Ann. Henri Poincaré 22 (2021), 1561–1627

1424-0637/21/051561-67
published online January 7, 2021
https://doi.org/10.1007/s00023-020-01011-2

Annales Henri Poincaré

Asymptotic Behavior of a Sequence of Conditional Probability Distributions and the Canonical Ensemble

Yu-Chen Cheng, Hong Qian and Yizhe Zhu

Abstract. The probability distribution of a function of a subsystem conditioned on the value of the function of the whole, in the limit when the ratio of their values goes to zero, has a limit law: It equals the unconditioned marginal probability distribution weighted by an exponential factor whose exponent is uniquely determined by the condition. We apply this theorem to explain the canonical equilibrium ensemble of a system in contact with a heat reservoir. Since the theorem only requires analysis at the level of the function of the subsystem and reservoir, it is applicable even without the knowledge of the composition of the reservoir itself, which extends the applicability of the canonical ensemble. Furthermore, we generalize our theorem to a model with strong interaction that contributes an additional term to the exponent, which is beyond the typical case of approximately additive functions. This result is new in both physics and mathematics, as a theory for the Gibbs conditioning principle for strongly correlated systems. A corollary provides a precise formulation of what a temperature bath is in probabilistic terms.

Mathematics Subject Classification. 60F05, 60F10, 82B05, 82B30.

1. Introduction

The canonical ensemble with mechanical energy distribution in an exponential form is the centerpiece of equilibrium statistical mechanics. It represents a *weight* for a microstate of a system in thermal equilibrium with its surrounding *heat bath* at a fixed temperature, where the bath is usually considered much larger in comparison. The theory has wide applications from condensed matter

H.Q. is partially supported by the Olga Jung Wan Endowed Professorship. Y.Z. is partially supported by NSF DMS-1949617.

physics to biophysical chemistry [4,11]. In textbooks, there are currently two heuristic justifications for the exponential factor. One is the original derivation by Boltzmann in 1877 based on an ideal gas [31], another is based on the notion of a large heat bath and a small system within, extensively discussed by Gibbs in his 1902 *magnum opus* [17]. After an extensive discussion of the properties of an *invariant measure* including demonstrating it has to be a function of the mechanical energy; however, Gibbs did not attempt to derive the canonical distribution; rather he simply stated that an exponential form "seems to represent the most simple case conceivable."

Boltzmann's derivation was based on the idea of *most probable frequency* under the constraint of given total energy. In the process he recognized the entropy $S = -N \sum_i f_i \log f_i$ from the multinomial distribution, where N is the number of gas molecules, and i represents a distinct molecule state with kinetic energy e_i. This derivation preceded both the modern theory of large deviations [8,37] as well as the principles of maximum entropy (MaxEnt) championed by Jaynes [20,29]. In terms of the contraction principle in the former, Boltzmann computed the large-deviation rate function for a sample frequency conditioned on a given sample mean of energy instead of obtaining the rate function for the random variable. This approach has now been made rigorous under the heading of the *Gibbs conditioning principle* [8,33]. MaxEnt, on the other hand, plays a pivotal role in information theory and machine learning [1,19]. In the 1980s, Boltzmann's logic was also rigorously developed into providing a connection between maximum entropy and conditional probability [39,43].

Gibbs' theory for the canonical distribution was based on the concept of *heat bath*. In [17], he noted that the distribution with the exponential form had "the property that when the system consists of parts with separate energies, the laws of the distribution in the phase of the separate parts are of the same nature." Having energy E_A for the microstate A of the small system and E_B for the microstate B of the heat bath, Gibbs assumed the phase-space distributions follow (i) additivity: $P(A, B) = P(A + B)$ (ii) independence: $P(A, B) = P(A)P(B)$. Under those two assumptions, the only possible probability distribution for A is exponential: $P(A) \propto e^{\lambda E_A}$. Furthermore, all small systems in contact with the same bath share the same parameter λ, which means they are of the "same nature." By assuming that every small system follows the *conjugate distribution laws* (a family of single parameter exponential priors), Khinchin [21] rigorously proved Gibbs' assertion of the common λ and further showed that it is determined by the given total energy.

The aim of the paper is to find a rigorous origin of the exponential weight itself for the canonical distribution from the standpoint of a heat bath. We were inspired by a very widely used derivation in standard statistical physics textbooks—based on Taylor's expansion of the entropy function of a heat bath [18,22,26]. The present work formulates this approach rigorously in probabilistic terms and then gives a proof. We indeed have obtained a rather general new mathematical theorem. The results can be applied back to particular scenarios in statistical physics under corresponding assumptions. Our theorems have clarified the notions of additivity, independency, and the vague "same natures

of systems." The last is actually a corollary of the existence and uniqueness of a single parameter in the exponential form of the canonical distribution, and independency is equivalent to additivity of energy functions of two systems during the map from a phase space to its corresponding energy space. We shall emphasize that independency of two systems is a special case of our theorem; the parameter then only depends on fluctuations of the heat bath but independent of the small system.

Our results are obtained based on two mathematical ideas: *conditional probability* and *asymptotics*. We use a *Gedankenexperiment* to illustrate the crucial role of the former—*conditional probability*—in our theorems: Let $Z := X + Y$, where X is a random variable for some function (e.g., energy) of a subsystem and Y describes the same quantity in the heat bath. If one is only interested in the *static statistics* of X, there is a way to set up an experiment: Let $Z(t)$ be a fluctuating total mechanical energy as a function of time, and its distribution has a support on $D \subseteq \mathbb{R}^+$, but one selects only those measurements for $X(t)$ that simultaneously have $Z(t) \in I \subseteq D$. In the language of mathematics, this thought experiment is about the conditional probability of $X(t)$ conditioned on the event $Z(t) \in I$. Why is this thought experiment regarding conditional probability very much in line with the physicist's picture of a canonical ensemble? The answer is in the idea of time-scale separation, which involves three different time scales. The first time scale is for the subsystem $X(t)$ to reach its equilibrium, the second time scale is to restrict the total system $Z(t)$ to be fluctuating inside a finite interval I, and the third time scale is for $Z(t)$ to reach its equilibrium. And the first one is much shorter than the second one, which is much shorter than the third one. Based on this framework of time-scale separation, the canonical ensemble is the statistical ensemble that represents the possible outcomes of the system of interest on the second time scale, i.e., when the subsystem has reached its equilibrium, but the total system is still "constrained" in a certain interval.

In fact, having its own stationary distribution of the total system (if it evolves long enough) is very significant for the theory of conditional probability for two reasons: (1) knowing the fluctuation of the large system is necessary to define the conditional probability mathematically and (2) to perturb the given condition of the total system to see how it has effects on the subsystem is the essence of our theory of the canonical distribution. In other words, even though the original problem is only about the behavior of $X(t)$ when $Z(t) \in I$, if we have more information of $Z(t)$ outside of I, we are able to seek a deeper understanding of the original problem. Not only for the canonical ensemble, this idea of treating a given constraint (parameter) as a variable with distribution has also been widely used in many other fields, for example, comparing the quenched and annealed invariance principles for the random conductance model [3], and in studying the initial-condition naturalness in the case of statistical mechanics [41].

Mathematically using conditional probability to understand Gibbs measure has a long history, see Lanford [23], Vasicek [40], Georgii [16], and Touchette

[38]. In particular, on the basis of Boltzmann's logic, using asymptotic conditional probability to describe the canonical ensemble has been well established through the Gibbs conditioning principle [8,33]. More discussion of this is provided in Sect. 2 for a contradistinction with our own work. In brief, the Gibbs conditioning principle addresses this question: Given a set $A \in \mathbb{R}$ and a constraint $Z_n \in A$, what are the limit points of the conditional probability

$$P(X_1 \leq x \mid Z_n \in A) \quad \text{as } n \to \infty \text{ ?} \tag{1.1}$$

In Eq. (1.1), $Z_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i are independent and identically distributed random variables (i.i.d. random variables). We can identify that (1.1) is very similar to our setup for the canonical distribution if we consider $Z_n := \frac{X_1}{n} + Y_n$, where $Y_n = \frac{1}{n}\sum_{i=2}^{n} X_i$ is the measurable function of the heat bath in our approach. However, Y_n in our setup could be defined in a much more general way: we only require that Y_n converges to some random variable Y in distribution (or the law of Y_n satisfies a large deviation principle) rather than has a special form as the sum of independent and identically distributed random variables. In other words, the present work is not a simple refinement of the Gibbs conditioning principle. Here, we give a concrete example to which our theorems can be applied but not the Gibbs conditioning principle: Let $\tilde{\zeta}_n = \xi_1 + \eta_n$ and $\eta_n = \sum_{i=2}^{n} \xi_i$, where $\{\xi_i\}_{i=1}^{n}$ are strongly correlated and not identically distributed, and let ζ_n be $\tilde{\zeta}_n$ with appropriate shifting and scaling such that ζ_n has a limiting distribution (or satisfies a large deviation principle). Subject to these conditions, the Gibbs conditioning principle would not be applicable to find the limit points of the conditional probability

$$P(\xi_1 \leq x \mid \zeta_n \in A) \quad \text{as } n \to \infty. \tag{1.2}$$

The present work will show that the canonical distribution in this non-i.i.d. example could still exist as a good approximation (Corollary 3.1) or the limiting distribution (Corollaries 3.2, 3.3) of the conditional probability (1.2). In fact, the setup for our theorems is very general in statistical mechanics: (i) a subsystem in contact with a relatively large heat bath, which is including but not limited to the model of a sum of many independent and identical subsystems, and (ii) the subsystem and the heat bath can have weak or strong interaction.

Back to Eq. (1.1), it seems that either using the Gibbs conditioning principle or using our approach to derive the canonical distribution, both sides are asking a very similar question: what is the asymptotic behavior of a conditional probability? However, based on the more general setup of the conditional probability, our approach to the asymptotic behavior of this conditional probability is very different from the Gibbs conditioning principle. For the Gibbs conditioning principle, it transforms the original problem to a sampling problem: what are the limit points of

$$\mathbb{E}\,[L_n \mid L_n \in \Gamma] \quad \text{as } n \to \infty \text{ ?} \tag{1.3}$$

In Eq. (1.3), $L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ is the corresponding empirical measure for Z_n and $\Gamma = \{\gamma : \int x\gamma(dx) \in A\}$ is the corresponding constraint of Z_n. In fact, even though this approach is called the "Gibbs" conditioning principle, its

logic exactly follows Boltzmann's derivation of the canonical ensemble. As a consequence of the Gibbs conditioning principle, it provides a mathematical foundation of why using the maximum entropy principle with certain constraint works to find the canonical distribution [39,43].

On the other hand, our approach is direct to find the asymptotic behavior of conditional probability (1.1) on the basis of two things: (i) a measurable function of the subsystem is asymptotically small relative to the function of the whole and (ii) the distribution of the measurable function of the heat bath converges to a limiting distribution by appropriate shifting and scaling. Intuitively, under this framework, the distribution of the measurable function of the subsystem shall consist of its unconditional distribution and a weight from a linear approximation of the limiting distribution of the measurable function of the heat bath. As we mentioned above, our approach follows Gibbs' theory for the canonical distribution, which involved the idea of "heat bath" that contributes a "bias" to the system. The common point of our approach and the Gibbs conditional principle is that both sides started with a very similar question of fundamental importance in statistical mechanics and adopted the concept of conditional probability to describe that problem. However, the method of solving the problem on each side has a very different philosophy, the Gibbs conditional principle is about counting statistics by Boltzmann's logic, and ours is inspired by the idea of a heat bath from Gibbs.

Besides the conditional probability, we adopt a very powerful mathematical technique in our theory: asymptotics. Indeed, asymptotics is not only a mathematical technique but also the essence of statistical mechanics. The purpose of statistical mechanics is to derive equilibrium properties of a macroscopic system with enormous numbers of molecules N and occupying a very large volume V, then that macroscopic equilibrium thermodynamics is an *emergent phenomenon* in the limiting case when $N \to \infty$ and $V \to \infty$. Following on from this concept, we shall show that the emergence of an exponential factor in the canonical ensemble is also a result of a limit law according to the probability theory. Take an analogy, our limit theorem is to the exponential form of the canonical distribution what the central limit theorem is to a normal distribution. As with every limit theorem, we have to define how our assumptions depend on n carefully. In our work, as n increases, a measurable function of the subsystem becomes "relatively small" compared with the total system. Based on this main assumption, we obtain two significant results: (i) For a sufficiently large n, a conditional distribution can be well approximated by its unconditional distribution weighted by an exponential factor, and (ii) a sequence of conditional distributions converges to a limit which is the unconditional distribution weighted by a unique exponential factor.

We obtain two theorems regarding the first result in Sect. 3.2, and they provide the existence of the canonical distribution when a system is contained in a finitely large total system (n is sufficiently large). Furthermore, we obtain two limit theorems regarding the second result in Sect. 3.3, and they provide the existence of a unique canonical distribution when the system is contained in an infinitely large total system ($n \to \infty$). In comparison with Sect. 3.3,

Sect. 3.2 only requires weaker conditions, but the exponential form in the canonical distribution may not be unique since there could be more than one sequence having the same asymptotic behavior. On the other hand, Sect. 3.3 requires stronger conditions, but it gives us a unique canonical distribution in the limit, and this distribution can be applied back to approximate the conditional probabilities for all finitely large n. This result can be regarded as an example that the limit theorems from probability predict the laws of nature. Here, we would like to quote from Anderson [2] "Starting with the fundamental laws and a computer, we would have to do two impossible things—solve a problem with infinitely many bodies, and then apply the result to a finite system—before we synthesized this behavior." Our idea echoes Anderson's view: To find the limiting behavior of a sequence of conditional probability distributions and apply it back to the distribution of a subsystem contained in a finitely large total system with some fluctuations, and this is how it is used as a scientific theory.

1.1. The Equivalence of Ensembles

Our work is another way to consider the theory of equivalence of ensembles. As far as we know, Khinchin's derivation of the canonical ensembles in 1949 [21] for a subsystem of a large isolated system by a local central limit theorem was the origin of the equivalence of ensembles. Then, Dobrushin and Tirozzi in 1977 [12] extended Khinchin's result from a classical idea gas to a Gibbs random field. In 1979, Martin-Löf [25,26] further related the microcanonical, canonical, and grand canonical ensembles in the thermodynamic limit when the volume of classical lattice systems tends to infinity. In the 1990s, beyond the scale of the central limit theorem, Deuschel et al. [10] and Georgii [15] showed the equivalence of ensembles on the scale of the large deviation principle. Tasaki [35] recently established the equivalence on the level of local states for large but finite quantum spin systems. A comprehensive introduction to infinite-volume Gibbs measures can be found in Chapter 6 in the textbook by Friedli and Velenik [14], and the discussion of the equivalence of ensembles is in Section 6.14.1.

Recently, a full survey of the equivalence of ensembles at the levels of thermodynamic, macrostates, and measures was presented by Touchette [38]. We shall note that discussions on the equivalence of ensembles at the thermodynamics level can also be traced back to the textbook on statistical mechanics by Hill [36]. In the book, Hill showed the thermodynamic equivalence of ensembles for systems having only a single most probable energy value. In Touchette's recent work, the equivalence was extended to other macrostates, e.g., the mean magnetization of a spin system. This extension was given by the superposition of a mixture of microcanonical ensembles to represent the canonical ensemble of macrostates. Under certain conditions, the equivalence at the macrostate level and the equivalence at the measure level are equivalent. In the language of modern probability, the correspondence between the equivalence at the macrostate level and the equivalence at the measure level

is by the Portmanteau theorem [5] for equivalent statements of weak convergence of measures. We shall emphasize that the conditional distribution of the state of a small subsystem converging to the canonical distribution becomes a corollary of the equivalence of ensembles between the microcanonical ensemble and canonical ensemble at the measure level based on the assumption that the state of a small subsystem is chosen at random with a *uniform distribution* in the large whole system.

The essential difference between our approach and the previous approaches for equivalence of ensembles is that we don't assume a uniform distribution of the state of a small subsystem in the large container. This assumption is equivalent to say that the heat bath (the large container − the subsystem) has to be considered as identical copies of the subsystems, which was usually given in the previous work for classical ideal gas systems or Gibbs random fields. In contradistinction to this assumption, our theorems treat the subsystem and its heat bath as two random variables via a measurable function, i.e., we only care about the effect of the "whole" heat bath on the subsystem with respect to that function.

We want to indicate that applying our mathematical theory to physics is new and original since it extends the applicability of the canonical ensemble: by the pushforward measure (up to a prefactor) via a measurable function, we can derive the canonical distribution of a subsystem without assuming a uniform structure of the whole system. For example, we can apply our results to approximate the distribution of certain measurable functions of a small defect within a material. We only require the subsystem (the defect) is small relative to its heat bath with respect to the values of the measurable functions, which is different from treating the heat bath as infinitely large n copies of the subsystem, interacting or not, in order to apply the microcanonical ensemble to the canonical ensemble. In biophysics, our theory can predict the distribution of side-chain conformational variations in protein structure [6,27,42]. Proteins in general have a non-uniform structure, so the canonical distribution of side-chain conformational variations can be justified by our theory but not the other approaches based on a uniform structure of the whole system.

We further generalize our theorem to a model when a subsystem and its heat bath have *strong interaction* (the function is not additive), which is beyond the weakly interacting system (the function is approximately additive). This result is new in both physics and mathematics, as a theory for the *Gibbs conditioning principle* for strongly correlated systems. The present work formulates our theory rigorously in probabilistic terms in Sect. 3, and then gives a proof in Sect. 4.

In Sect. 5.1, we apply our theory to concrete examples in statistical mechanics, under two situations when a subsystem and its heat bath are independent or strongly correlated. Since our theory also provides a sharp and precise bound of the convergence rate of conditional probabilities, we use it to approximate the conditional Poisson distribution in Sect. 5.2. To build a connection with the equivalence of ensembles using the techniques of the large deviation principle (LDP) [24] and the central limit theorem (CLT) [12], we

applied our theory back to particular scenarios when the heat bath can be treated as a sum of identical random variables. The LDP in Sect. 5.3 or the CLT in Sect. 5.4 gives us a convergence of a sequence of random variables for the heat bath. Nevertheless, we want to emphasize that our theory does not require the LDT or the CTL in general. By proper scaling and shifting, if there exists a convergence of the heat-bath random variable with a smooth limiting distribution, our theory is still applicable. In Sect. 5.5, we provide a precise formulation of what a temperature bath is in probabilistic terms.

1.2. Organization of the Paper

We provide some useful theorems and definitions and explain our motivation in this problem in Sect. 2. In Sect. 3, we state and explain our main results. Proofs of the main results are provided in Sect. 4. In Sect. 5, we present several applications of our main theorems.

Notations. Throughout the paper, we will adopt the notations $a_n = o(b_n)$ when $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$, and $a_n = O(b_n)$ when $|a_n/b_n|$ is bounded by some constant $C > 0$.

For a set Ω, we use $C(\Omega)$ to represent the set of all continuous real functions on Ω, $C_b(\Omega)$ to represent the set of all bounded continuous functions on Ω, and $C^k(\Omega)$ to represent the set of all functions with continuous derivatives of order k on Ω.

We sometimes use brief notations of probabilities in our proofs, e.g., $P_{X_n|Z_n}(x; I) = P(X_n = x \mid Z_n \in I)$. We always use X_n, Y_n, Z_n to denote sequences of random variables, whose definitions might change in different theorems, but we will give their exact definitions before stating the theorems.

2. Preliminaries

2.1. Maximum Entropy and Conditional Probability

We first recall the following classical results. Here, we don't specify the regularity conditions in the statements of the two theorems below. For more details, see the original references.

Theorem 2.1 [43]. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent and identically distributed (i.i.d.) random variables with continuous density $f(x)$, then under appropriate regularity conditions, we have*

$$\lim_{n\to\infty} P\left(X_1 \le x \mid S_n = n\mu + c_n\right) = P\left(X_1 \le x\right), \tag{2.1}$$

where $S_n := X_1 + X_2 + \cdots + X_n$, $\mu := \mathbb{E}[X_1]$, $s_n^2 := Var[S_n]$, and $c_n = O(s_n)$.

Theorem 2.2 [39]. *Let $\{X_n\}_{n\in\mathbb{N}}$ and S_n follow definitions in Theorem 2.1. Let $\alpha \in \mathbb{R}$ and let $f(x)$ be the density function of X_1, then under appropriate regularity conditions,*

$$\lim_{n\to\infty} P\left(X_1 \le x \mid S_n = n\alpha\right) = \left(\int_{-\infty}^{x} e^{\lambda s} f(s)\mathrm{d}s\right) / c(\lambda), \tag{2.2}$$

where

$$c(\lambda) = \mathbb{E}\left[e^{\lambda X_1}\right] < \infty \quad \textit{and} \quad \alpha = \left(\int x e^{\lambda x} f(x)\mathrm{d}x\right)/c(\lambda). \tag{2.3}$$

Note that the parameter λ is determined by the constraint

$$\alpha = \left(\int x e^{\lambda s} f(x)\mathrm{d}x\right)/c(\lambda), \tag{2.4}$$

and the density $g(x) = e^{\lambda x} f(x)/c(\lambda)$ maximizes the entropy relative to the density $f(x)$ of X_1 given by

$$H(X_1) = -\int g(x) \log \frac{g(x)}{f(x)}\mathrm{d}x, \tag{2.5}$$

with respect to the constraint that

$$\left(\int x g(x)\mathrm{d}x\right) = \alpha. \tag{2.6}$$

We see that Theorem 2.1 implies the convergence of the conditional probability distribution of X_1 to its unconditional distribution. In this case, the sum of X_i is conditioned on the scale of Gaussian fluctuations: $S_n = n\mu + c_n$, where $n\mu$ is the mean of S_n and c_n is in the order of standard deviation of S_n. On the other hand, we see that Theorem 2.2 implies the convergence of the conditional probability distribution of X_1 to the (normalized) product of its unconditional distribution and the maximal entropy distribution $e^{\lambda x}$. The parameter λ is determined by the condition $S_n = n\alpha$, which is on the scale of large deviations when $\alpha \neq \mathbb{E}[X_1]$.

Theorem 2.2 is a particular case of the *Gibbs conditioning principle*, which is the meta-theorem [9] regarding the conditional probability of X_i given on the *empirical measure* of an i.i.d. $\{X_i\}_{i=1}^n$

$$L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \tag{2.7}$$

belonging to some rare event such as

$$\int x L_n(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \alpha \quad \text{and} \quad \alpha \neq \mathbb{E}[X_1]. \tag{2.8}$$

Using the empirical measure defined in (2.7) conditioned on the rare event (2.8) to find the limit of conditional probability in Theorem 2.2 turns out to be equivalent to find the limit

$$\gamma^* := \lim_{n\to\infty} \mathbb{E}\left[L_n \mid L_n \in \Gamma\right], \quad \Gamma = \left\{\gamma : \int x\gamma(\mathrm{d}x) = \alpha\right\}. \tag{2.9}$$

By the Gibbs conditioning principle, under appropriate regularity conditions, γ^* minimizes the relative entropy

$$H(\gamma \mid \mu_X) := \int \mathrm{d}\gamma \log\left(\frac{\mathrm{d}\gamma}{\mathrm{d}\mu_X}\right),$$

where $\gamma \in \Gamma$ and μ_X is the law of X_1. In fact, this result implies the limit law derived in Theorem 2.2.

One of the most successful approaches to the Gibbs conditioning principle is through the theory of large deviations [9,33]. This approach involves *Sanov's theorem* [30] that provides the large-deviation rate function of the empirical measure induced by a sequence of i.i.d. random variables and the *contraction principle* [13] that describes how continuous mappings preserve the large deviation principle from one space to another space. In short, these theorems regarding counting and transformation in the theory of large deviations yield the Gibbs conditioning principle and provide the foundation of using the maximum entropy distribution under certain constraints to find the limit of a sequence of conditional probabilities.

2.2. Large Deviation Theory

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. absolutely integrable (i.e., $\mathbb{E}|X_1| < \infty$) real random variables with mean $\mu := \mathbb{E}[X_1]$, and let

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.10}$$

By the *weak law of large numbers*,

$$\overline{X}_n \xrightarrow{P} \mu \quad \text{when } n \to \infty. \tag{2.11}$$

That is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0. \tag{2.12}$$

To study the question how fast this probability tends to zero, Harald Cramér obtained the following theorem in 1938:

Theorem 2.3 (Cramér's theorem [7]). *Assume that*

$$A(\lambda) := \log \mathbb{E}\left[e^{\lambda X_1}\right] < \infty, \quad \lambda \in \mathbb{R}.$$

Then,

$$(i)\ \lim_{n\to\infty} \frac{1}{n} \log P\left(\overline{X}_n \geq y\right) = -\phi(y) \quad \textit{when } y > \mu,$$
$$(ii)\ \lim_{n\to\infty} \frac{1}{n} \log P\left(\overline{X}_n \leq y\right) = -\phi(y) \quad \textit{when } y < \mu,$$

where ϕ is defined by

$$\phi(y) := \sup_{\lambda\in\mathbb{R}} [y\lambda - A(\lambda)] \quad \textit{for } x \in \mathbb{R}. \tag{2.13}$$

The function A is called the *logarithmic moment generating function*. In the applications of the large deviation theory to statistical mechanics, A is also called the *free energy function* and the function ϕ is called the rate function of large deviations [37]. We can recognize that $\phi(y)$ is the *Legendre transform*

of $A(\lambda)$ (A is a convex function). Therefore, $\phi = A^*$ (the convex conjugate of A) and it leads to the following pair of reciprocal equations

$$\frac{\mathrm{d}A(\lambda)}{\mathrm{d}\lambda} = y \quad \text{if and only if} \quad \frac{\mathrm{d}\phi(y)}{\mathrm{d}y} = \lambda. \tag{2.14}$$

Now, we can apply this equivalence (2.14) to Theorem 2.2: The parameter λ of the maximum entropy distribution $e^{\lambda s}$ is implicitly solved by (2.4), which gives rise to λ determined by

$$\frac{\mathrm{d}\log \int e^{\lambda s} f(s)\mathrm{d}s}{\mathrm{d}\lambda} = \alpha. \tag{2.15}$$

By the definition of $A(\lambda)$ and (2.14) and (2.15), we have

$$\frac{\mathrm{d}A(\lambda)}{\mathrm{d}\lambda} = \alpha \quad \text{if and only if} \quad \frac{\mathrm{d}\phi(\alpha)}{\mathrm{d}\alpha} = \lambda. \tag{2.16}$$

Therefore, this result (2.16) shows that λ not only can be determined implicitly by the free energy function A but also can be founded explicitly by the rate function ϕ.

