Efficient Large Deviation Estimation Based on Importance Sampling

Arnaud Guyader, Hugo Touchette

Efficient large deviation estimation based on importance sampling

Arnaud Guyader*
Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris, France and CERMICS, École des Ponts ParisTech, France

Hugo Touchette†
Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa
(Dated: October 25, 2021)

We present a complete framework for determining the asymptotic (or logarithmic) efficiency of estimators of large deviation probabilities and rate functions based on importance sampling. The framework relies on the idea that importance sampling in that context is fully characterized by the joint large deviations of two random variables: the observable defining the large deviation probability of interest and the likelihood factor (or Radon–Nikodym derivative) connecting the original process and the modified process used in importance sampling. We recover with this framework known results about the asymptotic efficiency of the exponential tilting and obtain new necessary and sufficient conditions for a general change of process to be asymptotically efficient. This allows us to construct new examples of efficient estimators for sample means of random variables that do not have the exponential tilting form. Other examples involving Markov chains and diffusions are presented to illustrate our results.

Keywords: Rare events, importance sampling, large deviations, asymptotic efficiency

I. INTRODUCTION

Estimating the probability of rare events or fluctuations in random systems is an important problem arising in many applied fields, including engineering [1], where a rare event might represent a design failure, or chemistry, where changes between chemical species or polymer states arise from rare transitions in a free energy landscape [2–4]. In physical systems, the probability of rare fluctuations often has a large deviation form [5–8], owing to the interaction of many particles or the effect of thermal noise. In this case, the estimation of probabilities reduces to the estimation of *rate functions*, which determine the rate of decay of probabilities as a function of some parameter (e.g., volume, particle number, integration time or temperature) [8].

Rate functions are also important on their own, as they determine for equilibrium and nonequilibrium systems the onset of static and dynamical phase transitions [8–14], fluctuation symmetries [15–18], and in some cases the response to external perturbations [19]. As a result, they have been actively studied recently, especially for nonequilibrium systems describing particle transport processes [20–23] and diffusing particles [24–27], among other physical systems.

Traditionally, two statistical methods have been used to numerically estimate or sample large deviation probabilities: 1) *splitting* [28–33], also known as cloning in physics [34–37], which works by replicating events that "go in the direction" of the rare event of interest, and 2) *umbrella* or *importance sampling* (IS) [38–41], which works by modifying the process simulated so as to increase the likelihood of the event of interest and, ideally, to render it typical. The probability of that event is then computed via the likelihood factor or Radon–Nikodym derivative, which is the bridge connecting probabilities in the original and the modified processes.

In this paper, we consider the latter method with the aim of providing a complete framework for understanding the efficiency of IS when used to estimate large deviation probabilities and rate functions. For this purpose, we first review in Sec. II the basis of IS as applied to large deviation estimation, and then present the main results known about the efficiency of IS, which we illustrate with simple examples involving sums of random variables.

Most of these results were obtained by Bucklew and Sadowsky [41–44] (see also [45–47]) and are based on two basic but important observations. The first, found in any presentation of IS, is that, although it is necessary for an efficient change of process or "measure" in IS to render rare events typical, this is not sufficient, as we must also ensure that the IS probability estimator arising from the change of process has good variance properties [40]. The second observation, which is specific to large deviations, is that the notion of a "good" or an "efficient" change of process must be adapted to the exponentially decaying form of probabilities that we are trying to estimate. Thus, instead of seeking changes of process that achieve zero variance or a bounded relative error, which are too prohibitive, we must look for changes of process whose second moment decays exponentially with the largest rate possible [41]. This leads to the notion of *logarithmic efficiency* or *asymptotic efficiency*, defined in a precise way in the next section.

Following this review part of the paper, we present in Sec. III a new framework for determining and understanding whether a change of process is asymptotically efficient or not. The framework is itself based on large deviation theory and draws on the idea, recently put forward by one of us [48], that changes of processes and measures in general are completely characterized in the context of large deviation probabilities by the joint rate function of two random variables: 1) the random variable defining the rare event of interest, and 2) the Radon–Nikodym derivative, seen as a real random variable with respect to either the original or the modified process.

The resulting framework recovers results previously known about the efficiency of IS for large deviation estimation [41–45], but also extends them in two important ways. First, most of the results that have been derived in the past and that are now used in practice apply to a specific change of process known as the exponential tilting, the exponential family or the Esscher transform. By contrast, our formalism can be applied in principle to any change of process to determine whether that change is efficient and, if not, to understand in a clear way why this is so. Second, most works provide sufficient but not necessary conditions for asymptotic efficiency. For the exponential tilting, these conditions are based on the existence of so-called dominating points, related essentially to the convexity of rate functions and the convexity of the rare event set. They can be checked in many applications of interest, leading to efficient IS simulations, but they leave completely open the possibility that changes of process other than the exponential tilting can be asymptotically efficient. Indeed, the full characterization of such changes is still an open problem in IS as applied to large deviation estimation.

Here, we solve this problem by providing in Sec. III necessary and sufficient conditions for a change of process to be asymptotically efficient. We use these conditions in Sec. IV to revisit the efficiency of the exponential tilting, and then illustrate them with explicit examples of large deviations involving independent random variables and discrete-time Markov chains. From these, we also construct two intriguing examples of IS estimators that do not have the exponential tilting form and yet are asymptotically efficient, opening the way for more to be discovered. Applications to stochastic differential equations are finally presented to illustrate how our results can be applied beyond discrete-time models to estimate the large deviations of continuous-time Markov processes, commonly used as models of nonequilibrium systems.

II. IMPORTANCE SAMPLING OF LARGE DEVIATIONS

We define in this section the rare event or large deviation probabilities that we are interested in estimating using importance sampling and define the notion of asymptotic efficiency, used classically in the context of large deviations. Most of the results reviewed can be found in Bucklew's book [41], which follows the prior works [42–44].

A. Large deviation probabilities

The rare events that we consider are defined in a general way by considering two ingredients:

- A sequence $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ of random variables taking values on some space Λ_n with probability measure P_n;
- A function $M_n : \Lambda_n \to \mathcal{M}$, referred to as an *observable*.

Concretely, $\mathbf{X}_n$ represents the state of some system or process, P_n is the probabilistic model (the prior measure) that we have of that system, while M_n is some function of that system that can be observed or measured in some way. For example, $\mathbf{X}_n$ can be the microstate of an equilibrium system of n classical particles, in which case P_n is the ensemble (microcanonical, canonical, etc.) chosen to "weight" the microstates and M_n can represent the particles' energy. The system can also be a stochastic process, e.g., a Markov chain in discrete time, with $\mathbf{X}_n$ representing its path or trajectory over n time steps, P_n the probability measure over the trajectories, which defines the process, and M_n a function of the trajectories.

For simplicity, we consider the case where $X_i \in \mathbb{R}^d$, $d \geq 1$, so that $\Lambda_n = (\mathbb{R}^d)^n$, and $M_n(\mathbf{X}_n) \in \mathbb{R}^D$, $D \geq 1$, so that $\mathcal{M} \subset \mathbb{R}^D$. More general spaces can be used for both the process and the observable at the expense of introducing more complicated notations. For example, it is common in large deviation theory to consider $\mathcal{M}$ to be a Polish space to handle cases where M_n takes values in a function space, e.g., if M_n is an empirical distribution, as in Sanov's theorem [6].

Here, we limit ourselves to a setting where both $\mathbf{X}_n$ and M_n are finite-dimensional random variables, so as to simplify the notations. In fact, most of our results will be illustrated by considering simple examples where M_n is a sample mean of real random variables

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{1}$$

so that both $X_i \in \mathbb{R}$ and $M_n \in \mathbb{R}$. From these, it is easy to generalize to other processes and observables, including observables defined for Markov chains or even continuous-time Markov processes, as shown in Sec. IV.

Given $\mathbf{X}_n$, P_n and M_n, we are interested in estimating the probability

$$p_n \equiv P_n(M_n \in B), \tag{2}$$

where B is some measurable subset of $\mathcal{M}$ and P_n denotes, with a slight abuse of notation, the probability measure extended to M_n. As a particular case, we can set $B = [m, m + dm]$ to obtain, as is common in physics, the probability distribution of M_n with "discretization" dm. Our basic assumption is that this probability has a large deviation form with n, meaning that it decays exponentially with n and so describes a rare event that becomes rarer as n gets larger.

This decay of probabilities is encountered in many applications and can be expressed mathematically in different ways, depending on the level of generality adopted. Here, we say that $P_n(M_n \in B)$ has a large deviation form or satisfies, more precisely, the *large deviation principle* (LDP) if there exists a function $I_P : \mathcal{M} \to [0, \infty]$ such that

$$\lim_{n\to\infty} -\frac{1}{n} \log P_n(M_n \in B) = I_P(B), \tag{3}$$

where

$$I_P(B) = \inf_{m \in B} I_P(m). \tag{4}$$

The function I_P is called the *rate function* of M_n and is required to be lower semi-continuous, meaning that it has closed level sets. We assume, as is common in large deviation theory, that I_P is in fact a *good* rate function, meaning that it has compact level sets. This simplifies the analysis of large deviations, as it implies that the infimum in (4) is attained on at least one point in the closure $\bar{B}$ of B [6, Sec. 1.2]. It also means for $M_n \in \mathbb{R}^D$ that I_P is *coercive*, that is, $I_P(m) \to \infty$ as $\|m\| \to \infty$. The first assumption of our work is thus:

Assumption 1. *The observable M_n satisfies the LDP, in the sense of (3), with good rate function I_P such that $I_P(B) < \infty$.*

The limit in (3) is actually a simplification of the standard definition of the LDP found in the large deviation literature involving upper and lower bounds (see, e.g., [6, Sec. 1.2]). In using the definition above, we assume that B is a "good" set, called technically an I-continuity set [6, Sec. 1.2], such that

$$\inf_{m\in B} I_P(m) = \inf_{m\in B^\circ} I_P(m) = \inf_{m\in \bar{B}} I_P(m), \tag{5}$$

where B° represents the interior of B. In this case, the upper and lower bounds appearing in the standard definition of the LDP are the same, yielding the simple limit (3). This is a technical point, which is not important for physical or numerical applications.

Concretely, the LDP means again that the leading behavior of the distribution of M_n is a decaying exponential in n, with corrections in the exponential that are smaller than linear in n. This property is commonly summarized in large deviation theory by the asymptotic notation [5–8]

$$P_n(M_n \in [m, m+dm]) \asymp e^{-nI_P(m)}, \tag{6}$$

and applies whenever $I_P(m) > 0$. When $I_P(m) = 0$, the distribution of M_n either decays around m slower than exponentially in n or increases with n around that point. In many applications, $I_P(m)$ has only one zero, denoted in the following by m^*, so the latter case applies, yielding the law of large numbers

$$\lim_{n\to\infty} P_n(M_n \in [m^*, m^*+dm]) = 1 \tag{7}$$

or, more generally,

$$\lim_{n\to\infty} P_n(M_n \in B) = 1 \tag{8}$$

if $m^* \in B^\circ$. See [8] and references therein for cases where more than one zeros occur.

Probabilities having the LDP form are encountered in many applications of interest, including queues [1], hypothesis testing [7], and noisy detection systems [6]. In physics, the LDP is the basis of thermodynamics and describes, more generally, the fluctuations of equilibrium systems in the thermodynamic limit of large systems, which is a large deviation limit (see [8] for a review). The same exponential form of probabilities also arises in the context of nonequilibrium systems when considering systems perturbed by a small noise [49–52] as well as time-integrated functions or observables of Markov processes [53] modelling, for example, the fluctuating dynamics of mesoscopic diffusive systems [24–26] or many-particle transport processes [20–22]. In the latter case, the long-time limit is often combined with a low-noise limit describing the residual noise associated with a macroscopic (or hydrodynamic) limit where infinitely many interacting particles evolve in time over a substrate (e.g., a lattice) with boundary reservoirs [54].

B. Importance sampling

The simplest way to numerically estimate p_n in (2) is to sample M_n directly by generating multiple copies $\mathbf{X}_n^{(i)}$, $i = 1, 2, \ldots, N$, of the state from the probability measure P_n and by then counting the fraction of corresponding observable values $M_n^{(i)} = M_n(\mathbf{X}_n^{(i)})$ that fall in B:

$$\tilde{p}_n^N \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{M_n^{(i)} \in B}. \tag{9}$$

Since the random variables $\mathbf{1}_{M_n^{(i)} \in B}$ are independent and identically distributed (i.i.d.) Bernoulli with parameter p_n, it is easy to see that the estimator above, referred to as the direct or *crude Monte Carlo* (CMC) estimator, is unbiased in the sense that

$$\mathbb{E}_P[\tilde{p}_n^N] = p_n, \tag{10}$$

where $\mathbb{E}_P[\cdot]$ denotes the expectation with respect to P_n. Moreover, its variance is

$$\mathrm{Var}_P(\tilde{p}_n^N) = \mathbb{E}_P[(\tilde{p}_n^N - p_n)^2] = \frac{p_n(1-p_n)}{N} \tag{11}$$

and so decreases with N. However, since p_n becomes exponentially small as $n \to \infty$, the actual number of samples needed to accurately estimate this probability should be determined from the estimator's error relative to p_n, which can be approximated by

$$\frac{\sqrt{\mathrm{Var}_P(\tilde{p}_n^N)}}{p_n} \approx \frac{1}{\sqrt{N p_n}}. \tag{12}$$

As a result, we see that N must grow exponentially as $N \sim p_n^{-1} \asymp e^{nI_P(B)}$ in order for the relative error to be bounded in n, which is unachievable in practical simulations.

To overcome this problem, we resort to IS which works by sampling M_n not according to P_n but to a different probability measure Q_n, chosen to increase the likelihood that $M_n \in B$ [39–41]. To be consistent, Q_n must have support on all states that "hit" the event $\{M_n \in B\}$ with respect to P_n, which translates mathematically to requiring that $P_n \mathbf{1}_{M_n \in B}$, the restriction of P_n on $\{M_n \in B\}$, be absolutely continuous with respect to $Q_n \mathbf{1}_{M_n \in B}$ [40]. Here, we assume for simplicity that Q_n has the same support as P_n, so the two are equivalent in the sense of absolute continuity.

To estimate p_n, we now generate copies $\mathbf{X}_n^{(i)}$, $i = 1, 2, \ldots, N$, of the states according to Q_n,[1] compute the associated observable values $M_n^{(i)}$, $i = 1, 2, \ldots, N$, and construct the IS estimator as

$$\hat{p}_n^N \equiv \frac{1}{N} \sum_{i=1}^{N} L_n^{(i)} \mathbf{1}_{M_n^{(i)} \in B}, \tag{13}$$

where $L_n^{(i)} = L_n(\mathbf{X}_n^{(i)})$ and

$$L_n \equiv \frac{dP_n}{dQ_n} \tag{14}$$

is the *Radon–Nikodym derivative* of P_n with respect to Q_n. This derivative, also known as the likelihood factor, is included to ensure that the IS estimator remains unbiased, that is,

$$\mathbb{E}_Q[\hat{p}_n^N] = \int_{\Lambda_n} dQ_n(\mathbf{X}_n) \frac{dP_n}{dQ_n}(\mathbf{X}_n) \mathbf{1}_{M_n(\mathbf{X}_n) \in B} = \mathbb{E}_P[\mathbf{1}_{M_n \in B}] = p_n. \tag{15}$$

[1] We could identify the new copies with a different symbol, say $\tilde{\mathbf{X}}_n^{(i)}$, since they are generated from a different distribution and so represent a different random variable. Here, we keep $\mathbf{X}_n^{(i)}$ but always specify the distribution, P_n or Q_n, used. The same applies to the observable.

