \boldsymbol{\theta}_B). \tag{3.18}$$

where we note that $\nabla \equiv \left(\partial/\partial\theta^{(0)}, \partial/\partial\theta^{(1)}, \cdots, \partial/\partial\theta^{(n-1)}\right)$. The intuitive image of Bregman divergence in the one-dimensional case is shown in Figure 4. The Bregman divergence from $\rho(\boldsymbol{\theta}_A)$ to $\rho(\boldsymbol{\theta}_B)$ is defined by how far the potential function at $\boldsymbol{\theta}_A$ is from the tangent plane at $\boldsymbol{\theta}_B$. Eq. (3.18) is guaranteed to be always positive due to the convexity of ψ, and it also satisfies all three conditions of divergence written at the beginning of the previous subsection. Referring only to condition (3), the metric tensor $g_{i,j}$ corresponds to the Hessian of ψ. Eq. (3.17) is another way of writing the Fisher information matrix, which is a metric tensor of KL divergence.

The Bregman divergence leads to a dual structure. Legendre transformation on the previous coordinates $\boldsymbol{\theta}$ with a convex function ψ gives us a new coordinate

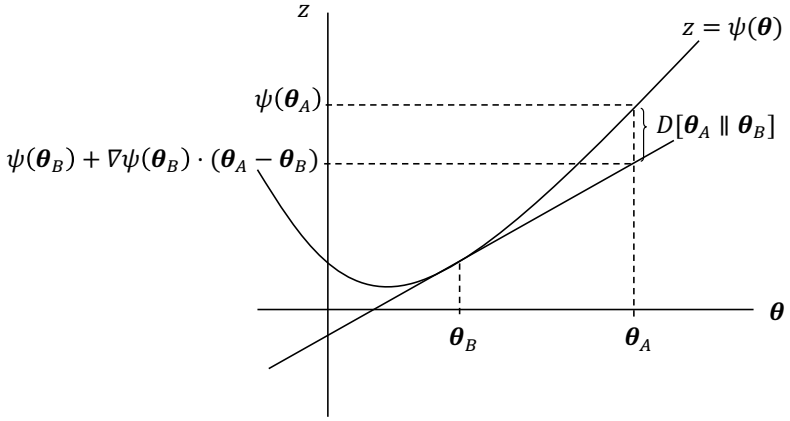

Figure 4: One-dimensional picture of the Bregman divergence derived from the convex function $\psi(\boldsymbol{\theta})$ and its tangent plane.

system $\boldsymbol{\theta}^*$ as

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}). \tag{3.19}$$

This dual coordinate $\boldsymbol{\theta}^*$ is often referred to as $\boldsymbol{\eta}$ in information geometry. By defining dual convex function for $\boldsymbol{\theta}^*$ as

$$\psi^*(\boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta}} \{\boldsymbol{\theta} \cdot \boldsymbol{\theta}^* - \psi(\boldsymbol{\theta})\}, \tag{3.20}$$

we obtain two coordinate systems. These are transformed to and from each other by Legendre transformation:

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}), \quad \psi^*(\boldsymbol{\theta}^*) = \boldsymbol{\theta} \cdot \boldsymbol{\theta}^* - \psi(\boldsymbol{\theta}) \tag{3.21}$$

$$\boldsymbol{\theta} = \nabla\psi^*(\boldsymbol{\theta}^*), \quad \psi(\boldsymbol{\theta}) = \boldsymbol{\theta}^* \cdot \boldsymbol{\theta} - \psi^*(\boldsymbol{\theta}^*) \tag{3.22}$$

The dual Bregman divergence is derived from the dual coordinate $\boldsymbol{\theta}^*$ and dual

convex function ψ^*,

$$\begin{aligned} D^*[\boldsymbol{\theta}_A^* \parallel \boldsymbol{\theta}_B^*] &= \psi^*(\boldsymbol{\theta}_A^*) - \psi^*(\boldsymbol{\theta}_B^*) - \nabla\psi^*(\boldsymbol{\theta}_B^*) \cdot (\boldsymbol{\theta}_A^* - \boldsymbol{\theta}_B^*) \\ &= \boldsymbol{\theta}_A \cdot \boldsymbol{\theta}_A^* - \psi(\boldsymbol{\theta}_A) - \boldsymbol{\theta}_B \cdot \boldsymbol{\theta}_B^* + \psi(\boldsymbol{\theta}_B) - \boldsymbol{\theta}_B \cdot (\boldsymbol{\theta}_A^* - \boldsymbol{\theta}_B^*) \\ &= \boldsymbol{\theta}_A \cdot \nabla\psi(\boldsymbol{\theta}_A) - \psi(\boldsymbol{\theta}_A) + \psi(\boldsymbol{\theta}_B) - \boldsymbol{\theta}_B \cdot \nabla\psi(\boldsymbol{\theta}_A) \\ &= \psi(\boldsymbol{\theta}_B) - \psi(\boldsymbol{\theta}_A) - \nabla\psi(\boldsymbol{\theta}_A) \cdot (\boldsymbol{\theta}_B - \boldsymbol{\theta}_A) \\ &= D[\boldsymbol{\theta}_B \parallel \boldsymbol{\theta}_A]. \end{aligned} \tag{3.23}$$

We have mentioned that there is no symmetry in divergence, but when we consider the dual coordinate, the dual structure is established: a Bregman divergence with a different order of two arguments becomes a dual Bregman divergence

$$D^*[\boldsymbol{\theta}_A^* \parallel \boldsymbol{\theta}_B^*] = D[\boldsymbol{\theta}_B \parallel \boldsymbol{\theta}_A],$$

which is the most interesting and important result of information geometry.

Next, we give a proof of the generalized Pythagorean theorem (GPT) and its geometric structure based on the Bregman divergence. Suppose there exist a flat coordinate system (the affine coordinate system) $\boldsymbol{\theta}$. We refer to the curved line

$$\boldsymbol{\theta}(u) = \boldsymbol{a}u + \boldsymbol{b}, \tag{3.24}$$

as geodesic line, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the constant vectors and u is the real parameter. For the dual coordinates $\boldsymbol{\theta}^*$, we also refer to the curved line

$$\boldsymbol{\theta}^*(u) = \boldsymbol{c}u + \boldsymbol{d}, \tag{3.25}$$

as dual geodesic line, where $\boldsymbol{c}$ and $\boldsymbol{d}$ are the constant vectors. Of course, the geodesic line and dual geodesic line are different and they are not straight lines from the point

of view of different coordinate systems. Two geodesic lines (or dual geodesic lines) are orthogonal if their tangent vectors are orthogonal to each other.

The GPT

$$D[P \parallel R] = D[P \parallel Q] + D[Q \parallel R], \tag{3.26}$$

holds when the geodesic line connecting P and Q and the dual geodesic line connecting Q and R are orthogonal. Let the coordinates of points P, Q and R be $\boldsymbol{\theta}_P$, $\boldsymbol{\theta}_Q$ and $\boldsymbol{\theta}_R$ respectively, and using the definition of Bregman divergence (3.18), we obtain

$$D[P \parallel Q] + D[Q \parallel R] - D[P \parallel R] = (\boldsymbol{\theta}_P - \boldsymbol{\theta}_Q) \cdot (\boldsymbol{\theta}_R^* - \boldsymbol{\theta}_Q^*). \tag{3.27}$$

The geodesic line connecting P and Q is written as

$$\boldsymbol{\theta}(u) = u\boldsymbol{\theta}_P + (1-u)\boldsymbol{\theta}_Q, \tag{3.28}$$

and

$$\frac{\partial}{\partial u}\boldsymbol{\theta}(u) = \boldsymbol{\theta}_P - \boldsymbol{\theta}_Q. \tag{3.29}$$

The dual geodesic line connecting Q and R is also written as

$$\boldsymbol{\theta}^*(u) = (1-u)\boldsymbol{\theta}_Q^* + u\boldsymbol{\theta}_R^* \tag{3.30}$$

and

$$\frac{\partial}{\partial u}\boldsymbol{\theta}^*(u) = \boldsymbol{\theta}_R^* - \boldsymbol{\theta}_Q^*. \tag{3.31}$$

Since these are orthogonal we get

$$(\boldsymbol{\theta}_P - \boldsymbol{\theta}_Q) \cdot (\boldsymbol{\theta}_R^* - \boldsymbol{\theta}_Q^*) = 0 \tag{3.32}$$

which means that the right-hand side of Eq. (3.27) is zero. We note that the dot

product in Eq. (3.32) cannot be defined without the Fisher metric g_{ij} mentioned earlier since it is an inner product on a statistical manifold which is a Riemannian manifold.

It is already known that the KL divergence is the Bregman divergence of the exponential family, where the coordinate system is $\boldsymbol{\theta}$ and the convex function is $\psi(\boldsymbol{\theta})$. The dual coordinate is

$$\boldsymbol{\theta}^* = \nabla\psi(\boldsymbol{\theta}) = -\langle\mathcal{H}|\rho(\boldsymbol{\theta})\rangle, \tag{3.33}$$

and the dual convex function is

$$\psi^*(\boldsymbol{\theta}^*) = -S[\rho(\boldsymbol{\theta})]. \tag{3.34}$$

The scaled KL divergence introduced in the previous subsection is also derived as a Bregman divergence when considering two canonical distributions. In this case, we take (the minus of) the free energy $-F(\alpha)$ of the canonical distribution as a convex function and α^{-1} as the coordinate system instead of $\alpha(= -\theta^{(0)})$. The Bregman divergence is as follows:

$$\begin{aligned}
D[\alpha_A^{-1} \parallel \alpha_B^{-1}] &= -F(\alpha_A) + F(\alpha_B) + \nabla_{\alpha_B^{-1}} F(\alpha_B) \cdot (\alpha_A^{-1} - \alpha_B^{-1}) \\
&= -F(\alpha_A) + F(\alpha_B) - S[\rho_{\mathrm{can}}(\alpha_B)](\alpha_A^{-1} - \alpha_B^{-1}) \\
&= -F(\alpha_A) + \langle H|\rho_{\mathrm{can}}(\alpha_B)\rangle - \alpha_A^{-1} S[\rho_{\mathrm{can}}(\alpha_B)] \\
&= -\alpha_A^{-1}\langle \log\rho_{\mathrm{can}}(\alpha_A)|\rho_{\mathrm{can}}(\alpha_B)\rangle + \alpha_A^{-1}\langle \log\rho_{\mathrm{can}}(\alpha_B)|\rho_{\mathrm{can}}(\alpha_B)\rangle \\
&= \alpha_A^{-1} D_{KL}[\rho_{\mathrm{can}}(\alpha_B) \parallel \rho_{\mathrm{can}}(\alpha_A)]
\end{aligned} \tag{3.35}$$

The right-hand-side of Eq. (3.35) is the scaled KL divergence itself. The dual coordinate is entropy $S[\rho_{\mathrm{can}}(\alpha)]$ and the dual convex function is energy $E = \langle H|\rho_{\mathrm{can}}(\alpha)\rangle$. Therefore, the condition of the GPT in Eq. (3.9) is in the form of a product of

difference of temperature and difference of entropy. This is a completely different geometrical structure to that of the bare KL divergence as you can see by comparing Figure 3 and Figure 12 in Appendix D.

The differences in structure is seen from considering local KL divergence for a canonical distribution. In the case of bare KL divergence,

$$\begin{aligned} D_{KL}[\rho_{\mathrm{can}}(\alpha) \parallel \rho_{\mathrm{can}}(\alpha + d\alpha)] &= \frac{1}{2} d\alpha \Delta E^2 d\alpha \\ &= \frac{1}{2} d\alpha \frac{d(-E)}{d\alpha} d\alpha \\ &= \frac{1}{2} d\alpha d(-E) \end{aligned} \tag{3.36}$$

where ΔE^2 is the Fisher information matrix. On the other hand, the scaled local KL divergence is also written as

$$\begin{aligned} \alpha^{-1} D_{KL}[\rho_{\mathrm{can}}(\alpha) \parallel \rho_{\mathrm{can}}(\alpha + d\alpha)] &= \frac{1}{2} d\alpha \frac{\Delta E^2}{\alpha} d\alpha \\ &= \frac{1}{2} d(\alpha^{-1}) \frac{dS}{d\alpha} d\alpha \\ &= \frac{1}{2} d(\alpha^{-1}) dS. \end{aligned} \tag{3.37}$$

The dual orthogonality, $d\alpha d(-E)$ is changed to $d(\alpha^{-1})dS$ by the scale transformation.