One of our main theorems (Theorem 3.4) can be applied to a particular type of heat bath as the sum of i.i.d. random variables (Theorem 5.3), then we directly show that λ is uniquely determined by the first derivative of the rate function ϕ given on the condition α. In this case, we apply the large deviation principle directly to the distribution of the heat bath

$$Y_n = \frac{1}{n}\sum_{i=2}^{n} X_i$$

rather than use the large deviation principle for the empirical measure

$$L_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

In fact, the former (our approach) actually follows Gibbs' logic of the canonical distribution through the heat bath method; The later (Gibbs conditioning principle) follows Boltzmann's logic of the canonical distribution through counting statistics. The reason to call the "Gibbs" conditioning principle was in order to comprehend Gibbs' prediction of the canonical distribution from a mathematical standpoint [33]; however, in our opinion, it is closer to the idea of Boltzmann's derivation of the canonical distribution.

From our perspective, choosing the maximum entropy distribution to approximate the conditional probability is a natural consequence of the emergence of $e^{\lambda x} f(x)$ when the finite subsystem is contained in an infinitely large system with a value far from its mean. In other words, (normalized) $e^{\lambda x} f(x)$ is the density of the limit of a sequence of conditional probabilities and it maximizes the relative entropy (2.5) as an inevitable corollary from the setup of the heat bath method. In comparison with the Gibbs conditioning principle, our logic provides a very different point of view of why the maximum entropy principle works to find the limit of conditional probabilities. Even though these two

approaches have very different philosophies, in terms of mathematics, they are connected by the reciprocal equations (2.14) through the Legendre transform.

2.3. Asymptotic Behavior of Probabilities

In order to define how "good" of an approximation of conditional probability is, we first need to decide which metric we would use in the space of measures. In what follows, let Ω denote a measurable space with σ-algebra $\mathcal{F}$ and let $\mathbb{P}$, $\mathbb{Q}$ denote two probability measures on $(\Omega, \mathcal{F})$.

Definition 2.1 (*KL-divergence*). For two probability distributions of a continuous random variable, $\mathbb{P}$ and $\mathbb{Q}$, the *KL-divergence* is defined by

$$D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q}) := \int_{-\infty}^{+\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x, \tag{2.17}$$

where p, q are the density functions of $\mathbb{P}, \mathbb{Q}$, respectively. For two probability distributions of a discrete random variable, $\mathbb{P}$ and $\mathbb{Q}$, the Kullback–Leibler divergence between them can be written as

$$D_{\mathrm{KL}}\left(\mathbb{P} \parallel \mathbb{Q}\right) = \sum_{k \in \Omega} P(k) \log \left(\frac{P(k)}{Q(k)} \right), \tag{2.18}$$

where P, Q are the probability mass functions of $\mathbb{P}, \mathbb{Q}$, respectively, and Ω is a countable space. By continuity arguments, the convention is assumed that $0 \log \frac{0}{q} = 0$ for $q \in \mathbb{R}$ and $p \log \frac{p}{0} = \infty$ for $p \in \mathbb{R} \backslash \{0\}$. Therefore, the KL-divergence can take values from zero to infinity.

Definition 2.2 (*total variation*). The total variation distance between two probability measures $\mathbb{P}, \mathbb{Q}$ on a sigma-algebra $\mathcal{F}$ is defined by

$$\delta(\mathbb{P}, \mathbb{Q}) := \sup_{A \in \mathcal{F}} |\mathbb{P}(A) - \mathbb{Q}(A)|.$$

It's well known that we have the following relation between KL-divergence and total variation by Pinsker's inequality [28]:

$$\delta(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{Q})}. \tag{2.19}$$

Definition 2.3 (*convergence of measures in total variation*). Given the above definition of total variation distance, let $\{\mathbb{P}_n\}_{n \in \mathbb{N}}$ be a sequence of measures on $(\Omega, \mathcal{F})$. The sequence is said to converge to a measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ in total variation distance if

$$\lim_{n \to \infty} \delta(\mathbb{P}_n, \mathbb{P}) = 0$$

and it is equivalent to

$$\lim_{n \to \infty} \sup_{\|f\|_\infty \leq 1} \left| \int f \mathrm{d}\mathbb{P}_n - \int f \mathrm{d}\mathbb{P} \right| = 0.$$

Definition 2.4 (*weak convergence of measures*). Let $\{\mathbb{P}_n\}_{n\in\mathbb{N}}$ be a sequence of probability measures on $(\Omega, \mathcal{F})$. We say that $\mathbb{P}_n$ converges *weakly* to a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ if

$$\lim_{n\to\infty} \int f \mathrm{d}\mathbb{P}_n = \int f \mathrm{d}\mathbb{P},$$

for all $f \in C_b(\Omega)$.

From the two definitions above, total variation convergence of measures always implies weak convergence of measures.

Definition 2.5 (*convergence in distribution*). A sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables is said to *convergence in distribution* to the random variable X if

$$\mu_{X_n} \to \mu_X \quad \text{weakly},$$

where μ_{X_n} is the law of X_n and μ is the law of X.

Even though the KL-divergence is not a metric, by the inequality (2.19), if the KL-divergence of one sequence of measures from another sequence of measures converges to zero, then the two sequences of measures have to converge to zero in total variation. So they must converge to zero weakly. Following this line of implication, in the present work, we start with defining the KL-divergence between two sequences of measures then understand what conditions guarantee it converges to zero. Once we have that, we will attain both strong convergence and weak convergence of the two sequences of measures to zero under those conditions.

Furthermore, we mention two classical theorems (see reference [32]) regarding the convergence of probability distributions which we will use in our proofs.

Theorem 2.4 (Berry–Esseen theorem). *Let X have mean zero, $\mathbb{E}[X^2] = \sigma^2$, and $\mathbb{E}|X|^3 < \infty$. Let $Z_n = (X_1 + \cdots + X_n)/\sqrt{n}\sigma$, where $X_1, \ldots, X_n$ are i.i.d. copies of X. Then, we have*

$$|P(Z_n < z) - P(G < z)| = O\left(\frac{\mathbb{E}|X|^3}{\sqrt{n}}\right) \tag{2.20}$$

for all $z \in \mathbb{R}$, where $G \sim N(0,1)$.

Theorem 2.5 (Slutsky's theorem). *Let $\{Z_n\}_{n\in\mathbb{N}}, \{W_n\}_{n\in\mathbb{N}}$ be sequences of random variables. If Z_n converges in distribution to a random variable X and W_n converges in probability to a constant c, then*

$$Z_n + W_n \to X + c \quad \textit{in distribution.} \tag{2.21}$$

Corollary 2.1. *Let X have mean zero, $\mathbb{E}[X^2] = \sigma^2$, and $\mathbb{E}|X|^3 < \infty$. For some finite $k \in \mathbb{N}$, let $W_n = (X_1 + \cdots + X_k)/\sqrt{n}\sigma$ and $Z_n = (X_{k+1} + \cdots + X_{n+k})/\sqrt{n}\sigma$, where $X_1, \ldots, X_{n+k}$ are i.i.d. copies of X. Let $\tilde{Z}_n = Z_n + W_n$, then we have*

$$\tilde{Z}_n \to G \quad \textit{in distribution}, \quad G \sim N(0,1). \tag{2.22}$$

Furthermore,

$$\left|P\left(\tilde{Z}_n < z\right) - P\left(G < z\right)\right| = O\left(\frac{\mathbb{E}|X|^3}{\sqrt{n}}\right) \tag{2.23}$$

for all $z \in \mathbb{R}$.

This corollary follows from Theorems 2.4 and 2.5. The proof is provided in Appendix 6.2.

3. Main Results

3.1. Setup

In Sect. 1 of introduction, we have already provided our philosophy of adopting the *conditional probability* to derive the canonical ensemble. In this section of the main results, we are going to rigorously show: when a measurable function of the subsystem is "small" relative to the whole system, the "canonical distribution" is a "good" approximation of that conditional distribution. For the sake of simplicity, we will use the terms: "subsystem," "heat bath," and "whole system" to represent a measurable function of those systems, respectively. Within this framework, we first need to define three things rigorously:

1. A relatively small subsystem.
2. Canonical probability distributions.
3. Good approximations.

For the definition of (1): a relatively small subsystem, we consider a sequence of conditional densities

$$f_{X|\tilde{Z}_n}(x; E_n), \quad E_n := \mu_n + I/\beta_n, \tag{3.1}$$

where $\tilde{Z}_n := X + \tilde{Y}_n$, X is a nonnegative continuous random variable and $\tilde{Y}_n$ is a sequence of continuous random variables, I is a finite interval and μ_n, β_n are positive sequences. Note that we here use $\tilde{Y}_n, \tilde{Z}_n$ instead of Y_n, Z_n because we will do transformations for $\tilde{Y}_n, \tilde{Z}_n$ later, so Y_n, Z_n will be used to define transformed $\tilde{Y}_n, \tilde{Z}_n$. The formula of E_n is to represent two kinds of transformations that we can do for the interval I: μ_n is the parameter of shifting and β_n is the parameter of scaling. Through different combinations of μ_n and β_n, the given condition of $\tilde{Z}_n$ will be on certain significant scales. For two examples,

1. Assume $\mu_n := \mathbb{E}[\tilde{Z}_n] = n\mu$, μ is a constant and $\beta_n = 1/\sqrt{n}$, then $\tilde{Z}_n$ is conditioned to be inside the interval $E_n = n\mu + \sqrt{n}I$. The interval E_n is then around $\mathbb{E}[\tilde{Z}_n]$ with a scale of the Gaussian fluctuations in central limit theorem.
2. Assume $\beta_n = 1/n$, then $\tilde{Z}_n$ is conditioned to be inside the interval $E_n = n\mu + nI$. The interval E_n is then around $\mathbb{E}[\tilde{Z}_n]$ with a scale of the large deviations.

In our theorems, we will assume that

$$\mathbb{E}[X^j] < \infty, \text{ for some finite } j, \quad \text{and} \quad \beta_n = o(1). \tag{3.2}$$

Therefore, the definition (3.1) of conditional densities is a sequence of densities for the nonnegative continuous random variable X with $\mathbb{E}[X^j] < \infty$ conditioned on the event $\tilde{Z}_n \in E_n$ with $E_n \to \infty$ ($\beta_n \to 0$). In this way, the positive sequence β_n characterizes that the subsystem is relatively "small" to the given condition of the whole system.

Then, we will extend our definition of a "small" subsystem to the case when we have discrete random variables. Consider a sequence of conditional probability functions

$$P\big(K = k \mid \tilde{H}_n \in E_n\big), \quad E_n = \mu_n + I/\beta_n, \tag{3.3}$$

where $\tilde{H}_n := K + \tilde{L}_n$, K is a nonnegative discrete random variables and we assume that

$$\mathbb{E}[K^j] < \infty, \text{ for some finite } j, \quad \text{and} \quad \beta_n = o(1), \tag{3.4}$$

and $\tilde{L}_n$ is a sequence of discrete random variables and $\tilde{H}_n := K + \tilde{L}_n$.

For the definition of (2): canonical probability distributions, we are introducing a general form of the canonical probability distribution as follows: Let I be the interval in the setup (3.1). We consider a sequence of functions $\zeta_n : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$, where $\mathcal{A}$ is the set of all finite intervals on $\mathbb{R}$. For the canonical probability distribution of a nonnegative continuous random variable X, its density can be represented by

$$\frac{f_X(x)e^{-\zeta_n(I;x)x}}{\displaystyle\int_{\mathbb{R}^+} f_X(x)e^{-\zeta_n(I;x)x}\mathrm{d}x} \quad \text{and} \quad 0 \le \zeta_n(I;x) < \infty, \text{ for all } x \in \mathbb{R}^+. \tag{3.5}$$

Consider a sequence of functions $\hat{\zeta}_n : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$. For the canonical probability distribution of a nonnegative discrete random variable K, it can be represented by

$$\frac{P(K=k)e^{-\hat{\zeta}_n(I;k)k}}{\sum_{k\in S} P(K=k)e^{-\hat{\zeta}_n(I;k)k}} \quad \text{and} \quad 0 \le \hat{\zeta}_n(I;k) < \infty, \text{ for all } k \in S, \tag{3.6}$$

where S is a set of the support of $P(K = k)$.

For the definition of (3): good approximations, "good" is defined by a sufficiently small distance of two distributions in total variation (2.19). In most of our results, we prove that two sequences of distributions converge to zero in KL-divergence, by Pinsker's inequality, it implies those two sequences converge to zero in total variation, i.e., one sequence is a good approximation of the other one.

3.2. Approximation of Conditional Probabilities

Based on the definitions of (1), (2), and (3) in the setup, we provide two approximation theorems to show the existence of the canonical distributions

as good approximations of conditional distributions when the subsystem is sufficiently small relative to the whole systems.

Based on the setup (3.1), let $X_n := \beta_n X$ and take $j = 2$ for the assumption (3.2), i.e.,

$$\mathbb{E}[X^2] < \infty, \quad \text{and} \quad \beta_n = o(1). \tag{3.7}$$

Let $a_n := \beta_n^2 \mathbb{E}[X^2]$, hence we have that

$$\mathbb{E}[X_n^2] = a_n, \ a_n = o(1). \tag{3.8}$$

Let $Y_n := \beta_n \left(\tilde{Y}_n - \mu_n\right)$ and $Z_n := X_n + Y_n$. Note that Y_n, Z_n are the linear transformations of $\tilde{Y}_n, \tilde{Z}_n$, respectively, and recall that $\tilde{Z}_n = X + \tilde{Y}_n$ and the parameters of the transformation, β_n, μ_n, are from $E_n = \mu_n + I/\beta_n$ in the conditional density (3.1). Since we assume I is a finite interval in (3.1), we can define it explicitly as $I = [h, h+\delta], \ h, \delta \in \mathbb{R}$ and $\delta > 0$.

Based on the definitions given above, let $\mathbb{P}_I^{(n)}$ be a sequence of probability measures with density functions

$$\frac{f_X(x) e^{-\beta_n \psi_n(I;\beta_n x)x}}{\displaystyle\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi_n(I;\beta_n x)x} \mathrm{d}x},$$
$$\psi_n(I;\beta_n x) := \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = \beta_n x\big)}{\partial y}\bigg|_{y=h}. \tag{3.9}$$

And let $\mathbb{Q}_I^{(n)}$ be a sequence of probability measures with density functions $f_{X|\tilde{Z}_n}(x; E_n)$.

Our first theorem for continuous random variables is as follows:

Theorem 3.1. *Assume there exist positive constants C_1, C_2, a positive sequence $b_n = o(1)$, and an open interval D such that the following holds:*

1. *For all $x \in \mathbb{R}^+$, $y \in \mathbb{R}$,*

$$\left|\frac{\partial^2 P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y^2}\right| \le C_1,$$
$$\left|\frac{\partial^2 \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y^2}\right| \le C_2. \tag{3.10}$$

2. *For all $x \in \mathbb{R}^+$ and every $[y, y+\delta] \subset D$, there exist positive constants δ_1, C_3 depending on y such that*

$$P\big(Y_n \in [y, y+\delta] \mid X_n = x\big) \ge \delta_1, \quad 0 \le \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y} \le C_3, \tag{3.11}$$

$$\big|P\big(Y_n \in [y, y+\delta] \mid X_n = x\big) - P\big(Y_n \in [y, y+\delta]\big)\big| \le b_n P\big(Y_n \in [y, y+\delta]\big). \tag{3.12}$$

3. For every $[z, z+\delta] \subset D$, there exists a positive constant δ_2 depending on z such that

$$P\big(Z_n \in [z, z+\delta]\big) \geq \delta_2. \tag{3.13}$$

Given an interval $I \subset D$, then

$$D_{KL}\left(\mathbb{P}_I^{(n)} \| \, \mathbb{Q}_I^{(n)}\right) = O(a_n + b_n), \tag{3.14}$$

and $\mathbb{P}_I^{(n)}$ satisfies the definition of the canonical probability distributions in (3.5).

Remark 3.1. By Pinsker's inequality, Theorem 3.1 implies that

$$\delta\left(\mathbb{P}_I^{(n)}, \mathbb{Q}_I^{(n)}\right) = O\left(\sqrt{a_n + b_n}\right).$$

Remark 3.2. Interpretations of Theorem 3.1 for statistical mechanics: the sequence $a_n = o(1)$ represents that the second moment of the function of the subsystem X scaled by the size of the given condition of the whole system asymptotically goes to zero. And the sequence $b_n = o(1)$ represents that X_n and Y_n are asymptotically independent. By our approximation theorem, using the canonical distribution to approximate the conditional distribution results in a very small error $O(\sqrt{a_n + b_n})$ when n is sufficiently large, i.e.,

1. The subsystem is small relative to the whole system.
2. The subsystem has weak interaction with its surrounding.

Note that these conditions (1) and (2) echo the physicist's setup of the canonical ensemble in statistical mechanics.

For Theorem 3.1, we require the condition 3.12 and the sequence b_n in that condition is asymptotic to zero. As Remark 3.2, it means that the subsystem and the heat bath are asymptotically independent. In the following corollary, we are going to extend Theorem 3.1 to the case when the subsystem X_n and its surrounding (the heat bath) Y_n are not asymptotically independent.

Recall that $I = [h, h+\delta]$, $h, \delta \in \mathbb{R}$ and $\delta > 0$ and $\mathbb{Q}_I^{(n)}$ is a sequence of probability measures with density functions $f_{X|\tilde{Z}_n}(x; E_n)$. Let $\hat{\mathbb{P}}_I^{(n)}$ be a sequence of probability measures with density functions

$$\frac{f_X(x) e^{-\beta_n \phi_n(I) x}}{\displaystyle\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \phi_n(I) x} \mathrm{d}x},$$

where

$$\begin{aligned} \phi_n(I) := & \left.\frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = 0\big)}{\partial y}\right|_{y=h} \\ & - \left.\frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = 0\big)}{\partial x}\right|_{y=h}. \end{aligned} \tag{3.15}$$

Corollary 3.1. *Assume there exist positive constants C_1, C_2, and an open interval D such that the following holds:*

1. *For all* $x \in \mathbb{R}^+$, $y \in \mathbb{R}$,
$$\left|\partial^{(2)} P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)\right| \leq C_1,$$
$$\left|\partial^{(2)} \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)\right| \leq C_2, \tag{3.16}$$
where $\partial^{(2)}$ *denotes all the second-order partial derivatives.*
2. *For all* $x \in \mathbb{R}^+$ *and every* $[y, y+\delta] \subset D$, *there exist positive constants* δ_1, C_3 *depending on* y *such that*
$$P\big(Y_n \in [y, y+\delta] \mid X_n = x\big) \geq \delta_1,$$
$$0 \leq \partial^{(1)} \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big) \leq C_3, \tag{3.17}$$
where $\partial^{(1)}$ *denotes all the first-order partial derivatives.*
3. *For every* $[z, z+\delta] \subset D$, *there exists a positive constant* δ_2 *depending on* z *such that*
$$P\big(Z_n \in [z, z+\delta]\big) \geq \delta_2. \tag{3.18}$$

Given an interval $I \subset D$, *then*
$$D_{KL}\left(\hat{\mathbb{P}}_I^{(n)} \| \, \mathbb{Q}_I^{(n)}\right) = O(a_n), \tag{3.19}$$
and $\hat{\mathbb{P}}_I^{(n)}$ *satisfies the definition of the canonical probability distributions in* (3.5).

The proof of Corollary 3.1 basically follows from the proof of Theorem 3.1, and we provide the details of the proof in Appendix 6.4.

Remark 3.3. Here, we want to emphasize the difference between Theorem 3.1 and Corollary 3.1: On the one hand, Corollary 3.1 requires a stronger condition that those partial derivatives in the conditions (3.16) and (3.17) have to be bounded both in the x and y directions; however, Theorem 3.1 only requires that the partial derivatives in the conditions (3.10) and (3.11) are bounded in the y direction. On the other hand, Corollary 3.1 does not require the condition (3.12) in Theorem 3.1, which is to define the asymptotic independence between X_n and Y_n. Based on the difference of those conditions, Theorem 3.1 and Corollary 3.1 give rise to distinct parameters of the exponential factors. The parameter of the exponential factor (3.15) in Corollary 3.1 includes one additional term which involves the partial derivative with respect to x.

Note that the parameter of the exponential factor (3.15) can be rewritten as
$$\phi_n(I) = \frac{\partial \log P\big(Y_n \in [y, y+\delta]\big)}{\partial y}\bigg|_{y=h} + \left(\frac{\partial \log C(x,y)}{\partial y} - \frac{\partial \log C(x,y)}{\partial x}\right)\bigg|_{(x=0, y=h)}, \tag{3.20}$$
where
$$C(x,y) = \frac{P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{P\big(Y_n \in [y, y+\delta]\big)}. \tag{3.21}$$

Corollary 3.1 with the parameter represented by (3.20) has a critical interpretation in statistical mechanics: For a system in contact with a heat bath, if the interaction are not weak (i.e., the correlation in mathematical terms does not approach zero), then the effect of this interaction will appear in the parameter of the exponential factor as the function of $C(x, y)$ in (3.20) for the canonical distribution. This result is different from the standard example in statistical mechanics: in the limit where the interaction goes to zero, the parameter only includes the effect of the fluctuations of heat bath (the first term on the right side of (3.20)) without any effect from the correlation (the second term on the right side of (3.20)).

Now we extend our approximation theorem to discrete random variables based on the setup (3.3). Recall that $\tilde{H}_n = K + \tilde{L}_n$ and $E_n = \mu_n + I/\beta_n$ defined in the conditional probability mass function (3.3). Take $j = 2$ for the assumption (3.4), i.e.,

$$\mathbb{E}[K^2] < \infty, \tag{3.22}$$

and by the definition (3.6), we have a set S such that

$$S := \{k \in \mathbb{R} : P(K = k) > 0\}. \tag{3.23}$$

Let $K_n := \beta_n K$ be a sequence of nonnegative discrete random variables and let $a_n := \beta_n^2 \mathbb{E}[K^2]$. By (3.22) and (3.23), we have that

$$\mathbb{E}[K_n^2] = a_n, \ a_n = o(1), \tag{3.24}$$

and a sequence of sets S_n such that

$$S_n := \{\beta_n k \in \mathbb{R} : P(K_n = \beta_n k) > 0\}. \tag{3.25}$$

By shifting with μ_n and scaling with β_n, we can define a linear transformation of $\tilde{L}_n$, $L_n := \beta_n \left(\tilde{L}_n - \mu_n\right)$, and let $H_n := K_n + L_n$. Furthermore, let Y_n be a sequence of continuous random variables and $Z_n := K_n + Y_n$. Based on the given definitions, our second theorem for discrete random variables is as follows:

Theorem 3.2. *Assume the following conditions hold:*

1. *All conditions in Theorem 3.1 hold for K_n, Y_n, Z_n on an open interval D.*
2. *There exists a set $D' \subset D$ and a positive sequence $c_n = o(1)$ such that for every interval $I' \subset D'$,*

$$\sup_{\beta_n k \in S_n} \left| P\big(K_n = \beta_n k \mid H_n \in I'\big) - P\big(K_n = \beta_n k \mid Z_n \in I'\big) \right| = O(c_n). \tag{3.26}$$

Given an interval $I \subset D'$, then

$$\sup_{k \in S} \left| P\big(K = k \mid \tilde{H}_n \in E_n\big) - B_n P(K = k) e^{-\beta_n \hat{\psi}_n(I;\beta_n k)k} \right| = O\left(c_n + \sqrt{a_n + b_n}\right), \tag{3.27}$$

where

$$\frac{1}{B_n} := \sum_{k\in S} P(K=k)e^{-\beta_n \psi_n(I;\beta_n k)k} \quad \textit{and}$$

$$\hat{\psi}_n(I;\beta_n k) := \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid K_n = \beta_n k\big)}{\partial y}\bigg|_{y=h},$$

and b_n is defined in condition (3.12) of Theorem 3.1. Furthermore,

$$B_n P(K=k)e^{-\beta_n \hat{\psi}_n(I;\beta_n k)k} \tag{3.28}$$

satisfies the definition of the canonical probability distribution in (3.6).

Note that the given assumption (1) in Theorem 3.2: all conditions in Theorem 3.1 hold for K_n, Y_n, Z_n on an open interval D, in which K_n is corresponding to X_n in Theorem 3.1, and all conditions defined for "all $x \in \mathbb{R}^+$" in Theorem 3.1 become defined for "all $x \in S_n$" for Theorem 3.2. In this way, even K_n is a sequence of discrete random variables, all conditions in Theorem 3.1 are well defined.