The variance, however, is modified to

$$\mathrm{Var}_Q(\hat{p}_n^N) = \mathbb{E}_Q[(\hat{p}_n^N - p_n)^2] = \frac{\mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}] - p_n^2}{N}, \tag{16}$$

which obviously depends on Q_n, leading to the relative variance

$$\frac{\mathrm{Var}_Q(\hat{p}_n^N)}{p_n^2} = \frac{\mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}]}{N p_n^2} - \frac{1}{N}. \tag{17}$$

The problem of IS is to determine which Q_n is *optimal*, that is, which choice achieves the smallest variance or relative variance, ideally smaller than the CMC variance obtained with $Q_n = P_n$, given some design or application-specific constraints on the class of Q_n that can be simulated.

If no constraints are imposed, then it is known that there is a zero-variance change of measure given by the restriction of P_n on the event of interest, that is, $Q_n \propto P_n \mathbf{1}_{M_n \in B}$. This measure, known in physics as the microcanonical ensemble [55], cannot be simulated in practice, however, because it involves a hard-to-implement constraint and, more importantly, because its normalization involves the probability that we are trying to estimate. As a result, other choices must be considered that are either approximations of the zero-variance solution (following, e.g., cross-entropy methods [56]) or that are optimal or efficient according to some bounding criterion on the variance or relative variance. For the purpose of estimating large deviations, a common criterion used is the *asymptotic (or logarithmic) efficiency* [39–41], discussed next.

C. Asymptotic efficiency

The notion of asymptotic efficiency is based on the relative variance of the IS estimator, as given by (17). Since p_n scales exponentially with n, so does generally the second moment $\mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}]$ with a scaling exponent given by

$$R_Q(B) \equiv \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}]. \tag{18}$$

We will provide in Section III A specific assumptions that ensure the existence of this limit (see Lemma 2). For now, we note that the LDP for p_n, combined with the positivity of the variance in (16), implies

$$R_Q(B) \leq 2 I_P(B). \tag{19}$$

When equality is achieved, we say that the IS measure Q_n or, more precisely, the sequence $(Q_n)_{n>0}$ of IS measures, is *asymptotically efficient*. This criterion is also referred to in the literature as *logarithmic efficiency* or *asymptotic optimality*.

It can be checked that CMC achieves $R_Q(B) = I_P(B)$ and so it is not asymptotically efficient, as expected, while the zero-variance choice $Q_n \propto P_n \mathbf{1}_{M_n \in B}$ is asymptotically efficient, since it has zero variance for all n. By comparison, an asymptotically efficient Q_n does not necessarily have zero variance – this is again too restrictive for our purpose – but is such that the term $\mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}]$ in the variance decays with the fastest exponential rate equal to $2I_P(B)$. When this happens, the ratio $\mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}]/p_n^2$ in (17) does not grow exponentially with n, which means that the number N_n of samples needed to have a fixed relative variance grows sub-exponentially in n. Hence, if Q_n is asymptotically efficient, we have

$$\lim_{n\to\infty} \frac{1}{n} \log N_n = 0. \tag{20}$$

This is often taken as a definition of asymptotic efficiency.

The asymptotic efficiency of Q_n is studied in most works [41–46] for the *exponential tilting* or *exponential family*, defined by

$$Q_n(d\mathbf{X}_n) = \frac{e^{n\langle k, M_n\rangle} P_n(d\mathbf{X}_n)}{\mathbb{E}_P[e^{n\langle k, M_n\rangle}]}, \tag{21}$$

where $k \in \mathbb{R}^D$ is a vector having the same dimension as M_n and $\langle \cdot, \cdot \rangle$ is the standard scalar product in $\mathbb{R}^D$. We will not review all the results known about this change of measure, which also corresponds in physics to the canonical ensemble [55]. For our purposes, two results are worth noting. The first, proved in [43], states that the exponential tilting is asymptotically efficient if B has a dominating point (see [41, Sec. 5.2] for a definition of this concept), which holds essentially when $I_P(m)$ is a convex function and B is a convex set. In that case, the value $k \in \mathbb{R}^D$ that must be chosen in (21) to achieve efficiency satisfies

$$\nabla \lambda_P(k) = \mu, \tag{22}$$

where $\mu \in \mathbb{R}^D$ is the dominating point of B and $\lambda_P(k)$ is the *scaled cumulant generating function* (SCGF), defined by

$$\lambda_P(k) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_P[e^{n\langle k, M_n\rangle}], \qquad k \in \mathbb{R}^D. \tag{23}$$

We refer to [43, Thm. 2] for the complete statement of this result, including the conditions underlying it, and [44, Thm. 3] for its application to Markov chains. We give some examples next to illustrate the relation (22), which comes from the application of the Gärtner–Ellis theorem and the fact, more precisely, that the rate function is given, according to this theorem, by the Legendre–Fenchel transform of the SCGF when the latter is differentiable; see [8, Sec. 4.4] for more details. In fact, the conditions underlying the efficiency of the exponential tilting are overall nothing but the conditions of the Gärtner–Ellis theorem.

The second result worth noting, also found in [43], is that the sample size N_n required for the IS estimator $\hat{p}_n^N$ to have a bounded relative variance grows according to

$$N_n \asymp e^{n[2I_P(B) - R_Q(B)]} \tag{24}$$

in the limit $n \to \infty$. This essentially follows from the result (17) for the relative variance of the IS estimator, in which the term $1/N$ can be neglected. In particular, $N_n \asymp e^{nI_P(B)}$ for CMC, as seen before, while $N_n \asymp e^{n0}$ if Q_n is asymptotically efficient, consistently with the limit (20) above and the fact again that constant relative variance is achieved in this case by increasing N_n sub-exponentially with n. In some cases, it turns out in fact that bounded relative variance is achieved with $N_n = O(\sqrt{n})$ when Q_n is asymptotically efficient [44, Sec. 5.4], leading to a drastic increase in simulation efficiency.

D. Examples

We close this section by illustrating the theory developed so far with simple examples involving the sample mean

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{25}$$

of a sequence $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ of i.i.d. random variables. The examples are presented briefly, since they appear in other works (see, e.g., [41]), and will be used again in the next sections to illustrate our new framework. More involved applications of IS related to large deviations have been considered in the context of random graphs [57–59], finance [60], escape pathways [61], and nonequilibrium systems [62–65], among other topics.

Example 1: We first consider a sample mean of standard Gaussian random variables, so that P_n is the product measure $\mathcal{N}(0,1)^{\otimes n}$, and look for the probability that $p_n = P_n(M_n \geq 1)$ by choosing $B = [1, \infty)$. This probability can be found directly from the fact that $M_n \sim \mathcal{N}(0, 1/n)$, leading to $p_n \asymp e^{-n/2}$. Alternatively, we can use Cramér's theorem [6, Sec. 2.2] to find that the SCGF is $\lambda_P(k) = k^2/2$, yielding $I_P(m) = m^2/2$ by Legendre–Fenchel transform and, therefore,

$$I_P(B) = \inf_{m \geq 1} I_P(m) = \frac{1}{2}. \tag{26}$$

This shows that the probability $P_n(M_n \geq 1)$ is dominated exponentially by the probability that M_n is close to 1, so only the boundary of B plays a role, as is common with large deviations.

With this result, it is natural to choose the IS measure to be a sequence of i.i.d. Gaussian random variables centered at 1, so that $Q_n = \mathcal{N}(1,1)^{\otimes n}$. It is clear that this change of measure makes $M_n = 1$ typical. Moreover, it can be checked by calculating $R_Q(B)$ directly from its definition (18) that this measure is asymptotically efficient with

$$R_Q(B) = 1 = 2I_P(B). \tag{27}$$

Alternatively, one can notice, following [41, Ex. 5.2.1], that the dominating point of $B = [1, \infty)$ is $\mu = 1$, which leads, with the equation $\lambda_P'(k) = \mu$, to $k = 1$. From (21), the exponentially-tilted measure that is asymptotically efficient is then

$$Q_n(d\mathbf{X}_n) = \frac{e^{nM_n} P_n(d\mathbf{X}_n)}{\mathbb{E}_P[e^{nM_n}]} = \left\{ \prod_{i=1}^{n} \frac{e^{-(X_i-1)^2/2}}{\sqrt{2\pi}} dX_i \right\}, \tag{28}$$

which is indeed the product measure $\mathcal{N}(1,1)^{\otimes n}$. Note that the Radon–Nikodym derivative of P_n with respect to Q_n is

$$L_n = L_n(\mathbf{X}_n) = \frac{dP_n}{dQ_n}(\mathbf{X}_n) = e^{-n(M_n - 1/2)}. \tag{29}$$

Therefore, in the end, the IS estimator that is asymptotically efficient is

$$\hat{p}_n^N = \frac{1}{N} \sum_{i=1}^{N} e^{-n(M_n^{(i)} - 1/2)} \mathbf{1}_{M_n^{(i)} \geq 1}, \tag{30}$$

where $\{M_n^{(i)}\}_{i=1}^N$ is an i.i.d. sample generated with Q_n. ■

This example can be generalized, obviously, to any $B = [b, \infty)$, $b > 0$, by choosing k in the exponential tilting such that $\lambda_P'(k) = b$ or, equivalently, $k = I'(b)$ by Legendre duality (see Sec. 3.5 of [8]). This gives $Q_n = \mathcal{N}(b,1)^{\otimes n}$ as the modified measure that changes the event $\{M_n \geq b\}$ from being rare to being typical. This is asymptotically efficient, as b is the dominating point of B. Choosing Q_n to concentrate *inside* B rather than at its boundary, that is, $Q_n = \mathcal{N}(b', 1)^{\otimes n}$ with $b' > b$, is not asymptotically efficient, although it does make $\{M_n \geq b\}$ typical.

As a variation of this example, we change the distribution of the X_i's to an exponential distribution. Other distributions, such as Bernoulli, uniform or Laplace, are treated in [41].

Example 2: Let the sequence $X_1, X_2, \ldots, X_n$ of i.i.d. random variables be distributed according to the exponential distribution $\mathcal{E}(1)$ with parameter 1, so that $P_n = \mathcal{E}(1)^{\otimes n}$. We consider again the sample mean M_n as an observable and $B = [b, \infty)$ with $b > 1$, so that $p_n = P_n(M_n \in B)$ is a rare event such that [8]

$$I_P(B) = b - 1 - \log b. \tag{31}$$

As shown in [41, Ex. 5.2.6], the asymptotically efficient exponential tilting associated with this problem is the product measure $Q_n = \mathcal{E}(1/b)^{\otimes n}$ of exponential distributions with mean $\mathbb{E}_Q[X_i] = \mathbb{E}_Q[M_n] = b$. This follows by noting that $\lambda_P(k) = -\log(1-k)$ for $k < 1$, from which we find $k = 1 - 1/b$ by solving $\lambda_P'(k) = b$. Equivalently, $k = I'(b) = 1 - 1/b$. ■

The last example is a classic one in IS showing that the exponential tilting is not always asymptotically efficient, in particular, when dealing with nonconvex sets B.

Example 3: Consider, as in the first example, a sequence of i.i.d. normal random variables with the same P_n and Q_n, but now take B to be the union of two disjoint sets, namely, $B = (-\infty, -b] \cup [1, \infty)$ with $b > 1$, so that the probability to estimate is

$$p_n = P_n(M_n \leq -b \text{ or } M_n \geq 1). \tag{32}$$

From Example 1, we still have $I_P(B) = 1/2$, since $M_n \leq -b$ is rarer than $M_n \geq 1$ for $b > 1$. The calculation of $R_Q(B)$ for this case can be found in [41, Ex. 5.2.13]. The result is $R_Q(B) = 1$ if $b \geq 3$ and $R_Q(B) < 1$ otherwise, implying that $Q_n = \mathcal{N}(1,1)^{\otimes n}$ is not asymptotically efficient if $b \in (1, 3)$. Note that this cannot be inferred from the dominating point result, since B is nonconvex and, as such, has no dominating point for any b. ■

The non-efficiency of Q_n in the last example is due to the fact that, although the probability that $M_n \leq -b$ is exponentially small, this rare event leads to exponentially large values of the likelihood factor in (30) that dramatically increase the variance of the IS estimator. In fact, it can be checked (see again [41, Ex. 5.2.13]) that for values of b close to -1, $R_Q(B)$ becomes negative, so that the second moment of the IS estimator can diverge with n as a result of the accumulation of many different likelihood factors that are exponentially large with n.

Other examples involving nonconvex sets B have been studied, in particular, by Glasserman and Wang [66], who show that IS based on the exponential tilting can perform worse than CMC and can even lead to $R_Q(B) = -\infty$, so one should be cautious about the generally-accepted idea that a good choice of IS measure is one that makes a rare event typical.

To avoid the case where $R_Q(B) = -\infty$, which is clearly not efficient, we assume here that $R_Q(B) > -\infty$. In fact, for the results to come, we need a slightly stronger assumption:

Assumption 2. *For some $\delta > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_Q[L_n^{2+\delta} \mathbf{1}_{M_n \in \bar{B}}] < \infty. \tag{33}$$

This condition implies with Hölder's inequality that, if $R_Q(B)$ exists, then $R_Q(B) > -\infty$ and, therefore, with Assumption 1 and (19), that $R_Q(B)$ is finite. It also ensures overall that we are not in a situation where the second moment of L_n has the correct exponential scaling in n, but its moment of order $2 + \delta$ behaves super-exponentially.

This type of "Lyapunov" condition often appears in large deviation theory in the context of Varadhan's integral lemma (see [6, Thm. 4.3.1]). Other weaker conditions can be defined (see [6, Thm. 4.3.1] and [46] in the context of IS), although they might be more difficult to check. In our case, we will see in the next section that the limit above can be re-expressed more naturally in terms of the steepness or coercivity of a rate function involving M_n and L_n (see Assumption 2′).

III. JOINT LDP APPROACH TO ASYMPTOTIC EFFICIENCY

The theory presented in the previous section can be applied to sample large deviations in an efficient way not just for i.i.d. sample means, as illustrated, but also for functionals of Markov chains, jump processes, and diffusions. One problem of the theory, however, is that it focuses almost exclusively on the exponential tilting, leaving aside the possibility that other changes of measure might also be asymptotically efficient. Moreover, the efficiency conditions that we have for the exponential tilting, based on the existence of a dominating point, are only sufficient conditions that cannot be applied to nonconvex problems, as illustrated in Example 3. In principle, one can determine the efficiency of an arbitrary Q_n by calculating the exponent $R_Q(B)$ [41], but this is very difficult to carry out in practice beyond the case of i.i.d. sample means and convex B.