3.3 Structural changes due to scaling divergence

More generally, we consider the divergence in the form of the KL divergence between $\rho(\boldsymbol{\theta}_A)$ and $\rho(\boldsymbol{\theta}_B)$ divided by the scalar quantity λ_B which depends on $\rho(\boldsymbol{\theta}_B)$, i.e.,

$\lambda_B^{-1} D_{KL}[\rho(\boldsymbol{\theta}_A) \parallel \rho(\boldsymbol{\theta}_B)]$. The condition under which the GPT holds is

$$\begin{aligned}
&\frac{1}{\lambda_C} D_{KL}[\rho(\boldsymbol{\theta}_A) \parallel \rho(\boldsymbol{\theta}_C)] = \frac{1}{\lambda_B} D_{KL}[\rho(\boldsymbol{\theta}_A) \parallel \rho(\boldsymbol{\theta}_B)] + \frac{1}{\lambda_C} D_{KL}[\rho(\boldsymbol{\theta}_B) \parallel \rho(\boldsymbol{\theta}_C)] \\
&\Leftrightarrow (S[\rho(\boldsymbol{\theta}_A)] - S[\rho(\boldsymbol{\theta}_B)]) \cdot \left(\frac{1}{\lambda_B} - \frac{1}{\lambda_C} \right) + (\nabla\psi(\boldsymbol{\theta}_A) - \nabla\psi(\boldsymbol{\theta}_B)) \cdot \left(\frac{\boldsymbol{\theta}_B}{\lambda_B} - \frac{\boldsymbol{\theta}_C}{\lambda_C} \right) = 0 \\
&\Leftrightarrow S[\rho(\boldsymbol{\theta}_A)] = S[\rho(\boldsymbol{\theta}_B)] \qquad \text{and} \qquad \frac{\boldsymbol{\theta}_B}{\lambda_B} = \frac{\boldsymbol{\theta}_C}{\lambda_C}. \qquad (3.38)
\end{aligned}$$

Fig. 5 shows an image of this situation. The KL divergence scaled by temperature,

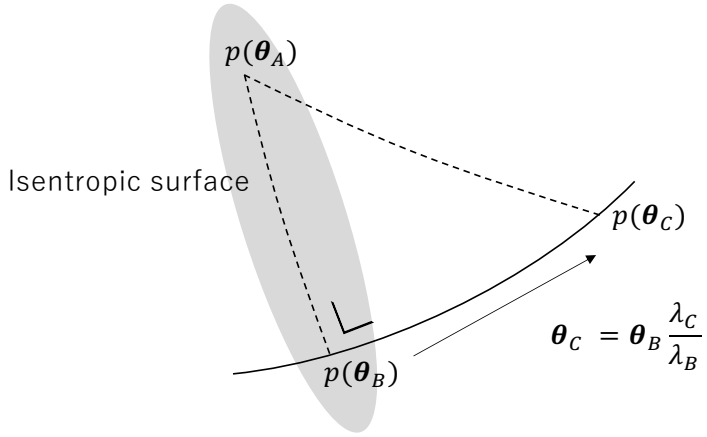

Figure 5: The visual image of Eq. (3.38). The GPT holds when the isentropic condition and the scaling condition $\boldsymbol{\theta}_B/\lambda_B = \boldsymbol{\theta}_C/\lambda_C$ are satisfied.

introduced in subsection 3.1, satisfies this latter condition $\boldsymbol{\theta}_B/\lambda_B = \boldsymbol{\theta}_C/\lambda_C$ because the second argument (distribution) of the divergence was restricted to the canonical distribution. Thus, only the isentropic condition is needed. The set of canonical distributions is a kind of special subspace that satisfies the condition $\boldsymbol{\theta}_B/\lambda_B = \boldsymbol{\theta}_C/\lambda_C$, and the scaled KL divergence is one of the examples of this geometric structure.

In the above setup, we scaled by the quantity associated with the second argument of KL divergence. Now, let us consider scaling with the quantity associated with the

first argument, i.e., $\lambda_A^{-1} D_{KL}[\rho(\boldsymbol{\theta}_A) \parallel \rho(\boldsymbol{\theta}_B)]$. The condition under which the GPT holds is

$$\begin{aligned}
&\frac{1}{\lambda_A} D_{KL}[\rho(\boldsymbol{\theta}_A) \parallel \rho(\boldsymbol{\theta}_C)] = \frac{1}{\lambda_A} D_{KL}[\rho(\boldsymbol{\theta}_A) \parallel \rho(\boldsymbol{\theta}_B)] + \frac{1}{\lambda_B} D_{KL}[\rho(\boldsymbol{\theta}_B) \parallel \rho(\boldsymbol{\theta}_C)] \\
&\Leftrightarrow (\psi(\boldsymbol{\theta}_C) - \psi(\boldsymbol{\theta}_B)) \cdot \left(\frac{1}{\lambda_A} - \frac{1}{\lambda_B} \right) + \left(\frac{\nabla\psi(\boldsymbol{\theta}_B)}{\lambda_B} - \frac{\nabla\psi(\boldsymbol{\theta}_A)}{\lambda_A} \right) \cdot (\boldsymbol{\theta}_C - \boldsymbol{\theta}_B) = 0 \\
&\Leftrightarrow \psi(\boldsymbol{\theta}_B) = \psi(\boldsymbol{\theta}_C) \qquad \text{and} \qquad \frac{\nabla\psi(\boldsymbol{\theta}_B)}{\lambda_B} = \frac{\nabla\psi(\boldsymbol{\theta}_A)}{\lambda_A}.
\end{aligned} \tag{3.39}$$

Fig. 6 shows an image of this situation. If we restrict to the space of canonical

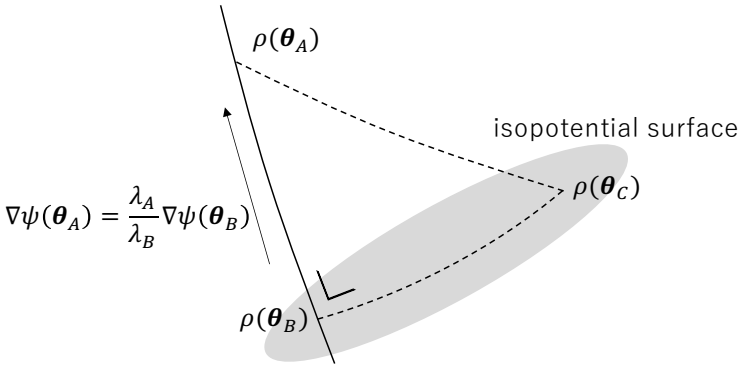

Figure 6: The visual image of Eq. (3.39). The GPT holds when the isopotential condition and the scaling condition $\nabla\psi(\boldsymbol{\theta}_B)/\lambda_B = \nabla\psi(\boldsymbol{\theta}_A)/\lambda_A$ are satisfied.

distributions, $\psi = -\alpha F(\alpha)$ and $\nabla\psi = -E$. Therefore, if $\rho(\boldsymbol{\theta}_A)$ and $\rho(\boldsymbol{\theta}_B)$ are canonical distributions and λ is taken by energy, only the isopotential condition remains. This is the KL divergence scaled by the inverse of the energy. For this scaled divergence, the space of canonical distributions is also a special subspace.

Scaling by a quantity related to the first argument of divergence appears when considering the "cost performance" of a system in contact with a finite heat bath. "Cost performance" is defined by the work obtained in ratio to the size of the finite

heat bath. Therefore, the KL divergence scaled by E^{-1} may be useful in considering the "cost performance" of heat engines in the future.

Table 1 summarizes the dual structure of the three divergences we have obtained so far. We use these three divergences depending on what we want to keep constant and what we want to optimize. First, the optimization of KL divergence leads to the maximum entropy principle under constant energy. Second, the optimization of KL divergence scaled by α^{-1} leads to the maximum work under constant entropy. Finally, the optimization of KL divergence scaled by E^{-1} leads to the maximum temperature under constant potential.

Table 1: Dual structure of the three divergences.

Bregman divergence	coordinates	convex function	dual coordinates	dual convex function
KL divergence	α	$-\alpha F(\alpha)$	$-E$	$-S$
KL divergence scaled by α^{-1}	α^{-1}	$-F(\alpha)$	S	E
KL divergence scaled by E^{-1}	E^{-1}	$-E^{-1}S$	$\alpha F(\alpha)$	α

3.4 The case of non-cyclic processes

The maximum work formulation of the generalized second law for a non-cyclic operation ($\Delta F(\alpha) \neq 0$) may also be obtained by the GPT based on the scaled KL divergences. The work inequality is written as Eq. (2.32),

$$W \geq \Delta F(\alpha) - \alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)].$$

When $\tilde{\beta}$ is the inverse effective temperature determined by the isentropic condition (2.35), i.e.,

$$S[\rho_{\mathrm{can},T}(\tilde{\beta})] = S[\rho_0],$$

the following inequality holds for any α,

$$\Delta F(\tilde{\beta}) - \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] \geq \Delta F(\alpha) - \alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] \quad (3.40)$$

which means that the inverse effective temperature maximizes the right-hand-side of Eq. (2.32).

We prove Eq. (3.40) as follows. First, we rewrite the difference between the two sides of Eq. (3.40) in terms of scaled KL divergences as,

$$\begin{aligned}
&\Delta F(\tilde{\beta}) - \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] - \Delta F(\alpha) + \alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\alpha)] \\
&= F_T(\tilde{\beta}) - F_0(\tilde{\beta}) + \tilde{\beta}^{-1} S[\rho_0] + F_0(\tilde{\beta}) - \langle H_0 | \rho_0 \rangle \\
&\qquad - F_T(\alpha) + F_0(\alpha) - \alpha^{-1} S[\rho_0] - F_0(\alpha) + \langle H_0 | \rho_0 \rangle \\
&= -\tilde{\beta}^{-1} \left\{ -S[\rho_0] - \tilde{\beta} F_T(\tilde{\beta}) + \tilde{\beta} \langle H_T | \rho_0 \rangle \right\} + \alpha^{-1} \left\{ -S[\rho_0] - \alpha F_T(\alpha) + \alpha \langle H_T | \rho_0 \rangle \right\} \\
&= -\tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\tilde{\beta})] + \alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\alpha)]. \qquad (3.41)
\end{aligned}$$

Then the next theorem holds:

Theorem 2. *When the isentropic condition (2.35) is given, the GPT based on the scaled KL divergences holds as*

$$\alpha^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\alpha)] = \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\tilde{\beta})] + \alpha^{-1} D_{KL}[\rho_{\mathrm{can},T}(\tilde{\beta}) \parallel \rho_{\mathrm{can},T}(\alpha)]. \quad (3.42)$$

Proof. Eq. (3.42) is derived straightforwardly,

$$\begin{aligned}
&\alpha^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\alpha)] - \tilde{\beta}^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\tilde{\beta})] - \alpha^{-1}D_{KL}[\rho_{\mathrm{can},T}(\tilde{\beta}) \parallel \rho_{\mathrm{can},T}(\alpha)] \\
&= -\alpha^{-1}S[\rho_0] - F_T(\alpha) + \langle H_T|\rho_0\rangle + \tilde{\beta}^{-1}S[\rho_0] + F_T(\tilde{\beta}) - \langle H_T|\rho_0\rangle \\
&\qquad + \alpha^{-1}S[\rho_{\mathrm{can},T}(\tilde{\beta})] + F_T(\alpha) - \langle H_T|\rho_{\mathrm{can},T}(\tilde{\beta})\rangle \\
&= -\alpha^{-1}S[\rho_0] + \tilde{\beta}^{-1}S[\rho_0] + F_T(\tilde{\beta}) + \alpha^{-1}S[\rho_{\mathrm{can},T}(\tilde{\beta})] - \langle H_T|\rho_{\mathrm{can},T}(\tilde{\beta})\rangle \\
&= -\alpha^{-1}S[\rho_0] + \tilde{\beta}^{-1}S[\rho_0] + \alpha^{-1}S[\rho_{\mathrm{can},T}(\tilde{\beta})] - \tilde{\beta}^{-1}S[\rho_{\mathrm{can},T}(\tilde{\beta})] \\
&= (\tilde{\beta}^{-1} - \alpha^{-1})\left(S[\rho_0] - S[\rho_{\mathrm{can},T}(\tilde{\beta})]\right) \\
&= 0,
\end{aligned} \tag{3.43}$$

where we used the relation,

$$F_T(\tilde{\beta}) - \langle H_T|\rho_{\mathrm{can},T}(\tilde{\beta})\rangle = -\tilde{\beta}^{-1}S[\rho_{\mathrm{can},T}(\tilde{\beta})], \tag{3.44}$$

and the isentropic condition Eq. (2.35). □

The GPT (3.42) is rearranged to yield the right-hand-side of Eq. (3.41),

$$\begin{aligned}
&-\tilde{\beta}^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\tilde{\beta})] + \alpha^{-1}D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},T}(\alpha)] \\
&= \alpha^{-1}D_{KL}[\rho_{\mathrm{can},T}(\tilde{\beta}) \parallel \rho_{\mathrm{can},T}(\alpha)] \geq 0,
\end{aligned} \tag{3.45}$$

guaranteeing the non-negativity of the left-hand-side of Eq. (3.41) and the validity of Eq. (3.40).