Remark 3.4. In Theorems 3.1 and 3.2, X and K are defined as a nonnegative random variable. In the following two points, we extend our approximation theorem to the case when X (or K) is bounded from below (shifting property) and the case when X (or K) is a nonpositive random variable (reflection property):

1. (Shifting property) Let X be a continuous random variable bounded from below. By change of variables, let $\hat{X_n} := \beta_n(X - C)$, where C is the finite lower bound, since $\beta_n = o(1)$, we still have
$$\mathbb{E}[\hat{X}_n^2] = o(1). \tag{3.29}$$
In addition, assume that the conditional probability
$$P\big(Y_n \in [y, y+\delta] \mid \hat{X}_n = x\big)$$
satisfies all of the conditions in Theorem 3.1, then we can apply Theorem 3.1 to obtain the canonical distribution for X. We call this the *shifting property* of the canonical distributions. For the discrete random variable K, its canonical probability distribution has this property as well. This *shifting property* can be interpreted as the extension of the cases restricted to nonnegative quantities (e.g., energy and number of molecules) for the canonical ensemble and the grand canonical ensemble in statistical mechanics: the canonical distribution can be generalized to represent the possible values of a function which is *bounded from below* of the subsystem in thermal equilibrium with the heat bath at a *positive temperature* (In Theorem 3.1, we choose the condition I such that $0 \le \psi_n(I;\beta_n x) < \infty$).
2. (Reflection property) Let X be a nonpositive continuous random variable. Assume the condition (3.11) in Theorem 3.1 becomes
$$P\big(Y_n \in [y, y+\delta] \mid X_n = x\big) \ge \delta_1,$$

$$-C_3 < \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y} \leq 0, \tag{3.30}$$

for all $x \in \mathbb{R}^-$. And assume all of the other conditions in Theorem 3.1 are satisfied, then Theorem 3.1 can be applied to an interval $I = [h, h+\delta] \subset D$ such that $-\infty < \psi_n(I; \beta_n x) \leq 0$, for all $x \in \mathbb{R}^-$. We call this *reflection property* of the canonical distributions. For the discrete random variable K, its canonical probability distribution has this property as well. Here is our interpretation of this *reflection property* for statistical mechanics: When a given condition I of the whole system gives rise to a negative parameter ($-\infty < \psi_n(I; \beta_n x) \leq 0$) in the exponential weight of the canonical distribution, our approximation theorem can be applied to the case of a nonpositive function of the subsystem. In combination with this property with the shifting property, the canonical distribution can represent the possible values of a function which is *bounded from above* of the subsystem in thermal equilibrium with the heat bath at a *negative temperature* (Here, we choose the condition I such that $-\infty < \psi_n(I; \beta_n x) \leq 0$).

3.3. Limit Theorems for Conditional Probabilities

In this section, we provide two limit theorems to show that a sequence of conditional distributions converges to a unique canonical distribution by appropriate shifting and scaling, where the convergence is also in a corresponding scaling of the KL-divergence of this sequence of conditional distributions from its limit distribution. In contrast to Sect. 3.2, here we obtain a unique canonical distribution at the appropriate scale when a system is conditioned on an infinitely large total system ($n \to \infty$). It is different from Sect. 3.2 in which we derive the canonical distribution for each finitely large n directly.

Recall that from Sect. 3.2, for a sufficiently large n, we know that $\mathbb{Q}_I^{(n)}$ with density function

$$f_{X|\tilde{Z}_n}(x; E_n)$$

can be well approximated by $\mathbb{P}_I^{(n)}$ with density function

$$\frac{f_X(x) e^{-\beta_n \psi_n(I;\beta_n x)x}}{\displaystyle\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi_n(I;\beta_n x)x} \mathrm{d}x} \quad \text{and}$$

$$\psi_n(I; \beta_n x) := \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = \beta_n x\big)}{\partial y}\bigg|_{y=h}. \tag{3.31}$$

Note that the parameter of the exponential function $\psi_n(I; \beta_n x)$ in (3.31) depends on n and x.

Through our limit theorems in this section, we show that the sequence of measures $\mathbb{Q}_I^{(n)}$ can be well approximated by a unique (sequence of) canonical distribution(s) with density function(s)

$$\frac{f_X(x)e^{-\lambda_n(I)x}}{\int_{\mathbb{R}^+} f_X(x)e^{-\lambda_n(I)x}\mathrm{d}x} \tag{3.32}$$

in one of the cases:

1. $\lambda_n(I) = \beta_n\psi(I)$, where $\beta_n = o(1)$, and $\psi : \mathcal{A} \to \mathbb{R}$ is a function such that $\mathcal{A}$ is the set of all finite intervals on $\mathbb{R}$, and $0 < \psi(I) < \infty$.
2. $\lambda_n(I) = \varphi(I)$, where $\varphi : \mathcal{A} \to \mathbb{R}$ is a function satisfying $0 < \varphi(I) < \infty$.

Note that $\psi(I)$ and $\varphi(I)$ are independent of x and n in comparison with $\psi_n(I;\beta_n x)$ in (3.31). One of the main ideas behind the proof of our limit theorems is as follows: Let $\tilde{\mathbb{P}}_I^{(n)}$ be a sequence of probability measures with density functions (normalized) $f_X(x)e^{-\beta_n\psi(I)x}$, and let $\mathbb{P}_I$ be a probability measure with density function (normalized) $f_X(x)e^{-\varphi(I)x}$. With D_{KL} defined as KL-divergence, Case (1) can be considered as

$$D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right) \to 0 \quad \text{as } n \to \infty; \tag{3.33}$$

Case (2) can be considered as

$$D_{\mathrm{KL}}\left(\mathbb{P}_I \parallel \mathbb{Q}_I^{(n)}\right) \to 0 \quad \text{as } n \to \infty. \tag{3.34}$$

Note that in Case (1), since $\beta_n = o(1)$, the sequence $\lambda_n(I) \to 0$ for any bounded $\psi(I)$. Therefore, we have to scale the distance D_{KL} by some function of β_n to guarantee the uniqueness of $\psi(I)$. More details are provided in Theorem 3.3.

Furthermore, we require stronger conditions than the conditions for (3.31) in order to apply Lemmas 4.2 and 4.3 to the proof of our limit theorems. Here is the essence of those two lemmas: under appropriate regularity conditions, the sequence $\lambda_n(I)$ in (3.32) is uniquely determined by a linear approximation of the following sequence

$$\log\left(\frac{f_{X|\tilde{Z}_n}(x;E_n)}{f_X(x)}\right). \tag{3.35}$$

Therefore, most of the conditions in our limit theorems are required to guarantee that (3.35) is well approximated by a linear function and the remainder term converges to zero fast enough.

Recall that $X_n := \beta_n X$, $Y_n := \beta_n\left(\tilde{Y}_n - \mu_n\right)$, and $Z_n := X_n + Y_n$, where β_n, μ_n are positive sequences and $\beta_n = o(1)$, and $E_n = \mu_n + I/\beta_n$, $I = [h, h+\delta]$, $h, \delta \in \mathbb{R}$ and $\delta > 0$.

Our first limit theorem for Case (1): $\lambda_n(I) = \beta_n\psi(I)$ is as follows

Theorem 3.3. *Consider a function $\psi : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ such that $0 < \psi(I) < \infty$ for the given interval I. Let $\tilde{\mathbb{P}}_I^{(n)}$ be a sequence of probability measures with density functions*

$$\frac{f_X(x)e^{-\beta_n\psi(I)x}}{\int_{\mathbb{R}^+} f_X(x)e^{-\beta_n\psi(I)x}\mathrm{d}x}. \tag{3.36}$$

Assume the following conditions hold:

1. *X is a nonconstant random variable with $\mathbb{E}[X^3] < \infty$ and $\dfrac{f_{X|\tilde{Z}_n}(x; E_n)}{f_X(x)}$ is uniformly bounded on $\mathbb{R}^+$.*
2. *$Y_n \to Y$ in distribution. The distribution function of Y is bounded on $\mathbb{R}^+$ and satisfies*

$$\log P\left(Y \in [y, y+\delta]\right) \in C^2(D) \quad \text{and} \quad 0 < \left.\frac{\partial \log P(Y \in [y, y+\delta])}{\partial y}\right|_h < \infty, \tag{3.37}$$

where D is an open interval containing h.

3. *There exists a sequence of functions $g_n : \mathbb{R} \to \mathbb{R}$ with $\left|g_n(x)e^{-\beta_n \xi x}\right|$ uniformly bounded on $\mathbb{R}^+$ for any $\xi > 0$ and $\mathbb{E}\left[g_n(X)^2\right] \to 0$ such that on $I_n = [0, d_n]$ with $d_n = O\left(\frac{1}{\beta_n}\right)$,*

$$\log\left(\frac{P\left(Y_n \in I - \beta_n x \mid X_n = \beta_n x\right)}{P\left(Z_n \in I\right)}\right) = \log\left(\frac{P\left(Y \in I - \beta_n x\right)}{P\left(Y \in I\right)}\right) + \beta_n g_n(x). \tag{3.38}$$

Then,

$$\lim_{n\to\infty} \frac{D_{KL}\left(\tilde{\mathbb{P}}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right)}{\beta_n^2} = 0 \quad \text{if and only if} \quad \psi(I) = \left.\frac{\partial \log P(Y \in [y, y+\delta])}{\partial y}\right|_h.$$

And $\tilde{\mathbb{P}}_I^{(n)}$ satisfies the definition of the canonical probability distributions in (3.5).

Our second limit theorem for Case (2): $\lambda_n(I) = \varphi(I)$ is as follows

Theorem 3.4. *Let $\varphi : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ be a function such that $0 < \varphi(I) < \infty$ for the given interval I. Let $\mathbb{P}_I$ be a probability measure with density function*

$$\frac{f_X(x)e^{-\varphi(I)x}}{\displaystyle\int_{\mathbb{R}^+} f_X(x)e^{-\varphi(I)x}dx}.$$

Assume the following conditions hold:

1. *X is a nonconstant random variable with $\mathbb{E}[X] < \infty$ and $\dfrac{f_{X|\tilde{Z}_n}(x; E_n)}{f_X(x)}$ is uniformly bounded on $\mathbb{R}^+$.*
2. *$Y_n \to \mu$ in probability, for some constant $\mu \notin I$. The sequence of laws of Y_n satisfies a large deviation principle with speed $1/\beta_n$ and rate function $\phi \in C^2(D)$, where D is an open interval containing I, and $-\infty < \phi'(y) < 0$ for all $y \in I$.*
3. *There exists a sequence of functions $r_n : \mathbb{R} \to \mathbb{R}$ with $\left|r_n(x)e^{-\xi x}\right|$ uniformly bounded on $\mathbb{R}^+$ for any $\xi > 0$ and $\mathbb{E}\left[r_n(X)^2\right] \to 0$ such that on*

$$I_n = [0, d_n] \text{ with } d_n = O\left(\tfrac{1}{\beta_n}\right),$$

$$\log\left(\frac{P\left(Y_n \in I - \beta_n x \mid X_n = \beta_n x\right)}{P\left(Z_n \in I\right)}\right) = \log\left(\frac{\exp\left[-\frac{1}{\beta_n}\phi\left(y^* - \beta_n x\right)\right]}{\exp\left[-\frac{1}{\beta_n}\phi\left(y^*\right)\right]}\right) + r_n(x), \tag{3.39}$$

$$y^* = \left\{y : \inf_{y \in I} \phi(y)\right\}. \tag{3.40}$$

Then,

$$\lim_{n\to\infty} D_{KL}\left(\mathbb{P}_I \parallel \mathbb{Q}_I^{(n)}\right) = 0 \quad \text{if and only if} \quad \varphi(I) = -\phi'(y^*).$$

And $\mathbb{P}_I$ satisfies the definition of the canonical probability distributions in (3.5).

The following is the discussion about the circumstances when the condition (3.38) (or the condition (3.39)) for Theorem 3.3 (or Theorem 3.4) could hold. Here, we only discuss the condition (3.38) but it is applied to the condition (3.39) as well. We can consider three circumstances

1. When $Y_n \to Y$ in distribution.
2. When $X_n \to 0$ in probability.
3. When X_n and Y_n are asymptotically independent.

Even though the condition (3.38) and the combination of circumstances (1)–(3) are not the exact same, they are very close; therefore, these three circumstances provide us an insight regarding three elements of the condition (3.38): Circumstance (1) means the heat bath has a limiting distribution; Circumstance (2) means the subsystem is relatively small in comparison with the whole system; Circumstance (3) means those two systems have weak interaction.

As Corollary 3.1 for the approximation theorem 3.1, we are going to extend our limit theorems to the case when X_n and Y_n are not asymptotically independent.

Corollary 3.2. *Let $G : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$ be a function such that $\mathcal{A}$ is the set of all finite intervals. For the given interval I, $G(I;0) = 1$ and $\log G(I,\xi) \in C^2(\mathbb{R}^+)$ with respect to ξ. Assume the condition* (3.38) *in Theorem 3.3 becomes*

$$\log\left(\frac{P\left(Y_n \in I - \beta_n x \mid X_n = \beta_n x\right)}{P\left(Z_n \in I\right)}\right) = \log\left(\frac{P\left(Y \in I - \beta_n x\right) \cdot G(I;\beta_n x)}{P\left(Y \in I\right)}\right) + \beta_n g_n(x), \tag{3.41}$$

and the other conditions in Theorem 3.3 hold. Then,

$$\lim_{n\to\infty} \frac{D_{KL}\left(\tilde{\mathbb{P}}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right)}{\beta_n^2} = 0 \quad \text{if and only if}$$

$$\psi(I) = \frac{\partial \log P(Y \in [y, y+\delta])}{\partial y}\bigg|_h - \frac{\partial \log G(I;\xi)}{\partial \xi}\bigg|_0.$$

And $\tilde{\mathbb{P}}_I^{(n)}$ *satisfies the definition of the canonical probability distributions in* (3.5).

Remark 3.5. The function G in (3.41) can be considered as an approximation:

$$\frac{P\left(Y_n \in I - \xi \mid X_n = \xi\right)}{P\left(Y_n \in I - \xi\right)} \approx G(I;\xi),$$

in which the left side is equivalent to the joint probability of X_n, Y_n divided by the products of their marginal probabilities. Therefore, G could represent an estimation of the correlation of X_n and Y_n; in information theory, the function G is closely related to the mutual information between X_n and Y_n.

Corollary 3.3. *Let* $R : \mathcal{A} \times \mathbb{R} \to \mathbb{R}$ *be a function, for the given interval* I, $R(I;0) = 1$ *and* $\log R(I,\xi) \in C^2(\mathbb{R}^+)$ *with respect to* ξ. *Assume the condition* (3.39) *in Theorem* 3.4 *becomes*

$$\log\left(\frac{P\left(Y_n \in I - \beta_n x \mid X_n = \beta_n x\right)}{P\left(Z_n \in I\right)}\right) = \log\left(\frac{\exp\left[-\frac{1}{\beta_n}\phi\left(y^* - \beta_n x\right)\right] \cdot \left(R(I;\beta_n x)\right)^{\frac{1}{\beta_n}}}{\exp\left[-\frac{1}{\beta_n}\phi\left(y^*\right)\right]}\right) + r_n(x), \tag{3.42}$$

$$y^* = \left\{y : \inf_{y \in I} \phi(y)\right\}, \tag{3.43}$$

and the other conditions in Theorem 3.4 *hold. Then,*

$$\lim_{n\to\infty} D_{KL}\left(\mathbb{P}_I \parallel \mathbb{Q}_I^{(n)}\right) = 0 \quad \textit{if and only if} \quad \varphi(I) = -\phi'(y^*) - \left.\frac{\partial \log R(I;\xi)}{\partial \xi}\right|_0.$$

And $\mathbb{P}_I$ *satisfies the definition of the canonical probability distributions in* (3.5).

Remark 3.6. The function R in (3.42) can be considered as an approximation:

$$\frac{P\left(Y_n \in I - \xi \mid X_n = \xi\right)}{P\left(Y_n \in I - \xi\right)} \approx \left(R(I,\xi)\right)^{\frac{1}{\beta_n}}.$$

In comparison with Corollary 3.2, when the sequence of laws of Y_n satisfies a large deviation principle with speed $1/\beta_n$, the correlation of the subsystem and its heat bath has to be in $O(R^{\frac{1}{\beta_n}})$ to contribute an additional term in the parameter of the exponential weight. Otherwise, if the correlation is just in $O(R)$ as the order in Corollary 3.2, then it has no influence on the canonical distribution.

The proof of Corollary 3.2 (Corollary 3.3) basically follows from the proof of Theorem 3.3 (Theorem 3.4). We provide the details of proof in Appendix 6.4.

As our approximation theorems in Sect. 3.2, we can extend our limit theorems to discrete random variables, random variables bounded below, and random variables bounded above as follows:

1. Discrete random variables: Theorems 3.3 and 3.4 can also be applied to the case when we have a nonnegative discrete random variable K, a sequence of discrete random variables $\tilde{L}_n$, and $\tilde{H}_n := K + \tilde{L}_n$. It is said that the sequence of conditional probabilities $P(K = k \mid \tilde{H}_n \in E_n)$ has a limit (by appropriate scaling)
$$\frac{P(K = k)e^{-\lambda_n(I)k}}{\sum_{k\in S} P(K = k)e^{-\lambda_n(I)k}}. \tag{3.44}$$
The case of $\lambda_n(I) = \beta_n\psi(I)$ follows from Theorem 3.3; The case of $\lambda_n(I) = \varphi(I)$ follows from Theorem 3.4. Furthermore, the probability function (3.44) satisfies the definition of the canonical probability distribution in (3.6).
2. Random variables bounded below: As Remark 3.4, we can extend those limit theorems to the case when X is bounded below. By change of variable, let $\hat{X}_n := \beta_n(X - C)$, where C is the finite lower bound, we still have
$$\mathbb{E}[(X - C)^j] < \infty,\ j = 1 \text{ or } 3. \tag{3.45}$$
Note that $j = 3$ is for Theorem 3.3 and $j = 1$ is for Theorem 3.4. In addition, assume
$$\log\left(\frac{P\left(Y_n \in I - \beta_n x \mid X_n = \beta_n x\right)}{P\left(Z_n \in I\right)}\right)$$
satisfies the condition of linear approximation in (3.38) and (3.39), for Theorems 3.3 and 3.4, respectively. Then, we can apply those limit theorems to obtain a unique canonical distribution of X. Therefore, as the point (1) in Remark 3.4, a unique canonical distribution derived by the limit of a sequence of conditional distributions has the "shifting property." For the discrete random variable K, its unique canonical distribution has this property as well.
3. Random variables bounded above: Let X be a nonpositive continuous random variable and the corresponding canonical distribution be a sequence of distributions with density functions
$$\frac{f_X(x)e^{-\lambda_n(I)x}}{\displaystyle\int_{\mathbb{R}^-} f_X(x)e^{-\lambda_n(I)x}\mathrm{d}x}, \quad -\infty < \lambda_n(I) < 0. \tag{3.46}$$
When $\lambda_n(I) = \beta_n\psi(I)$, Theorem 3.3 can be applied to an interval I such that $-\infty < \psi(I) < 0$; When $\lambda_n(I) = \varphi(I)$, Theorem 3.4 can be applied to an interval I such that $-\infty < \varphi(I) < 0$. Therefore, as the point (2) in Remark 3.4, a unique canonical distribution derived by the limit of a sequence of conditional distributions has the "reflection property." For the discrete random variable K, its unique canonical distribution has this property as well. This reflection property provides us an explanation of the possibility of *negative temperature*: For some given condition of the whole system which arises a *negative* parameter ($-\infty < \lambda_n(I) < 0$) in the exponential weight, a unique canonical distribution for a function

bounded from above of the subsystem emerges as the limit of a sequence of conditional distributions.

4. Proofs of Main Results

4.1. Proofs of Theorems 3.1 and 3.2

4.1.1. Proof of Theorem 3.1.

Proof. We first prove for the case: $\{x : f_{X_n}(x) > 0\} = \mathbb{R}^+$. In this case, $P(Z_n \in I \mid X_n = x)$ is well defined for all $x \in \mathbb{R}^+$. Let

$$I = [h, h+\delta] \subseteq D, \quad I - x := \{y - x : \ y \in I\},$$

with condition (3.13): for $I \subseteq D$, $P\big(Z_n \in [h, h+\delta]\big) \geq \delta_2$, we can derive the following conditional density by Bayes' theorem

$$\begin{aligned} f_{X_n|Z_n}(x; I) &= \frac{f_{X_n}(x) P(Z_n \in I \mid X_n = x)}{P(Z_n \in I)} \\ &= \frac{f_{X_n}(x) P(Y_n \in I - x \mid X_n = x)}{P(Z_n \in I)}, \quad \text{for } x \in \mathbb{R}^+. \end{aligned} \tag{4.1}$$

Note that $P(Y_n \in I - x \mid X_n = x) = P(Y_n \in [h - x, h + \delta - x] \mid X_n = x)$. Define

$$G_\delta(y, x) := P\big(Y_n \in [y, y+\delta] \mid X_n = x\big).$$

By Taylor expansion and condition (3.10), we can expand $G_\delta(h-x, x)$ at (h, x) to get

$$\begin{aligned} G_\delta(h - x, x) = G_\delta(h, x) &- \frac{\partial G_\delta(h, x)}{\partial y} x \\ &+ \frac{\partial^2 G_\delta(h - \alpha_n x, x)}{2 \partial y^2} x^2, \quad \text{for some } \alpha_n \in (0, 1). \end{aligned}$$

It implies that

$$\begin{aligned} &P\big(Y_n \in [h - x, h + \delta - x] \mid X_n = x\big) \\ &\quad = P\big(Y_n \in [h, h+\delta] \mid X_n = x\big) \\ &\qquad - \frac{\partial P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y}\bigg|_{y=h} \cdot x + r_n(x) x^2 \\ &\quad = P\big(Y_n \in [h, h+\delta] \mid X_n = x\big) \big[1 - \psi_n(I; x) \cdot x \\ &\qquad + \frac{r_n(x) x^2}{P\big(Y_n \in [h, h+\delta] \mid X_n = x\big)}\bigg] \\ &\quad = P\big(Y_n \in [h, h+\delta] \mid X_n = x\big) \bigg[e^{-\psi_n(I;x)x} - \frac{(\psi_n(I; x) x)^2 e^{-\gamma_n \cdot \psi_n(I;x)x}}{2} \\ &\qquad + \frac{r_n(x) x^2}{P\big(Y_n \in [h, h+\delta] \mid X_n = x\big)}\bigg] \end{aligned}$$

$$= P\big(Y_n \in [h, h+\delta] \mid X_n = x\big) \left[e^{-\psi_n(I;x)x} + k_n(x)x^2\right], \tag{4.2}$$

where

$$\psi_n(I;x) = \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y}\bigg|_{y=h}, \tag{4.3}$$

$$r_n(x) = \frac{1}{2}\frac{\partial^2 P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)}{\partial y^2}\bigg|_{y=h-\alpha_n x}, \tag{4.4}$$

$$k_n(x) = \frac{r_n(x)}{P\big(Y_n \in [h, h+\delta] \mid X_n = x\big)} - \frac{\psi_n(I;x)^2 e^{-\gamma_n \cdot \psi_n(I;x)x}}{2}, \tag{4.5}$$

and we have applied Taylor's expansion

$$e^{y_n} = 1 + y_n + \frac{(y_n)^2 e^{\gamma_n y_n}}{2}, \quad \text{for some } \gamma_n \in (0,1) \quad \text{and} \quad y_n := \psi_n(I;x)x$$

to the third equation in (4.2). Note that by condition (3.11),

$$0 \le \psi_n(I;x) \le C_3, \tag{4.6}$$

and by Conditions (3.10) and (3.11), for all $x \in \mathbb{R}^+$, $k_n(x)$ is uniformly bounded. Therefore, by the results of (4.1) and (4.2), for all $x \in \mathbb{R}^+$, we obtain that

$$f_{X_n|Z_n}(x;I) = \frac{f_{X_n}(x)P(Y_n \in I \mid X_n = x)(e^{-\psi_n(I;x)x} + k_n(x)x^2)}{P(Z_n \in I)}. \tag{4.7}$$

In the following, we will use brief notations

$$P_{Y_n|X_n}(I;x) := P(Y_n \in I \mid X_n = x), \quad P_{Z_n}(I) := P(Z_n \in I).$$

First, we let

$$A_n := \frac{1}{\displaystyle\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\mathrm{d}x}. \tag{4.8}$$

Since $\int_{\mathbb{R}^+} f_{X_n}(x)\mathrm{d}x = 1$, from (4.6), we have $\displaystyle\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\mathrm{d}x \le 1$, hence $A_n \ge 1$ for all $n \ge 1$. By definition $X_n = \beta_n X$, $\beta_n \to 0$, we also have

$$\lim_{n\to\infty}\frac{1}{A_n} = \lim_{n\to\infty}\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\mathrm{d}x = \lim_{n\to\infty}\int_{\mathbb{R}^+} f_X(x)e^{-\psi_n(I;\beta_n x)\beta_n x}\mathrm{d}x = 1$$

by the dominated convergence theorem, so A_n is uniformly bounded from above and from below.