We address these issues in this section by providing new conditions for a general change of measure Q_n to be asymptotically efficient. These conditions follow from two basic observations about the second moment $\mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}]$ that determines the efficiency of Q_n via (19). The first is that this moment involves both L_n and M_n, which means that it can be computed knowing the joint distribution of these two (generally correlated) random variables. The second is that, in many cases of interest, the Radon–Nikodym derivative L_n scales exponentially in n and has a distribution that satisfies the LDP [67]. Therefore, it is natural to study the efficiency of Q_n based on the joint large deviations of M_n and L_n, which is what we do in this section.

By reformulating the asymptotic efficiency criterion in terms of a joint LDP involving M_n and L_n, we derive necessary and sufficient conditions for a general Q_n to be asymptotically efficient. These conditions provide new insights into what makes a change of measure efficient. They show, in particular, that L_n does not have to be deterministic conditionally on M_n, which is the essential property of the exponential tilting that makes it asymptotically efficient. The fluctuations of L_n only need to be "bounded" or "controlled" in a precise way, suggesting new changes of measure, different from the exponential tilting, that are asymptotically efficient.

A. Joint large deviations

The idea of formulating a joint LDP for M_n and L_n follows the recent work of one of us [48]. As in that work, we consider L_n via the scaled log-likelihood or *action*, defined as

$$W_n \equiv -\frac{1}{n} \log L_n, \tag{34}$$

to account for the fact that L_n is expected to scale exponentially with n. The action is obviously a real random variable whose distribution can be determined in principle with respect to either P_n or Q_n. The couple (M_n, W_n) is thus a random variable taking values in the product space $\mathcal{M} \times \mathbb{R}$, where $\mathcal{M}$, the space of M_n, is again a subset of $\mathbb{R}^D$.

From now on, we assume the following:

Assumption 3.

(a) *(M_n, W_n) satisfies the LDP relative to P_n on $\mathcal{M} \times \mathbb{R}$ with good rate function J_P;*

(b) *(M_n, W_n) satisfies the LDP relative to Q_n on $\mathcal{M} \times \mathbb{R}$ with good rate function J_Q;*

(c) *J_P and J_Q have the same non-empty domain on $\mathcal{M} \times \mathbb{R}$, that is, the same set of values on which these functions are finite.*

(d) *$B \times \mathbb{R}$ is a good set for $(m, w) \mapsto 2w + J_Q(m, w)$, meaning that*

$$\inf_{(m,w)\in B\times\mathbb{R}} \{2w + J_Q(m,w)\} = \inf_{(m,w)\in B^\circ\times\mathbb{R}} \{2w + J_Q(m,w)\} = \inf_{(m,w)\in \bar{B}\times\mathbb{R}} \{2w + J_Q(m,w)\}\,. \tag{35}$$

In the last assumption, there is an abuse of language, since the function $2w + J_Q(m, w)$ is not necessarily a rate function.

The LDPs in Conditions (a)-(b) follow the rigorous definition given in Sec. II and mean in terms of the asymptotic notation that

$$P_n(M_n \in B, W_n \in C) \asymp \exp\left\{-n \inf_{m\in B, w\in C} J_P(m,w)\right\} \tag{36}$$

and

$$Q_n(M_n \in B, W_n \in C) \asymp \exp\left\{-n \inf_{m\in B, w\in C} J_Q(m,w)\right\} \tag{37}$$

for "good" sets $B \times C$. More concretely, we can also write

$$P_n(M_n \in [m, m + dm], W_n \in [w, w + dw]) \asymp e^{-nJ_P(m,w)} \tag{38}$$

and

$$Q_n(M_n \in [m, m + dm], W_n \in [w, w + dw]) \asymp e^{-nJ_Q(m,w)}. \tag{39}$$

As for Condition (c), it follows from our previous assumption that P_n and Q_n have the same support on Λ_n, so they also have the same support when pushed forward to $\mathcal{M} \times \mathbb{R}$ with the random variables $(M_n(\mathbf{X}_n), W_n(\mathbf{X}_n))$. This property is again not essential, but simplifies the derivation and analysis of our results.

We will show in Sec. IV how the two joint LDPs can be derived in practice using techniques from large deviation theory. The existence of these LDPs can be viewed as a strong assumption of our theory, but they are necessary, as will become clear, to fully understand the asymptotic efficiency of Q_n.

For now, we assume that two rate functions J_P and J_Q for (M_n, W_n) are given and proceed to relate them to $I_P(B)$ and $R_Q(B)$. To this end, it is important to note that, although J_P and J_Q are defined with respect to P_n and Q_n, respectively, both rate functions actually depend on Q_n, since they both involve the action W_n defined from L_n. The joint rate functions are also linked, since expectations with respect to Q_n are related to expectations with respect to P_n, and vice versa, via the identity

$$\mathbb{E}_Q[L_n(\mathbf{X}_n)\, f(\mathbf{X}_n)] = \mathbb{E}_P[f(\mathbf{X}_n)], \tag{40}$$

where f is any test function. This result was already used in (15) to show that the IS estimator $\hat{p}_n^N$ is unbiased, and implies the following large deviation result, referred to in physics as a *fluctuation relation* [18]:

Proposition 1 ([48, Prop. 2])**.** *Under Assumption 3, the two rate functions J_P and J_Q are such that*

$$J_P(m, w) = J_Q(m, w) + w \tag{41}$$

for all (m, w) in their domain.

This result simply follows by applying (40) with indicator functions to transform joint probability distributions as follows:

$$\begin{aligned} P_n(M_n \in dm, W_n \in dw) &= \mathbb{E}_P[\mathbf{1}_{M_n \in dm}\mathbf{1}_{W_n \in dw}] \\ &= \mathbb{E}_Q[e^{-nW_n}\mathbf{1}_{M_n \in dm}\mathbf{1}_{W_n \in dw}] \\ &= e^{-nw}Q_n(M_n \in dm, W_n \in dw). \end{aligned} \tag{42}$$

Here, we have used $L_n = e^{-nW_n}$ and the shorthand $M_n \in dm$ to mean $M_n \in [m, m+dm]$ (similarly for $W_n \in dw$). Taking the large deviation limit then yields (41). We refer to [48] for a rigorous presentation of this argument, based on the definition of the LDP and Assumption 3.

The existence of a joint LDP for (M_n, W_n) also implies that M_n and W_n satisfy the LDP individually. This "marginalization" of joint LDPs is covered in [48, Prop. 1] and follows in large deviation theory from the contraction principle [5–8], stated for convenience in Appendix B. The application of this principle to marginalize (viz., trace out) W_n, for example, gives the following representation of the rate function of M_n with respect to P_n:

$$I_P(m) = \inf_{w \in \mathbb{R}} J_P(m, w). \tag{43}$$

Therefore,

$$I_P(B) = \inf_{m \in B, w \in \mathbb{R}} J_P(m, w) = \inf_{m \in B, w \in \mathbb{R}} \{w + J_Q(m, w)\}, \tag{44}$$

where we have used Proposition 1 to obtain the second equality. Similar formulas apply with respect to Q_n, including

$$I_Q(m) = \inf_{w \in \mathbb{R}} J_Q(m, w), \tag{45}$$

which is the rate function of M_n associated with its LDP with respect to Q_n.

At this point, we formulate one more assumption needed to derive our main result:

Assumption 4. *There exists a unique, finite pair* (m^*, w^*) *such that* $J_Q(m^*, w^*) = 0$.

This assumption means concretely that the pair (M_n, W_n) satisfies the weak law of large numbers with respect to Q_n (see [68, Thm. 2.5]), that is,

$$\lim_{n \to \infty} Q_n\left(M_n \in [m^*, m^* + dm], W_n \in [w^*, w^* + dw]\right) = 1. \tag{46}$$

In this case, we say that (m^*, w^*) is the *typical value* or *concentration point* of (M_n, W_n) under Q_n. Since rate functions are positive, Assumption 4 and (45) imply $I_Q(m^*) = 0$, so that m^* is also the typical value of M_n with respect to Q_n. A good change of measure, as we have seen, should be such that $m^* \in \bar{B}$ to transform the event $\{M_n \in B\}$ from being rare under P_n to being typical under Q_n. In large deviation terms, this means

$$I_Q(B) \equiv \inf_{m \in B} I_Q(m) = I_Q(m^*) = 0. \tag{47}$$

This is the first step for constructing a good change of measure for IS – to make B typical. The next step is to ensure that Q_n is asymptotically efficient.

B. Efficiency results

We study the efficiency of Q_n from the result in (19) by expressing $R_Q(B)$ as a variational formula involving $J_Q(m,w)$, similarly to the formula (44) that we have for $I_P(B)$, and by then comparing these two formulas to infer conditions on $J_Q(m,w)$ that guarantee that $R_Q(B) = 2I_P(B)$. The first part is the subject of the next result.

Lemma 2. *Under Assumptions 1, 2 and 3, the second moment rate $R_Q(B)$ defined in (18) exists, is finite, and is given in terms of J_P and J_Q by*

$$R_Q(B) = \inf_{m\in B, w\in\mathbb{R}} \{w + J_P(m,w)\} = \inf_{m\in B, w\in\mathbb{R}} \{2w + J_Q(m,w)\}. \tag{48}$$

Proof. These variational representations of $R_Q(B)$ are a direct consequence of the Laplace principle for approximating exponential integrals, which is formulated in a rigorous way in large deviation theory via Varadhan's integral lemma [69]. For our purpose, we apply a version of that theorem found in [5, Thm. II.7.2] to $R_Q(B)$ as defined by (18). Given Assumption 3(d), to show that

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n\in B}] = \inf_{m\in B, w\in\mathbb{R}} \{2w + J_Q(m,w)\}, \tag{49}$$

it suffices to prove that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n\in \bar{B}}] \geq \inf_{m\in \bar{B}, w\in\mathbb{R}} \{2w + J_Q(m,w)\}, \tag{50}$$

and

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n\in B^\circ}] \leq \inf_{m\in B^\circ, w\in\mathbb{R}} \{2w + J_Q(m,w)\}. \tag{51}$$

Under Assumptions 2 and 3(b), to establish (50) (respectively (51)), one may adapt the proof of [5, Thm. II.7.2] detailed in Appendix B.2, item (a) (respectively (b)), replacing K (respectively G) with $\bar{B}\times\mathbb{R}$ (respectively $B^\circ\times\mathbb{R}$). Hence, $R_Q(B)$ defined as the limit in (18) exists. Moreover, Assumption 1 and (18) ensure that $R_Q(B) < \infty$, and Assumption 2 that $R_Q(B) > -\infty$, so that $R_Q(B)$ is finite. Finally, (41) shows the equivalence between both relations for $R_Q(B)$ in (48). □

The result of Lemma 2 complements the methods developed by Bucklew [41] for calculating $R_Q(B)$, which are based on generating functions rather than the joint large deviations of M_n and W_n. The advantage of our result is that it can be used with (44) to express the efficiency bound $R_Q(B) \leq 2I_P(B)$ as a variational inequality involving the rate function $J_Q(m,w)$:

$$\inf_{m\in B, w\in\mathbb{R}} \{2w + J_Q(m,w)\} \leq 2 \inf_{m\in B, w\in\mathbb{R}} \{w + J_Q(m,w)\}. \tag{52}$$

Therefore, *Q_n is asymptotically efficient if and only if J_Q is such that the inequality above is an equality.* The same inequality can be expressed in terms of J_P using (41), but this is not useful, since we want to characterize the efficiency of Q_n. Note, however, that the right-hand side of (52), although written with J_Q, does not actually depend on Q_n, since it is equal to $2I_P(B)$.

Our aim now is to find conditions on Q_n, and therefore on $J_Q(m,w)$, to have equality in (52). This is a non-trivial task, despite the simple form of this inequality, because the minimizers on either side need not be the same. Moreover, although J_Q is positive, w is not, so bounds based on the minimizer (m^*, w^*) of J_Q do not yield any useful conditions. Rather, such conditions are found by observing that w is the relevant variable in (52), since the minimizer over m is common

to both sides of this inequality, and that the unconstrained minimization over $w \in \mathbb{R}$ has the form of a Legendre–Fenchel transform.

Based on these observations, we define

$$I_Q^B(w) \equiv \inf_{m \in \bar{B}} J_Q(m, w). \tag{53}$$

This function of $w \in \mathbb{R}$ is positive, since $J_Q(m, w) \geq 0$, although it is not, as such, a rate function, since it does not necessarily have a zero. To understand this point, let us assume for simplicity that B is closed. In that case, note that the joint LDP for M_n and W_n with respect to Q_n implies the following LDP for the distribution of W_n conditioned on $M_n \in B$:

$$Q_n(W_n \in [w, w + dw] | M_n \in B) \asymp e^{-nI_Q(w|B)}, \tag{54}$$

where

$$I_Q(w|B) = \inf_{m \in B} J_Q(m, w) - \inf_{m \in B, w \in \mathbb{R}} J_Q(m, w) = I_Q^B(w) - I_Q(B) \tag{55}$$

The conditional distribution of W_n is normalized, so its rate function $I_Q(w|B)$ is a true rate function, in the sense that

$$\inf_{w \in \mathbb{R}} I_Q(w|B) = 0. \tag{56}$$

However, we see from (55) that, unless $I_Q(B) = 0$, we have $I_Q^B(w) > 0$, so the latter function is indeed not a true rate function in general.

The case that interests us is precisely the case where $I_Q(B) = 0$. That is, if $m^* \in \bar{B}$, then B is typical under Q_n, so that $I_Q(B) = 0$ and thus $I_Q(w|B) = I_Q^B(w)$. In that case, $I_Q^B(w)$ is interpreted as a conditional rate function having a zero (at w^* from Assumption 4). The converse is also true, leading us to the following result:

Lemma 3. *$I_Q^B(w^*) \geq 0$ with equality if and only if $m^* \in \bar{B}$.*

Proof. We have $I_Q^B(w^*) \geq 0$ by definition of rate functions. For the direct part, suppose that

$$0 = I_Q^B(w^*) = \inf_{m \in \bar{B}} J_Q(m, w^*). \tag{57}$$

As mentioned in the discussion before Assumption 1, since J_Q is a good rate function, the infimum is achieved on $\bar{B}$. By Assumption 4, this ensures that $m^* \in \bar{B}$.