3.5 Example 1: thermally isolated ideal gas

A simple example that illustrates the GPT is a thermally isolated ideal gas that is confined to a container with a movable wall. Suppose that initially the volume of gas is V_i and the temperature is β_i^{-1}. The wall is then adiabatically moved, expanding

the volume to V_f as work is extracted from the gas. This is a non-cyclic process.

The GPT for scaled KL divergences, in this case, is given by Eq. (3.42). We describe the initial state and the final canonical state using a Maxwellian velocity distribution ρ_{Max} as $\rho_0 = \rho_{\mathrm{Max}}(\beta_i, V_i)$ and $\rho_{\mathrm{can},T}(\alpha) = \rho_{\mathrm{Max}}(\alpha, V_f)$, where

$$\rho_{\mathrm{Max}}(\alpha, V) = \frac{1}{Z(\alpha, V)} \prod_{\nu=1}^{3N} \exp\left(-\alpha \frac{p_\nu^2}{2m}\right) \chi_V \tag{3.46}$$

where N is the number of particles and p_ν and m are momentum and mass of a particle, respectively. The indicator function χ_V serves to indicate the configuration space dependence of the distribution and Z is the partition function,

$$Z(\alpha, V) = \frac{V^N}{N!} \left(\frac{2\pi m}{h^2 \alpha}\right)^{3N/2} \tag{3.47}$$

where h represents Planck constant.

The GPT is then written as

$$\begin{aligned} &\alpha^{-1} D_{KL}[\rho_{\mathrm{Max}}(\beta_i, V_i) \parallel \rho_{\mathrm{Max}}(\alpha, V_f)] \\ &= \tilde{\beta}^{-1} D_{KL}[\rho_{\mathrm{Max}}(\beta_i, V_i) \parallel \rho_{\mathrm{Max}}(\tilde{\beta}, V_f)] + \alpha^{-1} D_{KL}[\rho_{\mathrm{Max}}(\tilde{\beta}, V_f) \parallel \rho_{\mathrm{Max}}(\alpha, V_f)], \end{aligned} \tag{3.48}$$

which holds for the effective temperature $\tilde{\beta}$ determined by the following condition,

$$\tilde{\beta} = \left(\frac{V_f}{V_i}\right)^{2/3} \beta_i. \tag{3.49}$$

The proof is straightforward using Eqs. (3.46) and (3.47) as

$$
\begin{aligned}
&\alpha^{-1} D_{KL}[\rho_{\mathrm{Max}}(\beta_i, V_i) \parallel \rho_{\mathrm{Max}}(\alpha, V_f)] \\
&\quad - \tilde{\beta}^{-1} D_{KL}[\rho_{\mathrm{Max}}(\beta_i, V_i) \parallel \rho_{\mathrm{Max}}(\tilde{\beta}, V_f)] - \alpha^{-1} D_{KL}[\rho_{\mathrm{Max}}(\tilde{\beta}, V_f) \parallel \rho_{\mathrm{Max}}(\alpha, V_f)] \\
&= \alpha^{-1} \log\left(\frac{Z(\alpha, V_f)}{Z(\beta_i, V_i)}\right) - \frac{3}{2} N \left(\alpha^{-1} - \beta_i^{-1}\right) - \tilde{\beta}^{-1} \log\left(\frac{Z(\tilde{\beta}, V_f)}{Z(\beta_i, V_i)}\right) \\
&\quad + \frac{3}{2} N \left(\tilde{\beta}^{-1} - \beta_i^{-1}\right) - \alpha^{-1} \log\left(\frac{Z(\alpha, V_f)}{Z(\tilde{\beta}, V_f)}\right) + \frac{3}{2} N \left(\alpha^{-1} - \tilde{\beta}^{-1}\right) \\
&= -\alpha^{-1} \log\left(Z(\beta_i, V_i)\right) - \tilde{\beta}^{-1} \log\left(Z(\tilde{\beta}, V_f)\right) + \tilde{\beta}^{-1} \log\left(Z(\beta_i, V_i)\right) + \alpha^{-1} \log\left(Z(\tilde{\beta}, V_f)\right) \\
&= \log\left(\frac{Z(\tilde{\beta}, V_f)}{Z(\beta_i, V_i)}\right) \left(\alpha^{-1} - \tilde{\beta}^{-1}\right) \\
&= \log\left(\frac{V_f^N \tilde{\beta}^{-3N/2}}{V_i^N \beta_i^{-3N/2}}\right) \left(\alpha^{-1} - \tilde{\beta}^{-1}\right) \\
&= \left\{\log\left(\frac{V_f}{V_i}\right)^N - \log\left(\frac{\tilde{\beta}}{\beta_i}\right)^{3N/2}\right\} \left(\alpha^{-1} - \tilde{\beta}^{-1}\right)
\end{aligned}
\tag{3.50}
$$

where we used the relation,

$$
\alpha_2^{-1} D_{KL}[\rho_{\mathrm{Max}}(\alpha_1, V_1) \parallel \rho_{\mathrm{Max}}(\alpha_2, V_2)] = \alpha_2^{-1} \log\left(\frac{Z(\alpha_2, V_2)}{Z(\alpha_1, V_1)}\right) - \frac{3}{2} N \left(\alpha_2^{-1} - \alpha_1^{-1}\right). \tag{3.51}
$$

In order for Eq. (3.50) to become zero, we must have

$$
\begin{aligned}
&\log\left(\frac{V_f}{V_i}\right)^N - \log\left(\frac{\tilde{\beta}}{\beta_i}\right)^{3N/2} = 0 \\
&\Leftrightarrow \left(\frac{V_f}{V_i}\right)^N = \left(\frac{\tilde{\beta}}{\beta_i}\right)^{3N/2}
\end{aligned}
\tag{3.52}
$$

which leads to the condition (3.49).

Eq. (3.48) guarantees that the final state of a Maxwellian velocity distribution with the effective temperature $\tilde{\beta}^{-1}$ gives the maximum work as discussed in this section. The condition of effective temperature (3.49) is the well-known polytropic process equation. Of course, it can also be derived from the isentropic condition and the Sackur-Tetrode equation for the entropy of an ideal gas.

3.6 Example 2: two-level quantum system

We apply the GPT to an optimization problem in nonequilibrium statistical mechanics. From the generalized maximum work formulation, the maximum work is extracted by an operation under which the final state is the canonical state with the final Hamiltonian and the effective temperature. The final canonical state with the effective temperature is the ideally optimal state to extract the maximum work. The ideally optimal state may be difficult to realize experimentally. It is important to figure out what is the "closest" (minimum divergence) state from the ideally optimal state in a set of realizable final states. The "closest" state is the experimentally optimal state to extract work that satisfies the orthogonality condition.

3.6.1 A spin-$1/2$ particle operated by a magnetic field

We consider a thermally isolated two-level quantum system as a simple example [34]. The application of nonequilibrium statistical mechanics to few-degrees-of-freedom quantum systems is a current topic of much interest. Time-dependent two-level quantum systems often appear in the context of matter-field interactions or nuclear magnetic resonance. Such systems have been used to demonstrate the validity of the fluctuation relations of modern nonequilibrium statistical mechanics [47, 49].

We consider a spin-$1/2$ particle embedded in a magnetic field that is controlled

[50]. The Hamiltonian of the system is

$$H_t = \frac{\hbar\omega}{2}\left(\cos\phi_t\sigma_z + \sin\phi_t\sigma_x\right), \tag{3.53}$$

where σ_i $(i = x, y, z)$ are the standard Pauli matrices and we choose the direction of the magnetic field restricted to the $x - z$ plane with ϕ_t the rotation-angle of the magnetic field around the y-axis at time t. The Hamiltonian is rewritten by using rotation operators in spin space,

$$H_t = \frac{\hbar\omega}{2}\mathrm{e}^{-\mathrm{i}\sigma_y\phi_t/2}\sigma_z\mathrm{e}^{\mathrm{i}\sigma_y\phi_t/2}. \tag{3.54}$$

The eigenvalues of the Hamiltonian are $E_0 = -\hbar\omega/2$ and $E_1 = \hbar\omega/2$, which we identify as the ground and excited state energies, respectively.

The spin-$1/2$ particle is externally operated from $t = 0$ to $t = T$. The operation is represented by the time-dependent angle in the Hamiltonian. We choose a cyclic operation, $\phi_0 = 0$ at $t = 0$ and $\phi_T = 2\pi$ at $t = T$. The angle increases from 0 to 2π monotonically and $\phi_\tau = \pi$ at $t = \tau$. The Hamiltonian is proportional to the Pauli matrix σ_z at $t = 0$. It is proportional to $-\sigma_z$ at $t = \tau$. When $t = T$ the Hamiltonian returns to its original form at $t = 0$.

We discuss how to extract the maximum work from a simple noneqilibrium initial state such as a pure excited state [32–34]. The maximum work can be obtained by the cyclic operation including two processes:

1. The short stabilization process for $t \in [0, \tau)$ in which the pure excited state becomes the ground state by changing H_0 to $H_\tau = -H_0$.

2. The restoration process for $t \in [\tau, T)$ to the original Hamiltonian, $H_T = H_0$, without any transition to the excited state.

These processes are analogous to those that have been discussed on how to extract

the maximum work from a noneqilibrium initial state in a thermodynamic system [32–34]. The first corresponds to the stabilization process to prevent spontaneous relaxation in a thermodynamic system. For simplicity we consider here the sudden process, $\tau \to 0$. The second is generally a quasi-static (isentropic) process without dissipation. After the sudden process, the initial state does not change and the Hamiltonian changes as $H_{0+} = -H_0$. The control parameter of our operation is the rate at which the field is rotated. We will find the optimal angular frequency to extract work. The solutions for finite frequencies correspond to dissipationless non-quasi-static processes.

The Schrödinger equation is

$$\mathrm{i}\hbar\frac{\partial}{\partial t}|\psi_t\rangle = H_t|\psi_t\rangle = \frac{\hbar\omega}{2}\mathrm{e}^{-\mathrm{i}\sigma_y\phi_t/2}\sigma_z\mathrm{e}^{\mathrm{i}\sigma_y\phi_t/2}|\psi_t\rangle. \tag{3.55}$$

We choose a simple linear time dependence for the rotation angle,

$$\phi_t = \pi + \Omega t \quad \text{for } 0 < t \leq T \tag{3.56}$$

where $\Omega = \pi/T$ is the adjustable angular frequency. Then, the Schrödinger equation is easily solved by using the rotating amplitude $\mathrm{e}^{\mathrm{i}\sigma_y\phi_t/2}|\psi_t\rangle$. The solution is written as,

$$|\psi_t\rangle = u^t|\psi_0\rangle \tag{3.57}$$

where the time evolution operator u^t is given as

$$u^t = \mathrm{e}^{-\mathrm{i}(\pi+\Omega t)\sigma_y/2}\mathrm{e}^{-\mathrm{i}(\omega\sigma_z-\Omega\sigma_y)t/2}\mathrm{e}^{\mathrm{i}\pi\sigma_y/2}. \tag{3.58}$$

3.6.2 A vector representation of a quantum state

We introduce a vector representation of a quantum state. This section 3.6.2 prepares for the optimization of the rotation-angle frequency discussed in the next section

3.6.3. A quantum state is represented as a point vector in parameter space in information geometry. The orthogonality condition from the GPT is written in terms of these vectors [16, 17].