Recall the definition of KL-divergence from (2.17), and the definitions of $\mathbb{P}_I^{(n)}$ and $\mathbb{Q}_I^{(n)}$ from (3.9), we have

$$D_{\mathrm{KL}}\left(\mathbb{P}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right) = \int_{\mathbb{R}^+} A_n f_X(x)e^{-\psi_n(I;\beta_n x)\beta_n x}\log\left(\frac{A_n f_X(x)e^{-\psi_n(I;\beta_n x)\beta_n x}}{f_{X|\tilde{Z}_n}(x, E_n)}\right)\mathrm{d}x$$

$$= \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{A_n f_{X_n}(x) e^{-\psi_n(I;x)x}}{f_{X_n|Z_n}(x,I)}\right) \mathrm{d}x \tag{4.9}$$

$$= \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{f_{X_n|Z_n}(x,I)}{A_n f_{X_n}(x) e^{-\psi_n(I;x)x}}\right) \mathrm{d}x \right|. \tag{4.10}$$

(4.9) is obtained by the change of variables $X_n = \beta_n X$ and the scale invariant property of the KL-divergence. (4.10) is true because the KL-divergence is nonnegative. With (4.1), the right-hand side in (4.10) can be written as

$$\begin{aligned}
&\left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)} \log\left(\frac{f_{X_n}(x) P_{Y_n|X_n}(I-x;x)}{P_{Z_n}(I)} \cdot \frac{1}{A_n f_{X_n}(x) e^{-\psi_n(I;x)x}}\right) \mathrm{d}x \right| \\
&= \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)} \log\left(\frac{P_{Y_n|X_n}(I-x;x)}{P_{Y_n|X_n}(I;x) e^{-\psi_n(I;x)x}} \cdot \frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I) A_n}\right) \mathrm{d}x \right| \\
&\le \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)} \log\left(\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)\, A_n}\right) \mathrm{d}x \right| \\
&\quad + \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)} \log\left(\frac{P_{Y_n|X_n}(I-x;x)}{P_{Y_n|X_n}(I;x) e^{-\psi_n(I;x)x}}\right) \mathrm{d}x \right|.
\end{aligned} \tag{4.11}$$

From the expression of $f_{X_n|Z_n}(x;I)$ in (4.7), we have the following identity

$$\begin{aligned}
1 &= \int_{\mathbb{R}^+} f_{X_n|Z_n}(x;I)\mathrm{d}x \\
&= \frac{\displaystyle\int_{\mathbb{R}^+} f_{X_n}(x) e^{-\psi_n(I;x)x} P_{Y_n|X_n}(I;x)\mathrm{d}x}{P_{Z_n}(I)} \\
&\quad + \frac{\displaystyle\int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 P_{Y_n|X_n}(I;x)\mathrm{d}x}{P_{Z_n}(I)}.
\end{aligned} \tag{4.12}$$

For the second term in (4.12), Conditions (3.10) and (3.11) imply that $P_{Y_n|X_n}(I;x)$ and $k_n(x)$ are uniformly bounded and condition (3.13) implies that $P_{Z_n}(I)$ is uniformly bounded from below. Then, by the assumption $\mathbb{E}[X_n^2] = a_n$, the first term in (4.12) satisfies

$$\begin{aligned}
&\left| \frac{\displaystyle\int_{\mathbb{R}^+} f_{X_n}(x) e^{-\psi_n(I;x)x} P_{Y_n|X_n}(I;x)\mathrm{d}x}{P_{Z_n}(I)} \right| \\
&= \left| \frac{\displaystyle\int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 P_{Y_n|X_n}(I;x)\mathrm{d}x}{P_{Z_n}(I)} - 1 \right| = 1 + O(a_n).
\end{aligned} \tag{4.13}$$

With condition (3.11), (4.13) implies

$$\frac{1}{P_{Z_n}(I)A_n} = \frac{\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\mathrm{d}x}{P_{Z_n}(I)} \leq \frac{1}{\delta_1} + O(a_n). \tag{4.14}$$

By Conditions (3.11) and (3.12): $\left|P_{Y_n|X_n}(I;x) - P_{Y_n}(I)\right| \leq b_n P_{Y_n}(I)$ with $b_n \to 0$, therefore

$$P_{Y_n}(I) \leq \frac{P_{Y_n|X_n}(I;x)}{1-b_n} \leq K_1 \tag{4.15}$$

for some constant $K_1 > 0$. With (4.14) and (4.15) and recalling the definition of A_n in (4.8), we have

$$\begin{aligned}
&\left|\frac{\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}P_{Y_n|X_n}(I;x)\mathrm{d}x}{P_{Z_n}(I)} - \frac{P_{Y_n}(I)}{P_{Z_n}(I)A_n}\right| \\
&= \left|\frac{\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\left(P_{Y_n|X_n}(I;x) - P_{Y_n}(I)\right)\mathrm{d}x}{P_{Z_n}(I)}\right| \\
&\leq \frac{\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\left|P_{Y_n|X_n}(I;x) - P_{Y_n}(I)\right|\mathrm{d}x}{P_{Z_n}(I)} \\
&\leq \frac{b_n P_{Y_n}(I)\int_{\mathbb{R}^+} f_{X_n}(x)e^{-\psi_n(I;x)x}\mathrm{d}x}{P_{Z_n}(I)} = \frac{b_n P_{Y_n}(I)}{P_{Z_n}(I)A_n} \\
&\leq K_1 b_n\left(\frac{1}{\delta_1} + O(a_n)\right) = O(b_n).
\end{aligned} \tag{4.16}$$

And similarly,

$$\left|\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)A_n} - \frac{P_{Y_n}(I)}{P_{Z_n}(I)A_n}\right| \leq \frac{b_n P_{Y_n}(I)}{P_{Z_n}(I)A_n} = O(b_n). \tag{4.17}$$

By the triangle inequality, from (4.13), (4.16), and (4.17), we have

$$\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)A_n} = 1 + O(a_n + b_n).$$

Since $\log(1+x) \leq x$ for all $x > -1$, for sufficiently large n, we have

$$\log\left(\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)A_n}\right) = O(a_n + b_n). \tag{4.18}$$

Note that the term $O(a_n+b_n)$ in (4.18) is independent of x. Therefore, for the first term in (4.11) we have

$$\begin{aligned}
&\left|\int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)A_n}\right) \mathrm{d}x\right| \\
&\quad \le \sup_x \left|\log\left(\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)A_n}\right)\right| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \mathrm{d}x \\
&\quad = \sup_x \left|\log\left(\frac{P_{Y_n|X_n}(I;x)}{P_{Z_n}(I)A_n}\right)\right| \cdot 1 = O(a_n+b_n). \qquad (4.19)
\end{aligned}$$

Define

$$\hat{G}_\delta(y,x) := \log P\big(Y_n \in [y, y+\delta] \mid X_n = x\big).$$

Then, by Taylor expansion and the conditions (3.10), (3.11), we can expand $\hat{G}_\delta(h-x,x)$ at (h,x) to get

$$\begin{aligned}
\hat{G}_\delta(h-x,x) &= \hat{G}_\delta(h,x) - \frac{\partial \hat{G}_\delta(h,x)}{\partial y} x \\
&\quad + \frac{\partial^2 \hat{G}_\delta(h-\hat{\alpha}_n x, x)}{2\partial y^2} x^2, \quad \text{for some } \hat{\alpha}_n \in (0,1), \\
&= \hat{G}_\delta(h,x) - \psi_n(I;x)x + q_n(x)x^2, \qquad (4.20)
\end{aligned}$$

where $q_n(x) := \dfrac{1}{2}\dfrac{\partial^2 \log P\big(Y_n \in [y,y+\delta] \mid X_n = x\big)}{\partial y^2}\Big|_{y=h-\hat{\alpha}_n x}$. Therefore, for the second term in (4.11), by (4.20), we can get

$$\begin{aligned}
&\log\left(\frac{P\big(Y_n \in [h-x, h+\delta-x] \mid X_n = x\big)}{P\big(Y_n \in [h, h+\delta] \mid X_n = x\big) e^{-\psi_n(I;x)x}}\right) \\
&\quad = \hat{G}_\delta(h-x,x) - \hat{G}_\delta(h,x) + \psi_n(I;x)x = q_n(x)x^2.
\end{aligned}$$

And by condition (3.10), for all $x \in \mathbb{R}^+$, there is a constant $K_2 > 0$ such that

$$|e^{-\psi_n(I;x)x} q_n(x)| \le K_2. \qquad (4.21)$$

In the following, we use a brief notation $P_{Y_n|X_n}\big(E_n - x; x\big) = P\big(Y_n \in [y, y+\delta] \mid X_n = x\big)$. By (4.21), and the uniform boundedness of A_n, and the assumption: $\mathbb{E}\big[X_n^2\big] = a_n$, the second term in (4.11) satisfies

$$\begin{aligned}
&\left|\int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{P_{Y_n|X_n}(I-x;x)}{P_{Y_n|X_n}(I;x)e^{-\psi_n(I;x)x}}\right) \mathrm{d}x\right| \\
&\quad = \left|\int_{\mathbb{R}^+} A_n e^{-\psi_n(I;x)x} f_{X_n}(x) q_n(x) x^2 \mathrm{d}x\right| \\
&\quad \le M\left|\int_{\mathbb{R}^+} f_{X_n}(x) e^{-\psi_n(I;x)x} q_n(x) x^2 \mathrm{d}x\right| \le M K_2 \mathbb{E}\big[X_n^2\big] = O(a_n). \qquad (4.22)
\end{aligned}$$

Combining (4.10), (4.11), (4.19) and (4.22),

$$D_{\mathrm{KL}}\left(\mathbb{P}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right) = O(a_n + b_n). \tag{4.23}$$

For the case $S_n := \{x : f_{X_n}(x) > 0\} \subset \mathbb{R}^+$, we can only define $P(Z_n \in I \mid X_n = x)$ on S_n. But we can still define the KL-divergence on $\mathbb{R}^+$ since the part of KL-divergence on $\mathbb{R}^+ \backslash S_n$ is 0. Therefore, in the same way as (4.9),

$$\begin{aligned} D_{\mathrm{KL}}\left(\mathbb{P}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right) &= \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{A_n f_{X_n}(x) e^{-\psi_n(I;x)x}}{f_{X_n|Z_n}(x,I)}\right) \mathrm{d}x \\ &= \int_{S_n} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{A_n f_{X_n}(x) e^{-\psi_n(I;x)x}}{f_{X_n|Z_n}(x,I)}\right) \mathrm{d}x \\ &= \left| \int_{S_n} A_n f_{X_n}(x) e^{-\psi_n(I;x)x} \log\left(\frac{f_{X_n|Z_n}(x,I)}{A_n f_{X_n}(x) e^{-\psi_n(I;x)x}}\right) \mathrm{d}x \right|. \end{aligned} \tag{4.24}$$

Then, we can follow every step from the step (4.10) in our proof for the case $\{x : f_{X_n}(x) > 0\} = \mathbb{R}^+$ to get

$$D_{\mathrm{KL}}\left(\mathbb{P}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right) = O(a_n + b_n). \tag{4.25}$$

Furthermore, let $\zeta_n(I;x) := \beta_n \psi_n(I; \beta_n x)$, by the condition (3.11), there is a constant $C > 0$ such that for all $x \in \mathbb{R}^+$, $0 \le \zeta_n(I;x) < C$. Therefore, $\mathbb{P}_I^{(n)}$ with the density function $A_n f_X(x) e^{-\beta_n \psi_n(I;\beta_n x)x}$ satisfies the definition of the canonical probability distributions in (3.5). □

4.1.2. Proof of Theorem 3.2. For a finite interval $I = [h, h+\delta]$, $h, \delta \in \mathbb{R}$ and $\delta > 0$, let

$$\hat{\mathbb{P}}_I^{(n)} = P(K_n = \beta_n k \mid Z_n \in I) \quad \text{and} \quad \hat{\mathbb{Q}}_I^{(n)} = B_n P(K_n = \beta_n k) e^{-\hat{\psi}(I;\beta_n k)\beta_n k},$$

where

$$\frac{1}{B_n} := \sum_{\beta_n k \in S_n} P(K_n = \beta_n k) e^{-\hat{\psi}(I;\beta_n k)\beta_n k}$$

and

$$\hat{\psi}_n(I;\beta_n k) := \left.\frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid K_n = \beta_n k\big)}{\partial y}\right|_{y=h}.$$

We first state the following lemma. The proof follows from the proof of Theorem 3.1 with the definition of KL-divergence for discrete probability distributions in (2.18).

Lemma 4.1. *Assume there exist positive constants $\delta_1, \delta_2, \{C_i, 1 \le i \le 3\}$, a sequence $b_n = o(1)$, and an open interval D such that the conditions (3.10)–(3.13) in Theorem 3.1 hold for K_n, Y_n, and Z_n. Then,*

$$D_{KL}\left(\hat{\mathbb{P}}_I^{(n)} \parallel \hat{\mathbb{Q}}_I^{(n)}\right) = O(a_n + b_n), \quad \textit{for every } I \subseteq D. \tag{4.26}$$

Now we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. All of the conditions in Theorem 3.1 hold for K_n, Y_n, Z_n by the assumptions, hence Lemma 4.1 can be applied. Therefore, we obtain the following relation between total variation and KL-divergence from (2.19): for every $I \subseteq D$,

$$\sup_{\beta_n k \in S_n} \left| P\big(K_n = \beta_n k \mid Z_n \in I\big) - B_n P(K_n = \beta_n k) e^{-\hat{\psi}_n(I;\beta_n k)\beta_n k} \right| \leq \delta\left(\hat{\mathbb{P}}_I^{(n)}, \hat{\mathbb{Q}}_I^{(n)}\right) \leq \frac{1}{2}\sqrt{D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_I^{(n)} \| \, \hat{\mathbb{Q}}_I^{(n)}\right)} = O\left(\sqrt{a_n + b_n}\right). \tag{4.27}$$

With (3.26) and (4.27), the conclusion (3.27) follows from the change of variable $K_n = \beta_n K$ and the triangle inequality:

$$\begin{aligned}
&\sup_{k \in S} \left| P\big(K = k \mid \tilde{H}_n \in E_n\big) - B_n P(K = k) e^{-\beta_n \hat{\psi}_n(I;\beta_n k)k} \right| \\
&= \sup_{\beta_n k \in S_n} \left| P\big(K_n = \beta_n k \mid H_n \in I\big) - B_n P(K_n = \beta_n k) e^{-\hat{\psi}_n(I;\beta_n k)\beta_n k} \right| \\
&\leq \sup_{\beta_n k \in S_n} \left| P\big(K_n = \beta_n k \mid Z_n \in I\big) - B_n P(K_n = \beta_n k) e^{-\hat{\psi}_n(I;\beta_n k)\beta_n k} \right| \\
&\quad + \sup_{\beta_n k \in S_n} \left| P\big(K_n = \beta_n k \mid H_n \in I\big) - P\big(K_n = \beta_n k \mid Z_n \in I\big) \right| \\
&= O(c_n + \sqrt{a_n + b_n}).
\end{aligned}$$

Furthermore, let $\hat{\zeta}_n(I;k) := \beta_n \hat{\psi}_n(I;\beta_n k)$. We can check that $0 \leq \hat{\zeta}_n(I;k) < C$ for all $k \in S$ and a constant $C > 0$. Therefore, $B_n P(K = k) e^{-\beta_n \hat{\psi}_n(I;\beta_n k)k}$ satisfies the definition of the canonical probability distributions in (3.6). □

4.2. Proofs of Theorems 3.3 and 3.4

Let X be a nonnegative continuous random variable and with $\mathbb{E}[X] < \infty$ and let Z_n be a sequence of real-valued continuous random variables. Given a Borel measurable set $E \in \mathcal{B}(\mathbb{R})$ and a function $\psi : \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ with $0 < \psi(E) < \infty$, let $\mathbb{P}_E$ be a probability measure with density function

$$A f_X(x) e^{-\psi(E)x}, \quad \frac{1}{A} := \int_{\mathbb{R}^+} f_X(x) e^{-\psi(E)x} \mathrm{d}x.$$

And let $\mathbb{Q}_E^{(n)}$ be a probability measure with density function $f_{X|Z_n}(x;E)$. We obtain the following lemma for the case (2) of the canonical distribution (3.32):

Lemma 4.2. *Assume the following conditions hold:*

1. *(Boundedness)* $\left|\dfrac{f_{X|Z_n}(x;E)}{f_X(x)}\right|$ *and* $\left|e^{-\xi x} \log\left(\dfrac{f_{X|Z_n}(x;E)}{f_X(x)}\right)\right|$, *for any* $\xi > 0$, *are uniformly bounded on* $\mathbb{R}^+$.
2. *(Linear approximation) There exist constants* $b, c \in \mathbb{R}$, $0 < c < \infty$, *and a sequence of functions* $q_n : \mathbb{R} \to \mathbb{R}$ *with*

$$\mathbb{E}\left[q_n(X)^2\right] = \gamma_n \to 0$$

such that on an interval $I_n = [0, d_n]$ *with* $d_n \to \infty$,

$$\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right) = b - cx + q_n(x). \tag{4.28}$$

Then,

$$\lim_{n\to\infty} D_{KL}\left(\mathbb{P}_E \parallel \mathbb{Q}_E^{(n)}\right) = 0 \quad \textit{if and only if} \quad c = \psi(E).$$

Furthermore, assume $\mathbb{E}[X^3] < \infty$ and X is not a constant random variable, let $\tilde{\mathbb{P}}_E^{(n)}$ be a probability measure with density function

$$\tilde{A}_n f_X(x) e^{-\beta_n \psi(E)x}, \quad \frac{1}{\tilde{A}_n} := \int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi(E)x} \mathrm{d}x, \tag{4.29}$$

in which $\beta_n > 0,\ \beta_n = o(1)$. We obtain the following lemma for the case (1) of the canonical distribution (3.32):

Lemma 4.3. *Assume the following conditions hold :*

1. *(Boundedness)* $\left|\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right|$ *and* $\left|e^{-\beta_n \xi x}\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right)\right|$, *for any* $\xi > 0$, *are uniformly bounded on* $\mathbb{R}^+$.
2. *(Linear approximation) There exist constants* $b, c \in \mathbb{R}$, $0 < c < \infty$, *and a sequence of functions* $q_n : \mathbb{R} \to \mathbb{R}$ *with* $\mathbb{E}\left[q_n(X)^2\right] \to 0$ *such that on* $I_n = [0, d_n]$ *with* $d_n = O\left(\frac{1}{\beta_n}\right)$,

$$\frac{1}{\beta_n}\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right) = b - cx + q_n(x). \tag{4.30}$$

Then,

$$\lim_{n\to\infty} \frac{D_{KL}\left(\tilde{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right)}{\beta_n^2} = 0 \quad \textit{if and only if} \quad c = \psi(E).$$

Remark 4.1. In particular, if we choose $Z_n = \beta_n X + \beta_n(\tilde{Y}_n - \mu_n)$, where $\tilde{Y}_n, \beta_n, \mu_n$ are given in the definitions in Sect. 3.2, and choose the Borel set E to be a finite interval I, by Eq. (3.1), those general results of Lemmas 4.2 and 4.3 for $f_{X|Z_n}(x, E)$ can be applied to $f_{X|\tilde{Z}_n}(x, E_n)$, which is the conditional density defined in Sect. 3.2.

4.2.1. Proof of Lemma 4.2.

Proof. Note that for any uniformly bounded function $|b_n(x)|$ on $\mathbb{R}^+$:

$$\begin{aligned}\left|\int_{\mathbb{R}^+\setminus I_n} f_X(x) b_n(x)\mathrm{d}x\right| &\le \|b_n(x)\|_\infty \int_{\mathbb{R}^+\setminus I_n} f_X(x)\mathrm{d}x = \|b_n(x)\|_\infty \mathbb{P}\left(X \ge d_n\right) \\ &\le \|b_n(x)\|_\infty \left(\frac{\mathbb{E}\left[X\right]}{d_n}\right) = O(\varepsilon_n),\end{aligned} \tag{4.31}$$

for a sequence $\varepsilon_n \to 0$ since $d_n \to \infty$ by condition (2) and $\mathbb{E}[X]$ is bounded by the assumption.

We first prove $c = \psi(E) \Rightarrow D_{\mathrm{KL}}\left(\mathbb{P}_E \parallel \mathbb{Q}_E^{(n)}\right) \to 0$.

By condition (2),

$$\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right) = b - \psi(E)x + q_n(x) \quad \text{on } I_n, \tag{4.32}$$

therefore, we have

$$\log\left(\frac{Af_X(x)e^{-\psi(E)x}}{f_{X|Z_n}(x;E)}\right) = \log A - b - q_n(x) \quad \text{on } I_n. \tag{4.33}$$

Since $\int_{\mathbb{R}^+} f_X(x)\mathrm{d}x = 1$, there exists a bounded closed set $D \subset \mathbb{R}^+$ such that $\int_D f_X(x)\mathrm{d}x > 0$. Hence,

$$\begin{aligned} A &= \frac{1}{\displaystyle\int_{\mathbb{R}^+} f_X(x)e^{-\psi(E)x}\mathrm{d}x} \le \frac{1}{\displaystyle\int_D f_X(x)e^{-\psi(E)x}\mathrm{d}x} \\ &\le \frac{1}{\displaystyle\inf_{x\in D}\left|e^{-\psi(E)x}\right|\int_D f_X(x)\mathrm{d}x} < \infty. \end{aligned} \tag{4.34}$$

Furthermore, we can derive

$$\begin{aligned} 1 = \int_{\mathbb{R}^+} f_{X|Z_n}(x;E)\mathrm{d}x &= \int_{I_n} f_X(x)\frac{f_{X|Z_n}(x;E)}{f_X(x)}\mathrm{d}x \\ &\quad + \int_{\mathbb{R}^+\setminus I_n} f_X(x)\frac{f_{X|Z_n}(x;E)}{f_X(x)}\mathrm{d}x \\ &= \int_{I_n} f_X(x)e^{b-\psi(E)x+q_n(x)}\mathrm{d}x + O(\varepsilon_n), \end{aligned} \tag{4.35}$$

in which the last equality is from Eq. (4.32), and the result of (4.31) applied to the uniformly bounded function $|b_n(x)| = \left|\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right|$ on $\mathbb{R}^+$ (condition (1)). Multiplying by e^{-b} on both sides in (4.35), we have

$$e^{-b} = \int_{I_n} f_X(x)e^{-\psi(E)x+q_n(x)}\mathrm{d}x + O(\varepsilon_n). \tag{4.36}$$

Then, we can apply Taylor's expansion to $e^{q_n(x)}$ to get

$$\begin{aligned} e^{-b} = \int_{I_n} f_X(x)e^{-\psi(E)x}\mathrm{d}x + \int_{I_n} f_X(x)e^{-\psi(E)x}\,(q_n(x) \\ + \frac{q_n(x)^2}{2}e^{\alpha_n\cdot q_n(x)}\Big)\,\mathrm{d}x + O(\varepsilon_n), \end{aligned} \tag{4.37}$$

for some sequence $\alpha_n \in (0,1)$. Note that we use the formula

$$e^y = 1 + y + \frac{e^{\alpha(y)\cdot y}}{2}y^2, \quad \alpha(y) \in (0,1)$$

and let $y = q_n(x)$, $\alpha_n = \alpha(q_n(x))$. Combined with Eq. (4.32) and condition (1), it implies there exists a constant $M > 1$ independent of n such that

$e^{-\psi(E)x+q_n(x)} \leq M$ for all $x \in I_n$. Since $\alpha_n \in (0,1)$ and $\psi(E) > 0$ in the assumption,

$$e^{-\psi(E)x+\alpha_n \cdot q_n(x)} \leq M^{\alpha_n} e^{-(1-\alpha_n)\psi(E)x} \leq M \quad \text{for all } x \in I_n. \tag{4.38}$$

Hence, $e^{-\psi(E)x+\alpha_n q_n(x)}$ is uniformly bounded on I_n. Then,

$$\begin{aligned}
&\int_{I_n} f_X(x) e^{-\psi(E)x} \left(\frac{q_n(x)^2}{2} e^{\alpha_n \cdot q_n(x)} \right) \mathrm{d}x \\
&\quad \leq \left\| \frac{e^{-\psi(E)x+\alpha_n \cdot q_n(x)}}{2} \right\|_\infty \int_{I_n} f_X(x) q_n(x)^2 \mathrm{d}x \\
&\quad \leq M \mathbb{E}\left[q_n(X)^2 \right] = O(\gamma_n),
\end{aligned} \tag{4.39}$$

where $O(\gamma_n) \to 0$ by condition (2). By Equations (4.37) and (4.39),

$$\begin{aligned}
e^{-b} &\leq \int_{I_n} f_X(x) e^{-\psi(E)x} \mathrm{d}x + \int_{I_n} f_X(x) e^{-\psi(E)x} q_n(x) \mathrm{d}x + O(\gamma_n) + O(\varepsilon_n) \\
&= \int_{\mathbb{R}^+} f_X(x) e^{-\psi(E)x} \mathrm{d}x - \int_{\mathbb{R}^+ \setminus I_n} f_X(x) e^{-\psi(E)x} \mathrm{d}x \\
&\quad + \int_{I_n} f_X(x) e^{-\psi(E)x} q_n(x) \mathrm{d}x + O(\gamma_n) + O(\varepsilon_n) \\
&= \frac{1}{A} + \int_{I_n} f_X(x) e^{-\psi(E)x} q_n(x) \mathrm{d}x + O(\gamma_n) + O(\varepsilon_n),
\end{aligned} \tag{4.40}$$

where the last equation is from the result of (4.31). And since A is bounded by the result (4.34), we have

$$A e^{-b} \leq 1 + \int_{I_n} A f_X(x) e^{-\psi(E)x} q_n(x) \mathrm{d}x + O(\gamma_n) + O(\varepsilon_n). \tag{4.41}$$