For the converse part, simply note that $m^* \in \bar{B}$ implies

$$I_Q^B(w^*) = \inf_{m \in \bar{B}} J_Q(m, w^*) = J_Q(m^*, w^*) = 0 \tag{58}$$

by Assumption 4. □

We are now ready to state our main result for the asymptotic efficiency of Q_n based on I_Q^B. The statement of the result uses the *subdifferential* $\partial I_Q^B(w^*)$ of I_Q^B at the point w^*, which is the set of values $k \in \mathbb{R}$ such that

$$I_Q^B(w) \geq I_Q^B(w^*) + k(w - w^*) \tag{59}$$

for all $w \in \mathbb{R}$. More information about subdifferentials can be found in Appendix A. For the proof and the interpretation of the result, we also need the *Legendre–Fenchel transform* of I_Q^B, defined by[2]

$$\lambda_Q^B(k) \equiv \sup_{w\in\mathbb{R}}\{kw - I_Q^B(w)\}, \qquad k \in \mathbb{R}. \tag{60}$$

This is a convex function of k, as also explained in Appendix A, such that

$$\lambda_Q^B(0) = -\inf_{w\in\mathbb{R}} I_Q^B(w) \leq 0. \tag{61}$$

Theorem 4. *Under Assumptions 1-4, Q_n is asymptotically efficient if and only if $I_Q^B(w^*) = 0$ (typicality condition) and $-2 \in \partial I_Q^B(w^*)$ (steepness condition).*

Proof. Suppose that Q_n is asymptotically efficient. Then the efficiency criterion $R_Q(B) = 2I_P(B)$ leads to equality in (52), which can be re-expressed as

$$\lambda_Q^B(-2) = 2\lambda_Q^B(-1) \tag{62}$$

with the definition of I_Q^B and its Legendre–Fenchel transform.

This relation constrains the graph of $\lambda_Q^B(k)$, as illustrated in Fig. 1: P_1 and P_2 show the points of λ_Q^B at $k = -1$ and $k = -2$, respectively, which are related in an affine way according to (62). Moreover, we know that $\lambda_Q^B(0) \leq 0$ from (61). From these two results, and the fact that $\lambda_Q^B(k)$ is convex, we conclude that $\lambda_Q^B(k)$ must be linear over $k \in [-2, 0]$, as no other convex function can pass through both P_1 and P_2 while intersecting the ordinate axis below 0. Hence,

$$\lambda_Q^B(k) = -\lambda_Q^B(-1)k, \qquad k \in [-2, 0]. \tag{63}$$

In particular, $\lambda_Q^B(0) = 0$, shown in Fig. 1b as the point P_0 at the origin. By (61), this implies that

$$0 = \inf_{w\in\mathbb{R}} I_Q^B(w) = \inf_{(m,w)\in\bar{B}\times\mathbb{R}} J_Q(m, w). \tag{64}$$

Since $\bar{B} \times \mathbb{R}$ is closed, the infimum is reached on the latter and, by Assumption 4, this is possible only at point (m^*, w^*), so that $m^* \in \bar{B}$ and $I_Q^B(w^*) = 0$ by Lemma 3.

The typicality condition is obtained from this result by noting, following the proof of Lemma 3, that the existence of P_0 implies with (61) that $I_Q^B(w^*) = 0$. The minimum w^* is unique by Assumption 4. The steepness condition, on the other hand, is obtained by using standard results of convex analysis, stated with references in Appendix A, to show that the linear part of λ_Q^B leads to I_Q^B having a cusp at its global minimum, characterized by more than one supporting lines, including one with slope 0 and one with slope -2. To simplify the proof, we will first assume that $I_Q^B(w)$ is a convex function of w and will then explain why the result also holds when I_Q^B is not convex.

Note first that w^* being a global minimum of I_Q^B means that $0 \in \partial I_Q^B(w^*)$. Using the duality result expressed in (A4), we then obtain $w^* \in \partial\lambda_Q^B(0)$. However, since $\lambda_Q^B(k)$ is linear for $k \in [-2, 0]$, we also have $w^* \in \partial\lambda_Q^B(k)$ for $k \in [-2, 0]$, which implies by applying (A4) again that

$$[-2, 0] \subset \partial I_Q^B(w^*). \tag{65}$$

Therefore, $-2 \in I_Q^B(w^*)$, which is the steepness condition.

[2] We use the same letter λ for the Legendre–Fenchel transform and for the SCGF in (23), since, as already mentioned, the Gärtner–Ellis theorem ensures that, under appropriate conditions, both functions coincide.

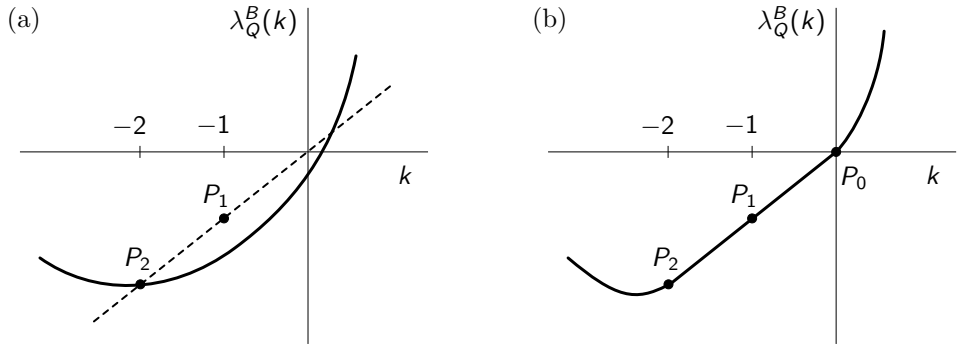

FIG. 1. (a) General $\lambda_Q^B(k)$. (b) Imposing $\lambda_Q^B(-2) = 2\lambda_Q^B(-1)$ implies, by convexity of $\lambda_Q^B(k)$, that this function passes through the origin P_0 and is linear for $k \in [-2, 0]$.

If I_Q^B is nonconvex, then the same argument applies by replacing I_Q^B in the duality result by its convex envelope $(I_Q^B)^{**}$, given by the Legendre–Fenchel transform of λ_Q^B or, equivalently, by the double Legendre–Fenchel transform of I_Q^B itself. We also have to note that a function and its convex envelope necessarily have the same global minima, if there are any, which means here that $I_Q^B(w^*)^{**} = I_Q^B(w^*) = 0$ and $0 \in \partial I_Q^B(w^*)^{**}$. Finally, where a function coincides with its convex envelope, the subdifferentials are the same, so that $\partial I_Q^B(w^*)^{**} = \partial I_Q^B(w^*)$. All these results are presented with references in Appendix A and imply, in the end, that the relation (65) holds at w^* even if I_Q^B is not convex.

To complete the proof, we consider the converse statement. We have again that, since I_Q^B has a global minimum at w^*, $0 \in \partial I_Q^B(w^*)$. By further assuming that $-2 \in \partial I_Q^B(w^*)$, we then have $[-2, 0] \subset \partial I_Q^B(w^*)$, since subdifferentials are closed convex sets. Hence, -1 also belongs to the subdifferential of $I_Q^B(w^*)$, which means with (59) that

$$I_Q^B(w) \geq I_Q^B(w^*) - (w - w^*) \tag{66}$$

and, therefore,

$$\inf_{w\in\mathbb{R}} \{w + I_Q^B(w)\} = w^* + I_Q^B(w^*) = w^*. \tag{67}$$

The same argument for -2 gives

$$\inf_{w\in\mathbb{R}} \{2w + I_Q^B(w)\} = 2w^* + I_Q^B(w^*) = 2w^*. \tag{68}$$

Consequently,

$$\inf_{w\in\mathbb{R}} \{2w + I_Q^B(w)\} = 2 \inf_{w\in\mathbb{R}} \{w + I_Q^B(w)\} = 2w^*, \tag{69}$$

which implies from (52) that Q_n is asymptotically efficient. □

C. Interpretation and special cases

We will see in the next section that our main result in Theorem 4 covers the efficiency of the exponential tilting as a special case. The important contribution of this theorem, compared to previous results, is the subdifferential condition, which guarantees that the second moment

$$F_n(B) \equiv \mathbb{E}_Q[L_n^2 \mathbf{1}_{M_n \in B}] = \int_{\mathcal{M}\times\mathbb{R}} e^{-2nw} \mathbf{1}_{m\in B}\, Q_n(dm, dw). \tag{70}$$

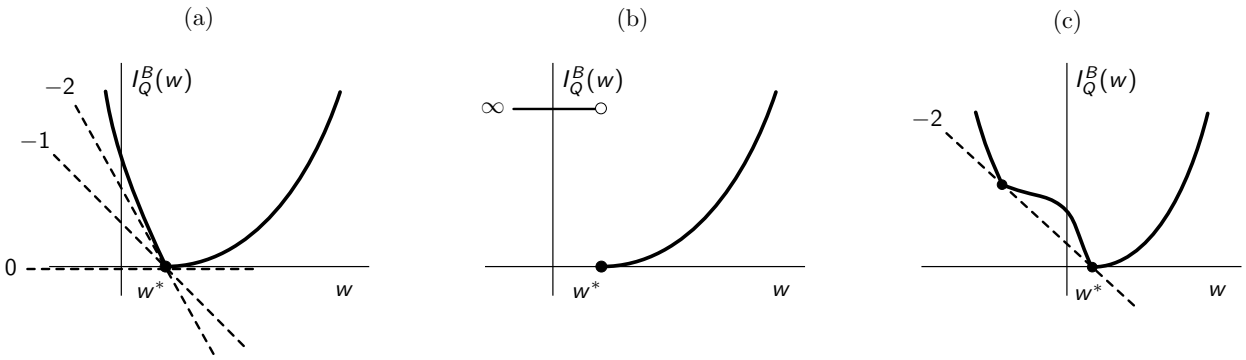

FIG. 2. (a) Efficiency condition for convex and left-differentiable $I_Q^B(w)$. (b) Asymptotically efficient $I_Q^B(w)$ diverging at the left of w^*. (c) Nonconvex $I_Q^B(w)$ that is also asymptotically efficient.

entering in the definition of $R_Q(B)$ is dominated by w^* and not by another rare value of the action smaller than w^* that would lead to an exponentially larger value of the likelihood factor L_n. If this condition is satisfied, in addition to the obvious condition that B be typical under Q_n, then Q_n is asymptotically efficient, which means that it can be used to sample $P_n(M_n \in B)$ with a sample size N_n according to (24) that is not exponentially large with n.

Comparing (67) and (68), we can also say that Q_n is asymptotically efficient if and only if w^* is the minimizer on both sides of the efficiency criterion

$$\inf_{w\in\mathbb{R}}\{2w + I_Q^B(w)\} = 2\inf_{w\in\mathbb{R}}\{w + I_Q^B(w)\}, \tag{71}$$

which follows from (52). In other words, Q_n is asymptotically efficient if and only if the IS estimator $\hat{p}_n^N$ and its second moment are dominated by the same typical value w^* of the action, yielding $I_P(B) = w^*$ and $R_Q(B) = 2w^*$.

This interpretation of the efficiency in terms of the typical value w^* does not mean altogether that the likelihood L_n or its action W_n does not fluctuate. This is a very important point. The subdifferential condition is only a condition about the "shape" of $I_Q^B(w)$ *below* the typical value w^*, which means that the fluctuations of W_n *above* w^* are not constrained in any way.

In many cases, we find that $I_Q^B(w)$ is a convex function and is left-differentiable at w^*. Then the subdifferential condition reduces to

$$I_Q^B(w^{*-})' \leq -2, \tag{72}$$

where $I_Q^B(w^{*-})'$ is the left-derivative of I_Q^B at w^*. This result is illustrated in Fig. 2a and explains why we refer to the subdifferential condition as a "steepness" condition. Obviously, if $I_Q^B(w)$ is convex and is differentiable at its minimum, then

$$I_Q^B(w^{*-})' = I_Q^B(w^*)' = 0 \tag{73}$$

and so Q_n is not asymptotically efficient. This offers a simple test that can be used in practice to identify non-efficient IS measures.

In general, I_Q^B might not be left-differentiable at its minimum or be convex, contrary to λ_Q^B which is convex by definition. This is why we need to state the steepness condition using the concept of subdifferentials. For instance, if I_Q^B diverges for $w < w^*$, as illustrated in Fig. 2b, then we have efficiency, since $(-\infty, 0] \subset \partial I_Q^B(w^*)$, so that $-2 \in \partial I_Q^B(w^*)$. This case arises when W_n has no possible values below w^* and so $W_n \geq w^*$. On the other hand, if I_Q^B is not convex, then

we have efficiency if I_Q^B has a supporting line at w^* with slope smaller than or equal to -2, as shown in Fig. 2c. This follows from the interpretation of subdifferentials, explained in Appendix A. Equivalently, we have efficiency if the left-derivative of the convex envelope of I_Q^B at w^* is smaller than or equal to -2. This case will be illustrated in the next section by revisiting the Gaussian sample mean and nonconvex set B studied before.

With this geometric interpretation of efficiency, it is now clear that the likelihood factor L_n does not have to be a deterministic function of M_n or become so in the limit $n \to \infty$ to efficiently estimate $P_n(M_n \in B)$, as often stated in studies of IS. The likelihood can fluctuate jointly with M_n so long as the fluctuations of the action W_n conditioned on $M_n \in B$ are sufficiently suppressed below the typical value $W_n = w^*$, that is, so long as $I_Q^B(w)$ is steep enough below w^*. The right steepness of I_Q^B is not constrained in any way because values $W_n > w^*$ are exponentially suppressed in the integral (70), which determines the efficiency of Q_n. Only values $W_n < w^*$ can affect the efficiency, as L_n is then exponentially larger than the typical value $L_n^* = e^{-nw^*}$. In other words, accumulating likelihood factors that are exponentially *smaller* than L_n^* does not influence the efficiency of IS, but accumulating factors that are exponentially *larger* than L_n^* *too frequently* does.

To close this section, we note that if there is a finite w such that $-2 - \delta \in \partial I_Q^B(w)$ for some $\delta > 0$, then under Assumption 3(b) the limit (33) in our Assumption 2 must be finite. As a result, we can rephrase that assumption in a more geometric and practical way as follows:

Assumption 2′. *$I_Q^B(w)$ must be coercive enough so that it has a point whose subdifferential contains a value strictly smaller than -2.*

If $I_Q^B(w)$ is a convex and differentiable function, then this amounts to saying that there is a point whose slope is strictly smaller than -2.

IV. EXAMPLES

We illustrate in this section our results with simple examples to show how $I_Q^B(w)$ is calculated in practice and how its steepness determines the efficiency of IS estimators. We begin by considering the exponential tilting as a general change of measure, and then revisit the Gaussian and exponential i.i.d. sample means, introduced in Sec. II, as particular cases of that change of measure for which $I_Q^B(w)$ can be computed exactly. We also construct a variation of the exponential sample mean that shows that an IS estimator can be asymptotically efficient without having the exponential tilting form. This is an important result of this section.

We close the section with two examples related to Markov chains and stochastic differential equations to illustrate the generality of our formalism, to explain how the likelihood factor is defined for Markov processes, and to point out minor changes of notation when dealing with continuous-time processes. These examples should serve as a template to study the large deviations of more physical systems modelled by Markov processes in the context, for example, of nonequilibrium systems driven in steady-states and stochastic thermodynamics.