A quantum state is written as a 2×2 density matrix that is expanded as a linear combination of four Hermitian matrixes: the unit matrix I and the three Pauli matrixes σ_i $(i = x, y, z)$. Any state except pure states is written as a positive definite density matrix,

$$\rho = \frac{1}{2} I - \frac{1}{2} \hat{\eta} \cdot \sigma \tanh |\eta| \tag{3.59}$$

where η is a three dimensional real vector, $\hat{\eta} = \eta/|\eta|$ is its unit vector, and $\sigma = (\sigma_x, \sigma_y, \sigma_z)$. The condition of total probability is obvious, $\mathrm{Tr}[\rho] = 1$. The positive definiteness is guaranteed from the range, $|\tanh |\eta|| < 1$, since the eigenvalues of ρ are $(1 \pm \tanh |\eta|)/2$. A density matrix corresponding to a pure state is positive semi-definite. We consider the pure state in the limit of $|\tanh |\eta|| \to 1$ for $|\eta| \to \infty$. The pure ground state of H_0 is $\rho = (I - \sigma_z)/2$ where $\hat{\eta} = (0, 0, 1)$ and the pure excited state is $\rho = (I + \sigma_z)/2$ where $\hat{\eta} = (0, 0, -1)$. Hereafter, we call the above representation based on a linear combination of four Hermitian matrixes as the M-representation.

The parameter vector η naturally appears in an exponential-type representation (E-representation),

$$\rho = \frac{\mathrm{e}^{-\eta \cdot \sigma}}{Z}. \tag{3.60}$$

where

$$Z = \mathrm{Tr}[\mathrm{e}^{-\eta \cdot \sigma}] = 2 \cosh |\eta|. \tag{3.61}$$

The equivalence between two representations is confirmed by the Euler-like formula,

$$\mathrm{e}^{-\eta \cdot \sigma} = \cosh |\eta| I - \hat{\eta} \cdot \sigma \sinh |\eta|. \tag{3.62}$$

We write the nonequilibrium initial state at time $t = 0$ as

$$\rho_0 = \frac{1}{2}I - \frac{1}{2}\hat{\eta}_0 \cdot \sigma \tanh|\eta_0|. \tag{3.63}$$

We choose $\hat{\eta}_0 = (0, -\sin\varphi, -\cos\varphi)$ ($|\varphi| < \pi/2$) for the nonequilibrium initial state. From the condition ($|\varphi| < \pi/2$), the initial energy $E_0 = \mathrm{Tr}[\rho_0 H_0]$ is positive so that work is extractable. We chose the x component $\hat{\eta}_{0,x} = 0$ for simplicity. We may be able to adjust $\hat{\eta}_{0,x}$ by an additional operation around the y-axis.

From the solution of the Schrödinger equation, the time evolution of the density matrix is given,

$$\rho_t = u^t \rho_0 u^{-t} = \frac{1}{2}I - \frac{1}{2}\hat{\eta}_t \cdot \sigma \tanh|\eta_t|, \tag{3.64}$$

where

$$\eta_t = \frac{1}{2}\mathrm{Tr}[\sigma u^t (\eta_0 \cdot \sigma) u^{-t}]. \tag{3.65}$$

We used the orthogonality of Pauli matrixes, $\mathrm{Tr}[\sigma_i \sigma_j] = 2\delta_{i,j}$. The norm $|\eta_t|$ is constant in time under the unitary time evolution,

$$|\eta_t| = \sqrt{\frac{1}{2}\mathrm{Tr}[(u^t(\eta_0 \cdot \sigma)u^{-t})^2]} = |\eta_0|. \tag{3.66}$$

The final state at time $t = T$ is written as

$$\rho_T = \frac{1}{2}I - \frac{1}{2}\hat{\eta}_T \cdot \sigma \tanh|\eta_T|. \tag{3.67}$$

We note that the time period T may also be considered as the control parameter in our operation, since the angular frequency $\Omega = \pi/T$. We will discuss the optimal final state to extract work in the section 3.6.3.

The canonical state at time t ($t \in (0, T]$) with the inverse of the effective

temperature $\tilde{\beta}$ is written in E-representation as

$$\rho_{\mathrm{can},t} = \frac{\mathrm{e}^{-\tilde{\beta}H_t}}{Z_{\mathrm{can}}} = \frac{\mathrm{e}^{-\eta_{\mathrm{can},t}\cdot\sigma}}{Z_{\mathrm{can}}} \tag{3.68}$$

where $\eta_{\mathrm{can},t} = \mathrm{Tr}[\sigma\tilde{\beta}H_t]/2$ and

$$Z_{\mathrm{can}} = \mathrm{Tr}[\mathrm{e}^{-\tilde{\beta}H_t}] = 2\cosh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right). \tag{3.69}$$

The partition function Z_{can} is constant in time, since the trace is preserved under the rotation operation. The initial canonical state before the sudden process is the same as the final canonical state because of the cyclic operation. The initial/final state is written in E-representation as

$$\rho_{\mathrm{can},0/T} = \frac{\mathrm{e}^{-\eta_{\mathrm{can},0/T}\cdot\sigma}}{Z_{\mathrm{can}}} \tag{3.70}$$

where $0/T$ means 0 or T, $|\eta_{\mathrm{can},0/T}| = \tilde{\beta}\hbar\omega/2$ and $\hat{\eta}_{\mathrm{can},0/T} = (0,0,1)$. Similarly, the initial/final canonical state is written in M-representation as

$$\rho_{\mathrm{can},0/T} = \frac{1}{2}I - \frac{1}{2}\hat{\eta}_{\mathrm{can},0/T}\cdot\sigma\tanh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right). \tag{3.71}$$

We note that the final canonical state does not depend on the controllable time period T, since $\phi_T = 2\pi$ for any T.

We briefly discuss the isentropic condition to determine the effective temperature, $S[\rho_{\mathrm{can},T}(\tilde{\beta})] = S[\rho_0]$. The entropy of the initial state is calculated by using both E-

and M-representations,

$$\begin{aligned} S[\rho_0] &= -\mathrm{Tr}[\rho_0 \log \rho_0] \\ &= -\mathrm{Tr}[(\frac{1}{2}I - \frac{1}{2}\hat{\eta}_0 \cdot \sigma \tanh|\eta_0|)(-\log(2\cosh|\eta_0|)I - \eta_0 \cdot \sigma)] \\ &= \log(2\cosh|\eta_0|) - |\eta_0| \tanh|\eta_0|. \end{aligned} \tag{3.72}$$

Similarly, the entropy of the final canonical state is

$$S[\rho_{\mathrm{can},T}(\tilde{\beta})] = \log(2\cosh|\eta_{\mathrm{can},T}|) - |\eta_{\mathrm{can},T}| \tanh|\eta_{\mathrm{can},T}| \tag{3.73}$$

where $|\eta_{\mathrm{can},T}| = \tilde{\beta}\hbar\omega/2$. The isentropic condition is simply rewritten as $|\eta_{\mathrm{can},T}| = |\eta_0|$. The effective temperature is determined from the initial condition,

$$\frac{\tilde{\beta}\hbar\omega}{2} = |\eta_0|. \tag{3.74}$$

3.6.3 Optimization of the rotation rate

We consider the quantum mechanical version of the work identity Eq. (2.31) for the effective temperature $\tilde{\beta}$,

$$W = \tilde{\beta}^{-1} D_{KL}[\rho_T \parallel \rho_{\mathrm{can},T}(\tilde{\beta})] - \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] \tag{3.75}$$

where $\Delta F(\tilde{\beta}) = 0$ for a cyclic operation is used. The quantum relative entropy is defined as the following quantum-mechanical analog of the KL divergence,

$$D_{KL}[\rho_A \parallel \rho_B] = \langle \log \rho_A | \rho_A \rangle - \langle \log \rho_B | \rho_A \rangle \tag{3.76}$$

where the bra-ket notation is defined as

$$\langle A|B\rangle = \mathrm{Tr}[AB]. \tag{3.77}$$

We used the same notation $D_{KL}[\rho_A \parallel \rho_B]$ for the quantum relative entropy since Eq. (3.76) is the same as Eq. (2.24).

The extractable work is the negative of the work done on the system,

$$-W = W_{\max} - \tilde{\beta}^{-1} D_{KL}[\rho_T \parallel \rho_{\mathrm{can},T}(\tilde{\beta})] \tag{3.78}$$

where the maximum extractable work is defined as

$$W_{\max} = \tilde{\beta}^{-1} D_{KL}[\rho_0 \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] \geq 0. \tag{3.79}$$

Now we try to find the optimal operation with the angular frequency $\Omega^* = \pi/T^*$ for which the final state satisfies the minimum scaled KL divergence,

$$T^* = \arg\min_T \tilde{\beta}^{-1} D_{KL}[\rho_T \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]. \tag{3.80}$$

where we replaced the final canonical state with the initial canonical state to make clear the absence of T dependence of the final canonical state, since $H_T = H_0$ for any T.

First we consider the validity condition of the GPT $\Delta = 0$ [51], where

$$\Delta = \tilde{\beta}^{-1} D_{KL}[\rho_T \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] - \tilde{\beta}^{-1} D_{KL}[\rho_T \parallel \rho_{T^*}] - \tilde{\beta}^{-1} D_{KL}[\rho_{T^*} \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]. \tag{3.81}$$

The KL divergence may be divided into the sum of the negative entropy and the cross entropy,

$$D_{KL}[\rho_A \parallel \rho_B] = -S[\rho_A] + S_{\mathrm{C}}[\rho_A, \rho_B], \tag{3.82}$$

where the cross entropy is

$$S_{\mathrm{C}}[\rho_A, \rho_B] = -\mathrm{Tr}[\rho_A \log \rho_B]. \tag{3.83}$$

After substituting the above sum into Δ,

$$\begin{aligned}\Delta = \tilde{\beta}^{-1} S_{\mathrm{C}}[\rho_T, \rho_{\mathrm{can},0}(\tilde{\beta})] - \tilde{\beta}^{-1} S_{\mathrm{C}}[\rho_T, \rho_{T^*}] \\ + \tilde{\beta}^{-1} S_{\mathrm{C}}[\rho_{T^*}, \rho_{T^*}] - \tilde{\beta}^{-1} S_{\mathrm{C}}[\rho_{T^*}, \rho_{\mathrm{can},0}(\tilde{\beta})].\end{aligned} \tag{3.84}$$

where we used $S[\rho_{T^*}] = S_{\mathrm{C}}[\rho_{T^*}, \rho_{T^*}]$. Similar to the entropy, the cross entropy between state A and state B is calculated by using both E- and M-representations,

$$\begin{aligned}S[\rho_A, \rho_B] &= -\mathrm{Tr}[\rho_A \log \rho_B] \\ &= \log\left(2\cosh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right)\right) - \frac{\tilde{\beta}\hbar\omega}{2}\tanh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right)\hat{\eta}_A \cdot \hat{\eta}_B\end{aligned} \tag{3.85}$$

where we chose $|\eta_A| = |\eta_B| = \tilde{\beta}\hbar\omega/2$. After substituting the above expression of the cross entropy, we obtain,

$$\Delta = \frac{\hbar\omega}{2}\tanh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right)(\hat{\eta}_{T^*} - \hat{\eta}_T) \cdot (\hat{\eta}_{\mathrm{can},0} - \hat{\eta}_{T^*}). \tag{3.86}$$

Quantum states are represented as vectors such as η_T. The validity condition of the GPT is written as the orthogonality condition $(\hat{\eta}_{T^*} - \hat{\eta}_T) \cdot (\hat{\eta}_{\mathrm{can},0} - \hat{\eta}_{T^*}) = 0$ in the vector space. In information geometry, a state is represented as a vector in a space of parameters. The "distance" (square of length in the sense of the Pythagorean theorem) between two states is measured through the divergence. In nonequilibrium statistical mechanics, the energy-dimensional "distance" between two states is measured through the scaled KL divergence.

Suppose $\hat{\eta}_{T^*} = \hat{\eta}_{\mathrm{can},0}$, then the GPT is valid globally. The vector representation

of the ideally optimal state to extract the maximum work is $\hat{\eta}_{\mathrm{can},0}$. However, if $\hat{\eta}_{0,y} \neq 0$, then $\hat{\eta}_T$ cannot be the ideally optimal state $\hat{\eta}_{\mathrm{can},0}$ for any T in general, since the rotation is limited around the y-axis. Although any realizable final states cannot be the ideally optimal state, we can figure out what is the "closest" (minimum divergence) state from the ideally optimal state in a set of realizable final states. The "closest" state is the experimentally optimal state to extract work and satisfies the local orthogonal condition (see Figure 7).