Using the inequality $\log(1+x) \leq x$ for all $x > -1$, we find a bound

$$\log A - b \leq \int_{I_n} A f_X(x) e^{-\psi(E)x} q_n(x) \mathrm{d}x + O(\gamma_n) + O(\varepsilon_n). \tag{4.42}$$

Furthermore, by condition (1), $\left| e^{-\psi(E)x} \log \left(\frac{f_{X|Z_n}(x;E)}{f_X(x)} \right) \right|$ is uniformly bounded on $\mathbb{R}^+$, so we can check that

$$\left| e^{-\psi(E)x} \log \left(\frac{A f_X(x) e^{-\psi(E)x}}{f_{X|Z_n}(x;E)} \right) \right| \text{ is uniformly bounded on } \mathbb{R}^+ \text{as well.} \tag{4.43}$$

Recall that

$$\log \left(\frac{A f_X(x) e^{-\psi(E)x}}{f_{X|Z_n}(x;E)} \right) = \log A - b - q_n(x) \text{ on } I_n$$

by (4.33). With the result of (4.42), we can get

$$\begin{aligned}
&D_{\mathrm{KL}} \left(\mathbb{P}_E \parallel \mathbb{Q}_E^{(n)} \right) \\
&\quad = \int_{I_n} A f_X(x) e^{-\psi(E)x} \log \left(\frac{A f_X(x) e^{-\psi(E)x}}{f_{X|Z_n}(x;E)} \right) \mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&+\int_{\mathbb{R}^+\setminus I_n} Af_X(x)e^{-\psi(E)x}\log\left(\frac{Af_X(x)e^{-\psi(E)x}}{f_{X|Z_n}(x;E)}\right)\mathrm{d}x\\
&=\int_{I_n} Af_X(x)e^{-\psi(E)x}\left(\log A-b\right)\mathrm{d}x-\int_{I_n} Af_X(x)e^{-\psi(E)x}q_n(x)\mathrm{d}x+O(\varepsilon_n)\\
&=(\log A-b)-\int_{\mathbb{R}^+\setminus I_n} Af_X(x)e^{-\psi(E)x}\left(\log A-b\right)\mathrm{d}x\\
&\quad-\int_{I_n} Af_X(x)e^{-\psi(E)x}q_n(x)\mathrm{d}x+O(\varepsilon_n)\\
&=\log A-b+O(\varepsilon_n)-\int_{I_n} Af_X(x)e^{-\psi(E)x}q_n(x)\mathrm{d}x+O(\varepsilon_n)\\
&=O(\gamma_n)+O(\varepsilon_n),
\end{aligned}
\tag{4.44}
$$

where the $O(\varepsilon_n)$ terms are from the result of (4.31) applied to the bounded function (4.43). Therefore, by (4.34) and (4.44), we get

$$
D_{\mathrm{KL}}\left(\mathbb{P}_E\parallel\mathbb{Q}_E^{(n)}\right)\to 0.
$$

Next we prove

$$
D_{\mathrm{KL}}\left(\mathbb{P}_E\parallel\mathbb{Q}_E^{(n)}\right)\to 0\ \Rightarrow\ c=\psi(E). \tag{4.45}
$$

By condition (2), there exists a constant $\hat{b}$ and a sequence of functions $\hat{q}_n(x)$ such that

$$
\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right)=b-cx+q_n(x)\quad\text{on } I_n. \tag{4.46}
$$

Similar to the derivation of (4.33), we have

$$
\log\left(\frac{\hat{A}f_X(x)e^{-cx}}{f_{X|Z_n}(x;E)}\right)=\log\hat{A}-b-q_n(x)\quad\text{on } I_n, \tag{4.47}
$$

where

$$
\hat{A}=\frac{1}{\displaystyle\int_{\mathbb{R}^+} f_X(x)e^{-cx}dx}<\infty, \tag{4.48}
$$

which can be proved by a similar approach as in (4.34). Then, following the previous proof from (4.35) to (4.44), we can get

$$
D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_E\parallel\mathbb{Q}_E^{(n)}\right)\to 0, \tag{4.49}
$$

where $\hat{\mathbb{P}}_E$ is a probability measure with density function $\hat{A}f_X(x)e^{-cx}$. By the assumption (4.45), we also know

$$
D_{\mathrm{KL}}\left(\mathbb{P}_E\parallel\mathbb{Q}_E^{(n)}\right)\to 0. \tag{4.50}
$$

By Pinsker's inequality (2.19), we have that the total variation distance denoted by $\delta(\cdot,\cdot)$ satisfies

$$
\delta(\hat{\mathbb{P}}_E,\mathbb{Q}_E^{(n)})\to 0\quad\text{and}\quad\delta(\mathbb{P}_E,\mathbb{Q}_E^{(n)})\to 0. \tag{4.51}
$$

Then, by the triangle inequality, $\delta(\hat{\mathbb{P}}_E, \mathbb{P}_E) = 0$. It implies

$$\int_0^x \left(\hat{A} f_X(s) e^{-cs} ds - A f_X(s) e^{-\psi(E)s}\right) \mathrm{d}s = 0, \quad \text{for all } x \in \mathbb{R}^+.$$

Hence,

$$\hat{A} f_X(x) e^{-cx} = A f_X(x) e^{-\psi(E)x} \quad \text{holds almost everywhere on } \mathbb{R}^+.$$

Since $\hat{A}$ and A are both independent of x and there exists an interval such that $f_X(x) > 0$, we obtain $c = \psi(E)$. □

4.2.2. Proof of Lemma 4.3.

Proof. Note that for any uniform bounded function $|b_n(x)|$ on $\mathbb{R}^+$:

$$\left|\int_{\mathbb{R}^+\setminus I_n} f_X(x) b_n(x) \mathrm{d}x\right| \le \|b_n(x)\|_\infty \int_{\mathbb{R}^+\setminus I_n} f_X(x) \mathrm{d}x = \|b_n(x)\|_\infty \mathbb{P}(X \ge d_n)$$
$$\le \|b_n(x)\|_\infty \left(\frac{\mathbb{E}\left[X^3\right]}{d_n^3}\right) = O(\beta_n^3), \tag{4.52}$$

where the existence of $O(\beta_n^3)$ is due to $d_n = O(\frac{1}{\beta_n})$ by condition (2) and $\mathbb{E}[X^3] < \infty$ by the assumption.

We first prove $c = \psi(E) \Rightarrow \dfrac{D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right)}{\beta_n^2} \to 0$. By condition (2),

$$\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right) = \beta_n\left(b - \psi(E)x + q_n(x)\right) \quad \text{on } I_n, \tag{4.53}$$

Therefore, we have

$$\log\left(\frac{A_n f_X(x) e^{-\beta_n \psi(E)x}}{f_{X|Z_n}(x;E)}\right) = \log A_n - \beta_n b - \beta_n q_n(x) \quad \text{on } I_n. \tag{4.54}$$

Following the proof in (4.34), for each n, we have

$$A_n = \frac{1}{\displaystyle\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi(E)x} \mathrm{d}x} < \infty, \tag{4.55}$$

and we can check that $\lim_{n\to\infty} \displaystyle\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n\psi(E)x}\mathrm{d}x \to 1$, hence, A_n is uniformly bounded.

We can apply a similar proof as for Lemma 4.2 to Eq. (4.54). Substituting b by $\beta_n b$, $\psi(E)x$ by $\beta_n\psi(E)x$, $q_n(x)$ by $\beta_n q_n(x)$ and A by A_n, then every step from Eq. (4.35) to Eq. (4.44) follows. Therefore, we can get.

$$D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right) = O(\beta_n^2\gamma_n) + O(\beta_n^3),$$

where the $O(\beta_n^2\gamma_n)$ term follows from the derivation of the $O(\gamma_n)$ term in Lemma 4.2, the $O(\beta_n^3)$ term follows from Eq. (4.52) and the derivation of the

$O(\varepsilon_n)$ term in Lemma 4.2. It implies

$$\frac{D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right)}{\beta_n^2} = O(\gamma_n) + O(\beta_n) \to 0.$$

Next we prove $\dfrac{D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right)}{\beta_n^2} \to 0 \;\Rightarrow\; c = \psi(E)$. Similar to the proof for Lemma 4.2, we can show

$$\frac{D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right)}{\beta_n^2} \to 0, \tag{4.56}$$

where $\hat{\mathbb{P}}_E^{(n)}$ is a probability measure with density function $\hat{A}_n f_X e^{-\beta_n c x}$. By the assumption, we also know

$$\frac{D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_E^{(n)} \parallel \mathbb{Q}_E^{(n)}\right)}{\beta_n^2} \to 0. \tag{4.57}$$

Therefore, by Pinsker's inequality, we have that the total variation distance denoted by $\delta(\cdot,\cdot)$ satisfies

$$\frac{1}{\beta_n}\delta(\hat{\mathbb{P}}_E, \mathbb{Q}_E^{(n)}) \to 0 \quad \text{and} \quad \frac{1}{\beta_n}\delta(\tilde{\mathbb{P}}_E, \mathbb{Q}_E^{(n)}) \to 0. \tag{4.58}$$

Then, by the triangle inequality, $\frac{1}{\beta_n}\delta(\hat{\mathbb{P}}_E, \tilde{\mathbb{P}}_E) \to 0$. It implies

$$\lim_{n\to\infty} \frac{1}{\beta_n}\left(\int_0^x \hat{A}_n f_X(s) e^{-\beta_n c s}\mathrm{d}s - \int_0^x A_n f_X(s) e^{-\beta_n \psi(E) s}\mathrm{d}s\right) = 0, \quad \text{for all } x \in \mathbb{R}^+. \tag{4.59}$$

We can apply Taylor's expansion to $e^{-\beta_n c s}$ and $e^{-\beta_n \psi(E) s}$ to get

$$e^{-\beta_n c s} = 1 - \beta_n c s + O(\beta_n^2 s^2) \quad \text{and} \quad e^{-\beta_n \psi(E) s} = 1 - \beta_n \psi(E) s + O(\beta_n^2 s^2). \tag{4.60}$$

By the results of (4.60), Eq. (4.59) can be written as

$$\lim_{n\to\infty} \int_0^x \frac{1}{\beta_n}\left(\tilde{A}_n - A_n - \left(\tilde{A}_n c - A_n \psi(E)\right)\beta_n s + O(\beta_n^2 s^2)\right) f_X(s)\mathrm{d}s = 0, \quad \text{for all } x \in \mathbb{R}^+. \tag{4.61}$$

Since $\mathbb{E}[X^2] < \infty$ from the fact $\mathbb{E}[X^3] < \infty$, we know $\displaystyle\int_0^x s^2 f_X(s)\mathrm{d}s < \infty$ on $\mathbb{R}^+$. Therefore, the $O(\beta_n^2 s^2)$ in Eq. (4.61) can be dropped and we obtain

$$\lim_{n\to\infty} \int_0^x \frac{1}{\beta_n}\left(\tilde{A}_n - A_n - \left(\tilde{A}_n c - A_n \psi(E)\right)\beta_n s\right) f_X(s)\mathrm{d}s = 0, \quad \text{for all } x \in \mathbb{R}^+. \tag{4.62}$$

By the dominated convergence theorem,

$$\hat{A}_n = \frac{1}{\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n c x} \mathrm{d}x} \to 1 \quad \text{and} \quad A_n = \frac{1}{\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi_n(E) x} \mathrm{d}x} \to 1. \tag{4.63}$$

Also we have

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{\beta_n} \left(\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi_n(E) x} \mathrm{d}x - \int_{\mathbb{R}^+} f_X(x) e^{-\beta_n c x} \mathrm{d}x \right) \\ &= \lim_{n\to\infty} \frac{1}{\beta_n} \left(\int_{\mathbb{R}^+} \left((c - \psi(E)) \beta_n x + O(\beta_n^2 x^2) \right) f_X(x) \mathrm{d}x \right) \\ &= (c - \psi(E)) \, \mathbb{E}[X] + \lim_{n\to\infty} O(\beta_n \mathbb{E}[X^2]) = (c - \psi(E)) \, \mathbb{E}[X], \end{aligned} \tag{4.64}$$

where in the first equality we apply Taylor's expansion (4.60) again. By (4.63) and (4.64), we have

$$\begin{aligned} \lim_{n\to\infty} \frac{\tilde{A}_n - A_n}{\beta_n} &= \lim_{n\to\infty} \frac{1}{\beta_n} \left(\frac{\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi_n(E) x} \mathrm{d}x - \int_{\mathbb{R}^+} f_X(x) e^{-\beta_n c x} \mathrm{d}x}{\int_{\mathbb{R}^+} f_X(x) e^{-\beta_n c x} \mathrm{d}x \int_{\mathbb{R}^+} f_X(x) e^{-\beta_n \psi_n(E) x} \mathrm{d}x} \right) \\ &= (c - \psi(E)) \, \mathbb{E}[X]. \end{aligned} \tag{4.65}$$

Therefore, from (4.62) and (4.65),

$$\lim_{n\to\infty} \int_0^x \left(\tilde{A}_n c - A_n \psi(E) \right) s f_X(s) \mathrm{d}s = (c - \psi(E)) \int_0^x s f_X(s) \mathrm{d}s, \quad \text{for all } x \in \mathbb{R}^+. \tag{4.66}$$

Therefore, we can apply the results of (4.65) and (4.66) to Eq. (4.62) to get

$$(c - \psi(E)) \int_0^x \mathbb{E}[X] f_X(s) \mathrm{d}s = (c - \psi(E)) \int_0^x s f_X(s) \mathrm{d}s, \quad \text{for all } x \in \mathbb{R}^+. \tag{4.67}$$

Since X is not a constant random variable by our assumption, (4.67) is only true when $c = \psi(E)$. □

4.2.3. Proof of Theorem 3.3.

Proof. The proof follows from Lemma 4.3. By the condition (2) in Theorem 3.3, we have

$$\begin{aligned} &\log \left(\frac{P\left(Y_n \in I - \beta_n x \mid X_n = \beta_n x\right)}{P\left(Z_n \in I\right)} \right) \\ &= \log \left(\frac{P\left(Y \in I - \beta_n x\right)}{P\left(Y \in I\right)} \right) + g_n(x) \\ &= -\frac{\partial \log P(Y \in [y, y+\delta])}{\partial y} \bigg|_h \beta_n x + O(\beta_n^2 x^2) + \beta_n g_n(x). \end{aligned} \tag{4.68}$$

We now check whether all conditions in Lemma 4.3 are satisfied:

1. (Boundedness): $\left|\frac{f_{X|Z_n}(x;I)}{f_X(x)}\right| = \left|\frac{f_{X|\tilde{Z}_n}(x;E_n)}{f_X(x)}\right|$, which is uniformly bounded on $\mathbb{R}^+$ by the condition (2) in Theorem 3.3. And from (4.68), for any $\xi > 0$,

$$\left|e^{-\beta_n\xi x}\log\left(\frac{f_{X|Z_n}(x;I)}{f_X(x)}\right)\right| = \left|e^{-\beta_n\xi x}\log\left(\frac{P\left(Y_n \in I-\beta_n x \mid X_n=\beta_n x\right)}{P\left(Z_n\in I\right)}\right)\right|$$
$$\leq \left|e^{-\beta_n\xi x}O(\beta_n x+\beta_n^2x^2)\right| + \left|e^{-\beta_n\xi x}g_n(x)\right|,$$

   where the first term is uniformly bounded on $\mathbb{R}^+$, and the second term is uniformly bounded on $\mathbb{R}^+$ by the condition (3) in Theorem 3.3.
2. (Linear approximation): Following (4.68), we have

$$\frac{1}{\beta_n}\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right) = \frac{1}{\beta_n}\log\left(\frac{P\left(Y_n \in I-\beta_n x \mid X_n=\beta_n x\right)}{P\left(Z_n\in I\right)}\right)$$
$$= -\frac{\partial\log P(Y\in[y,y+\delta])}{\partial y}\bigg|_h x + O(\beta_n x^2) + g_n(x)$$

   on $I_n=[0,d_n]$ with $d_n = O\left(\frac{1}{\beta_n}\right)$. Therefore, we obtain

$$c = \frac{\partial\log P(Y\in[y,y+\delta])}{\partial y}\bigg|_h \quad\text{and}\quad q_n(x) = O(\beta_n x^2)+g_n(x)$$

   and we can check that $\mathbb{E}[q_n(X)^2]\to 0$ since $\mathbb{E}[g_n(X)^2]\to 0$ by the condition (3).

Therefore, applying Lemma 4.3, we have

$$\lim_{n\to\infty}\frac{D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right)}{\beta_n^2} = 0 \quad\text{if and only if}\quad \psi(I) = \frac{\partial\log P(Y\in[y,y+\delta])}{\partial y}\bigg|_h.$$

Furthermore, since $0 < \frac{\partial\log P(Y\in[y,y+\delta])}{\partial y}\Big|_h < C$ for a constant $C>0$, $\tilde{\mathbb{P}}_I^{(n)}$ satisfies the definition of the canonical probability distributions in (3.5). □

4.2.4. Proof of Theorem 3.4.

Proof. The proof follows from Lemma 4.2. By the condition (2) in Theorem 3.4, we have

$$\log\left(\frac{P\left(Y_n \in I-\beta_n x \mid X_n=\beta_n x\right)}{P\left(Z_n\in I\right)}\right) = \log\left(\frac{\exp\left[-\frac{1}{\beta_n}\phi\left(y^*-\beta_n x\right)\right]}{\exp\left[-\frac{1}{\beta_n}\phi\left(y^*\right)\right]}\right) + r_n(x)$$
$$= \phi'(y^*)x + O(\beta_n x^2) + r_n(x). \tag{4.69}$$

To check that all conditions are satisfied:

1. (Boundedness): $\left|\frac{f_{X|Z_n}(x;I)}{f_X(x)}\right| = \left|\frac{f_{X|\tilde{Z}_n}(x;E_n)}{f_X(x)}\right|$, which is uniformly bounded on $\mathbb{R}^+$ by the condition (1) in Theorem 3.4. And by (4.69), for any $\xi>0$,

$$\left|e^{-\xi x}\log\left(\frac{f_{X|Z_n}(x;I)}{f_X(x)}\right)\right| = \left|e^{-\xi x}\log\left(\frac{P\left(Y_n\in I-\beta_n x \mid X_n=\beta_n x\right)}{P\left(Z_n\in I\right)}\right)\right|$$
$$\leq \left|e^{-\xi x}O(x+\beta_n x^2)\right| + \left|e^{-\xi x}r_n(x)\right|,$$

where the first term is uniformly bounded on $\mathbb{R}^+$, and the second term is uniformly bounded on $\mathbb{R}^+$ by the condition (3) in Theorem 3.4.

2. (Linear approximation): As follows from (4.69), we have

$$\log\left(\frac{f_{X|Z_n}(x;E)}{f_X(x)}\right) = \log\left(\frac{P\left(Y_n\in I-\beta_n x \mid X_n=\beta_n x\right)}{P\left(Z_n\in I\right)}\right)$$
$$= \phi'(y^*)x + O(\beta_n x^2) + r_n(x)$$

on $I_n=[0,d_n]$ with $d_n = O\left(\frac{1}{\beta_n}\right)$. Hence, we obtain

$$c=-\phi'(y^*) \quad \text{and} \quad q_n(x) = O(\beta_n x^2) + r_n(x),$$

and we can check that $\mathbb{E}[q_n(X)^2]\to 0$ since $\mathbb{E}[r_n(X)^2]\to 0$ by the condition (3).

Therefore, by applying Lemma 4.2, we have

$$\lim_{n\to\infty} D_{\mathrm{KL}}\left(\mathbb{P}_I \parallel \mathbb{Q}_I^{(n)}\right) = 0 \quad \text{if and only if} \quad \varphi(I) = -\phi'(y^*).$$

Since $0<-\phi'(y^*)<\infty$, $\mathbb{P}_I$ satisfies the definition of the canonical probability distributions in (3.5). □

5. Applications

5.1. Gibbs Measure on the Phase Space

Definition 5.1. Consider a probability space $(\Omega,\mathcal{F},\mathbb{P})$, let $\mathbf{V}=(V_1,V_2,...,V_n):\Omega\to\mathbb{R}^n$ be a measurable function and let π_1,π_2, and π be three projection maps defined on $\mathbb{R}^n$ such that

$$\pi_1(\mathbf{V})=\mathbf{U}=(V_1,V_2,...,V_k), \quad \pi_2(\mathbf{V})=\mathbf{W}=(V_{k+1},V_{k+2}...,V_n), \quad \pi(\mathbf{V})=\mathbf{V}. \tag{5.1}$$

Assume there exist measurable functions $e_1:\mathbb{R}^k\to\mathbb{R}^+$, $e_2:\mathbb{R}^{n-k}\to\mathbb{R}^+$, and $e:\mathbb{R}^n\to\mathbb{R}^+$ such that

$$(e_1\circ\pi_1)(\mathbf{V})=e_1(\mathbf{U}), \quad (e_2\circ\pi_2)(\mathbf{V})=e_2(\mathbf{W}), \quad (e\circ\pi)(\mathbf{V})=e(\mathbf{V}).$$

Therefore, random variables and induced measures can be defined through the following maps:

$$(\Omega,\mathcal{F},\mathbb{P})\xrightarrow{\mathbf{V}}(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n),\mu)\xrightarrow{\pi_1}(\mathbb{R}^k,\mathcal{B}(\mathbb{R}^k),\nu_1)\xrightarrow{e_1}(\mathbb{R}^+,\mathcal{B}(\mathbb{R}^+),\lambda_1)$$
$$(\Omega,\mathcal{F},\mathbb{P})\xrightarrow{\mathbf{V}}(\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n),\mu)\xrightarrow{\pi_2}(\mathbb{R}^{n-k},\mathcal{B}(\mathbb{R}^{n-k}),\nu_2)\xrightarrow{e_2}(\mathbb{R}^+,\mathcal{B}(\mathbb{R}^+),\lambda_2).$$

Definition 5.2. Let $e_1\circ\pi_1, e_2\circ\pi_2$, and $e\circ\pi$ be the functions given in Definition 5.1. Define $e_1\circ\pi_1$ and $e_2\circ\pi_2$ to be *additive* on $\mathbf{V}$ if

$$e_1\circ\pi_1(\mathbf{V}) + \ e_2\circ\pi_2(\mathbf{V}) = e\circ\pi(\mathbf{V}). \tag{5.2}$$

Theorem 5.1. *Suppose $e_1 \circ \pi_1$ and $e_2 \circ \pi_2$ are additive on $\mathbf{V}$ and suppose $e_1(\mathbf{U}), e_2(\mathbf{W})$ are continuous nonnegative independent random variables. Denote $X := e_1(\mathbf{U})$, $Y := e_2(\mathbf{W})$, $Z := e(\mathbf{V})$ and let $I = [h, h+\delta]$ be a finite interval. Assume the following conditions hold:*

1. $\mathbb{E}[X^2] = \varepsilon_n^2$, *where* $\varepsilon_n \to 0$.
2. *For all* $y \in \mathbb{R}^+$, *there exists a nonnegative integrable function* $\Gamma \in C^2(\mathbb{R}^+)$ *such that*

$$P\left(Y \in [y, y+\delta]\right) = \frac{\displaystyle\int_y^{y+\delta} \Gamma(s)\mathrm{d}s}{\displaystyle\int_{\mathbb{R}^+} \Gamma(s)\mathrm{d}s} \quad \text{and} \quad \left|\frac{\partial^2 \log P\big(Y \in [y, y+\delta]\big)}{\partial y^2}\right| < \infty. \tag{5.3}$$

3. $I \subset supp(\Gamma)$ *and* $\Gamma'(y) \geq 0$, *for* $y \in I$.

Then, we have

$$\sup_{S \in \mathcal{B}(\mathbb{R}^+)} \left| \mathbb{P}\left(e_1(\mathbf{U}) \in S \mid Z \in I\right) - \int_{e_1(\mathbf{u}) \in S} A e^{-\psi(I) e_1(\mathbf{u})} \nu_1(\mathrm{d}\mathbf{u}) \right| = O(\varepsilon_n), \tag{5.4}$$

where $\psi(I) = \left.\dfrac{\partial \log \displaystyle\int_y^{y+\delta} \Gamma(s)\mathrm{d}s}{\partial y}\right|_{y=h}$.

Proof. Since the functions $e_1 \circ \pi_1, e_2 \circ \pi_2$ are *additive* on $\mathbf{V}$, we have

$$\begin{aligned} X + Y &= e_1(\mathbf{U}) + e_2(\mathbf{W}) = (e_1 \circ \pi_1)(\mathbf{V}) + (e_2 \circ \pi_2)(\mathbf{V}) \\ &= (e \circ \pi)(\mathbf{V}) = e(\mathbf{V}) = Z. \end{aligned}$$

Since $X + Y = Z$ and they are corresponding to X_n, Y_n, Z_n in Theorem 3.1, respectively, it suffices to show that all the conditions in Theorem 3.1 are satisfied for X, Y, and Z.