A. Exponential tilting

We consider for simplicity the case where $M_n \in \mathbb{R}$, since our goal is not to prove the efficiency of the exponential tilting in the most general setting but to illustrate our formalism based on $I_Q^B(w)$. The change of measure that we consider, as already defined in (21), is thus

$$Q_n(d\mathbf{X}_n) = \frac{e^{nkM_n} P_n(d\mathbf{X}_n)}{\mathbb{E}_P[e^{nkM_n}]}, \tag{74}$$

where $P_n(d\mathbf{X}_n)$ is, as before, our prior measure or model of $\mathbf{X}_n$ and k is now a real parameter.

The use of this exponential change of measure to study the large deviations of M_n requires, as mentioned in Sec. II, that the SCGF $\lambda_P(k)$ of M_n, as defined in (23), exists and is differentiable in k. In this case, it follows from the Gärtner-Ellis theorem [6, Thm. 2.3.6] that $I_P(m)$ is a strictly convex function, given by the Legendre–Fenchel transform of $\lambda_P(k)$, which means that it has a unique minimum and zero, denoted by $\bar{m}$, corresponding to the typical value of M_n under P_n. Thus, $I_P(\bar{m}) = 0$, which translates by Legendre duality into $\lambda_P'(0) = \bar{m}$.

In most applications, $I_P(m)$ is found to be a smooth (differentiable) function of m, so we will assume this property in this section to simplify the analysis. The particular form of the exponential change of measure then implies (see [68, Thm. 2.4]) that

$$I_Q(m) = I_P(m) - km + \lambda_P(k), \tag{75}$$

so $I_Q(m)$ is smooth by assumption. It is also a good rate function, since $I_P(m)$ itself, as obtained from the Gärtner–Ellis theorem, is a good rate function.

With these results, we now apply our formalism by noting that the action W_n of the exponential tilting is

$$W_n = kM_n - c_n(k), \tag{76}$$

where

$$c_n(k) = \frac{1}{n} \log \mathbb{E}_P[e^{nkM_n}]. \tag{77}$$

This is a deterministic function of M_n that we write as $W_n = f_n(M_n)$, which implies that $J_Q(m, w)$ is defined only on the line $w = f_n(m)$ and is equal to $I_Q(m)$ on that line. The appearance of n in this contraction appears a priori to be a problem; however, we show in Appendix B that, since $c_n(k) \to \lambda_P(k)$ and J_Q is a good rate function, f_n can actually be replaced by the limit function $f(m) = km - c(k)$, where $c(k) = \lambda_P(k)$, consistently with (75) and (41). As a result,

$$I_Q^B(w) = \begin{cases} \inf\limits_{m \in \bar{B}} I_Q(m) & w = f(m) \\ \infty & \text{otherwise.} \end{cases} \tag{78}$$

Having found I_Q^B, we now consider the rare event $M_n \in B \equiv [b, \infty) = \bar{B}$ with $b > \bar{m}$. To make this event typical under Q_n, we fix k so that the typical value m^* of M_n "hits" the boundary b. This is achieved by setting $k > 0$ such that $\lambda_P'(k) = b$, leading to $I_Q(b) = 0$ [55] and

$$\inf_w J_Q(b, w) = J_Q(b, f(b)) = I_Q(b) = 0. \tag{79}$$

This shows that we have a unique, typical pair $(m^*, w^*) = (b, f(b))$ under Q_n. Since $m^* = b \in \bar{B}$, we then have $I_Q^B(w^*) = 0$ by Lemma 3, so the first condition for efficiency in Theorem 4 is met.

To check the second condition, note that, since we have $k > 0$ to achieve $m^* = b > \bar{m}$, $J_Q(m, w)$ is not finite on B when $w < w^*$, as shown in Fig. 3a, so that $I_Q^B(w) = \infty$ for all $w < w^*$. Thus, I_Q^B is infinitely steep below w^*, which is sufficient, as mentioned before, to conclude that Q_n is asymptotically efficient. Above w^*, we see instead that $J_Q(m, w)$ is finite on B, so $I_Q^B(w)$ is also finite for $w > w^*$. In fact, since $J_Q(m, w)$ has a unique zero at (m^*, w^*), we have $0 < I_Q^B(w) < \infty$ for $w > w^*$, showing overall that $I_Q^B(w)$ has the shape shown in Fig. 2b, associated once again with a Q_n that is asymptotically efficient.

The same argument can be used to show that Q_n is not asymptotically efficient if m^* is chosen inside B, that is, such that $m^* > b$. In this case, $I_Q^B(w)$ still has a zero at $w^* = f(m^*)$, but it does

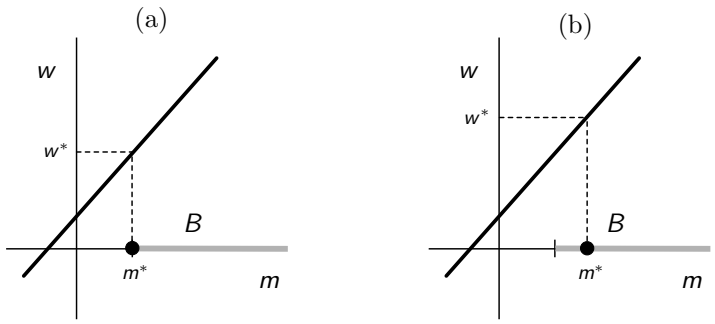

FIG. 3. Line $w = f(m)$ in the (m, w) plane on which $J_Q(m, w)$ is defined for the exponential change of measure. (a) Asymptotically efficient Q_n for which the typical value m^* of M_n is chosen on the boundary of B. (b) Non-efficient Q_n associated with m^* in the interior of B.

not diverge on the left of w^* because the line $w = f(m)$ on which $J_Q(m, w)$ is finite does not "go out" of B when w goes below w^*; see Fig. 3b. Since $I_Q(m)$ is assumed to be smooth, $I_Q^B(w)$ must therefore have a smooth minimum in the vicinity of w^* with zero derivative as in (73), implying that Q_n is not asymptotically efficient.

Of course, the steepness condition could be satisfied in this case if $I_Q(m)$ had a steep-enough corner at m^*, but this would violate our assumption that $I_Q(m)$ is smooth, which is what is observed again in many applications.[3] With this assumption, the exponential tilting is therefore asymptotically efficient, as proved, if it "hits" B on its boundary b rather than in the interior of B. This can be generalized to $M_n \in \mathbb{R}^D$ by requiring that Q_n "hit" the dominating point of B, which is usually on the boundary of B; see [43] for details.

All these results apply obviously if we change the rare event to $B = (-\infty, b]$ with $b < \bar{m}$, in which case $k < 0$. The efficiency of Q_n is also direct if we consider the infinitesimal set $B = [b, b + dm]$. Then k must be chosen such that $\lambda_P'(k) = b$ to achieve $m^* = b$, as mentioned before, which fixes $w^* = f(b)$ as the only action for which $I_Q^B(w)$ is finite. Thus, $I_Q^B(w) = 0$ for $w = w^*$ and ∞ otherwise, which is obviously asymptotically efficient. This is a common case considered in physics, where the focus is usually on computing the rate function $I_P(m)$ rather than the probability $P_n(M_n \in B)$. In this case, one performs many simulations with different values of k to estimate the probability of small, contiguous "windows" $[b, b + dm]$, which are converted with the large deviation limit (3) into points of $I_P(m)$ [70].

Such a use of the exponential tilting in simulations is asymptotically efficient, as just shown, if $I_P(m)$ is a convex differentiable function and B is a convex set. We have already seen in Sec. II that the exponential tilting can be non-efficient if B is nonconvex. By revisiting this example below, we will see that the problem in this case comes from the steepness condition controlling the asymptotic variance. On the other hand, the exponential tilting can also be non-efficient if $I_P(m)$ is nonconvex. The problem in this case is not the steepness of I_Q^B, and so the variance, but the fact that not all values of M_n can be made typical by varying the tilting parameter k. This is related in physics to the nonequivalence of statistical ensembles; see [48] for more details.

[3] A corner in $I_P(m)$ or $I_Q(m)$ signals physically a dynamical phase transition in the fluctuations of M_n. Here, we assume, for simplicity, that no such phase transition occurs. Note that a corner in the function $I_Q^B(w)$ is not related to a dynamical phase transition, since this function is obtained by conditioning. It can have a corner, as the example of the exponential tilting shows, regardless of whether $I_P(m)$ or $I_Q(m)$ is smooth.

B. Gaussian sample mean

We now revisit the Gaussian sample mean studied in Sec. II to show how the efficiency of Q_n can be ascertained by calculating I_Q^B explicitly using standard techniques from large deviation theory. This example also provides a further illustration of the exponential change of measure.

The setting is the same as in Example 1: P_n is the product measure $\mathcal{N}(0,1)^{\otimes n}$ of n i.i.d. standard normal random variables, M_n is their sample mean, and we look for the probability that $p_n = P_n(M_n \geq 1)$, so that $B = [1, \infty)$, leading to the rate exponent shown in (26).

We consider as the change of measure the product measure $Q_n = \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ associated with n i.i.d. Gaussian random variables with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The action for this change of measure, which is more general than the one considered in Example 1, can be expressed as

$$W_n = \frac{\mu}{\sigma^2} M_n + \left(\frac{\sigma^2 - 1}{2\sigma^2} \right) C_n - \frac{\mu^2}{2\sigma^2} - \log \sigma, \tag{80}$$

where M_n is the sample mean and

$$C_n = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \tag{81}$$

is the sample second moment. Since both M_n and C_n involve i.i.d. random variables, we can use Cramér's theorem to find their joint rate function $K_Q(m, c)$ as the Legendre–Fenchel transform of the joint SCGF with respect to Q_n:

$$\lambda_Q(k, \gamma) = \log \mathbb{E}_Q[e^{kX + \gamma X^2}]. \tag{82}$$

For $X \sim Q = \mathcal{N}(\mu, \sigma^2)$, we find

$$\lambda_Q(k, \gamma) = \frac{k^2\sigma^2/2 + \mu(\gamma\mu + k)}{1 - 2\gamma\sigma^2} - \frac{1}{2} \log\left(1 - 2\gamma\sigma^2\right) \tag{83}$$

for $1 - 2\gamma\sigma^2 > 0$. Accordingly,

$$K_Q(m, c) = \sup_{k,\gamma} \{km + \gamma c - \lambda_Q(k, \gamma)\} = \frac{\sigma^2 \log\left(\frac{\sigma^2}{c - m^2}\right) + c + \mu^2 - 2\mu m - \sigma^2}{2\sigma^2}, \tag{84}$$

which is defined for $m^2 < c$ by the Cauchy-Schwarz inequality. Changing variables from (M_n, C_n) to (M_n, W_n), we then deduce

$$J_Q(m, w) = K_Q(m, c(m, w)), \tag{85}$$

where

$$c(m, w) = \frac{2w\sigma^2 - 2m\mu + \mu^2 + \sigma^2 \log \sigma^2}{\sigma^2 - 1}. \tag{86}$$

This holds if $\sigma \neq 1$. If $\sigma = 1$, then W_n is only a function of M_n,

$$W_n = f(M_n) = \mu M_n - \frac{\mu^2}{2}, \tag{87}$$

similarly to the exponential change of measure and, therefore,

$$J_Q(m, w) = \begin{cases} (m - \mu)^2/2 & w = f(m) \\ \infty & \text{otherwise,} \end{cases} \tag{88}$$

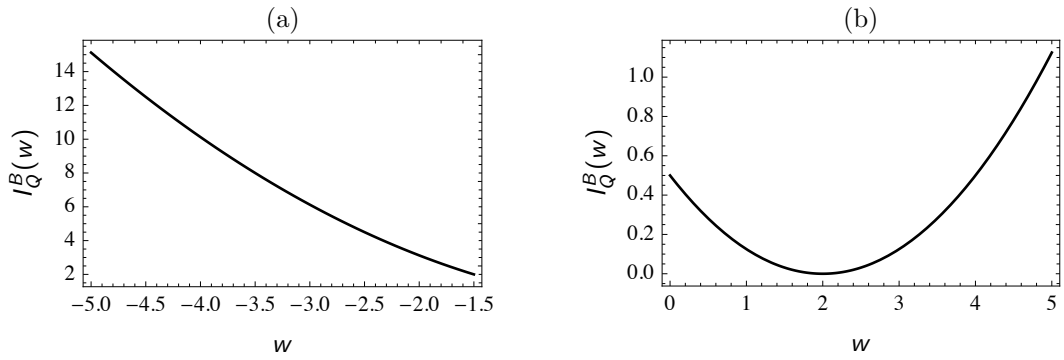

FIG. 4. $I_Q^B(w)$ for the Gaussian sample mean for (a) $\mu = -1$ and $\sigma = 1$ and (b) $\mu = 2$ and $\sigma = 1$. Note that only the finite part of $I_Q^B(w)$ is shown.

given that $I_Q(m) = (m-\mu)^2/2$.

The efficiency of Q_n is determined by computing $I_Q^B(w)$ from these explicit rate functions for various values of μ and σ. We begin with $\sigma = 1$ and consider three cases for μ, noting that $m^* = \mu$ and $w^* = f(m^*) = \mu^2/2$:

- $\mu < 1$: Q_n is not asymptotically efficient in this case simply because $m^* \notin \bar{B}$. This is confirmed by calculating $I_Q^B(w)$ from the contraction (53). The result is shown for $\mu = -1$ in Fig. 4a: it does not have a zero, so Q_n is indeed not asymptotically efficient.

- $\mu \geq 1$: The calculation of $I_Q^B(w)$ gives in this case

$$I_Q^B(w) = \begin{cases} (w/\mu - \mu/2)^2/2 & w \geq \mu - \mu^2/2 \\ \infty & \text{otherwise.} \end{cases} \tag{89}$$

  For $\mu = 1$, we have efficiency, since $I_Q^B(w)$ has its zero at $w^* = 1/2$, implying $m^* \in \bar{B}$, and diverges left of w^*, so it is infinitely steep, as in Fig. 2b. This is also confirmed by the fact that Q_n is the exponential change of measure with $m^* = 1$ at the boundary of B. For $\mu > 1$, $I_Q^B(w)$ also has its zero at w^* but $I_Q^B(w^{*-})' = 0$, so it is not steep left of w^*, as shown in Fig. 4b.

These results show overall that Q_n is asymptotically efficient for $\sigma^2 = 1$ if and only if $\mu = 1$.

For $\sigma^2 \neq 1$, the contraction of $J_Q(m, w)$ leading to $I_Q^B(w)$ is more complicated to solve, since the minimization on $m \in B$ is further constrained by $m^2 < c$ in the transformation (85). This results in a tedious constrained minimization problem, which can easily be solved numerically, however, to plot $I_Q^B(w)$ for any μ and $\sigma^2 \neq 1$. Two representative solutions are shown in Fig. 5 and confirm our expectations from Theorem 4. On the one hand, if $\mu < 1$, then Q_n is not asymptotically efficient since $m^* \notin \bar{B}$, as reflected by the fact that $I_Q^B(w)$ has no zero (Fig. 5a). On the other hand, if $\mu \geq 1$, then $m^* \in \bar{B}$, so $I_Q^B(w)$ has a zero, but the rate function is not steep left of that zero, so Q_n is still not asymptotically efficient (Fig. 5b). This applies whether $\mu = 1$ or $\mu > 1$, which means in the end that Q_n is not asymptotically efficient when $\sigma^2 \neq 1$.