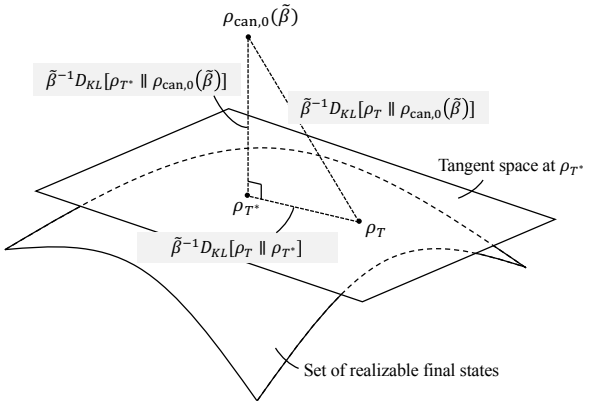

Figure 7: The GPT locally holds when the geodesic line connecting ρ_{T^*} and $\rho_{\mathrm{can},0}(\tilde{\beta})$ is orthogonal to the tangent space at ρ_{T^*}. For any ρ_T in the neighborhood of ρ_{T^*}, ρ_T is considered in the tangent space of ρ_{T^*}.

The validity condition of the local GPT becomes the following local orthogonal condition between the tangent vector at ρ_{T^*} and the geodesic line connecting $\rho_{\mathrm{can},0}(\tilde{\beta})$ and ρ_{T^*},

$$\lim_{T\to T^*} \frac{\Delta}{T^* - T} \propto \frac{d\hat{\eta}_{T^*}}{dT^*} \cdot (\hat{\eta}_{\mathrm{can},0} - \hat{\eta}_{T^*}) = \frac{d\hat{\eta}_{T^*}}{dT^*} \cdot \hat{\eta}_{\mathrm{can},0} \propto \frac{d\hat{\eta}_{T^*,z}}{dT^*} = 0 \tag{3.87}$$

where we used $d(\hat{\eta}_T \cdot \hat{\eta}_T)/dT = d1/dT = 0$ and $\hat{\eta}_{\mathrm{can},0} = (0, 0, 1)$. $\hat{\eta}_{T,z}$ is written as

$$\hat{\eta}_{T,z} = \frac{1}{2}\mathrm{Tr}[\sigma_z(u^T \hat{\eta}_0 \cdot \sigma u^{-T})] \tag{3.88}$$

where

$$u^T = \mathrm{e}^{-\mathrm{i}\pi\sigma_y}\mathrm{e}^{-\mathrm{i}(\omega T\sigma_z - \pi\sigma_y)/2}\mathrm{e}^{\mathrm{i}\pi\sigma_y/2}. \tag{3.89}$$

The calculation of $d\hat{\eta}_{T,z}/dT$ is straightforward but tedious. We show only the final result,

$$\hat{\eta}_{T,z} = \left\{1 - \cos\left(\sqrt{\omega^2T^2 + \pi^2}\right)\right\} \frac{\pi}{\omega^2T^2 + \pi^2}\left(-\omega T\hat{\eta}_{0,y} + \pi\hat{\eta}_{0,z}\right) - \hat{\eta}_{0,z} \tag{3.90}$$

and

$$\begin{aligned}\frac{d\hat{\eta}_{T,z}}{dT} = \frac{2\pi\omega}{\omega^2T^2 + \pi^2}\sin\left(\frac{\sqrt{\omega^2T^2 + \pi^2}}{2}\right)\Bigg\{&\cos\left(\frac{\sqrt{\omega^2T^2 + \pi^2}}{2}\right)\frac{\omega T}{\sqrt{\omega^2T^2 + \pi^2}}(-\omega T\hat{\eta}_{0,y} + \pi\hat{\eta}_{0,z}) \\ &+ \sin\left(\frac{\sqrt{\omega^2T^2 + \pi^2}}{2}\right)\left(-\frac{2\omega T}{\omega^2T^2 + \pi^2}\left(-\omega T\hat{\eta}_{0,y} + \pi\hat{\eta}_{0,z}\right) - \hat{\eta}_{0,y}\right)\Bigg\}\end{aligned} \tag{3.91}$$

There are clearly two ways of solving $d\hat{\eta}_{T^*,z}/dT^* = 0$ that yield two sets of solutions for T^*. One way is when $\sin(\sqrt{\omega^2T^{*2} + \pi^2}/2) = 0$ so that $\omega T^* = \sqrt{4n^2 - 1}\pi$ $(n = 1, 2, ...)$. For these T^*, the scaled KL divergence $\tilde{\beta}^{-1}D_{KL}[\rho_{T^*} \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]$ takes the constant value since $\hat{\eta}_{T^*,z} = -\hat{\eta}_{0,z}$ and the divergence only depends on $\hat{\eta}_{T^*,z}$ i.e.,

$$\tilde{\beta}^{-1}D_{KL}[\rho_{T^*} \parallel \rho_{\mathrm{can},0}(\tilde{\beta})] = \frac{\hbar\omega}{2}\tanh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right)\left(1 - \hat{\eta}_{T^*,z}\right) = \frac{\hbar\omega}{2}\tanh\left(\frac{\tilde{\beta}\hbar\omega}{2}\right)\left(1 + \hat{\eta}_{0,z}\right) \tag{3.92}$$

This value also corresponds to the quasi-static process of $T \to \infty(\Omega \to 0)$ as illustrated in Figures 8 and 9.

Another set of solutions is obtained when the terms inside the curly brackets in

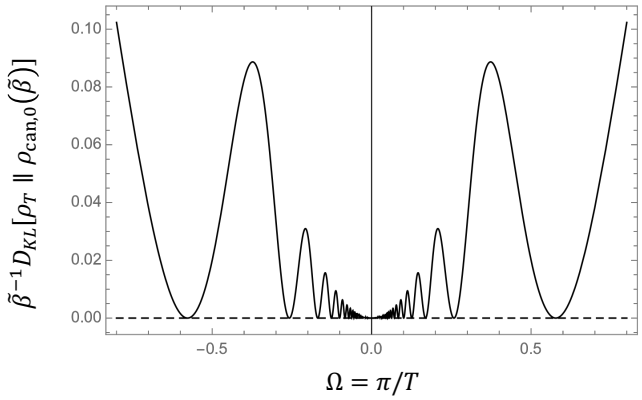

Figure 8: The scaled KL divergence $\tilde{\beta}^{-1}D_{KL}[\rho_T \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]$ versus $\Omega(=\pi/T)$ for $\varphi = 0$ in $\hat{\eta}_0$, $|\eta_0| = 1$ and $\omega = \hbar = 1$. The dashed line shows the level of the divergence for the quasi-static process, $\Omega = 0$. The local minimums are realized for $T = \sqrt{4n^2 - 1}\pi/\omega$ $(n = 1, 2, ...)$ and the local maximums are realized for the second series of solutions.

Eq. (3.91) sum to zero. (In this case a simple expression for T^* is not available.) The scaled KL divergences $\tilde{\beta}^{-1}D_{KL}[\rho_{T^*} \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]$ at these T^* become local maximum or local minimum according to the initial conditions as illustrated in Figures 8 and 9. For $\varphi = 0$ in $\hat{\eta}_0$ (see Figure 8), the divergences of the second series take local maximum. However, for $\varphi = \pi/3$ (see Figure 9), the divergence takes the globally minimum, 0.024792... at $\Omega^* = \pi/T^* = 0.358859...$ that is one of the solutions of the second series. This interesting result tells us that we can extract more work than the quasi-static process by choosing the optimal parameter Ω^* in the case of $\hat{\eta}_{0,y} \neq 0$ [52]. The local GPT or the generalized projection theorem gives us candidates of the realizable optimal state that is "closest" to the ideal optimal state.

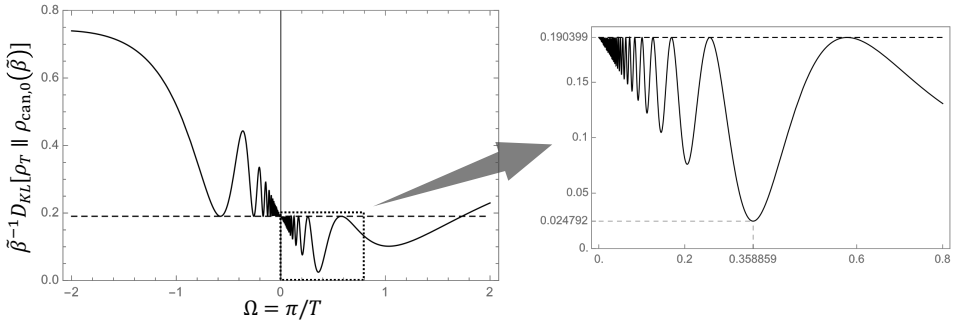

Figure 9: The scaled KL divergence $\tilde{\beta}^{-1}D_{KL}[\rho_T \parallel \rho_{\mathrm{can},0}(\tilde{\beta})]$ versus $\Omega(=\pi/T)$ for $\varphi = \pi/3$ in $\hat{\eta}_0$, $|\eta_0| = 1$ and $\omega = \hbar = 1$. The dashed line shows the level of the divergence for the quasi-static process, $\Omega = 0$. The value of the globally minimum is 0.024792... that is much lower than the level of the quasi-static process, $\Omega = 0$.

4 Concluding remarks

In this thesis, we presented the generalized second law and its alternative information-geometric foundation. The generalized second law for a transition between nonequilibrium states gives a perspective on some situations involving the interplay of energy and information. We have presented the generalization for both thermally isolated systems and systems in contact with a heat reservoir. As shown in Section 2, the former case is more fundamental since a total isolated system may be divided into a system of interest coupled to a larger system that acts as a reservoir.

In Section 3, we introduced "thermodynamic distance" as an energy-dimensional KL divergence scaled by temperature. The change in "thermodynamic distance" gives us the dissipative work. The GPT based on three "thermodynamic distances" gives us a geometric interpretation of the maximum work formulation of the generalized second law. From the nonequilibrium initial state, the canonical state with the

effective temperature is the closest "point" in "thermodynamic distance." Since our new concept of "thermodynamic distance" has an important role in nonequilibrium statistical mechanics, it can also be called "thermodynamic divergence."

The most important result in this paper is that the geometric structure based on the KL divergence scaled by a variable temperature is different from that based on the bare KL divergence. Scaling the divergence with variable temperature completely changes the geometrical structure, such as how to take the coordinate systems and the orthogonal conditions in the sense of the GPT. Conversely, scaled divergences appear by taking different coordinates and convex functions. We would have failed to recognize this result if we had only considered isothermal systems, as many other studies have done. For isothermal processes with constant temperature $\mathcal{T}$, multiplying by temperature simply acts as a constant scale factor. It does not change the geometrical structure. The validity condition of the GPT based on the bare KL divergences is an isoenergetic condition. On the other hand, for a variable temperature, the validity condition of the GPT based on the energy-scaled KL divergences is an isentropic condition.

The geometrical interpretation of the generalized maximum work formulation gives us a systematic method in figuring out a protocol to realize the optimal operation. In this thesis, we applied the geometrical interpretation of the generalized maximum work formulation to a simple two-level quantum system. The optimal cyclic operation to extract work from a nonequilibrium state was determined by minimalizing the scaled KL divergence between the final state and the final canonical state.

In Appendix A, we extended the information geometrical interpretation based on the scaled KL divergence to the case of Tsallis statistics. In the framework of Tsallis's q-functions, Amari-Ohara's q-divergence is a Bregman divergence, and the GPT holds. By scaling this q-divergence with temperature again, a geometrical structure appears in which temperature and q-entropy are dual coordinates. From

the GPT for the scaled q-divergence, the maximum q-work is obtained.

From section 3 onwards, we discussed only adiabatic processes in thermally isolated Hamiltonian systems in order to highlight the information-geometrical foundation of thermodynamics. We excluded any phenomenological assumptions such as the existence of a heat reservoir when considering its relation to information geometry. However, we would still need high and low temperature heat reservoirs to consider a heat engine [33]. Since the information geometry is flexibly applicable for any parametrized probability distributions, the scaled KL divergence would play an important role in the heat engine. One may consider scaling the KL divergence by a quantity with power dimension, such as temperature divided by time. The GPT based on this new scaled KL divergence would give us important knowledge for optimizing the power. We will discuss this problem in the near future.