1. For all $y \in \mathbb{R}^+$, since $\Gamma(y) \in C^2(\mathbb{R}^+)$, $\left|\dfrac{\partial^2 P\left(Y \in [y, y+\delta]\right)}{\partial^2 y}\right|$ exists and is bounded on $\mathbb{R}^+$.
   And $\left|\dfrac{\partial^2 \log P\left(Y \in [y, y+\delta]\right)}{\partial^2 y}\right|$ is bounded on $\mathbb{R}^+$ by (5.3). Therefore, (3.10) holds.
2. Since $I \subset supp(\Gamma)$, there exists $\delta_1 > 0$ such that $P\left(Y \in I\right) \geq \delta_1$. And we can derive

$$\left.\frac{\partial \log P(Y \in [y, y+\delta])}{\partial y}\right|_{y=h} = \frac{\Gamma(h+\delta) - \Gamma(h)}{\displaystyle\int_h^{h+\delta} \Gamma(s)\mathrm{d}s}. \tag{5.5}$$

   Again, since $I \subset supp(\Gamma)$, and the nonnegative function $\Gamma(y) \in C^2(\mathbb{R}+)$, $\Gamma'(y) \geq 0$, for $y \in I$, we can check that there exists a positive constant C such that

$$0 \le \frac{\partial \log P\big(Y \in [y, y+\delta]\big)}{\partial y} \le C \quad \text{for } [y, y+\delta] \subset I, \tag{5.6}$$

hence (3.11) holds for $D = I$. Furthermore, since X and Y are independent, $b_n = 0$. Therefore, (3.12) holds.

3. Since X and Y are supported on $\mathbb{R}^+$, there exists $\delta_2 > 0$ such that

$$P\left(Z \in [z, z+\delta]\right) \ge \delta_2 \quad \text{for } [y, y+\delta] \subset \mathbb{R}^+,$$

hence (3.13) holds.

Therefore, all of the conditions hold for $D = I$ in Theorem 3.1, we can apply it with $a_n = \varepsilon_n^2$, $b_n = 0$, and Pinsker's inequality (2.19) to get

$$\sup_{S \in \mathcal{B}(\mathbb{R}^+)} \left| \mathbb{P}\left(X \in S \mid Z \in I\right) - \int_{x \in S} A e^{-\psi(I)x} f_X(x) \mathrm{d}x \right| = O(\varepsilon_n), \tag{5.7}$$

where $\psi(I) = \left. \dfrac{\partial \log \displaystyle\int_y^{y+\delta} \Gamma(s)\mathrm{d}s}{\partial y} \right|_{y=h}$. Then, applying a change of variables

$$\int_{x \in S} A e^{-\psi(I)x} f_X(x)\mathrm{d}x = \int_{x \in S} A e^{-\psi(I)x} \lambda_1(\mathrm{d}x) = \int_{e_1(\mathbf{u}) \in S} A e^{-\psi(I) e_1(\mathbf{u})} \nu_1(\mathrm{d}\mathbf{u}) \tag{5.8}$$

to (5.7), we obtain Eq. (5.4). It completes the proof. □

In statistical mechanics, the induced measure $\nu_1(\mathrm{d}\mathbf{u})$ in phase space is often considered as the Lebesgue measure $\mathrm{d}\mathbf{u}$ normalized by the total volume of the phase space Λ (Here, we assume it is finite). Therefore, for the random vector $\mathbf{U}$, we have its density

$$\hat{A} e^{-\psi(I) e_1(\mathbf{u})} \quad \text{with respect to } \mathrm{d}\mathbf{u}, \tag{5.9}$$

where $\hat{A} = A/\Lambda$ is the corresponding normalization factor.

The assumption $\nu_1(\mathrm{d}\mathbf{u}) = \mathrm{d}\mathbf{u}/\Lambda$ for the phase space has already implied that all microstates are *equally probable* when the system is unconstrained. It is a reasonable *prior* probability for $\mathbf{U}$ by a symmetry of a physical system when we do not have any previous information about it. For the random variable X (e.g., energy), its density $f_X(x)$ is referred to *prior* probability for X when it is unconstrained. Based on the principle of equal a priori probabilities of microstates in the phase space, we can show that $f_X(x) = \gamma(x)/\Lambda$, where $\gamma(x)$ is the Lebesgue measure of the surface area of microstates when the energy is fixed on x (i.e., $e_1(\mathbf{U}) = x$). This can be verified by

$$\begin{aligned} \int_{x \in S} f_X(x)\mathrm{d}x = \int_{e_1(\mathbf{u}) \in S} \nu_1(\mathrm{d}\mathbf{u}) &= \frac{1}{\Lambda} \int_{e_1(\mathbf{u}) \in S} \mathrm{d}\mathbf{u} \\ &= \frac{1}{\Lambda} \int_{x \in S} \gamma(x)\mathrm{d}x \quad \text{for all } S \in \mathcal{B}(\mathbb{R}^+). \end{aligned} \tag{5.10}$$

Note that $\gamma(x)$ is also known as the *structure function of* X. In Theorem 5.1, we also make the same assumption for Y: $f_Y(y) = \Gamma(y)/\Lambda$, where $\Gamma(y)$ is the *structure function of* Y.

Therefore, the density of X can be written as

$$\hat{A}e^{-\psi(I)x}\gamma(x) \quad \text{with respect to } dx, \tag{5.11}$$

which can be interpreted as a uniform prior biased by an exponential weight $e^{-\psi(I)x}$ when the system is conditioned on some extra information. Note that Eq. (5.9) is known as the density of Gibbs measure on the phase space and Eq. (5.11) is known as the density of Gibbs measure on the energy of the system [16].

In the work of Khinchin [21], he assumed *conjugate distribution laws* for all systems. It is said that

$$f_X(x) = \frac{e^{-\alpha x}\gamma(x)}{\int e^{-\alpha s}\gamma(s)\mathrm{d}s} \quad \text{and} \quad f_Y(y) = \frac{e^{-\alpha y}\Gamma(y)}{\int e^{-\alpha s}\Gamma(s)\mathrm{d}s} \tag{5.12}$$

for some constant α. Those priors are more general than the uniform prior and they have some nice properties, e.g., for a proper α, it may guarantee integrability of $e^{-\alpha s}\gamma(s)$ when $\gamma(s)$ itself is not integrable. However, we can show that the choice of $e^{-\alpha x}$ term does not have influence on our results. Here is the proof sketch: Suppose $\delta = o(1)$,

$$\begin{aligned}\hat{\psi}(I) := \frac{\partial \log P(Y \in [y, y+\delta])}{\partial y}\bigg|_{y=h} &= \frac{\partial \log \int_y^{y+\delta} \Gamma(s)e^{-\alpha s}\mathrm{d}s}{\partial y}\bigg|_{y=h} \\ &\approx \frac{\partial \log \int_y^{y+\delta} \Gamma(s)\mathrm{d}s}{\partial y}\bigg|_{y=h} - \alpha = \psi(I) - \alpha.\end{aligned} \tag{5.13}$$

By (5.12) and (5.13), we have

$$Af_X(x)e^{-\hat{\psi}(I)x} = \hat{A}\gamma(x)e^{-\alpha x}e^{-(\psi(I)-\alpha)x} = \hat{A}\gamma(x)e^{-\psi(I)x}. \tag{5.14}$$

Therefore, to choose priors as the structure functions multiplied by the exponential functions $e^{-\alpha x}$ for integrability is irrelevant to Gibbs measure. Indeed, it is the extra information (condition) giving rise to the exponential weight in Gibbs measure and the parameter of the exponential function is determined by

$$\psi(I) = \frac{\partial \log \int_y^{y+\delta} \Gamma(s)\mathrm{d}s}{\partial y}\bigg|_{y=h},$$

where $\int_y^{y+\delta} \Gamma(s)ds$ represents the volume of microstates in the shell between y and $y+\delta$. The logarithm of it is known as the entropy of Y, denoted by $S_Y(y)$, hence we have

$$\psi(I) = \frac{\partial S_Y(y)}{\partial y}\bigg|_{y=h}. \tag{5.15}$$

By Eq. (5.15), we can identify $\frac{1}{\psi(I)}$ as the temperature defined in statistical mechanics [18,22].

Remark 5.1. We can extend Theorem 5.1 to the model that the subsystem and its heat bath have strong interaction defined by non-additivity of energy functions in statistical mechanics. Assume there exists a measurable function $e_3 : \mathbb{R}^n \to \mathbb{R}^+$ such that

$$(e_3 \circ \pi)(\mathbf{V}) = e_3(\mathbf{V}),$$

which means that this energy function e_3 could depend on the whole vector $\mathbf{V} = (V_1, V_2, ..., V_n)$ in the phase space. And suppose

$$e_1 \circ \pi_1(\mathbf{V}) + e_2 \circ \pi_2(\mathbf{V}) + e_3 \circ \pi(\mathbf{V}) = e \circ \pi(\mathbf{V}),$$

in which the existence of the extra term $e_3 \circ \pi(\mathbf{V})$ means that $e_1 \circ \pi_1$ and $e_2 \circ \pi_2$ are not *additive* on $\mathbf{V}$ by Definition 5.1. Denote that $R := e_3(\mathbf{V})$. Recall that $\mathbf{V} = (\mathbf{U}, \mathbf{W})$ and $X = e_1(\mathbf{U}) = (e_1 \circ \pi_1)(\mathbf{V})$, $Y = e_2(\mathbf{W}) = (e_1 \circ \pi_2)(\mathbf{V})$, $Z = e(\mathbf{V}) = (e \circ \pi)(\mathbf{V})$. Then, we have

$$X + Y + R = Z. \tag{5.16}$$

In statistical mechanics, R is known as the *interaction energy* caused by interaction between the subsystem and its heat bath. Based on this setup, we can define a new random variable $\hat{Y} := Y + R$, but $X, \hat{Y}$ are no longer independent since the random variable R may depend on both $\mathbf{U}, \mathbf{W}$ in the phase space. If we modify the condition (5.3) in Theorem 5.1 to guarantee the existence and boundedness of

$$\begin{aligned} &\left|\partial^{(k)} P\big(\hat{Y} \in [y, y+\delta] \mid X = x\big)\right| \text{ and} \\ &\left|\partial^{(k)} \log P\big(\hat{Y} \in [y, y+\delta] \mid X = x\big)\right|, \text{ for } k = 0, 1, 2, \end{aligned} \tag{5.17}$$

in which the partial derivatives are with respect to both x and y, then we are able to apply Corollary 3.1 to this model. As the result (3.20) in Corollary 3.1, this model with strong interaction will give rise to a new parameter $\phi(I)$ of the exponential weight which involves two terms: one is from fluctuations of the energy of the "new" heat bath $\hat{Y}$ (it combines the energy of the heat bath Y without interaction and the interaction energy R), and the other one is from the correlation of X and $\hat{Y}$.

5.2. Integer-Valued Random Variables and Conditional Poisson Distributions

In Theorem 5.2, we will show a limiting behavior of a sequence of conditional probabilities for a nonnegative integer-valued random variables K, which is conditioned on $K + \tilde{L}_n$, $\tilde{L}_n$ is a sequence of sums of i.i.d random variables ξ_i. This sequence of conditional probabilities has the same limiting behavior as its unconditional probability $P(K = k)$ weighted by an exponential factor. The most important result of this theorem is that the parameter of this exponential factor determined by a normal distribution rather than the distribution of ξ_i. By this result, we provide a very simple formula with an approximation error to approximate an intractable problem in calculating the conditional probability

of an integer-valued random variable. And we give an example 1 to show an approximation formula for calculating the conditional probability of a Poisson random variable conditioned on the sum of that Poisson random variable with another Poisson random variable.

Theorem 5.2. *Let K be a nonnegative integer-valued random variable with $\mathbb{E}[K] < \infty$. Let $\tilde{L}_n = \sum_{i=1}^{n} \xi_i$, where $\{\xi_i\}_{i=1}^{n}$ are nonnegative i.i.d. random variables. K and $\tilde{L}_n$ are independent and denote $\tilde{H}_n := K + \tilde{L}_n$. Let $\mu = \mathbb{E}[\xi_i]$, $\sigma^2 = Var(\xi_i)$ and assume $\mathbb{E}[(\xi_i - \mu)^3] < \infty$. And let*

$$B_n = \sum_{k=0}^{\infty} \frac{1}{P(K=k)\exp\left(\frac{-\psi(I)k}{\sqrt{n}}\right)},$$

$$\psi(I) = \left.\frac{\partial \log P\big(Y \in [y, y+\delta]\big)}{\partial y}\right|_{y=-h}, \quad \text{and } Y \sim N(0, \sigma^2).$$

For every fixed finite interval $I = [-h, -h+\delta]$, $h, \delta \in \mathbb{R}^+$, $-h+\delta \le 0$, and $2\delta/\sigma^2 < \psi(I)$,

$$\sup_k \left| P(K=k \mid \tilde{H}_n \in n\mu + \sqrt{n}I) - B_n P(K=k)\exp\left(\frac{-\psi(I)k}{\sqrt{n}}\right)\right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{5.18}$$

Proof. Let $K_n := \frac{K}{\sqrt{n}}$, $L_n := \frac{\tilde{L}_n - n\mu}{\sqrt{n}}$ and $H_n := \frac{\tilde{H}_n - n\mu}{\sqrt{n}}$. We have $K_n + L_n = H_n$. By the Central Limit Theorem, L_n converges in distribution to Y. Furthermore, since $(\xi_i - \mu)$ has finite second and third moments, by Berry–Esseen Theorem 2.4,

$$\sup_k \left| P_{L_n}\left(I - \frac{k}{\sqrt{n}}\right) - P_Y\left(I - \frac{k}{\sqrt{n}}\right)\right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{5.19}$$

Since $\mathbb{E}[K_n] \to 0$, we have K_n converges to 0 in probability. By Slutsky's Theorem 2.5, H_n converges to Y in distribution. By Corollary 2.1, we can also get

$$P_{H_n}(I) = P_Y(I) + O\left(\frac{1}{\sqrt{n}}\right). \tag{5.20}$$

By (5.19) and (5.20),

$$\begin{aligned} P_{K|\tilde{H}_n}\left(k; n\mu + \sqrt{n}I\right) = P_{K|H_n}(k; I) &= P_K(k)\frac{P_{L_n}\left(I - \frac{k}{\sqrt{n}}\right)}{P_{H_n}(I)} \\ &= P_K(k)\frac{P_Y\left(I - \frac{k}{\sqrt{n}}\right) + O\left(\frac{1}{\sqrt{n}}\right)}{P_Y(I) + O\left(\frac{1}{\sqrt{n}}\right)} \\ &= P_K(k)\frac{P_Y\left(I - \frac{k}{\sqrt{n}}\right)}{P_Y(I)} + O\left(\frac{1}{\sqrt{n}}\right), \end{aligned} \tag{5.21}$$

in which we use the fact $Y \sim N(0,\sigma^2)$ and $P(-h \le Y \le -h+\delta)$ is bounded from below. Moreover, since $P_K(k) \le 1$, the term $O\left(\frac{1}{\sqrt{n}}\right)$ in (5.21) is independent of k. Let $\tilde{Y}_n \sim N(n\mu, n\sigma^2)$ and $\tilde{Z}_n := K + \tilde{Y}_n$. Then, we have

$$K_n + Y_n = Z_n, \quad \text{where} \quad Y_n := \frac{\tilde{Y}_n - n\mu}{\sqrt{n}} \text{ and } Z_n := \frac{\tilde{Z}_n - n\mu}{\sqrt{n}}. \tag{5.22}$$

Note that $Y_n = Y \sim N(0,\sigma^2)$ and Z_n converges in distribution to Y. Similar to (5.21),

$$P_{K|\tilde{Z}_n}\left(k; n\mu + \sqrt{n}I\right) = P_K(k)\frac{P_Y\left(I - \frac{k}{\sqrt{n}}\right)}{P_Y\left(I\right)} + O\left(\frac{1}{\sqrt{n}}\right). \tag{5.23}$$

Applying the triangle inequality to (5.21) and (5.23), we finally obtain

$$\sup_k \left| P_{K|\tilde{H}_n}\left(k; n\mu + \sqrt{n}I\right) - P_{K|\tilde{Z}_n}\left(k; n\mu + \sqrt{n}I\right)\right| = O\left(\frac{1}{\sqrt{n}}\right). \tag{5.24}$$

Now, it remains to show that the convergence rate of

$$\sup_k \left| P_{K|\tilde{Z}_n}\left(k; n\mu + \sqrt{n}I\right) - B_n P_K(k)\exp\left(\frac{-\psi(I)k}{\sqrt{n}}\right)\right|. \tag{5.25}$$

Then, it suffices to show that all the conditions in Theorem 3.1 are satisfied for K_n, Y_n, Z_n, then we can apply Theorem 3.2.

First, we can check that $\mathbb{E}[K_n^2] = a_n,\ a_n = o(1)$:

$$\mathbb{E}[K_n^2] = \frac{1}{n}\mathbb{E}[K^2] = O\left(\frac{1}{n}\right). \tag{5.26}$$

Second, by change of variables,

$$P_{K|\tilde{H}_n}\left(k; n\mu + \sqrt{n}I\right) = P_{K_n|H_n}\left(\frac{k}{\sqrt{n}}; I\right). \tag{5.27}$$

And we can define the set S in terms of the value for K as below:

$$S = \{k : k \in \mathbb{N},\ \mathbb{P}(K = k) > 0\}$$

such that for all $k \in S$, $P(K_n = \frac{k}{\sqrt{n}}) > 0$. Choose $d > 0$ such that $I = [-h, -h+\delta] \subseteq D = (-d, 0)$. Below, we follow every steps in Theorem 3.1 with slight modifications:

1. For all $y \in \mathbb{R}$, $Y_n = Y \sim N(0,\sigma^2)$, by the formula of the density of normal distribution, we have

$$\frac{\partial^2 P\big(Y \in [y, y+\delta]\big)}{\partial y^2} = f_Y'(y+\delta) - f_Y'(y) \tag{5.28}$$

and

$$\frac{\partial^2 \log P\big(Y \in [y, y+\delta]\big)}{\partial y^2} = \frac{f_Y'(y+\delta) - f_Y'(y)}{P\big(Y \in [y, y+\delta]\big)} - \left(\frac{f_Y(y+\delta) - f_Y(y)}{P\big(Y \in [y, y+\delta]\big)}\right)^2, \tag{5.29}$$

so we can check (5.28) exist and are uniformly bounded. For (5.29), we modify the boundedness slightly and the details of proof are provided in Appendix 6.3. Therefore, (3.10) with a slight modification holds.

2. Since $Y_n = Y \sim N(0, \sigma^2)$, there exist positive constants δ_1 and C depending on y such that $P\big(Y \in [y, y+\delta]\big) \geq \delta_1$ and $0 \leq \dfrac{\partial \log P\big(Y \in [y, y+\delta]\big)}{\partial y} \leq C$ for every $[y, y+\delta] \subset D$. Therefore, (3.11) holds. Since K_n and Y_n are independent, we have $b_n = 0$. Therefore, (3.12) holds.
3. Since $Z_n \to Y$ in distribution where $Y \sim N(0, \sigma^2)$, there exists $\varepsilon_n(z) \to 0$ such that
$$P\big(Z_n \in [z, z+\delta]\big) = P\big(Y \in [z, z+\delta]\big) + \varepsilon_n(z).$$
Since $P\big(Y \in [z, z+\delta]\big)$ is bounded from below for $[z, z+\delta] \subset D$, there exists a positive constant $\delta_2(z)$ such that $P\big(Z_n \in [z, z+\delta]\big) \geq \delta_2 > 0$ for all $[z, z+\delta] \subset D$. Then, the second inequality in (3.13) holds.

To apply Theorem 3.2, we then obtain

$$\sup_{k \in S} \left| P_{K_n|Z_n}\left(\frac{k}{\sqrt{n}}; I\right) - B_n P_{K_n}\left(\frac{k}{\sqrt{n}}\right) \exp\left(-\psi(I)\frac{k}{\sqrt{n}}\right) \right| = O\left(\frac{1}{n}\right), \quad (5.30)$$

where

$$\psi(I) = \frac{\partial \log P_Y\big([y, y+\delta]\big)}{\partial y}\bigg|_{y=-h} \quad \text{and} \quad Y \sim N(0, \sigma^2). \quad (5.31)$$

By change of variable, we then obtain

$$\sup_{k} \left| P_{K|\tilde{Z}_n}\big(k; n\mu + \sqrt{n}I\big) - B_n P_K(k) \exp\left(\frac{-\psi(I)k}{\sqrt{n}}\right) \right| = O\left(\frac{1}{n}\right), \quad (5.32)$$

where

$$B_n = \frac{1}{\sum_{k \in S} P_{K_n}(k/\sqrt{n}) \exp\left(-\psi(I)k/\sqrt{n}\right)} = \frac{1}{\sum_k P_K(k) \exp\left(-\psi(I)k/\sqrt{n}\right)}.$$

By applying triangle inequality to (5.24) and (5.32), we can obtain (5.18) in the theorem.

Finally, we apply Theorem 5.2 to a concrete example.

Example 1. Let $\lambda, \mu > 0$ be two constants. Consider two independent random variables $K \sim \text{Pois}(\lambda)$ and $\tilde{L}_n \sim \text{Pois}(n\mu)$. Let $\tilde{H}_n := K + \tilde{L}_n$. For every fixed finite interval I which follows from Theorem 5.2, we can show that

$$\sup_{k} \left| P\left(K = k \mid \tilde{H}_n \in n\mu + \sqrt{n}I\right) - B_n P(K = k) \exp\left(\frac{-\psi(I)k}{\sqrt{n}}\right) \right| = O\left(\frac{1}{\sqrt{n}}\right),$$

where $B_n = \displaystyle\sum_{k=0}^{\infty} \frac{1}{P(K=k) \exp\left(\frac{-\psi(I)k}{\sqrt{n}}\right)}$ and

$$\psi(I) = \frac{\partial \log P\big(Y \in [y, y+\delta]\big)}{\partial y}\bigg|_{y=-h}, \quad Y \sim N(0, \mu).$$

Proof. By the property of Poisson random variables, we can decompose $\tilde{L}_n$ as $\tilde{L}_n = \sum_{i=1}^{n} \xi_i$, where $\{\xi_i, 1 \leq n\}$ are independent Poisson random variables with mean μ and variance μ. We can check that all conditions are satisfies in Theorem 5.2. Hence, Theorem 5.2 can be applied. □

5.3. Emergence of Temperature (Conditioned on the Scale of Large Deviations)

In this section, we define the parameter $\frac{1}{\varphi(I)}$ in the exponential function $e^{-\varphi(I)x}$ as the *temperature* of the canonical distribution. Consider a sequence of conditional probabilities for a function of a subsystem represented by X in contact with its heat bath represented by $\tilde{Y}_n = \sum_{i=2}^{n} X_i$, where X_i are i.i.d. and X_i has the same distribution as X, and X_i, X are independent. Suppose that the total energy $\tilde{Z}_n = X + \tilde{Y}_n$ is conditioned on the scale of large deviations from its mean, we will show that the *temperature* $\frac{1}{\varphi(I)}$ is an emergent parameter uniquely determined by the rate function of $\frac{\tilde{Y}_n}{n}$.

Definition 5.3. Let X be a nonnegative and nonconstant continuous random variable with $\mathbb{E}[X^4] < \infty$, and let $\tilde{Y}_n := \sum_{i=2}^{n} X_i$, where all random variables in $\{X_i\}_{i=2}^{n} \cup \{X\}$ are i.i.d. Denote $\tilde{Z}_n := X + \tilde{Y}_n$. Consider an interval $I = [d, d+\delta]$, $d \in \mathbb{R}$, $\delta > 0$ with $\mathbb{E}[X] \notin I$, and a function $\varphi : I \to \mathbb{R}$ such that $0 < \varphi(I) < \infty$. Let $\mathbb{P}_I$ be a probability measure with density function $Af_X(x)e^{-\varphi(I)x}$, where

$$\frac{1}{A} = \int_{\mathbb{R}^+} f_X(x) e^{-\varphi(I)x} \mathrm{d}x.$$

Let $\mathbb{Q}_I^{(n)}$ be a sequence of probability measures with density functions $f_{X|\tilde{Z}_n}(x; nI)$.

Theorem 5.3. *Denote* $Y_n := \frac{\tilde{Y}_n}{n}, X_n := \frac{X}{n}$, *and* $Z_n := X_n + Y_n$, *and let* $I - \frac{x}{n} = \{y - \frac{x}{n},\ y \in I\}$. *Assume the following conditions hold:*

1. $\left| \dfrac{f_{X|Z_n}(x; I)}{f_X(x)} \right|$ *is uniformly bounded on* $\mathbb{R}^+$.
2. $|\log P_{Y_n}(I) - \log P_{Z_n}(I)|$ *converges to a finite constant as* $n \to \infty$.
3. *There exists a function* $\phi(y) \in C^2(D)$, *where* D *is an open interval containing* I, *with* $-\infty < \phi'(y) < 0$, *for* $y \in I$, *such that*

$$\log P_{Y_n}\left(I - \frac{x}{n}\right) = -n\phi\left(y^* - \frac{x}{n}\right) + s_n\left(I - \frac{x}{n}\right), \quad \text{for } I - \frac{x}{n} \subset D, \tag{5.33}$$

where $y^* = \{y : \inf_{y \in I} \phi(y)\}$, $\left| \dfrac{s_n(I - \frac{x}{n}) - s_n(I)}{s_n(I)} \right| = O\left(\dfrac{x}{n}\right)$, *and* $|s_n(I')| = o(n)$ *for all* $I' \subset D$.