C. Nonconvex B

We can use the results of the previous section to understand the efficiency of Q_n in Example 3, presented earlier to show that the exponential tilting can be non-efficient when B is nonconvex. The

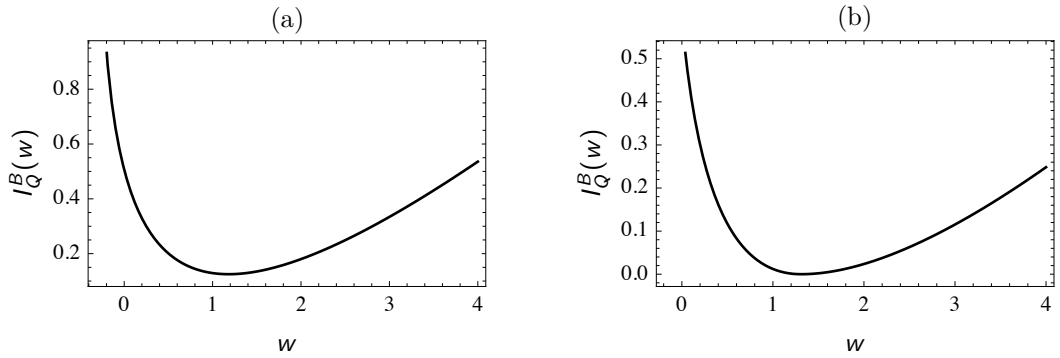

FIG. 5. $I_Q^B(w)$ for the Gaussian sample mean for (a) $\mu = 0$ and $\sigma = 2$ and (b) $\mu = 1$ and $\sigma = 2$. Note that only the finite part of $I_Q^B(w)$ is shown.

setting is the same as in the previous section, except that B is now chosen to be $B = (-\infty, -b] \cup [1, \infty)$ with $b > 1$. We also consider $\mu = 1$ and $\sigma^2 = 1$, which leads to

$$I_Q^B(w) = \begin{cases} (w - 1/2)^2/2 & w \geq 1/2 \text{ or } w \leq -b - 1/2 \\ \infty & \text{otherwise.} \end{cases} \tag{90}$$

This function is plotted in Fig. 6. It has one zero at $w^* = 1/2$, confirming that $m^* = 1 \in \bar{B}$ and a supporting line joining the two extremal points of $I_Q^B(w)$ at $w = -b - 1/2$ and $w = w^* = 1/2$, as shown in the figure, whose slope is $-(b + 1)/2$. This is the supporting line with smallest slope, so $-2 \in \partial I_Q^B(w^*)$ if and only if $b \geq 3$, confirming the result of [41] announced in Example 3.

We can generalize this result by calculating $I_Q^B(w)$ for $\mu \neq 1$ to conclude that there is no other efficient parameters and, thus, that Q_n is actually efficient if and only if $\mu = 1$ and $b \geq 3$. This follows by considering three cases:

- $\mu \leq -b$. Then

$$I_Q^B(w) = \begin{cases} (w/\mu - \mu/2)^2/2 & w \leq \mu - \mu^2/2 \text{ or } w \geq -\mu b - \mu^2/2 \\ \infty & \text{otherwise} \end{cases} \tag{91}$$

  From this result, it can be checked that $I_Q^B(w^{*-})' = 0$ if $\mu < -b$, so Q_n is not asymptotically efficient. Then, if $\mu = -b$, one has $-2 \in \partial I_Q^B(w^*)$ if and only if $0 < b \leq 1/3$, which contradicts our assumption that $b > 1$.

- $-b < \mu < 1$: Then $I_Q^B(w)$ does not have a zero, as expected from the fact that $m^* \notin \bar{B}$, so Q_n is not asymptotically efficient.

- $\mu > 1$: Then

$$I_Q^B(w) = \begin{cases} (w/\mu - \mu/2)^2/2 & w \geq \mu - \mu^2/2 \text{ or } w \leq -\mu b - \mu^2/2 \\ \infty & \text{otherwise} \end{cases} \tag{92}$$

  In this case $I_Q^B(w^*) = 0$ at $w^* = \mu^2/2$, but $I_Q^B(w^{*-})' = 0$.

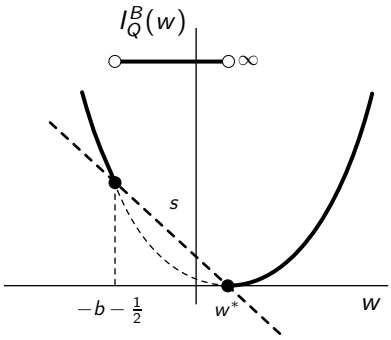

FIG. 6. Rate function $I_Q^B(w)$ for the Gaussian sample mean and nonconvex set B.

D. Partial exponential tilting

It is clear from the form of the exponential measure (21) that an i.i.d. measure remains i.i.d. when tilted by an additive functional M_n, which explains why such a change of measure is almost always considered when dealing with i.i.d. sample means. In principle, other changes of measure that are not independent, not identically distributed or both could be considered and proved efficient within the formalism developed here.

As a simple example, it can be checked for the Gaussian sample mean that changing all but one of the random variables is still asymptotically efficient, even though the resulting Q_n is only a *partial* exponential tilting (only $n-1$ random variables are tilted). This arises because the action of the partial exponential tilting differs from the action of the full exponential tilting by a sub-extensive term in n that does not influence the large deviations of the action at the scale (or speed) n.

Surprisingly, this argument cannot be generalized to all i.i.d. sample means and, in particular, not to the sample mean of exponential random variables considered in Example 2. In this case, we have seen that the product measure $Q_n = \mathcal{E}(\theta)^{\otimes n}$ of exponentials with parameter θ, which changes the mean of all the random variables from 1 to $1/\theta$, is asymptotically efficient if $\theta = 1/b$. This can be checked by calculating $I_Q^B(w)$ explicitly, following the calculations of the previous sections or from the fact that Q_n is the exponential tilting.

The surprising result comes if we change the first $n-1$ random variables from $\mathcal{E}(1)$ to $\mathcal{E}(\theta)$, but keep the last one as $X_n \sim \mathcal{E}(1)$. The action then is

$$W_n = (1-\theta)\frac{n-1}{n}M_{n-1} + \frac{n-1}{n}\log\theta = (1-\theta)c_n M_{n-1} + c_n \log\theta, \tag{93}$$

where M_{n-1} is the sample mean of the first $(n-1)$ random variables and $c_n = (n-1)/n$. Similarly, we can write

$$M_n = \frac{n-1}{n}M_{n-1} + \frac{X_n}{n} = c_n M_{n-1} + T_n, \tag{94}$$

defining the new random variable $T_n = X_n/n$, which is independent of M_{n-1}.

From these expressions, we find the joint rate function $J_Q(m,w)$ of M_n and W_n by noting that $M_{n-1} \sim \Gamma(n-1,(n-1)\theta)$, so this random variable satisfies the LDP with rate function

$$I_\Gamma(y) = \theta y - 1 - \log(\theta y), \tag{95}$$

for $y \geq 0$, whereas $T_n \sim \mathcal{E}(n)$, so it satisfies the LDP with rate function

$$I_\mathcal{E}(t) = t \tag{96}$$

also for $t \geq 0$. Both are good rate functions. From (93), (94) and the fact that $c_n \to 1$, we can then use the contraction principle presented in Appendix B to express $J_Q(m, w)$ as

$$J_Q(m, w) = \inf_{\substack{w=(1-\theta)y+\log\theta \\ m=y+t \\ y,t\geq 0}} I_\Gamma(y) + I_{\mathcal{E}}(t). \tag{97}$$

In the latter, we have $m \leftrightarrow M_n \geq 0$, $w \leftrightarrow W_n$, $y \leftrightarrow M_{n-1} \geq 0$, and $t \leftrightarrow T_n \geq 0$. The solutions to the constraints are

$$y = \frac{w - \log\theta}{1-\theta} \geq 0 \qquad \text{and} \qquad t = m - \frac{w - \log\theta}{1-\theta} \geq 0, \tag{98}$$

leading to $J_Q(m, w) = \infty$ if one of these constraints is not satisfied and

$$J_Q(m, w) = m - w - \log\frac{w - \log\theta}{1-\theta} - 1 \tag{99}$$

otherwise. This rate function is good and has a single zero at $m^* = 1/\theta$ and

$$w^* = \frac{1-\theta}{\theta} + \log\theta. \tag{100}$$

At this point, we determine the asymptotic efficiency of Q_n, as before, by computing $I_Q^B(w)$ for different cases for θ:

- $\theta > 1/b$: In this case, we do not even need to calculate $I_Q^B(w)$: $m^* = 1/\theta \leq 1 < b$, so that $m^* \notin \bar{B}$, implying that Q_n is not asymptotically efficient.
- $0 < \theta < 1/b$: A direct calculation based on the fact that the map $m \mapsto J_Q(m, w)$ is increasing gives in this case

  $$I_Q^B(w) = \frac{w - \log\theta}{1-\theta} - w - \log\frac{w - \log\theta}{1-\theta} - 1 \tag{101}$$

  for all $w \in [(1-\theta)b + \log\theta, w^*]$. From this result, we find $I_Q^B(w^{*-})' = 0$, so that Q_n is once again not asymptotically efficient.
- $\theta = 1/b$: A similar calculation as before now yields

  $$I_Q^B(w) = b - w - \log\frac{w - \log\theta}{1-\theta} - 1 = b - w - \log\frac{w + \log b}{1 - 1/b} - 1, \tag{102}$$

  for all $w \in (\log\theta, w^*]$. As a result,

  $$I_Q^B(w^{*-})' = -1 - \frac{1}{b-1}, \tag{103}$$

  which is smaller than -2 if and only if $b \leq 2$.

The conclusion is that Q_n is asymptotically efficiency if and only if $\theta = 1/b$ and $b \leq 2$, so the partial exponential tilting does not have the same efficiency as the full exponential tilting.

This result is special to the exponential distribution because X_n/n in this case has large deviations at the same scale as M_n and W_n, and so affects both random variables in the contraction (97). By contrast, if we choose $X_n \sim \mathcal{N}(\mu, \sigma^2)$, then it can be checked that X_n/n satisfies the LDP at the scale n^2 so adding or removing a Gaussian random variable from a sample mean has no effect on its large deviations. The same applies to sample means of bounded random variables and, more generally, random variables whose distribution decays faster than exponentially.

E. Markov chains

We move away from i.i.d. models to consider discrete-time Markov chains evolving on a set Ω. We assume, for simplicity, that Ω is finite and that the transition kernel $p(x, y) = P(X_{i+1} = y|X_i = x)$ is homogeneous and defines an ergodic Markov chain. Starting with an initial distribution $\rho(x) = P(X_1 = x)$, the probability model is then expressed as

$$P_n(X_1, \ldots, X_n) = \rho(X_1)p(X_1, X_2) \cdots p(X_{n-1}, X_n) \tag{104}$$

for all $\mathbf{X}_n = (X_1, \ldots, X_n) \in \Omega^n$.

The observable M_n is still a function of the configuration $\mathbf{X}_n$, now interpreted as a trajectory in discrete time, from which we define the rare event probability $p_n = P_n(M_n \in B)$. We assume as before that M_n satisfies the LDP with respect to P_n with good rate function I_P and consider a change of model Q_n to sample p_n with the IS estimator (13). The choice of Q_n depends, as always, on the observable considered. For additive functionals having the general form

$$M_n = \frac{1}{n} \sum_{i=1}^{n-1} g(X_i, X_{i+1}), \tag{105}$$

Q_n is usually chosen to be another ergodic Markov chain with transition kernel $q(x, y)$, absolutely continuous with respect to $p(x, y)$, so that

$$Q_n(X_1, \ldots, X_n) = \rho(X_1)q(X_1, X_2) \cdots q(X_{n-1}, X_n), \tag{106}$$

using the same initial distribution. In this case, the action simply is

$$W_n = \frac{1}{n} \sum_{i=1}^{n-1} \log \frac{q(X_i, X_{i+1})}{p(X_i, X_{i+1})}, \tag{107}$$

so both M_n and W_n are additive functionals of the Markov chain.

With this property, the joint large deviations of M_n and W_n can be obtained by standard techniques from large deviation theory (see, e.g., [6, Sec. 3.1]). Define the joint SCGF of M_n and W_n with respect to Q_n as

$$\lambda_Q(k, \gamma) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_Q[e^{nkM_n + n\gamma W_n}]. \tag{108}$$

It is known that this function is given by the logarithm of the principal eigenvalue $\zeta_Q(k, \gamma)$ of the so-called *tilted transition matrix*, defined by

$$q_{k,\gamma}(x, y) = q(x, y)e^{kg(x,y)+\gamma h(x,y)}, \tag{109}$$

where $h(x, y) = \log(q(x, y)/p(x, y))$. Thus,

$$\lambda_Q(k, \gamma) = \log \zeta_Q(k, \gamma). \tag{110}$$

From this result, the rate $J_Q(m, w)$ is then found from the Gärtner–Ellis theorem by taking the Legendre–Fenchel transform of $\lambda_Q(k, \gamma)$, similarly to the Gaussian sample mean example. From there, we find $I_Q^B(w)$, as before, by minimising $J_Q(m, w)$ on $m \in B$ and use, finally, this function to determine the efficiency of Q_n. These steps can be implemented analytically for small Markov chains with a few states, while larger chains can be dealt with numerically using standard eigenvalue packages.

In most applications, Q_n is chosen to be the exponential tilting, which for a Markov chain and additive M_n is known to be another Markov chain with transition kernel

$$q(x,y) = \frac{e^{kg(x,y)} r_k(y)}{r_k(x)\zeta_P(k)} p(x,y), \tag{111}$$

where $\zeta_P(k)$ is the dominant eigenvalue of the tilted matrix

$$p_k(x,y) = p(x,y)e^{kg(x,y)}, \tag{112}$$

and r_k is the associated (right) eigenvector. In this case, it easy to verify that W_n is given by (76) modulo boundary terms involving $r_k(X_1)$ and $r_k(X_n)$, which can be neglected as they do not play a role in the large deviations of W_n when Ω is finite.

The Markov kernel (111) has been discussed in many contexts, including queuing theory [40], simulations [41], and statistical physics [71], and can be seen as a generalization of Doob's h-transform arising in "bridge-like" conditionings of Brownian motion and other Markov processes; see [71, Sec. 4.2] for details.