A Extension to the q-geometry

A.1 Tsallis's and Amari-Ohara's q-divergence

In 1988, Tsallis introduced his q-entropy [38, 40],

$$\tilde{S}_q[\rho] \equiv \frac{1}{1-q}\langle \rho^{q-1} - 1|\rho\rangle. \tag{A.1}$$

The q-entropy satisfies axioms of Shannon–Khinchin of entropy except the additivity. The q-entropy is rewritten as

$$\tilde{S}_q[\rho] = \langle \log_q\left(\frac{1}{\rho}\right)|\rho\rangle \tag{A.2}$$

where we used the q-logarithmic form of Tsallis's functions for convenience. The q-logarithmic function is defined as

$$\log_q(x) \equiv \frac{x^{1-q}-1}{1-q}, \tag{A.3}$$

and its inverse function, the q-exponential function, is defined as

$$\exp_q(x) \equiv \{1 + (1-q)x\}^{\frac{1}{1-q}}. \tag{A.4}$$

In this Appendix, the subscript q always means Tsallis's functions [39, 41]. When $q \to 1$, they become normal logarithmic and exponential functions.

Tsallis introduced his q-divergence as

$$\tilde{D}_q[\rho_A||\rho_B] \equiv -\langle \log_q\left(\frac{\rho_B}{\rho_A}\right)|\rho_A\rangle, \tag{A.5}$$

which is known to be the same as α-divergence up to a constant factor where

$\alpha = 1 - 2q$ [42, 43]. Amari and Ohara, However, introduced their normalized q-divergence later by the conformal transformation of Tsallis's one, which led to the dual flat structure as a Bregman divergence. Amari-Ohara's q-divergence is defined as

$$\begin{aligned} D_q[\rho_A \parallel \rho_B] &= \frac{1}{h_q(\rho_A)} \tilde{D}_q[\rho_A \parallel \rho_B] \\ &= \frac{1}{(1-q)h_q(\rho_A)} \left(1 - \langle \rho_A^q \rho_B^{1-q} \rangle\right) \end{aligned} \tag{A.6}$$

where $h_q(\rho) = \langle 1|\rho^q\rangle$ is a conformal factor. This divergence (A.6) is derived naturally by considering the q-exponential family which consists of q-exponential distributions in the form of

$$\rho_q(x;\boldsymbol{\theta}) = \exp_q\left\{\boldsymbol{\theta}\cdot\mathcal{H}(x) - \psi_q(\boldsymbol{\theta})\right\}. \tag{A.7}$$

By taking $\boldsymbol{\theta}$ as a coordinate and $\psi_q(\boldsymbol{\theta})$ as a convex function and calculating the Bregman divergence

$$D[\boldsymbol{\theta}_A \parallel \boldsymbol{\theta}_B] = \psi_q(\boldsymbol{\theta}_A) - \psi_q(\boldsymbol{\theta}_B) - \nabla\psi_q(\boldsymbol{\theta}_B)\cdot(\boldsymbol{\theta}_A - \boldsymbol{\theta}_B), \tag{A.8}$$

we obtain Amari-Ohara's q-divergence (A.6). Amari-Ohara's q-entropy is also adjusted by the conformal transformation as

$$S_q[\rho] \equiv \frac{1}{h_q(\rho)}\tilde{S}_q[\rho]. \tag{A.9}$$

A.2 maximum q-work formulation

Using the above framework, we will consider the maximum q-work formulation. The q-work is the work with respect to the escort distribution, ρ_{es}, as

$$W_q \equiv \langle H_T|\rho_{\mathrm{es},T}\rangle - \langle H_0|\rho_{\mathrm{es},0}\rangle. \tag{A.10}$$

The escort density is a probability distribution. It is proportional to the density raised to the power of q,

$$\rho_{\mathrm{es}}(x) \equiv \frac{\rho^q(x)}{h_q(\rho)} \tag{A.11}$$

where $h_q(\rho)$ serves as the normalization constant which does not depend on time in Hamiltonian dynamics. By using the escort distribution, Amari-Ohara's q-divergence and q-entropy are rewritten respectively as

$$D_q[\rho_A \parallel \rho_B] = \langle \log_q(\rho_A)|\rho_{\mathrm{es},A}\rangle - \langle \log_q(\rho_B)|\rho_{\mathrm{es},A}\rangle \tag{A.12}$$

and

$$S_q[\rho] = -\langle \log_q \rho|\rho_{\mathrm{es}}\rangle. \tag{A.13}$$

It should be noted that all expectation values change to those with the escort distribution.

In order to calculate this q-work, we first check the correspondence between q-canonical distribution and q-exponential distribution. The q-canonical distribution is

$$\rho_{q-\mathrm{can}}(\alpha) = \frac{1}{Z_q(\alpha)} \exp_q\left(-\alpha H(x)\right), \tag{A.14}$$

which is calculated as follows:

$$\begin{aligned}
\rho_{q-\mathrm{can}}(x;\alpha) &= \frac{1}{Z_q(\alpha)}\left\{1+(1-q)\left(-\alpha H(x)\right)\right\}^{\frac{1}{1-q}} \\
&= \left\{Z_q^{q-1}(\alpha)+(1-q)\left(-\alpha Z_q^{q-1}(\alpha)H(x)\right)\right\}^{\frac{1}{1-q}} \\
&= \left\{1+(1-q)\left(-\alpha Z_q^{q-1}(\alpha)H(x)+\frac{Z_q^{q-1}(\alpha)-1}{1-q}\right)\right\}^{\frac{1}{1-q}} \\
&= \exp_q\left(-\alpha Z_q^{q-1}(\alpha)H(x)+\frac{Z_q^{q-1}(\alpha)-1}{1-q}\right) \\
&= \rho_q(x;\boldsymbol{\theta}_{\mathrm{can}}(\hat{\alpha}))
\end{aligned} \tag{A.15}$$

where in the last line we put following notations:

$$\mathcal{H}^{(0)} = -H(x) \tag{A.16}$$

$$\hat{\alpha} = \alpha Z_q^{q-1}(\alpha) \tag{A.17}$$

$$\boldsymbol{\theta}_{\mathrm{can}}(\hat{\alpha}) = (\hat{\alpha}, 0, \cdots, 0) \tag{A.18}$$

$$\psi_q(\boldsymbol{\theta}_{\mathrm{can}}(\hat{\alpha})) = \frac{1-Z_q^{q-1}(\alpha)}{1-q} = -\log_q(Z_q^{-1}(\alpha)). \tag{A.19}$$

Another way of writing the q-canonical distribution,

$$\rho_{q-\mathrm{can}}(\alpha) = \exp_q\left(\hat{\alpha}(F_q(\alpha)-H(x))\right), \tag{A.20}$$

gives the q-free energy as

$$\begin{aligned}
F_q(\alpha) &= -\frac{1}{\hat{\alpha}}\psi_q(\boldsymbol{\theta}_{\mathrm{can}}(\hat{\alpha})) = \frac{1}{\hat{\alpha}}\log_q(Z_q^{-1}(\alpha)) = \frac{Z_q^{1-q}(\alpha)}{\alpha}\frac{Z_q^{q-1}(\alpha)-1}{1-q} \\
&= \frac{1}{\alpha}\frac{1-Z_q^{1-q}(\alpha)}{1-q} = -\frac{1}{\alpha}\log_q(Z_q(\alpha)).
\end{aligned} \tag{A.21}$$

We then obtain a q-work identity similar to Eq. (2.31),

$$W_q = \Delta F_q(\alpha) + \frac{1}{\hat{\alpha}} D_q[\rho_T \parallel \rho_{q-\mathrm{can},T}(\alpha)] - \frac{1}{\hat{\alpha}} D_q[\rho_0 \parallel \rho_{q-\mathrm{can},0}(\alpha)]. \tag{A.22}$$

The proof of Eq. (A.22) is given by a straightforward calculation:

$$\begin{aligned}
W_q &= \langle H_T|\rho_{\mathrm{es},T}\rangle - \langle H_0|\rho_{\mathrm{es},0}\rangle \\
&= \Delta F_q(\alpha) - \hat{\alpha}^{-1}\left\langle \hat{\alpha}\left(F_{q,T}(\alpha) - H_T\right)|\rho_{\mathrm{es},T}\right\rangle + \hat{\alpha}^{-1}\left\langle \hat{\alpha}\left(F_{q,0}(\alpha) - H_0\right)|\rho_{\mathrm{es},0}\right\rangle \\
&\qquad - \hat{\alpha}^{-1} S_q[\rho_T] + \hat{\alpha}^{-1} S_q[\rho_0] \\
&= \Delta F_q(\alpha) - \hat{\alpha}^{-1}\left\langle \log_q\left(\rho_{q-\mathrm{can},T}(\alpha)\right)|\rho_{\mathrm{es},T}\right\rangle + \hat{\alpha}^{-1}\langle \log_q(\rho_T)|\rho_{\mathrm{es},T}\rangle \\
&\qquad + \hat{\alpha}^{-1}\left\langle \log_q\left(\rho_{q-\mathrm{can},0}(\alpha)\right)|\rho_{\mathrm{es},0}\right\rangle - \hat{\alpha}^{-1}\langle \log_q(\rho_0)|\rho_{\mathrm{es},0}\rangle \\
&= \Delta F_q(\alpha) + \hat{\alpha}^{-1} D_q[\rho_T \parallel \rho_{q-\mathrm{can},T}(\alpha)] - \hat{\alpha}^{-1} D_q[\rho_0 \parallel \rho_{q-\mathrm{can},0}(\alpha)]
\end{aligned} \tag{A.23}$$

where we used the conservation of q-entropy in Hamiltonian system.

The non-negativity of Amari-Ohara's q-divergence gives the inequality

$$W_q \geq \Delta F_q(\alpha) - \frac{1}{\hat{\alpha}} D_q[\rho_0 \parallel \rho_{q-\mathrm{can},0}(\alpha)] \tag{A.24}$$

which holds for any α. If it is a cyclic operation, $\Delta F_q = 0$, the maximum work is

$$W_q \geq -\frac{1}{\hat{\alpha}} D_q[\rho_0 \parallel \rho_{q-\mathrm{can},0}(\alpha)] \equiv \mathcal{W}_{q,LB}(\alpha). \tag{A.25}$$

A.3 The scaled q-divergence

Let us now consider cyclic operations for simplicity. Of course, the argument for non-cyclic operation is essentially the same. We want to find the maximum value of the right-hand side of Eq. (A.25) to get the maximum work that can be achieved,

just as we did in Section 2 and 3. Condition for the derivative to be zero, i.e.,

$$\frac{\partial}{\partial\alpha}\mathcal{W}_{q,LB}(\alpha) = 0, \tag{A.26}$$

gives the iso-q-entropic condition

$$S_q[\rho_{q-\mathrm{can},0}(\beta_{\mathrm{eff}})] = S_q[\rho_0] \tag{A.27}$$

where β_{eff} is the effective inverse temperature.

We, however, derive this iso-q-entropic condition by using the concept of scaled q-divergence and its GPT. We introduce the scaled q-divergence between ρ and $\rho_{q-\mathrm{can}}(\alpha)$ as $\hat{\alpha}^{-1}D_q[\rho \parallel \rho_{q-\mathrm{can}}(\alpha)]$. Note that the scaling variable is $\hat{\alpha}$, not α. The minimization of this scaled q-divergence means the maximization of the q-work in Eq. (A.25).