Then,

$$D_{KL}\left(\mathbb{P}_I \parallel \mathbb{Q}_I^{(n)}\right) \to 0 \qquad \text{if and only if} \qquad \varphi(I) = -\phi'(y^*),$$
$$y^* = \left\{ y : \inf_{y \in I} \phi(y) \right\}. \tag{5.34}$$

Remark 5.2. The conditions (1)–(3) formulated in Theorem 5.3 are technical, so we would like to characterize and verbally describe the underlying meaning and interpretation of them: The condition (1) can be written as

$$\left|\frac{f_{X|Z_n}(x;I)}{f_X(x)}\right| = \left|\frac{f_{X_n,Y_n}\left(\frac{x}{n}, I-\frac{x}{n}\right)}{f_{X_n}\left(\frac{x}{n}\right) f_{Y_n}\left(I-\frac{x}{n}\right)}\right| \quad \text{is uniformly bounded on } \mathbb{R}^+,$$

in which the right-hand side is related to the correlation of X_n and Y_n, therefore, this condition means that the interaction between X_n and Y_n is regulated; The condition (1) is corresponding to the setup that X_n is small relative to Z_n (hence the distributions of Y_n and Z_n have the same asymptotic behavior), specifically, that finite constant can be chosen to be zero (we provide a more general condition in this theorem); The condition (3) means that Y_n converges to a constant satisfying the large deviation principle with the rate function ϕ and the remainder term s_n.

Proof. The proof of Theorem 5.3 is just the application of Theorem 3.4, so we will show that all conditions in Theorem 3.4 are satisfied. First, condition (1) in Theorem 3.4 follows from condition (1), and $\mathbb{E}[X^4] < \infty$ is assumed in this theorem. Second, condition (2) in Theorem 3.4 follows from (i) $Y_n \to \mathbb{E}[X]$ in probability by the law of large numbers, (ii) $\mathbb{E}[X] \notin I$ by Definition 5.3, and (iii) the condition (3) in this theorem.

Third, since I is closed and contained in an open interval D, there exists a constant $d \in \mathbb{R}^+$ such that $I - \dfrac{x}{n} \subset D$ for $x \in [0, nd]$. Therefore, by condition (3),

$$\log P_{Y_n}\left(I-\frac{x}{n}\right) = -n\phi\left(y^*-\frac{x}{n}\right) + s_n\left(I-\frac{x}{n}\right), \quad y^* = \left\{y : \inf_{y\in I}\phi(y)\right\}. \tag{5.35}$$

Since $\left[y^*, y^*-\frac{x}{n}\right] \subseteq D$ and $\phi \in C^2(D)$, by Taylor's expansion,

$$\phi\left(y^*-\frac{x}{n}\right) = \phi(y^*) - \phi'(y^*)\frac{x}{n} + O\left(\frac{x^2}{n^2}\right) \quad \text{for all } x \in [0, nd]. \tag{5.36}$$

By conditions (1) and (3), there exists a sequence $\varepsilon_n \to 0$ and a constant k such that

$$\log P_{Z_n}(I) = \log P_{Y_n}(I) + k + \varepsilon_n = -n\phi(y^*) + s_n(I) + k + \varepsilon_n. \tag{5.37}$$

By condition (3), we have

$$\left|s_n\left(I-\frac{x}{n}\right) - s_n(I)\right| = |s_n(I)|\, O\left(\frac{x}{n}\right) = O(\delta_n x), \tag{5.38}$$

in which $\delta_n \to 0$. By the results of (5.35), (5.36), (5.37), and (5.38), we obtain

$$\log\left(\frac{P_{Y_n}\left(I-\frac{x}{n}\right)}{P_{Z_n}(I)}\right) = \log\left(\frac{\exp\left[-n\phi\left(y^*-\frac{x}{n}\right)\right]}{\exp\left[-n\phi(y^*)\right]}\right) + O\left(\frac{x^2}{n}\right) + O(\delta_n x) + \varepsilon_n \quad \text{on} \quad I_n = [0, nd]. \tag{5.39}$$

Let $r_n(x) := O\left(\frac{x^2}{n}\right) + O(\delta_n x) + \varepsilon_n$, we can check that (i) $\left|r_n(x)e^{-\xi x}\right|$ uniformly bounded on $\mathbb{R}^+$ for any $\xi > 0$, and (ii) $\mathbb{E}\left[r_n(X)^2\right] \to 0$ since $\mathbb{E}\left[X^4\right] < \infty$ by Definition 5.3. Hence, $r_n(x), d_n, \phi$ satisfy condition (3) in Theorem 3.4. Therefore, we have checked that all of the conditions in Theorem 3.4 hold, then we can apply it to get

$$D_{\mathrm{KL}}\left(\mathbb{P}_I \parallel \mathbb{Q}_I^{(n)}\right) \to 0 \quad \text{if and only if} \quad \varphi(I) = -\phi'(y^*). \tag{5.40}$$

□

By Cramér's Theorem 2.3, the existence of the function $\phi(y)$ in condition (3) is from the existence of the rate function of $Y_n = \sum_{i=1}^{n-1} X_i/n$. Let set $D_\phi := \{y \in \mathbb{R} : \phi(y) < \infty\}$ and we can choose $D = \mathrm{int}\,(D_\phi)$. By the properties of rate functions in Appendix 6.1, we have

$$\phi(y) \in C^2(D) \quad , \quad \phi(y) \text{ is convex on } D, \tag{5.41}$$

and $-\infty < \phi'(y) < 0$ for $y \in I \subset D$ if the interval I is chosen on the left side of the mean of Y_n. By Cramér's Theorem, the rate function satisfies

$$\log P_{Y_n}\left(I - \frac{x}{n}\right) = -n\phi\left(y^* - \frac{x}{n}\right) + o(n), \quad \text{for } I - \frac{x}{n} \subset D. \tag{5.42}$$

Comparing (5.42) with condition (3), Theorem 5.3 requires an explicit form of the remainder:

$$\log P_{Y_n}\left(I - \frac{x}{n}\right) = -n\phi\left(y^* - \frac{x}{n}\right) + s_n\left(I - \frac{x}{n}\right), \quad \text{for } I - \frac{x}{n} \subset D, \tag{5.43}$$

where $\left|\frac{s_n(I - \frac{x}{n}) - s_n(I)}{s_n(I)}\right| = O\left(\frac{x}{n}\right)$, and $|s_n(I')| = o(n)$ for all $I' \subset D$. This stronger condition guarantees the "if and only" if statement (5.34).

The following is our discussion on the connection between Theorem 5.3 and Van Campenhout and Cover's Theorem 2.2. In Theorem 5.3, if the condition is on the scale of large deviations, then the conditional density

$$f_{X|\tilde{Z}_n}(x; nI), \ n\mu \notin nI$$

can be approximated by the (normalized) product of its unconditional density $f_X(x)$ and an exponential function $e^{-\lambda x}$. This parameter $\lambda = \phi'(y^*)$ is unique and determined by the first derivative of the rate function evaluated at $y^* = \inf_{y \in I} \phi(y)$. It implies that we are able to find λ directly from the rate function without using the maximum entropy principle. Furthermore, by the pair of reciprocal equations (2.14):

$$\phi'(y^*) = \lambda \quad \text{if and only if} \quad A'(\lambda) = y^*, \tag{5.44}$$

which means the parameter λ we find by the derivative of the rate function (left side of (5.44)) is also the solution of the derivative of the free energy function A under the constraint $= y^*$ (right side of (5.44)).

Therefore, using the maximum entropy principle under the first moment constraint to find good approximations of conditional density (Van Campenhout and Cover's approach) is a natural consequence of the emergent behavior

of

$$\log\left(\frac{f_{X|\tilde{Z}_n}(x;nI)}{f_X(x)}\right). \tag{5.45}$$

And this emergent behavior gives rise to a large deviation function that uniquely determines the parameter of the exponential weight. As we discussed in Sect. 2, we apply the large deviation principle directly to the distribution of a the heat bath

$$Y_n = \frac{\tilde{Y}_n}{n} = \frac{1}{n}\sum_{i=2}^{n} X_i.$$

On the other hand, the Gibbs conditioning principle uses the large deviation principle for empirical measures

$$L_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}.$$

Denote that $X_1 := X$. Then, the limit problem of the sequence of probability measures $\mathbb{Q}_I^{(n)}$ with density functions

$$f_{X|Z_n}(x;I), \quad \text{where } Z_n = X + Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and the limit problem of the sequence of empirical measures

$$\mathbb{E}\left[L_n \mid L_n \in \Gamma\right], \quad \text{where } L_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i} \text{ and } \Gamma = \left\{\gamma : \int x\gamma(\mathrm{d}x) \in I\right\}$$

are just two sides of the same coin. Eventually, they both give arise to a limit as a canonical distribution with the density

$$f_X(x)e^{-\lambda x}.$$

In conclusion, our approach generates λ by the large-deviation rate function of the heat bath Y_n and the Gibbs conditioning principle solves λ by minimizing the relative entropy which is the large-deviation rate function of sampling. These two approaches are connected by the reciprocal equations (5.44) through the Legendre transform.

5.4. Emergence of Temperature (Conditioned on the Scale of Gaussian Fluctuations)

Similar to Sect. 5.3, in this section, we define the parameter $\frac{1}{\beta_n\psi(I)}$ in the exponential function $e^{-\beta_n\psi(I)x}$ as the *temperature* of the canonical distribution and consider a sequence of conditional probabilities for a function of a subsystem represented by X in contact with its heat bath represented by $\tilde{Y_n} = \sum_{i=2}^{n} X_i$, X_i are i.i.d. and X_i has a same distribution as X, and X, X_i are independent. In comparison with Sect. 5.3, here we suppose that the total energy $\tilde{Z}_n := X + \tilde{Y_n}$ is conditioned on the scale of Gaussian fluctuations. We will show that the *temperature* $\frac{1}{\beta_n\psi(I)}$ is an emergent parameter uniquely determined by a normal distribution $N(0,\sigma^2)$, where σ^2 is the variance of X.

Definition 5.4. Let X be a nonnegative and nonconstant continuous random variable with $\mathbb{E}[X^4] < \infty$, and let $\mu = \mathbb{E}[X]$, σ^2 be the variance of X. Let $\tilde{Y}_n = \sum_{i=2}^n X_i$, where all random variables in $\{X_i\}_{i=2}^n \cup \{X\}$ are i.i.d. Denote $\tilde{Z}_n := X + \tilde{Y}_n$. For an interval $I = [d, d+\delta]$, $d \in \mathbb{R}, \delta > 0$ and a function $\psi : I \to \mathbb{R}$ such that $0 < \psi(I) < \infty$. Let $\mathbb{P}_I^{(n)}$ be a sequence of probability measures with density functions $A_n f_X(x) e^{-\frac{\psi(I)}{\sqrt{n}}x}$, where

$$\frac{1}{A_n} = \int_{\mathbb{R}^+} f_X(x) e^{-\frac{\psi(I)}{\sqrt{n}}x} \mathrm{d}x$$

and let $\mathbb{Q}_I^{(n)}$ be a sequence of probability measures with density functions $f_{X|\tilde{Z}_n}(x; n\mu + \sqrt{n}I)$.

Theorem 5.4. *Denote* $Y_n = \frac{\tilde{Y}_n - (n-1)\mu}{\sqrt{n}}$, $X_n = \frac{X}{\sqrt{n}}$, $Z_n = X_n + Y_n$, *and let* $I - \frac{x}{\sqrt{n}} = \left\{ y - \frac{x}{\sqrt{n}},\ y \in I \right\}$. *Assume the following conditions hold:*

1. $\left| \frac{f_{X|Z_n}(x; I)}{f_X(x)} \right|$ *is uniformly bounded on* $\mathbb{R}^+$.
2. $Y_n \to Y$ *in distribution and* $\left. \frac{\partial \log P(Y \in [y, y+\delta])}{\partial y} \right|_{y=d} > 0$, $Y \sim N(0, \sigma^2)$.
3. *There exists a sequence of functions* $g_n : \mathbb{R} \to \mathbb{R}$ *with*

$$\left| g_n(x) e^{-\frac{\xi}{\sqrt{n}}x} \right| \text{ uniformly bounded on } \mathbb{R}^+, \text{ for any } \xi > 0, \quad \text{and} \quad \mathbb{E}\left[g_n(X)^2 \right] \to 0$$

such that

$$\log \left(\frac{P\left(Y_n \in I - \frac{x}{\sqrt{n}} \right)}{P(Z_n \in I)} \right) = \log \left(\frac{P\left(Y \in I - \frac{x}{\sqrt{n}} \right)}{P(Y \in I)} \right) + \frac{g_n(x)}{\sqrt{n}} \text{ on } I_n, \quad (5.46)$$

in which $I_n = [0, d_n]$ *with* $d_n = O(\sqrt{n})$.

Then,

$$nD_{KL}\left(\mathbb{P}_I^{(n)} \parallel \mathbb{Q}_I^{(n)} \right) \to 0 \quad \text{if and only if} \quad \psi(I) = \left. \frac{\partial \log P(Y \in [y, y+\delta])}{\partial y} \right|_{y=d}. \quad (5.47)$$

Remark 5.3. As Remark 5.2, the conditions (1)–(3) formulated in Theorem 5.4 are technical, so we would like to characterize and verbally describe the underlying meaning and interpretation of them: As Theorem 5.3, the condition (1) means that the interaction between X_n and Y_n is regulated; The condition (2) follows from the central limit theorem and we need to choose the interval $I = [d, d+\delta] \subset \mathbb{R}^-$ such that the partial derivative term is positive; The condition (3) combines the setup $X_n \to 0$ in probability and $Y_n \to Y$ in distribution, furthermore, the remainder term has a special form $\frac{g_n(x)}{\sqrt{n}}$.

The proof of Theorem 5.4 is just the application of Theorem 3.3. We can check that all of the conditions in Theorem 3.3 are satisfied. Here, we want to further discuss Eq. (5.46) in condition (3):

As the proof for Theorem 5.2, by Corollary 2.1 of Berry–Esseen theorem and Slutsky's theorem, we have

$$\log\left(\frac{P\left(Y_n \in I - \frac{x}{\sqrt{n}}\right)}{P\left(Z_n \in I\right)}\right) = \log\left(\frac{P\left(Y \in I - \frac{x}{\sqrt{n}}\right)}{P\left(Y \in I\right)}\right) + O\left(\frac{1}{\sqrt{n}}\right) \quad \text{on } I_n. \tag{5.48}$$

However, it only guarantees the convergence of $\mathbb{P}_I^{(n)}$ and $\mathbb{Q}_I^{(n)}$ in $\|\cdot\|_\infty$ by Theorem 5.2. Compare Eq. (5.48) with condition (2), Theorem 5.4 requires an explicit form of the remainder:

$$\log\left(\frac{P\left(Y_n \in I - \frac{x}{\sqrt{n}}\right)}{P\left(Z_n \in I\right)}\right) = \log\left(\frac{P\left(Y \in I - \frac{x}{\sqrt{n}}\right)}{P\left(Y \in I\right)}\right) + \frac{g_n(x)}{\sqrt{n}} \quad \text{on } I_n, \tag{5.49}$$

and $\mathbb{E}[g_n(X)^2] \to 0$. This explicit form of remainder guarantees the "if and only" if statement (5.47).

We now discuss the connection between Theorem 5.4 and Zabell's Theorem 2.1. If the condition is on the scale of Gaussian fluctuations, Theorem 2.1 only tells us that the sequence of conditional distributions $F_{X|\tilde{Z}_n}(x; n\mu + \sqrt{n}I)$ should converge to its unconditional distribution $F_X(x)$. By our theorem 5.4, we have an explicit formula for the canonical distribution to approximate the conditional distribution well:

$$F_{X|\tilde{Z}_n}(x; n\mu + \sqrt{n}I) \approx \int_{-\infty}^{x} A_n f_X(s) e^{-\frac{\psi(I)}{\sqrt{n}}s} \mathrm{d}x,$$

for a sufficiently large n, and it converges to $F_X(x)$ as $n \to \infty$ which is consistent with Zabell's Theorem 2.1. In addition, the parameter $\frac{\psi(I)}{\sqrt{n}}$ of the canonical distribution is uniquely determined if we require that the approximation is "good" enough, i.e., the KL-divergence of the conditional distribution from the canonical distribution converges to zero in the rate $o\left(\frac{1}{n}\right)$.

5.5. Mathematical Definitions of the Heat Bath

In Sect. 3, we provided two limit theorems of a sequence of conditional probabilities to derive a unique canonical distribution as an emergent phenomenon. In Theorem 3.3, the emergent parameter in the exponential weight is uniquely determined by the limiting distribution of the heat bath $Y_n \to Y$ (note that in Theorem 3.3, Y_n follows from the appropriate shifting and scaling of the original heat bath $\tilde{Y}_n$) evaluated on the interval $I = [h, h + \delta]$ such as

$$\psi(I) = \left.\frac{\partial \log P(Y \in [y, y + \delta])}{\partial y}\right|_h.$$

Similarly, in Theorem 3.4, the emergent parameter in the exponential weight is uniquely determined by the large-deviation rate function of the heat bath $Y_n \to \mu$ (note that in Theorem 3.4, Y_n follows from the appropriate shifting

and scaling of the original heat bath $\tilde{Y}_n$) evaluated on the interval $I = [h, h+\delta]$ such as

$$\varphi(I) = -\phi(y^*),$$

where ϕ is the rate function of Y_n and $y^* = \{y : \inf_{y \in I} \phi(y)\}$.

If we choose an interval $I' \subset I$, the parameter in the exponential weight may depend on I' in both of the limit theorems. However, since I' is just a subinterval of I, we expect that a well-defined *heat bath* should give rise to an *invariant temperature* of the canonical distribution by giving a constant parameter in the exponential weight no matter what subinterval I' we choose for it. In this section, we discuss two cases that follow from Theorems 3.3 and 3.4, respectively. Given a finite interval I, we first define the *subinterval invariant property* of a sequence of conditional distributions, then we provide three equivalent properties: (1) the subinterval invariant property of a sequence of conditional distributions (2) the invariant temperature property of the canonical distribution (3) the heat-bath property. Based on the equivalence of these three properties, we truly define the concept of "heat bath" in the language of mathematics.

Recall that X, $\tilde{Y}_n$, and $\tilde{Z}_n := X + \tilde{Y}_n$, are random variables from the definitions in Sect. 3. By proper shifting and scaling, let $X_n := \beta_n X$, $Y_n := \beta_n \left(\tilde{Y}_n - \mu_n\right)$, and $Z_n := X_n + Y_n$, where μ_n, β_n are positive sequences and $\beta_n = o(1)$.

For a finite interval $I = [h, h+\delta],\ h \in \mathbb{R}$ and $\delta > 0$, let $\mathbb{Q}_I^{(n)}$ be a sequence of probability measures with density functions $f_{X|Z_n}(x; I)$. The sequence of conditional probability measures $\mathbb{Q}_I^{(n)}$ represents our setup for the canonical ensemble, which should have a "nice" property such that the limiting behaviors of $\mathbb{Q}_{I'}^{(n)}$ and $\mathbb{Q}_I^{(n)}$ are the same for all subintervals $I' \subset I$. Hence, we define this "nice" property as follows:

Definition 5.5. Note that $\delta(\cdot, \cdot)$ represents the total variation distance of two probability measures. For any given interval $I' \subset I$,

$$\frac{\delta\left(\mathbb{Q}_{I'}^{(n)}, \mathbb{Q}_I^{(n)}\right)}{\alpha_n} \to 0, \tag{5.50}$$

in which we take $\alpha_n = \beta_n$ for Theorem 3.3, and $\alpha_n = 1$ for Theorem 3.4. Then, we say that the sequence of conditional probability measures $\mathbb{Q}_I^{(n)}$ has the *subinterval invariant property* on the interval I.

We start with our first theorem which follows Theorem 3.3. Recall that in Theorem 3.3, Y is a random variable such that $Y_n \to Y$ in distribution.

Theorem 5.5. *For a given interval $I' = [h', h'+\delta'],\ h' \in \mathbb{R}$, $\delta' > 0$, and $I' \subset I$, and a function $\psi : I' \to \mathbb{R}$, let $\tilde{\mathbb{P}}_{I'}^{(n)}$ be a sequence of probability measures with*

density functions

$$\frac{f_X(x)e^{-\beta_n\psi(I')x}}{\int_{\mathbb{R}^+} f_X(x)e^{-\beta_n\psi(I')x}\mathrm{d}x}, \tag{5.51}$$

where

$$\psi(I') = \left.\frac{\partial \log P(Y \in [y, y+\delta'])}{\partial y}\right|_{h'}. \tag{5.52}$$

Assume all of the conditions in Theorem 3.3 hold, then the following three statement are equivalent:

1. $\mathbb{Q}_I^{(n)}$ *has the subinterval invariant property on the interval* I.
2. $\tilde{\mathbb{P}}_{I'}^{(n)}$ *has a unique parameter (the invariant temperature property) such as*

$$\psi(I') = \psi(I) \quad \textit{for all} \quad I' \subset I.$$

3. $Y_n \to Y$ *in distribution and* Y *is a random variable with a distribution function*

$$P(Y \in [h', h'+\delta']) = \alpha(\delta')e^{\psi(I)h'} \quad \textit{for all } [h', h'+\delta'] \subset I, \tag{5.53}$$

where $\alpha : \mathbb{R}^+ \to \mathbb{R}$ *is a function.*

Proof. Since all of the conditions in Theorem 3.3 hold for all intervals $I' \subset I$ with (5.52), we can obtain that

$$\lim_{n\to\infty} \frac{D_{\mathrm{KL}}\left(\tilde{\mathbb{P}}_{I'}^{(n)} \parallel \mathbb{Q}_{I'}^{(n)}\right)}{\beta_n^2} = 0, \quad \text{for all } I' \subset I. \tag{5.54}$$

To prove ((1) $\Rightarrow$ (2) $\Rightarrow$ (3)): Assume the invariant temperature property holds, by applying the triangle inequality and Pinsker's inequality to Eq. (5.54) and the assumption of the subinterval invariant property (5.50) with $\alpha_n = \beta_n$, we have that

$$\frac{\delta\left(\tilde{\mathbb{P}}_{I'}^{(n)}, \tilde{\mathbb{P}}_{I}^{(n)}\right)}{\beta_n} \to 0, \quad \text{for all } I' \subset I. \tag{5.55}$$

Following every step from (4.59) to (4.67) in the proof 4.2.2 for Lemma 4.3, we can get

$$\psi(I') = \psi(I), \quad \text{for all } I' \subset I. \tag{5.56}$$

By (5.52) and (5.56), we have

$$\left.\frac{\partial \log P(Y \in [y, y+\delta'])}{\partial y}\right|_{h'} \equiv \psi(I), \quad \text{for all } [h', h'+\delta'] \subset I, \tag{5.57}$$

which implies Y has a distribution

$$P(Y \in [h', h'+\delta']) = \alpha(\delta')e^{\psi(I)h'}, \quad \text{for all } [h', h'+\delta'] \subset I,$$

with some function $\alpha : \mathbb{R}^+ \to \mathbb{R}$.

To prove ((3) ⇒ (2) ⇒ (1)): By the assumption (3) that

$$P(Y \in [h', h'+\delta']) = \alpha(\delta')e^{\psi(I)h'}, \quad \text{for all } [h', h'+\delta'] \subset I,$$

with some function $\alpha : \mathbb{R}^+ \to \mathbb{R}$, and Eq. (5.52), we can obtain that

$$\psi(I') = \psi(I), \quad \text{for all} \quad I' \subset I, \tag{5.58}$$

therefore, it implies

$$\frac{\delta\left(\tilde{\mathbb{P}}_{I'}^{(n)}, \tilde{\mathbb{P}}_{I}^{(n)}\right)}{\beta_n} = 0, \quad \text{for all } I' \subset I. \tag{5.59}$$

By applying the triangle inequality and Pinsker's inequality to (5.59) and (5.54), we have

$$\frac{\delta\left(\mathbb{Q}_{I'}^{(n)}, \mathbb{Q}_{I}^{(n)}\right)}{\beta_n} \to 0, \quad \text{for all } I' \subset I. \tag{5.60}$$

□

Next, we continue our analysis based on Theorem 3.4. Recall that in Theorem 3.4, $Y_n \to \mu$, for some constant μ, in probability and the sequence of laws of Y_n satisfies a large deviation principle with speed $1/\beta_n$ and rate function ϕ. The rate function $\phi \in C^2(D)$, where D is an open interval containing I, and

$$-\infty < \phi'(y) < 0, \text{ for all } y \in I. \tag{5.61}$$

Theorem 5.6. *For a given interval $I' = [h', h'+\delta']$, $h', \delta' \in \mathbb{R}$, $\delta' > 0$, and $I' \subset I$, and a function $\varphi : I' \to \mathbb{R}$, let $\mathbb{P}_{I'}$ be a probability measure with density function*

$$\frac{f_X(x)e^{-\varphi(I')x}}{\displaystyle\int_{\mathbb{R}^+} f_X(x)e^{-\varphi(I')x}\mathrm{d}x}, \tag{5.62}$$

where

$$\varphi(I') = -\phi'(\hat{y}^*), \quad \hat{y}^* = \left\{ y : \inf_{y \in I'} \phi(y) \right\}. \tag{5.63}$$

Assume all of the conditions in Theorem 3.4 hold, then the following three statements are equivalent:

1. *$\mathbb{Q}_I^{(n)}$ has subinterval invariant property on the interval I.*
2. *$\mathbb{P}_{I'}$ has a unique parameter (invariant temperature property) such as*

$$\varphi(I') = \varphi(I) \quad \text{for all} \quad I' \subset I.$$

3. *Let ϕ be the large-deviation rate function of Y_n. ϕ is a linear function such as*

$$\phi(y) = \phi'(y^*)y + c, \quad \text{for all } y \in I, \tag{5.64}$$

where $y^* = \{y : \inf_{y \in I} \phi(y)\}$ and c is some constant.