As a simple application, let us consider the symmetric binary Markov chain with $X_i \in \{0,1\}$ and transition matrix

$$p = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}, \tag{113}$$

where $\alpha \in (0,1)$. Considering the observable to be the sample mean

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{114}$$

which gives the fraction of 1's in $\mathbf{X}_n$, we can formulate two different changes of process that are asymptotically efficient. The first is the exponential tilting in (111), for which $\zeta_P(k)$ and r_k can be computed analytically as the principal eigenvalue and eigenvector of

$$p_k = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha e^k & (1-\alpha)e^k \end{pmatrix}, \tag{115}$$

obtained from (112) using $g(x,y) = x$. The asymptotic efficiency of the resulting Markov chain is determined by the previous results on the exponential change of measure (see Section IV A), and follows again from the fact that W_n is a function of M_n. Details can be found in [44, Thm. 3].

Surprisingly, the exponential tilting is not the only modified Markov chain for which W_n is a function of M_n. We can also take the transpose of the 2×2 matrix p_k above, which has the same principal eigenvalue as p_k, and normalize the rows to obtain the transition matrix

$$q = \begin{pmatrix} \frac{1-\alpha}{F_0} & \frac{\alpha e^k}{F_0} \\ \frac{\alpha}{F_1} & \frac{(1-\alpha)e^k}{F_1} \end{pmatrix}, \tag{116}$$

where $F_0 = 1-\alpha+\alpha e^k$ and $F_1 = \alpha + (1-\alpha)e^k$. It can be checked that the action induced by this transition matrix, which is obviously different from (111), is

$$W_n = kM_n - (1-M_n)\log F_0 - M_n \log F_1, \tag{117}$$

modulo unimportant boundary terms, so that W_n is an affine function of M_n. Consequently, we have found another example of process that is potentially asymptotically efficient and yet is not the

exponential tilting. The difference is that the value of k in (116) fixing the typical value $M_n = b$ under Q_n is not specified by the relation $\lambda'_P(k) = b$, which is special to the exponential tilting. This is not important for simulations, as we only need in practice a parameter that can be varied to fix any typical value of M_n, whatever the relation between the two.

In principle, other efficient changes of process could be constructed using, for example, higher-order Markov chains or even non-Markovian processes whose measure $Q_n(X_1, \dots, X_n)$ does not factorize as a product of transition probabilities. Very little, unfortunately, is known about non-Markovian processes and their large deviations [72]. The main reason for considering the exponential tilting is that it is known to be a Markov chain when the underlying measure P_n is Markovian and the observable M_n is additive in time [71, 73]. If one considers, for instance, the square of a sample mean as the observable M_n, then the exponential tilting is not Markovian.

F. Diffusion processes

The application of our results to Markov processes evolving in continuous time follows the examples above with minor changes of notations and techniques developed in large deviation theory to deal with this class of processes. For this reason, we do not cover this class in details, but only indicate the main changes involved, focusing as a specific example on diffusion processes $(X_t)_{t\geq 0}$ in $\mathbb{R}$, described by the following stochastic differential equation (SDE):

$$dX_t = F(X_t)dt + \sigma(X_t)dB_t. \tag{118}$$

Here, B_t is a Brownian motion in $\mathbb{R}$, while F and σ are two real functions of X_t, known as the drift and the noise amplitude, respectively. Many different observables can be defined in the context of SDEs, depending on the application and large deviation limit (low-noise or long-time) considered. We can consider, for example,

$$M_T = \frac{1}{T}\int_0^T f(X_t)dt \tag{119}$$

as a generalisation of the sample means studied before, which leads us to the problem of estimating the probability $P_T(M_T \in B)$ in the limit $T \to \infty$, where P_T is the probability measure of the process X_t over the time interval $[0, T]$ induced by the SDE (118).

Contrary to discrete-time Markov chains, we cannot write down any explicit expression for P_T; however, there is an explicit expression for the Radon–Nikodym derivative associated with a change of process if we consider that process to result from a change of drift. That is to say, change the drift F in (118) to obtain a new SDE

$$dX_t = G(X_t)dt + \sigma(X_t)dB_t, \tag{120}$$

which defines a new law for $(X_t)_{t=0}^T$ that we denote by Q_T. Then the action of this process, as compared to the original one, is obtained from Girsanov's theorem [74, Sec. 6.4], which states that

$$L_T = \frac{dP_T}{dQ_T} = \exp\left(\int_0^T c(X_t)dB_t - \frac{1}{2}c(X_t)^2 dt\right) \tag{121}$$

where

$$c(x) = \frac{F(x) - G(x)}{\sigma(x)}. \tag{122}$$

Consequently,

$$W_T = -\frac{1}{T}\log\frac{dP_T}{dQ_T} = \frac{1}{2T}\int_0^T c(X_t)^2 dt - \frac{1}{T}\int_0^T c(X_t) dB_t. \tag{123}$$

Both M_T and W_T are functions of the trajectory $(X_t)_{t=0}^T$ with law Q_T over $[0,T]$.

From this result, the joint large deviations of M_T and W_T with respect to Q_T can be obtained, similarly to Markov chains, by solving a spectral problem in which the transition matrix is replaced by the infinitesimal generator of X_t [53]. As for Markov chains, the notion of exponential change of measure can also be defined for continuous-time processes and involves spectral elements related to the large deviations of M_T with respect to P_T. This is fully explained in [71].

As a simple illustration of the exponential change of measure, consider the Ornstein–Uhlenbeck process in $\mathbb{R}$, defined by

$$dX_t = -\gamma X_t dt + \sigma dB_t, \tag{124}$$

where $\gamma > 0$ and $\sigma > 0$. Moreover, let us take

$$M_T = \frac{1}{T}\int_0^T X_t dt \tag{125}$$

as the observable, which represents the area of X_t per unit time. In this case, it can be shown (see [71, Sec. 6] for the full calculation) that the exponential change of measure associated with X_t, corresponding to the process version of (21), is another SDE with drift $G(x) = -\gamma(x-m)$ and noise amplitude σ. For this new process, the typical value of M_T is clearly m, so the process is asymptotically efficient for estimating the large deviation probability of $M_T \in B$ with $B = [m,\infty)$, $B = (-\infty, m]$ or $B = [m, m+dm]$. In all cases, we find from (123) that the typical value of W_T under Q_T is

$$w^* = \frac{\gamma^2 m^2}{2\sigma^2}, \tag{126}$$

which is the known rate function $I_P(m)$ of M_T with respect to P_T.

This result is a diffusion analog of the Gaussian sample mean studied before, for which we found that the exponential tilting is another Gaussian with translated mean. Here, we see that a Gaussian process tilted with the sample mean is a Gaussian process having the same variance but a different mean. It can be checked that, as for the i.i.d. Gaussian sample mean, this is the only efficient change of measure in the class of Gaussian processes with linear drift. If we change the friction coefficient γ to another value, in addition to adding a constant to change the mean, then the process is no longer asymptotically efficient for the same reason that changing the variance in the Gaussian sample mean is not efficient. The calculations for the SDE are more complicated, but the results are similar.

Applications of IS for diffusions have been studied in statistical physics [62] as well as more applied areas such as finance and queueing theory, focusing invariably on the exponential change of measure [40]. In future works, it would be interesting to apply our formalism to study other IS measures for Markov processes, such as the one proposed in [63–65], to determine their efficiency and to see, in the end, if there is any gain from not using the exponential tilting, which is difficult to construct in practice, since it involves the solution of a spectral problem whose knowledge is equivalent to solving the large deviation problem [75]. Another important problem is to determine whether our formalism can be applied to study the efficiency of IS in the low-noise limit of SDEs, which is extensively used in physics, chemistry and engineering to study rare transition pathways [49–52]. The IS method itself can be applied in this limit (see, e.g., [61]), but it is not clear to what extent our assumptions hold.

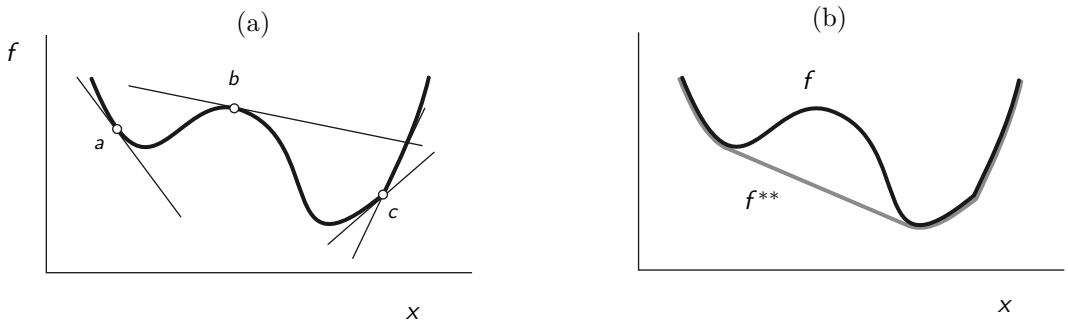

FIG. 7. (a) Function $f(x)$ with a unique supporting line at the point a, no supporting line at the point b, and many supporting lines at the point c, leading to $\partial f(a) = \{f'(a)\}$, $\partial f(b) = \emptyset$, and $\partial f(c) = [f'(c^-), f'(c^+)]$. (b) Function $f(x)$ and its convex envelope $f^{**}(x)$.

Appendix A: Convex analysis

We collect in this section basic results of convex analysis used in the paper in relation to the rate function $I_Q^B(w)$, defined in (53), and its Legendre–Fenchel transform $\lambda_Q^B(k)$, defined in (60). Both are functions of a single real variable, so we state the necessary results only for this simple case. We assume further that all convex functions are proper closed convex functions. For more general results and proofs, we refer to [76–78].

1. Subdifferentials

Let $f : \mathbb{R} \to \bar{\mathbb{R}}$ be a real function taking values in the set of extended reals $\bar{\mathbb{R}}$. The *subdifferential* $\partial f(x)$ of f at the point x is the set of all values $k \in \mathbb{R}$ such that

$$f(y) \geq f(x) + k(y - x) \tag{A1}$$

for all $y \in \mathbb{R}$ [76, Sec. 23]. Put differently, and as illustrated in Fig. 7a, $\partial f(x)$ is the set of slopes of all possible supporting lines of f at x. If f has not supporting line at x, then $\partial f(x) = \emptyset$. We will see next that this may happen when f is nonconvex.

For convex functions, subdifferentials exist everywhere in the domain of $f(x)$, except possibly at boundary points [76, Thm. 23.4]. For this class of functions, we have in fact $\partial f(x) = [f'(x^-), f'(x^+)]$, where $f'(x^-)$ is the left-derivative and $f'(x^+)$ the right-derivative [76, Thm. 24.3]. If these are equal, f is differentiable at x so that $\partial f(x) = \{f'(x)\}$ [76, Thm. 25.1]. In all cases, $\partial f(x)$ is a closed convex interval [76, p. 215].

2. Legendre–Fenchel transforms

The *Legendre–Fenchel transform* of f is the real function defined by

$$f^*(k) = \sup_{x \in \mathbb{R}} \{kx - f(x)\}, \qquad k \in \mathbb{R}. \tag{A2}$$

This function is also called the *dual* or *conjugate* of f and has the property of being convex [76, Thm. 12.2]. The *double dual* or *biconjugate* of f is the Legendre–Fenchel of f^*:

$$f^{**}(x) = \sup_{k \in \mathbb{R}} \{kx - f^*(k)\}. \tag{A3}$$

This is also a convex function, corresponding to the convex envelope or convex hull of f [77, Thm. 11.1], as illustrated in Fig. 7b.

With this geometric interpretation of f^{**}, it is natural to say that x is a *convex point* of f if $f(x) = f^{**}(x)$ and a *nonconvex point* of f if $f(x) \neq f^{**}(x)$. An important result proved in [68, Lem. 4.1] is that the set of convex points coincides with the set of points admitting supporting lines, except possibly at boundary points. With this proviso, we then have $f(x) = f^{**}(x)$ if and only if $\partial f(x) \neq \emptyset$. This is illustrated in Fig. 7a. The same result also implies that, if $f(x) = f^{**}(x)$, then $\partial f(x) = \partial f^{**}(x)$.

In this paper, we deal with rate functions, which always have at least one global minimum. Denoting one such minimizer by x^*, we then have $0 \in \partial f(x^*)$. Hence, x^* is a convex point such that $f(x^*) = f^{**}(x^*)$ and $\partial f(x^*) = \partial f^{**}(x^*)$.

3. Duality

The proof of our main result, Theorem 4, is based on another important result about convex functions stating (see [76, Cor. 23.5.1] or [77, Prop. 11.3]) that

$$k \in \partial f(x) \iff x \in \partial f^*(k). \tag{A4}$$

This property expresses a form of duality or conjugacy between the slopes of f and the slopes of f^*, illustrated in Fig. 8a. From this result, it is easy to see that convex, affine parts of f correspond to cusps of f^*, and vice versa, as shown in Fig. 8b.

The duality in (A4) also holds for f^{**}, since this function is convex and is the Legendre–Fenchel transform of f^*. Therefore,

$$k \in \partial f^{**}(x) \iff x \in \partial f^*(k). \tag{A5}$$

This result implies that f^* has a cusp also when f is nonconvex, as shown in Fig. 8, since f^{**} is affine where f is nonconvex. Thus, f^* has a cusp either if f is affine or f is nonconvex.

Since subdifferentials of f and f^{**} match at convex points, it is also clear from (A5) that the first duality (A4) holds locally at these points even if f is not globally convex. We use this result in this paper when dealing with the subdifferential of I_Q^B at its global minimum w^*, which is a convex point, as mentioned. In this case, the first duality result can be applied at that point even though I_Q^B might be nonconvex at other points, as in Fig. 2c or Fig. 6.

Appendix B: Contraction principle

The contraction principle is an important result in large deviation theory relating the rate functions of random variables that can be mapped to one another. Let $(A_n)_{n>0}$ be a sequence of random variables satisfying the LDP with good rate function I_A and let $(B_n)_{n>0}$ be another sequence such that $B_n = f(A_n)$ with f continuous. Then $(B_n)_{n>0}$ also satisfies the LDP with good rate function

$$I_B(b) = \inf_{a:f(a)=b} I_A(a). \tag{B1}$$

See [6, Thm. 4.2.1] for details.