The GPT based on scaled q-divergences holds as the following theorem:

Theorem 3. *If the iso-q-entropic condition is satisfied, i.e.,*

$$S_q[\rho_{q-\mathrm{can},0}(\beta_{\mathrm{eff}})] = S_q[\rho_0],$$

then the GPT holds,

$$\begin{aligned}&\hat{\alpha}^{-1}D_q[\rho_0 \parallel \rho_{q-\mathrm{can},0}(\alpha)]\\ &\qquad = \hat{\beta}_{\mathrm{eff}}^{-1}D_q[\rho_0 \parallel \rho_{q-\mathrm{can},0}(\beta_{\mathrm{eff}})] + \hat{\alpha}^{-1}D_q[\rho_{q-\mathrm{can},0}(\beta_{\mathrm{eff}}) \parallel \rho_{q-\mathrm{can},0}(\alpha)].\end{aligned} \tag{A.28}$$

Proof. The derivation is straightforward,

$$\begin{aligned}
&\hat{\alpha}^{-1} D_q[\rho_0 \parallel \rho_{q\text{-can},0}(\alpha)] \\
&\qquad - \hat{\beta}_{\text{eff}}^{-1} D_q[\rho_0 \parallel \rho_{q\text{-can},0}(\beta_{\text{eff}})] - \hat{\alpha}^{-1} D_q[\rho_{q\text{-can},0}(\beta_{\text{eff}}) \parallel \rho_{q\text{-can},0}(\alpha)] \\
&= -\hat{\alpha}^{-1} S_q[\rho_0] - F_{q,0}(\alpha) + \langle H_0 | \rho_{\text{es},0} \rangle + \hat{\beta}_{\text{eff}}^{-1} S_q[\rho_0] + F_{q,0}(\beta_{\text{eff}}) - \langle H_0 | \rho_{\text{es},0} \rangle \\
&\qquad + \hat{\alpha}^{-1} S_q[\rho_{q\text{-can},0}(\beta_{\text{eff}})] + F_{q,0}(\alpha) - \langle H_0 | \rho_{\text{es},q\text{-can},0}(\beta_{\text{eff}}) \rangle \\
&= -\hat{\alpha}^{-1} S_q[\rho_0] + \hat{\beta}_{\text{eff}}^{-1} S_q[\rho_0] + F_{q,0}(\beta_{\text{eff}}) + \hat{\alpha}^{-1} S_q[\rho_{q\text{-can},0}(\beta_{\text{eff}})] - \langle H_0 | \rho_{\text{es},q\text{-can},0}(\beta_{\text{eff}}) \rangle.
\end{aligned} \tag{A.29}$$

Substituting the following relation into Eq. (A.29),

$$F_{q,0}(\beta_{\text{eff}}) - \langle H_0 | \rho_{\text{es},q\text{-can},0}(\beta_{\text{eff}}) \rangle = -\hat{\beta}_{\text{eff}}^{-1} S_q[\rho_{q\text{-can},0}(\beta_{\text{eff}})], \tag{A.30}$$

we obtain

$$\begin{aligned}
&\hat{\alpha}^{-1} D_q[\rho_0 \parallel \rho_{q\text{-can},0}(\alpha)] \\
&\qquad - \hat{\beta}_{\text{eff}}^{-1} D_q[\rho_0 \parallel \rho_{q\text{-can},0}(\beta_{\text{eff}})] - \hat{\alpha}^{-1} D_q[\rho_{q\text{-can},0}(\beta_{\text{eff}}) \parallel \rho_{q\text{-can},0}(\alpha)] \\
&= -\hat{\alpha}^{-1} S_q[\rho_0] + \hat{\beta}_{\text{eff}}^{-1} S_q[\rho_0] + \hat{\alpha}^{-1} S_q[\rho_{q\text{-can},0}(\beta_{\text{eff}})] - \hat{\beta}_{\text{eff}}^{-1} S_q[\rho_{q\text{-can},0}(\beta_{\text{eff}})] \\
&= \left(\hat{\beta}_{\text{eff}}^{-1} - \hat{\alpha}^{-1} \right) \left(S_q[\rho_0] - S_q[\rho_{q\text{-can},0}(\beta_{\text{eff}})] \right) \\
&= 0,
\end{aligned} \tag{A.31}$$

where we used the isentropic condition (A.27) in the last line. □

From the non-negativity of q-divergence, Eq. (A.28) gives the following inequality

$$\hat{\alpha}^{-1} D_q[\rho_0 \parallel \rho_{q\text{-can},0}(\alpha)] \geq \hat{\beta}_{\text{eff}}^{-1} D_q[\rho_0 \parallel \rho_{q\text{-can},0}(\beta_{\text{eff}})] \tag{A.32}$$

which is valid for any α. The right-hand-side of Eq. (A.32) provides the greatest lower bound of W_q consistent with the iso-q-entropic condition (A.27).

Theorem 3 shows that the theory in the case of q-divergence is completely parallel to that of the KL divergence which is discussed in Section 3. This is because the natural Bregman divergence based on the q-exponential distribution is Amari-Ohara's q-divergence. As for the geometric structure, we also find that the scaled q-divergence leads to the different geometric structure from bare q-divergence. While the bare q-divergence is a Bregman divergence with $\psi_q = -\hat{\alpha}F_q(\alpha)$ as a convex function, the scaled q-divergence is a Bregman divergence with $-F_q(\alpha)$ as a convex function. Its coordinate system is $\hat{\alpha}^{-1}$ and its dual coordinate system is $-S_q$. This difference is shown in Figs.10 and 11. Figure 10 illustrates the GPT for bare q-divergences

$$D_q[\rho \parallel \rho_{q-\mathrm{can}}(\alpha)] = D_q[\rho \parallel \rho_{q-\mathrm{can}}(\beta)] + D_q[\rho_{q-\mathrm{can}}(\beta) \parallel \rho_{q-\mathrm{can}}(\alpha)] \quad \text{(A.33)}$$

which holds under the isoenergetic condition

$$\langle H|\rho_{\mathrm{es},q-\mathrm{can}}(\beta)\rangle = \langle H|\rho\rangle \quad \text{(A.34)}$$

On the other hand, Figure 11 illustrates the GPT (A.28) for scaled q-divergence.

B Non-negativity of the KL divergence

We consider two probability distributions $\rho(x)$ and $\sigma(x)$. The non-negativity of the KL divergence $D_{KL}[\rho \parallel \sigma]$ is easy to demonstrate using the elementary inequality of $\log x \leq x - 1$ or $-\log x \geq 1 - x$. We have

$$\begin{aligned} D_{KL}[\rho \parallel \sigma] &= -\left\langle \log \frac{\sigma}{\rho} | \rho \right\rangle \geq \left\langle \left(1 - \frac{\sigma}{\rho}\right) | \rho \right\rangle \\ &= \langle 1|\rho\rangle - \langle 1|\sigma\rangle = 0 \end{aligned} \quad \text{(B.1)}$$

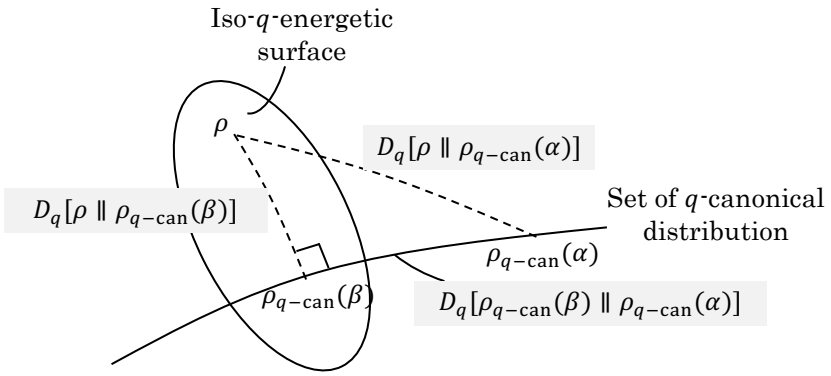

Figure 10: The image of the GPT based on the q-divergences. Subscripts indicating time are omitted.

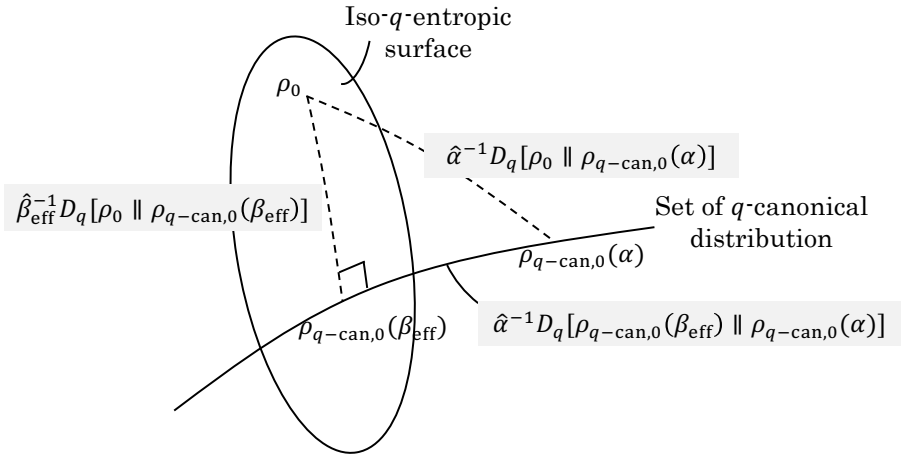

Figure 11: The image of the GPT based on the scaled q-divergences.

Note that equality occurs only when $\sigma(x)/\rho(x) = 1$ for all x, i.e., when the two distributions are identical.

For the quantum case involving density matrices ρ and σ we consider their

orthonormal decompositions: $\rho = \sum_i p_i |\phi_i\rangle\langle\phi_i|$ and $\sigma = \sum_j q_j |\psi_j\rangle\langle\psi_j|$. Then we have

$$\begin{aligned} D_{KL}[\rho \parallel \sigma] &= \mathrm{Tr}\,[\rho \log \rho] - \mathrm{Tr}\,[\rho \log \sigma] \\ &= \sum_i p_i \log p_i - \sum_i p_i \sum_j P_{ij} \log q_j, \end{aligned} \tag{A.4}$$

where $P_{ij} = \langle\phi_i|\psi_j\rangle\langle\psi_j|\phi_i\rangle = |\langle\phi_i|\psi_j\rangle|^2$. It is clear that $P_{ij} \geq 0$, $\sum_i P_{ij} = 1$ and $\sum_j P_{ij} = 1$. (As a matrix P_{ij} is doubly stochastic.)

Now, we let $r_i = \sum_j P_{ij} q_j$. (The r_i is regarded as a probability distribution.) Since the logarithm is a concave function, the logarithm of q_j is less than or equal to the tangential line at $q_j = r_i$, $\log q_j \leq \log r_i + (q_j - r_i)/r_i$. By taking the expectation value with respect to P_{ij} we obtain the Jensen-like inequality, $\sum_j P_{ij} \log q_j \leq \log r_i$. This tells us that

$$D_{KL}[\rho \parallel \sigma] \geq \sum_i p_i \log \frac{p_i}{r_i}. \tag{A.5}$$

The right-hand-side here is proved to be greater than or equal to zero by the same argument we used for the classical probability distribution above. Thus, $D_{KL}[\rho \parallel \sigma] \geq 0$.

Our discussion here largely follows that in Chapter 11 of Nielsen and Chuang [53]. For alternative proofs, one can consult Cover and Thomas [54].

C Concavity of $\mathcal{W}_{LB}$ with respect to the temperature

To determine the effective inverse temperature we use the fact that $\mathcal{W}_{LB}(\alpha)$ has one global maximum as a function of α. This is most easily seen from the concavity of $\mathcal{W}_{LB}$ as a function of the temperature, α^{-1}. To show this we use the facts

(known from thermodynamics and also holding for a canonical distribution) that the derivative of the Helmholtz free energy with respect to temperature is minus the entropy and the derivative of the entropy with respect to temperature is positive.

We have,

$$\begin{aligned}\frac{\partial}{\partial\alpha^{-1}}\mathcal{W}_{LB}(\alpha) &= \frac{\partial}{\partial\alpha^{-1}}F_T(\alpha) + S[\rho_0] \\ &= -S[\rho_{\mathrm{can},T}(\alpha)] + S[\rho_0]. \end{aligned} \tag{C.1}$$

Then,

$$\frac{\partial^2}{(\partial\alpha^{-1})^2}\mathcal{W}_{LB}(\alpha) = -\frac{\partial}{\partial\alpha^{-1}}S[\rho_{\mathrm{can},T}(\alpha)]. \tag{C.2}$$

The derivative of the entropy with respect to temperature is proportional to the heat capacity or equivalently to the variance of the energy, which is positive. Thus we have

$$\frac{\partial^2}{(\partial\alpha^{-1})^2}\mathcal{W}_{LB}(\alpha) < 0, \tag{C.3}$$

which means that $\mathcal{W}_{LB}(\alpha)$ is a concave function of α^{-1}.

Suppose $\partial_{\alpha^{-1}}\mathcal{W}_{LB}(\alpha) = 0$ at $\alpha^{-1} = \tilde{\beta}^{-1}$. The concavity of $\mathcal{W}_{LB}(\alpha)$ tells us that $\partial_{\alpha^{-1}}\mathcal{W}_{LB}(\alpha)$ is a strictly monotonic function (of α^{-1}) so that $\partial_{\alpha^{-1}}\mathcal{W}_{LB}(\alpha)$ cannot be zero for any other value of α^{-1} than $\tilde{\beta}^{-1}$. Since α^{-1} and α are in one-to-one correspondence, $\mathcal{W}_{LB}(\alpha)$ has then only one global maximum as a function of α; the one at $\alpha = \tilde{\beta}$.