Proof. Since all of the conditions in Theorem 3.4 hold for all intervals $I' \subset I$ with (5.63), we can obtain that

$$\lim_{n\to\infty} D_{\mathrm{KL}}\left(\mathbb{P}_{I'} \parallel \mathbb{Q}_{I'}^{(n)}\right) = 0, \quad \text{for all } I' \subset I. \tag{5.65}$$

We first show ((1) ⇒ (2) ⇒ (3)). The proof of ((1) ⇒ (2)) follows from the proof of ((1) ⇒ (2)) in Theorem 5.5, then we can get

$$\varphi(I') = \varphi(I), \quad \text{for all} \quad I' \subset I. \tag{5.66}$$

By (5.63) and (5.66),

$$\phi'(\hat{y}^*) = \phi'(y^*), \quad \text{for all } \hat{y}^* = \left\{ y : \inf_{y\in I'} \phi(y) \right\} \text{ with } I' \subset I.$$

With the assumption (5.61): $-\infty < \phi'(y) < 0$, for all $y \in I$, and the properties of the rate function ϕ in Appendix 6.1, we have that

$$\phi'(y) \equiv \phi'(y^*), \quad \text{for all } y \in I,$$

which implies

$$\phi(y) = \phi'(y^*)y + c, \quad \text{for all } y \in I, \tag{5.67}$$

where c is some constant.

Next we prove ((3) ⇒ (2) ⇒ (1)). Eq. (5.64) implies

$$\phi'(y) \equiv \phi'(y^*), \quad \text{for all } y \in I,$$

then we can obtain

$$\phi'(\hat{y}^*) = \phi'(y^*), \quad \text{for all } \hat{y}^* = \left\{ y : \inf_{y\in I'} \phi(y) \right\} \text{ with } I' \subset I.$$

With (5.63), it implies

$$\varphi(I') = \varphi(I) \quad \text{for all } I' \subset I.$$

Then, the proof of ((2) ⇒ (1)) follows from the proof of ((2) ⇒ (1)) in Theorem 5.5. □

Remark 5.4. The formula (5.53) for the third property (it is called the heat-bath property) in Theorem 5.5 provides the precise formulation of what a heat bath is in probabilistic terms when the heat bath Y_n converges to Y on the scale corresponding to Theorem 3.3; Similarly, the formula (5.64) for the third property in Theorem 5.6 provides the precise formulation of what a heat bath is in probabilistic terms when the heat bath Y_n converges to a constant μ on the scale corresponding to Theorem 3.4. Through these formulations and the equivalence of the three properties: (1) the subinterval invariant property (2) the invariant temperature property (3) the heat-bath property, we really define an invariant temperature bath mathematically.

Acknowledgements

We thank Krzysztof Burdzy, Amir Dembo, Ken Dill, Hao Ge, Hugo Touchette, Jan Van Campenhout, Yue Wang, and Ying-Jen Yang for many helpful discussions.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

6. Appendix

6.1. Properties of the Large-Deviation Rate Function

We include the following properties from [8]. Let $\mathcal{L}$ be the law of X_1, let $\mu := \mathbb{E}[X_1]$ and $\sigma^2 := \mathrm{Var}(X_1)$ and assume that $\sigma > 0$. Let

$$y_- := \inf(\mathrm{supp}(\mathcal{L})), \quad y_+ := \sup(\mathrm{supp}(\mathcal{L}))$$

and ϕ be the function defined in Theorem 2.3. Define

$$\mathcal{D}_\phi := \{y \in \mathbb{R} : \phi(y) < \infty\} \quad \text{and} \quad \mathcal{U}_\phi := int(\mathcal{D}_\phi). \tag{6.1}$$

Then, the following holds:

1. $\phi(y)$ is convex and lower semi-continuous.
2. $0 \le \phi(y) \le \infty$ for all $y \in \mathbb{R}$.
3. $\phi(y) = 0$ if and only if $y = \mu$.
4. $\mathcal{U}_\phi = (y_-, y_+)$ and $\phi(y)$ is infinitely differentiable on $\mathcal{U}_\phi$.
5. $\phi''(y) > 0$ on $\mathcal{U}_\phi$ and $\phi''(\mu) = 1/\sigma^2$.

6.2. Proof of Corollary 2.1

Proof. (2.22) follows from Theorem 2.5 since $Z_n \to G$ in distribution and $W_n \to 0$ in probability. (2.23) basically follows from the proof for Berry–Esseen Theorem (see for example Theorem 2.2.8. in [34]). We include a sketch of the proof here.

Let ϕ_Y be the characteristic function of a random variable Y and $\varepsilon = \mathbb{E}|X|^3/\sqrt{n}$. To prove (2.23), following every step in the proof given in [34], it suffices to show that

$$\int_{|t|<c/\varepsilon} \frac{|\phi_{\tilde{Z}_n}(t) - \phi_G(t)|}{1+|t|} \mathrm{d}t = O(\varepsilon), \tag{6.2}$$

for some small constant c. We can show that

$$\begin{aligned} |\phi_{\tilde{Z}_n}(t) - \phi_G(t)| &= \left| \exp\left[\frac{-t^2}{2} \left(\frac{n+k}{n} \right) + O\left(\varepsilon |t|^3 \left(\frac{n+k}{n} \right) \right) \right] - \exp(-t^2/2) \right| \\ &= O\left(\frac{t^2}{n} \exp(-t^2/4) \right) + O\left(\varepsilon |t|^3 \exp(-t^2/4) \right). \end{aligned} \tag{6.3}$$

Inserting this to (6.2), after integration, the first term in (6.3) has order $O(\frac{1}{n})$ and the second term has order $O(\varepsilon)$. It completes the proof. □

6.3. Proof of the Boundedness of Eq. (5.29)

Denote that

$$A(y) := \frac{\partial^2 \log P\big(Y \in [y, y+\delta]\big)}{\partial y^2} = \frac{f'_Y(y+\delta) - f'_Y(y)}{P\big(Y \in [y, y+\delta]\big)} - \left(\frac{f_Y(y+\delta) - f_Y(y)}{P\big(Y \in [y, y+\delta]\big)}\right)^2. \tag{6.4}$$

We can recognize that

$$A(h - \hat{\alpha}_n x) = 2q_n(x),$$

in which the function $q_n(x)$ is defined in Eq. (4.20) for the proof of Theorem 3.1.

In the entire proof of Theorem 3.1, the only place that we use the condition (3.10) regarding uniformly bounded $A(y)$ when $y \in \mathbb{R}$ is just for the proof of Eq. (4.21) to show that $\exp(-\psi(I)x) \cdot q_n(x)$ is uniformly bounded on $x \in \mathbb{R}^+$. Therefore, instead of proving uniformly bounded $A(y)$ in the condition (3.10), it suffices to show the uniform boundedness of $\exp(-\psi(I)) \cdot q_n(x)$: there exists a constant C such that

$$|\exp(-\psi(I)x) \cdot A(h - \hat{\alpha}_n x)| \le C, \quad \hat{\alpha}_n \in (0,1), \quad \text{for all } x \in \mathbb{R}^+. \tag{6.5}$$

By the mean value theorem and the formula of the density of normal distribution, we can show that there exists $\hat{y}$, $\hat{y} \in (y, y+\delta)$ such that the first term on the right side of (6.4) can be written as

$$\begin{aligned}\frac{f'_Y(y+\delta) - f'_Y(y)}{P\big(Y \in [y, y+\delta]\big)} &= (y+\delta)\exp\left[\frac{-y^2+\hat{y}^2}{2\sigma^2}\right]\left(\exp\left[\frac{-2y\delta-\delta^2}{2\sigma^2}\right] - 1\right) \\ &\quad + \delta \exp\left[\frac{-y^2+\hat{y}^2}{2\sigma^2}\right] \\ &= (y+\delta)\exp\left[\frac{(\hat{y}-y-\delta)(\hat{y}+y+\delta)}{2\sigma^2}\right] \\ &\quad - y\exp\left[\frac{(\hat{y}-y)(\hat{y}+y)}{2\sigma^2}\right]. \end{aligned} \tag{6.6}$$

Recall $y < \hat{y} < y+\delta$. When $\hat{y}+y+\delta \in [0, 2h+\delta]$, (6.6) is uniformly bounded. When $\hat{y}+y+\delta < 0$, we can further have

$$\left|\frac{f'_Y(y+\delta) - f'_Y(y)}{P\big(Y \in [y, y+\delta]\big)}\right| \le (h+\delta)\exp\left[\frac{-\delta(2y+\delta)}{2\sigma^2}\right] + h\exp\left[\frac{-\delta y}{\sigma^2}\right].$$

Therefore,

$$\begin{aligned}&\exp(-\psi(I)x)\left|\frac{f'_Y(y+\delta) - f'_Y(y)}{P\big(Y \in [y, y+\delta]\big)}\right| \\ &\quad \le [(h+\delta)\exp(-\delta^2/2\sigma^2) + h] \cdot \exp\left[\frac{-\delta y}{\sigma^2} - \psi(I)x\right]. \end{aligned} \tag{6.7}$$

By plugging in $y = h - \hat{\alpha}_n x$, $\hat{\alpha}_n \in (0,1)$ in (6.7), since we have $2\delta/\sigma^2 < \psi(I)$ from the assumptions in Theorem 5.2 , we can check the terms on the right-hand side in (6.7) is uniformly bounded when $x \in \mathbb{R}^+$.

The second term on the right side of (6.4) can be written as

$$\left(\frac{f_Y(y+\delta) - f_Y(y)}{P\big(Y \in [y, y+\delta]\big)}\right)^2 = \exp\left[\frac{-y^2+\hat{y}^2}{\sigma^2}\right]\left(\exp\left[\frac{-2y\delta-\delta^2}{2\sigma^2}\right]-1\right)^2. \quad (6.8)$$

When $y + \hat{y} \in [0, 2h+\delta]$, the right-hand side above is uniformly bounded. When $y + \hat{y} < 0$, from (6.8) we have

$$\begin{aligned}
&\exp(-\psi(I)x)\left(\frac{f_Y(y+\delta) - f_Y(y)}{P\big(Y \in [y, y+\delta]\big)}\right)^2 \\
&\le \exp\left[\frac{(\hat{y}-y)(\hat{y}+y)}{\sigma^2} - \psi(I)x\right]\left(\exp\left[\frac{-2y\delta-\delta^2}{2\sigma^2}\right]-1\right)^2 \\
&\le \exp(-\psi(I)x)\left(\exp\left[\frac{-2y\delta-\delta^2}{2\sigma^2}\right]-1\right)^2 \\
&= \exp(-\psi(I)x)\left(\exp\left[\frac{-2y\delta-\delta^2}{\sigma^2}\right] - 2\exp\left[\frac{-2y\delta-\delta^2}{2\sigma^2}\right]+1\right). \quad (6.9)
\end{aligned}$$

By plugging in $y = h - \hat{\alpha}_n x$, $\hat{\alpha}_n \in (0,1)$ in (6.9), since we have $2\delta/\sigma^2 < \psi(I)$, we can check the terms on the right hand is uniformly bounded when $x \in \mathbb{R}^+$. Therefore, combining the estimates in two parts, (6.5) is uniformly bounded for all $x \in \mathbb{R}^+$.

6.4. Proof of Corollaries 3.1, 3.2, and 3.3

6.4.1. Proof of Corollary 3.1. This proof basically follows the proof in Sect. 4.1.1 for Theorem 3.1, so we only provide the details of the difference here. For the derivation of Eq. 4.2, we do Taylor's expansion with respect to x and y for this corollary, so we will get Eqs. (4.3)–(4.5) as following:

$$\begin{aligned}
\phi_n(I) &= \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = 0\big)}{\partial y}\bigg|_{y=h} \\
&\quad - \frac{\partial \log P\big(Y_n \in [y, y+\delta] \mid X_n = 0\big)}{\partial x}\bigg|_{y=h}, \\
r_n(x) &= \frac{1}{2}\frac{\partial^2 P\big(Y_n \in [y, y+\delta] \mid X_n = \xi\big)}{\partial y^2}\bigg|_{y=h-\alpha_n x,\ \xi=\alpha_n x} \\
&\quad + \frac{1}{2}\frac{\partial^2 P\big(Y_n \in [y, y+\delta] \mid X_n = \xi\big)}{\partial \xi^2}\bigg|_{y=h-\alpha_n x,\ \xi=\alpha_n x} \\
&\quad - \frac{\partial^2 P\big(Y_n \in [y, y+\delta] \mid X_n = \xi\big)}{\partial \xi \partial y}\bigg|_{y=h-\alpha_n x,\ \xi=\alpha_n x}. \quad (6.10)
\end{aligned}$$

With the remaining term

$$k_n(x) = \frac{r_n(x)}{P\big(Y_n \in [h, h+\delta] \mid X_n = 0\big)} - \frac{\phi_n(I)^2 e^{-\gamma_n \cdot \phi_n(I)x}}{2},$$

for some $\alpha_n, \gamma_n \in (0,1)$. Then, we obtain

$$f_{X_n|Z_n}(x;I) = \frac{f_{X_n}(x) P(Y_n \in I \mid X_n = 0)(e^{-\phi(I)x} + k_n(x)x^2)}{P(Z_n \in I)}, \quad \text{for } x \in \mathbb{R}^+. \tag{6.11}$$

Based on these new expressions of Eqs. (4.3)–(4.5), Eq. (4.11) becomes

$$\begin{aligned}
&\left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\phi_n(I)} \log\left(\frac{f_{X_n}(x) P_{Y_n|X_n}\big(I-x;x\big)}{P_{Z_n}\big(I\big)} \cdot \frac{1}{A_n f_{X_n}(x) e^{-\phi_n(I)x}} \right) \mathrm{d}x \right| \\
&= \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\phi_n(I)} \log\left(\frac{P_{Y_n|X_n}\big(I-x;x\big)}{P_{Y_n|X_n}\big(I;0\big) e^{-\phi_n(I)x}} \cdot \frac{P_{Y_n|X_n}\big(I;0\big)}{P_{Z_n}\big(I\big) A_n} \right) \mathrm{d}x \right| \\
&\le \left| \log\left(\frac{B_n}{A_n} \right) \right| + \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\phi_n(I)} \log\left(\frac{P_{Y_n|X_n}\big(I-x;x\big)}{P_{Y_n|X_n}\big(I;0\big) e^{-\phi_n(I)x}} \right) \mathrm{d}x \right|,
\end{aligned} \tag{6.12}$$

where

$$A_n := \frac{1}{\int_{\mathbb{R}^+} f_{X_n}(x) e^{-\phi(I)x} dx} \quad \text{and} \quad B_n := \frac{P_{Y_n|X_n}\big(I;0\big)}{P_{Z_n}\,(I)}. \tag{6.13}$$

From the expression of $f_{X_n|Z_n}\big(x;I\big)$ in (6.11), we have the following identity

$$1 = \int_{\mathbb{R}^+} f_{X_n|Z_n}\big(x;I\big)\mathrm{d}x = \frac{B_n}{A_n} + B_n \int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 \mathrm{d}x. \tag{6.14}$$

Equation (6.14) implies

$$\log\left(\frac{B_n}{A_n} \right) = \log\left(1 - B_n \int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 \mathrm{d}x \right). \tag{6.15}$$

Now it remains to show

$$\left| B_n \int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 \mathrm{d}x \right| \tag{6.16}$$

is small for large n.

By the conditions in Corollary 3.1, $P_{Z_n}(I) \ge \delta_2 > 0$, hence there exists a constant $M_1 > 0$ such that

$$B_n = \frac{P_{Y_n|X_n}(I;0)}{P_{Z_n}(I)} \le M_1. \tag{6.17}$$

And since $k_n(x)$ is uniformly bounded as proof 4.1.1 for Theorem 3.1, with the assumption $\mathbb{E}[X_n^2] = a_n$, we can derive that

$$\left| B_n \int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 \mathrm{d}x \right| \le M_1 \cdot \sup |k_n(x)| \cdot \mathbb{E}[X_n^2] = O(a_n). \tag{6.18}$$

Recall (6.15), since $\log(1+x) \le x$ for all $x > -1$, for sufficiently large n, we have

$$\log\left(\frac{B_n}{A_n}\right) = \log\left(1 - B_n \int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 \mathrm{d}x\right)$$
$$\le B_n \int_{\mathbb{R}^+} f_{X_n}(x) k_n(x) x^2 \mathrm{d}x = O(a_n), \tag{6.19}$$

which gives us that the first term in (6.11) is in order $O(a_n)$.

The second term in (6.12) is also in order $O(a_n)$ which follows from the steps (4.20)–(4.22) in Sect. 4.1.1. Therefore, by the definition of KL-divergence (2.17) and Bayes' theorem for conditional probability and the inequality (6.12), we finally obtain

$$\begin{aligned} & D_{\mathrm{KL}}\left(\hat{\mathbb{P}}_I^{(n)} \parallel \mathbb{Q}_I^{(n)}\right) \\ &= \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\phi_n(I)} \log\left(\frac{f_{X_n|Z_n}(x;I)}{A_n f_{X_n}(x) e^{-\phi_n(I)x}}\right) \mathrm{d}x \right| \\ &= \left| \int_{\mathbb{R}^+} A_n f_{X_n}(x) e^{-\phi_n(I)} \log\left(\frac{f_{X_n}(x) P_{Y_n|X_n}(I-x;x)}{P_{Z_n}(I)} \cdot \frac{1}{A_n f_{X_n}(x) e^{-\phi_n(I)x}}\right) \mathrm{d}x \right| \\ &= O(a_n). \end{aligned} \tag{6.20}$$

6.4.2. Proof of Corollaries 3.2 and 3.3. For the proof of Corollary 3.2, since $\log G(I;0) = 0$ and $\log G(I;\xi) \in C(\mathbb{R}^+)$ with respect to ξ, we can do Taylor's expansion for it at zero to get

$$\log G(I;\beta_n x) = \left.\frac{\partial \log G(I;\xi)}{\partial \xi}\right|_0 \beta_n x + O(\beta_n^2 x^2) \quad \text{for } x \in \mathbb{R}^+. \tag{6.21}$$

And similarly, for the proof of Corollary 3.3, since $\log R(I;0) = 0$ and $\log R(I;\xi) \in C(\mathbb{R}^+)$ with respect to ξ, we can do Taylor's expansion for it at zero to get

$$\log\left(R(I;\beta_n x)\right)^{\frac{1}{\beta_n}} = \frac{1}{\beta_n}\left(\left.\frac{\partial \log R(I;\xi)}{\partial \xi}\right|_0 \beta_n x + O(\beta_n^2 x^2)\right) \quad \text{for } x \in \mathbb{R}^+. \tag{6.22}$$

Then, the proof of Corollary 3.2 follows from the proof given in Sect. 4.2.3 with an additional linear term $\frac{\partial \log G(I;0)}{\partial \xi}\beta_n x$ in (4.68), and the proof of Corollary 3.3 follows from the proof in Sect. 4.2.4 with an additional linear term $\frac{\partial \log R(I;0)}{\partial \xi} x$ in (4.69).

References

[1] Ackley, D.H., Hinton, G.E., Sejnowski, R.J.: A learning algorithm for Boltzmann machines. Cogn. Sci. **9**(1), 147–169 (1985)

[2] Anderson, P.W.: More is different. Science **177**(4047), 393–396 (1972)

[3] Barlow, M., Burdzy, K., Timár, Á.: Comparison of quenched and annealed invariance principles for random conductance model. Probab. Theory Rel. Fields **164**(3–4), 741–770 (2016)

[4] Bialek, W.: Biophysics: Searching for Principles. Princeton University Press, Princeton (2012)

[5] Billingsley, P.: Convergence of Probability Measures. Wiley, New York (2013)

[6] Butterfoss, G.L., Hermans, J.: Boltzmann-type distribution of side-chain conformation in proteins. Protein Sci. **12**(12), 2719–2731 (2003)

[7] Cramér, H.: A century with probability theory: some personal recollections. Ann. Probab. **4**(4), 509–546 (1976)

[8] Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications. Applications of Mathematics. Springer, Berlin (1988)

[9] Dembo, A., Zeitouni, O.: Refinements of the Gibbs conditioning principle. Probab. Theory Rel. Fields **104**(1), 1–14 (1996)

[10] Deuschel, J.D., Stroock, D.W., Zessin, H.: Microcanonical distributions for lattice gases. Commun. Math. Phys. **139**(1), 83–101 (1991)

[11] Dill, K.A., Bromberg, S.: Molecular driving forces: statistical thermodynamics in biology. Physics, and nanoscience. Garland Sci. (2012)

[12] Dobrushin, R.L., Tirozzi, B.: The central limit theorem and the problem of equivalence of ensembles. Commun. Math. Phys. **54**(2), 173–192 (1977)

[13] Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time. IV. Commun. Pure Appl. Math. **36**(2), 183–212 (1983)

[14] Friedli, A., Velenik, Y.: Statistical Mechanics of Lattice Systems: A Concrete Mathematical introduction. Cambridge University Press, Cambridge (2017)

[15] Georgii, H.O.: The equivalence of ensembles for classical systems of particles. J. Stat. Phys. **80**(5–6), 1341–1378 (1995)

[16] Georgii, H.O.: Gibbs Measures and Phase Transitions. De Gruyter Studies in Mathematics. De Gruyter, Berlin (2011)

[17] Gibbs, J.W.: Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics. C. Scribner's sons, New York (1902)

[18] Huang, K.: Statistical Mechanics. Wiley, New York (1975)

[19] Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**(4), 620 (1957)

[20] Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)

[21] Khinchin, A.I.: Mathematical Foundations of Statistical Mechanics. Courier Corporation, North Chelmsford (1949)

[22] Landau, L.D., Lifshitz, E.M.: Statistical Physics (Course of Theoretical Physics, vol 5). Pergamon Press, Oxford (1958)

[23] Lanford, O.E.: Entropy and equilibrium states in classical statistical mechanics. In: Lenard, A. (ed.) Statistical Mechanics and Mathematical Problems, pp. 1–113. Springer, Berlin, Heidelberg (1973)

[24] Lewis, J.T., Pfister, C.E., Sullivan, W.G.: Entropy, concentration of probability and conditional limit theorems. Markov Process. Rel. Fields **1**(3), 319–386 (1995)

[25] Martin-Löf, A.: The equivalence of ensembles and the Gibbs phase rule for classical lattice systems. J. Stat. Phys. **20**(5), 557–569 (1979)

[26] Martin-Löf, A.: Statistical Mechanics and the Foundations of Thermodynamics. Springer, Berlin (1979)

[27] Miao, Z., Cao, Y.: Quantifying side-chain conformational variations in protein structure. Sci. Rep. **6**(1), 1–10 (2016)

[28] Pinsker, M.S.: Information and Information Stability of Random Variables and Processes. Holden-Day Series in Time Series Analysis. Holden-Day, San Francisco (1964)

[29] Pressé, S., Ghosh, K., Lee, J., Dill, K.A.: Principles of maximum entropy and maximum caliber in statistical physics. Rev. Mod. Phys. **85**(3), 1115 (2013)

[30] Sanov, I.N.: On the probability of large deviations of random variables. Technical report, North Carolina State University. Department of Statistics (1958)

[31] Sharp, K., Matschinsky, F.: Translation of Ludwig Boltzmann's paper 'On the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium'. Entropy **17**(4), 1971 (2015)

[32] Shorack, G.R.: Probability for Statisticians. Springer Texts in Statistics. Springer, Berlin (2017)

[33] Stroock, D.W., Zeitouni, O.: Microcanonical distribution, Gibbs states, and the equivalence of ensembles. In: Kesten, H., Durrett, R. (eds.) Random Walks Brownian Motion, and Interacting Particle Systems: A Festschrift in Honor of Frank Spitzer, volume 28 of Progress in Probability, pp. 399–424. Springer, New York (2012)

[34] Tao, T.: Topics in Random Matrix Theory. Graduate Studies in Mathematics. American Mathematical Society, Providence (2012)

[35] Tasaki, H.: On the local equivalence between the canonical and the microcanonical ensembles for quantum spin systems. J. Stat. Phys. **172**(4), 905–926 (2018)

[36] Terrell, L.H.: Statistical Mechanics: Principles and Selected Applications. Dover Publications, Mineola (1987)

[37] Touchette, H.: The large deviation approach to statistical mechanics. Phys. Rep. **478**(1–3), 1–69 (2009)

[38] Touchette, H.: Equivalence and nonequivalence of ensembles: thermodynamic, macrostate, and measure levels. J. Stat. Phys. **159**(5), 987–1016 (2015)

[39] Van Campenhout, J., Cover, T.: Maximum entropy and conditional probability. IEEE Trans. Inf. Theory **27**(4), 483–489 (1981)

[40] Vasicek, O.A.: A conditional law of large numbers. Ann. Probab. **8**(1), 142–147 (1980)

[41] Wallace, D.: Naturalness and emergence. http://philsci-archive.pitt.edu/15757/ (2019)

[42] Xiang, Z., Steinbach, P.J., Jacobson, M.P., Friesner, R.A., Honig, B.: Prediction of side-chain conformations on protein surfaces. Proteins Struct. Funct. Bioinform. **66**(4), 814–823 (2007)

[43] Zabell, S.L.: Rates of convergence for conditional expectations. Ann. Probab. **8**(5), 928–941 (1980)

Yu-Chen Cheng and Hong Qian
Department of Applied Mathematics
University of Washington
Seattle WA 98195-3925
USA
e-mail: yuchench@uw.edu;
hqian@uw.edu

Yizhe Zhu
Department of Mathematics
University of California, San Diego
La Jolla CA 92093
USA
e-mail: yiz084@ucsd.edu

Communicated by Christian Maes.
Received: March 7, 2020.
Accepted: December 21, 2020.