D The generalized Pythagorean theorem based on the KL divergence

The importance of the GPT based on the scaled KL divergence is shown in Sections 3 and 4. Here, we will describe the GPT based on the bare (non-scaled) KL divergence for comparison. The geometric structure based on the bare KL divergence is different from the geometric structure based on the scaled KL divergence. We also derive the well-known principle of maximum entropy. While its proof usually just uses the non-negativity of the KL divergence [54,55], it is also be derived from the GPT [17]. In this Appendix, for brevity, we suppress the subscripts representing time.

Using bare KL divergences gives us the following theorem;

Theorem 4. *Under the isoenergetic condition, i.e.,*

$$\langle H|\rho\rangle = \langle H|\rho_{\rm can}(\beta)\rangle, \tag{D.1}$$

the GPT based on the KL divergence holds,

$$D_{KL}[\rho \parallel \rho_{\rm can}(\alpha)] = D_{KL}[\rho \parallel \rho_{\rm can}(\beta)] + D_{KL}[\rho_{\rm can}(\beta) \parallel \rho_{\rm can}(\alpha)]. \tag{D.2}$$

Proof. This theorem is easy to confirm as,

$$\begin{aligned}
&D_{KL}[\rho \parallel \rho_{\rm can}(\alpha)] - D_{KL}[\rho \parallel \rho_{\rm can}(\beta)] - D_{KL}[\rho_{\rm can}(\beta) \parallel \rho_{\rm can}(\alpha)] \\
&= -S[\rho] - \alpha(F(\alpha) - \langle H|\rho\rangle) - \{-S[\rho] - \beta(F(\beta) - \langle H|\rho\rangle)\} \\
&\qquad\qquad - \{\beta(F(\beta) - \langle H|\rho_{\rm can}(\beta)\rangle) - \alpha(F(\alpha) - \langle H|\rho_{\rm can}(\beta)\rangle)\} \\
&= \alpha\langle H|\rho\rangle - \beta\langle H|\rho\rangle - \alpha\langle H|\rho_{\rm can}(\beta)\rangle + \beta\langle H|\rho_{\rm can}(\beta)\rangle \\
&= (\beta - \alpha)\left(\langle H|\rho_{\rm can}(\beta)\rangle - \langle H|\rho\rangle\right) = 0
\end{aligned} \tag{D.3}$$

where we used the isoenergetic condition of (D.1) in the last line. □

The geometric image of the GPT based on KL divergences is illustrated in Figure 12. The (geodesic) line connecting ρ and $\rho_{\mathrm{can}}(\beta)$ on the isoenergetic surface is orthogonal to the parametric line of canonical distributions in terms of the KL divergence.

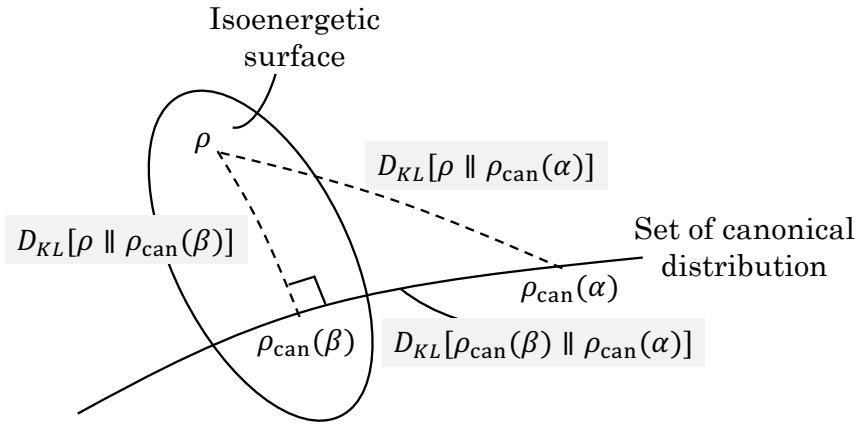

Figure 12: The image of the GPT based on the KL divergences.

By substituting the following expressions of the KL divergences into the GPT, Eq. (D.2),

$$D_{KL}[\rho \parallel \rho_{\mathrm{can}}(\alpha)] = -S[\rho] - \alpha(F(\alpha) - \langle H|\rho\rangle) \tag{D.4}$$

$$D_{KL}[\rho_{\mathrm{can}}(\beta) \parallel \rho_{\mathrm{can}}(\alpha)] = -S[\rho_{\mathrm{can}}(\beta)] - \alpha(F(\alpha) - \langle H|\rho_{\mathrm{can}}(\beta)\rangle), \tag{D.5}$$

Eq. (D.2) is rewritten as

$$\begin{aligned} &-S[\rho] - \alpha(F(\alpha) - \langle H|\rho\rangle) \\ &\quad = D_{KL}[\rho \parallel \rho_{\mathrm{can}}(\beta)] - S[\rho_{\mathrm{can}}(\beta)] - \alpha(F(\alpha) - \langle H|\rho_{\mathrm{can}}(\beta)\rangle). \end{aligned} \tag{D.6}$$

From the isoenergetic condition Eq. (D.1),

$$-S[\rho] = D_{KL}[\rho \parallel \rho_{\mathrm{can}}(\beta)] - S[\rho_{\mathrm{can}}(\beta)]. \tag{D.7}$$

Using the non-negativity of the KL divergence in the right-hand-side, we obtain

$$S[\rho] \leq S[\rho_{\mathrm{can}}(\beta)]. \tag{D.8}$$

This inequality holds for any probability distribution ρ satisfying the isoenergetic condition (D.1). The canonical distribution has the maximum entropy among isoenergetic distributions. This property is the well-known principle of maximum entropy.

References

[1] Tasaki H 2005 Thermodynamics (Baifukan)

[2] Tasaki H 2008 Statistical mechanics I and II, (Baifukan)

[3] Maxwell J C 1888 *Theory of Heat* (London and New York: Longmans) (Reprinted by Dover Publications, New York, 2001)

[4] Leff H S and Rex A F 2003 *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing* (Bristol and Philadelphia: Institute of Physics Publishing)

[5] Maruyama K, Nori F and Vedral V 2009 *Rev. Mod. Phys.* **81** 1–23

[6] Landauer R 1961 *IBM J. Res. Dev.* **5** 183–191

[7] Bennett C H 1987 *Sci. Am.* **257** 108–116

[8] Sagawa T and Ueda M 2008 *Phys. Rev. Lett.* **100** 080403

[9] Sagawa T and Ueda M 2009 *Phys. Rev. Lett.* **102** 250602

[10] Bérut A, Arakelyan A, Petrosyan A, Ciliberto S, Dillenschneider R and Lutz E 2012 *Nature* **483** 187–189

[11] Parrondo J, Horowitz J and Sagawa T 2015 *Nature Physics* **11** 131–139 and references therein

[12] Deffner S and Jarzynski C 2013 *Phys. Rev. X* **3** 041003

[13] Park J, Kim K-H, Sagawa T and Kim S 2013 *Phys. Rev. Lett.* **111** 230402

[14] Bera M, Riera A, Lewenstein M and Winter A 2017 *Nature Communications* **8** 2180

[15] Ren L-H, and Fan H 2017 *Phys. Rev. A* **96** 042304

[16] Amari S and Nagaoka H 2000 *Methods of Information Geometry* (vol 191 of Translations of Mathematical Monographs) (Providence: American Mathematical Society / Oxford: Oxford University Press) Translated from the 1993 Japanese original by Daishi Harada

[17] Amari S 2016 *Information Geometry and Its Applications* (Japan: Springer)

[18] Weinhold F 1975 *J. Chem. Phys.* **63** 2479

[19] Ruppeiner G 1979 *Phys. Rev. A* **20** 1608

[20] Crooks G E 2007 *Phys. Rev. Lett.* **99.10** 100602

[21] Sivak D A and Crooks G E 2012 *Phys. Rev. Lett.* **108.19** 190602

[22] Schlögl F 1985 *Z. Phys. B* **59** 449

[23] Salamon P, Nulton D J and Ihrig E 1984 *J. Chem. Phys.* **80** 436

[24] Salamon P and Berry R S 1983 *Phys. Rev. Lett.* **51** 1127

[25] Ito S 2018 *Phys. Rev. Lett.* **121.3** 030605, and references therein

[26] Ito S, Oizumi M and Amari S 2020 *Phys. Rev. Research* **2.3** 033048, and references therein.

[27] Sekimoto K 1997 *J. Phys. Soc. Jpn.* **66** 1234–1237

[28] Sekimoto K 1998 *Prog. Theor. Phys. Supp.* **130** 17

[29] Seifert U, 2012 *Rep. Prog. Phys.***75(12)** 126001

[30] Ito S 2019 *arXiv preprint* arXiv:1908.09446

[31] Nakamura T, Hasegawa H H and Driebe D J 2019 *J. Phys. Commun.* **3** 015015

[32] Hasegawa H H, Ishikawa J, Takara K and Driebe D 2010 *Phys. Lett. A* **374** 1001–1004

[33] Takara K, Hasegawa H H and Driebe D 2010 *Phys. Lett. A* **375** 88–92

[34] Ishikawa J, Takara K, Hasegawa H H and Driebe D 2014 *Entropy* **16** 3471–3481

[35] Hasegawa H H, Nakamura T and Driebe D 2017 *Chaos* **27** 104606

[36] Esposito M, Lindenberg K and Van den Broeck C 2010 *New J. Phys.* **12** 013013

[37] Esposito M and Van den Broeck C 2011 *Europhys. Lett.* **95** 40004

[38] Tsallis C 1988 *J. Stat. Phys.* **52** 479

[39] Tsallis C, Mendes R S and Plastino A R 1998 *Physica A* **261** 534

[40] Abe S 2000 *Phys. Lett. A* **275** 250

[41] Abe S and Okamoto Y (ed) 2001 *Nonextensive Statistical Mechanics and Its Applications* (Springer)

[42] Suyari H 2006 *Physica A* **368** 63

[43] Suyari H and Wada T 2008 *Physica A* **387** 71

[44] Ohara A 2007 *Phys. Lett. A* **370** 184

[45] Amari S and Ohara A 2011 *Entropy* **13** 01170

[46] de Abreu R and Guerra V 2012 *Am. J. Phys.* **80** 627

[47] Ribeiro W, Landi G and Semiao F 2016 *Am. J. Phys.* **84** 948

[48] When the initial state is a canonical state, the effective temperature is the temperature of the canonical state from the isentropic condition. The divergence in the right-hand-side of Eq. (2.37) vanishes so that no work is extractable for any cyclic operation. This argument is applicable to an isothermal system obtained by dividing a thermally isolated total system into the system of interest and a reservoir with a canonical distribution as discussed in section 2.3. We note that the microcanonical state is usually regarded as the equilibrium state in a thermally isolated system. As is well known though, the expectation value of any physical observable for a canonical state is approximated by that of the corresponding microcanonical state in a many-body thermodynamic system in the sense of the law of large numbers.

[49] Batalhao T *et al.* 2014 *Phys. Rev. Lett.* **113** 140601

[50] Messiah A 2014 *Quantum mechanics: Two volumes bound as one.* (New York: Dover)

[51] The temperature has already been fixed as the effective temperature so the scale factor is common for three divergences in Eq. (3.81). Thus, the validity condition of the GPT of the scaled KL divergences is the same as that of the bare KL divergences here.

[52] For $\varphi = 0$ the state is the ground state after the sudden change at $t = 0$. The former set of operations including the quasi-static operation keeps the ground state. The maximum work is realized for the former set of operations. For $\varphi = \pi/3$, the state has higher energy than the ground state after the sudden change at $t = 0$. The former set of operations including the quasi-static operation keep this higher energy. On the other hand some quick operations in the latter set induce transitions to make the energy of the state lower. As a

result, we can extract more work than the quasi-static process by the optimal operation in the latter set.

[53] Nielsen M and Chuang I 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)

[54] Cover T and Thomas J 2006 *Elements of Information Theory* (Hoboken: Wiley-Interscience)

[55] Pauli W 2000 *Pauli lectures on physics. vol. 4: Statistical mechanics* (MIT press, Cambridge 1973 and reprinted by Dover in 2000)