Instead of considering a single continuous function f as the contraction, one can also consider a sequence $(f_n)_{n>0}$ of continuous functions. In this case, the contraction principle also applies

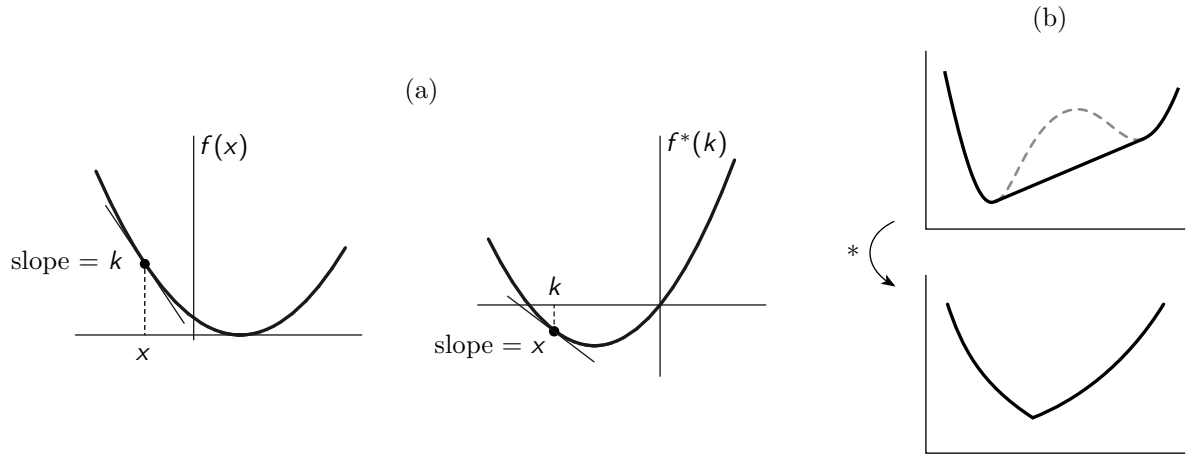

FIG. 8. (a) Illustration of the duality between the slopes of $f(x)$ and the slopes of its Legendre–Fenchel transform $f^*(k)$. (b) Functions with affine or nonconvex parts give rise to a Legendre–Fenchel transform having a cusp.

provided that f_n is "close enough" to f with respect to P_n. To be more precise, let $\mathcal{A}$ denote the space of A_n and define

$$\Gamma_{n,\delta} = \{a \in \mathcal{A} : \|f_n(a) - f(a)\| > \delta\} \tag{B2}$$

as the set of points for which f_n differs from f by at least $\delta > 0$ with respect to any metric $\|\cdot\|$ on $\mathcal{B}$, the space of B_n. Then, according to [6, Cor. 4.2.21], $B_n = f_n(A_n)$ satisfies the LDP with good rate function I_B given by (B1) with f as the contraction if, for all $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \log P_n(\Gamma_{n,\delta}) = -\infty. \tag{B3}$$

This condition only means that the probability that f_n differs from f decreases faster than exponentially with n in the large deviation limit. This is met in most cases when f_n is smooth and I_A is a good rate function.

Two particular applications of this result are considered in the paper.

Example 4: Consider two real random variables A_n and B_n related by the simple rescaling $B_n = c_n A_n$ with $c_n \to 1$ as $n \to \infty$. Here, the limit function is the identity $f(a) = a$, so one expects A_n and B_n to have the same rate function. This is verified by noting that, for every $M > 0$, there exists $n_0 = n_0(M, \delta)$ such that for all $n \geq n_0$, one has $\Gamma_{n,\delta} \subseteq (-\infty, -M] \cup [M, \infty)$. Therefore, from the definition of the LDP, we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log P_n(\Gamma_{n,\delta}) \leq - \inf_{|a|\geq M} I_A(a). \tag{B4}$$

But, since the rate function I_A of A_n is good, it is coercive, so that

$$\lim_{|a|\to\infty} I_A(a) = \infty. \tag{B5}$$

Therefore, the limit on the left-hand side of (B4) must give $-\infty$, implying $I_B(b) = I_A(b)$ from the condition (B3).

Example 5: Let $B_n = f(A_n) + c_n$ with $c_n \to c$. Then the rate function of B_n is obtained from (B1) with the contraction $B_n = f(A_n) + c$. This follows trivially because the distance between $f(a) + c_n$ and $f(a) + c$ is constant in a. Since $c_n \to c$, there must be an n beyond which $|c_n - c| < \delta$, leading to $P_n(\Gamma_{n,\delta}) = 0$, so the condition (B3) is also satisfied.

These results also hold if $\Gamma_{n,\delta}$ is defined on a subset of $\mathcal{A}$, since any restriction or constraint on A_n can be included in the definition of f_n. This arises, for example, when considering the contraction of $J_Q(m,w)$ to $I_Q^B(w)$, which involves the restriction $m \in B$.

ACKNOWLEDGMENTS

We are greatly indebted to Julien Reygner for valuable comments and insightful suggestions on the first version of the paper, which led to some technical modifications in Assumption 2, Assumption 3, and Eq. (53) in this version. We also thank Grégoire Ferré and Gabriel Stoltz for carefully reading the paper. A.G. thanks Maxime Sangnier for fruitful discussions during the writing of this paper. H.T. is supported by Stellenbosch University (Establishment Funds) and the National Research Foundation of South Africa (Grant No. 96199).

---

* arnaud.guyader@upmc.fr
† htouchet@alum.mit.edu, htouchette@sun.ac.za

[1] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*, Stochastic Modeling Series (Chapman and Hall, London, 1995).
[2] D. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge University Press, Cambridge, 2004).
[3] W. E, W. Ren, and E. Vanden-Eijnden, "Minimum action method for the study of rare events," Comm. Pure Appl. Math. **57**, 637–656 (2004).
[4] T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations: A Mathematical Perspective* (Imperial College Press, London, 2010).
[5] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics* (Springer, New York, 1985).
[6] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. (Springer, New York, 1998).
[7] F. den Hollander, *Large Deviations*, Fields Institute Monograph (AMS, Providence, RI, 2000).
[8] H. Touchette, "The large deviation approach to statistical mechanics," Phys. Rep. **478**, 1–69 (2009).
[9] J. P. Garrahan, R. L. Jack, V. Lecomte, E. Pitard, K. van Duijvendijk, and F. van Wijland, "Dynamical first-order phase transition in kinetically constrained models of glasses," Phys. Rev. Lett. **98**, 195702 (2007).
[10] J. P. Garrahan and I. Lesanovsky, "Thermodynamics of quantum jump trajectories," Phys. Rev. Lett. **104**, 160601 (2010).
[11] C. P. Espigares, P. L. Garrido, and P. I. Hurtado, "Dynamical phase transition for current statistics in a simple driven diffusive system," Phys. Rev. E **87**, 032115 (2013).
[12] G. Bunin, Y. Kafri, and D. Podolsky, "Cusp singularities in boundary-driven diffusive systems," J. Stat. Phys. **152**, 112–135 (2013).
[13] P. Tsobgni Nyawo and H. Touchette, "A minimal model of dynamical phase transition," Europhys. Lett. **116**, 50009 (2016), arxiv:1611.07707.
[14] A. Lazarescu, "Generic dynamical phase transition in one-dimensional bulk-driven lattice gases with exclusion," J. Phys. A: Math. Theor. **50**, 254004 (2017).
[15] G. Gallavotti and E. G. D. Cohen, "Dynamical ensembles in nonequilibrium statistical mechanics," Phys. Rev. Lett. **74**, 2694–2697 (1995).
[16] J. Kurchan, "Fluctuation theorem for stochastic dynamics," J. Phys. A: Math. Gen. **31**, 3719–3729 (1998).
[17] J. L. Lebowitz and H. Spohn, "A Gallavotti-Cohen-type symmetry in the large deviation functional for stochastic dynamics," J. Stat. Phys. **95**, 333–365 (1999).
[18] R. J. Harris and G. M. Schütz, "Fluctuation theorems for stochastic dynamics," J. Stat. Mech. **2007**, P07020 (2007).
[19] M. Baiesi, C. Maes, and B. Wynants, "Fluctuations and response of nonequilibrium states," Phys. Rev.

Lett. **103**, 010602 (2009).
[20] B. Derrida, "Non-equilibrium steady states: Fluctuations and large deviations of the density and of the current," J. Stat. Mech. **2007**, P07023 (2007).
[21] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim, "Stochastic interacting particle systems out of equilibrium," J. Stat. Mech. **2007**, P07014 (2007).
[22] R. J. Harris and H. Touchette, "Large deviation approach to nonequilibrium systems," in *Nonequilibrium Statistical Physics of Small Systems: Fluctuation Relations and Beyond*, Reviews of Nonlinear Dynamics and Complexity, Vol. 6, edited by R. Klages, W. Just, and C. Jarzynski (Wiley-VCH, Weinheim, 2013) pp. 335–360.
[23] J. P. Garrahan, "Aspects of non-equilibrium in classical and quantum systems: Slow relaxation and glasses, dynamical large deviations, quantum non-ergodicity, and open quantum dynamics," Physica A **504**, 130–154 (2018).
[24] K. Sekimoto, *Stochastic Energetics*, Lect. Notes. Phys., Vol. 799 (Springer, New York, 2010).
[25] U. Seifert, "Stochastic thermodynamics, fluctuation theorems and molecular machines," Rep. Prog. Phys. **75**, 126001 (2012).
[26] U. Seifert, "Stochastic thermodynamics: From principles to the cost of precision," Physica A **504**, 176–191 (2018).
[27] S. Ciliberto, "Experiments in stochastic thermodynamics: Short history and perspectives," Phys. Rev. X **7**, 021051 (2017).
[28] F. Cérou and A. Guyader, "Adaptive multilevel splitting for rare event analysis," Stoch. Anal. Appl. **25**, 417–443 (2007).
[29] T. Dean and P. Dupuis, "Splitting for rare event simulation: A large deviation approach to design and analysis," Stoch. Proc. Appl. **119**, 562–587 (2009).
[30] F. Cérou, A. Guyader, T. Lelièvre, and D. Pommier, "A multiple replica approach to simulate reactive trajectories," J. Chem. Phys. **134**, 054108 (2011).
[31] F. Cérou, B. Delyon, A. Guyader, and M. Rousset, "On the asymptotic normality of adaptive multilevel splitting," SIAM J. Uncertainty Quant. **7**, 1–30 (2019).
[32] F. Cérou, A. Guyader, and M. Rousset, "Adaptive multilevel splitting: Historical perspective and recent results," Chaos **29**, 043108 (2019).
[33] C.-E. Bréhier and T. Lelièvre, "On a new class of score functions to estimate tail probabilities of some stochastic processes with adaptive multilevel splitting," Chaos **29**, 033126 (2019).
[34] P. Grassberger, "Go with the winners: A general Monte Carlo strategy," Comp. Phys. Comm. **147**, 64–70 (2002).
[35] C. Giardina, J. Kurchan, and L. Peliti, "Direct evaluation of large-deviation functions," Phys. Rev. Lett. **96**, 120603 (2006).
[36] V. Lecomte and J. Tailleur, "A numerical approach to large deviations in continuous time," J. Stat. Mech. **2007**, P03004 (2007).
[37] L. Angeli, S. Grosskinsky, A. M. Johansen, and A. Pizzoferrato, "Rare event simulation for stochastic dynamics in continuous time," J. Stat. Phys. **176**, 1185–1210 (2019).
[38] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," Journal of Computational Physics **23**, 187–199 (1977).
[39] S. Juneja and P. Shahabuddin, "Rare-event simulation techniques: An introduction and recent advances," (Elsevier, Amsterdam, 2006) Chap. 11, pp. 291–350.
[40] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*, Stochastic Modelling and Applied Probability (Springer, New York, 2007).
[41] J. A. Bucklew, *Introduction to Rare Event Simulation* (Springer, New York, 2004).
[42] J. S. Sadowsky and J. A. Bucklew, "Large deviations theory techniques in Monte Carlo simulation," in *Proceedings of the 1989 Winter Simulation Conference*, edited by E. A. MacNair, K. J. Musselman, and P. Heidelberger (ACM, New York, 1989) pp. 505–513.
[43] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient Monte Carlo estimation," IEEE Trans. Info. Th. **36**, 579–588 (1990).
[44] J. A. Bucklew, P. Ney, and J. S. Sadowsky, "Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains," J. Appl. Prob. **27**, 44–59 (1990).
[45] H.-J. Schlebusch, "On the asymptotic efficiency of importance sampling techniques," IEEE Trans. Info. Th. **39**, 710–715 (1993).

[46] A. B. Dieker and M. Mandjes, "On asymptotically efficient simulation of large deviation probabilities," Adv. Appl. Prob. **37**, 539–552 (2005).

[47] B. Efron and D. Traux, "Large deviations theory in exponential families," Ann. Math. Stat. **39**, 1402–1424 (1968).

[48] H. Touchette, "Asymptotic equivalence of probability measures and stochastic processes," J. Stat. Phys. **170**, 962–978 (2018).

[49] M. Cottrell, J.-C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," IEEE Trans. Aut. Cont. **28**, 907– 920 (1983).

[50] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Grundlehren der Mathematischen Wissenschaften, Vol. 260 (Springer, New York, 1984).

[51] R. Graham, "Macroscopic potentials, bifurcations and noise in dissipative systems," in *Noise in Nonlinear Dynamical Systems*, Vol. 1, edited by F. Moss and P. V. E. McClintock (Cambridge University Press, Cambridge, 1989) pp. 225–278.

[52] D. G. Luchinsky, P. V. E. McClintock, and M. I. Dykman, "Analogue studies of nonlinear systems," Rep. Prog. Phys. **61**, 889–997 (1998).

[53] H. Touchette, "Introduction to dynamical large deviations of Markov processes," Physica A **504**, 5–19 (2018).

[54] L. Bertini, A. De Sole, D. Gabrielli, G. Jona-Lasinio, and C. Landim, "Macroscopic fluctuation theory," Rev. Mod. Phys. **87**, 593–636 (2015).

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

[56] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method* (Springer, New York, 2004).

[57] A. Engel, R. Monasson, and A. K. Hartmann, "On large deviation properties of Erdös–Rényi random graphs," J. Stat. Phys. **117**, 387–426 (2004).

[58] A. K. Hartmann, "Large-deviation properties of largest component for random graphs," Eur. J. Phys. B **84**, 627–634 (2011).

[59] T. Dewenter and A. K. Hartmann, "Large-deviation properties of resilience of power grids," New J. Phys. **17**, 015005 (2015).

[60] P. Guasoni and S. Robertson, "Optimal importance sampling with explicit formulas in continuous time," Finance Stoch. **12**, 1–19 (2008).

[61] E. Vanden-Eijnden and J. Weare, "Rare event simulation of small noise diffusions," Comm. Pure Appl. Math. **65**, 1770–1803 (2012).

[62] A. Kundu, S. Sabhapandit, and A. Dhar, "Application of importance sampling to the computation of large deviations in nonequilibrium processes," Phys. Rev. E **83**, 031119 (2011).

[63] K. Klymko, P. L. Geissler, J. P. Garrahan, and S. Whitelam, "Rare behavior of growth processes via umbrella sampling of trajectories," Phys. Rev. E **97**, 032123 (2018).

[64] S. Whitelam, "Sampling rare fluctuations of discrete-time Markov chains," Phys. Rev. E **97**, 032122 (2018).

[65] D. Jacobson and S. Whitelam, "Direct evaluation of dynamical large-deviation rate functions using a variational ansatz," Phys. Rev. E **100**, 052139 (2019).

[66] P. Glasserman and Y. Wang, "Counterexamples in importance sampling for large deviations probabilities," Ann. Appl. Prob. **7**, 731–746 (1997).

[67] A. Puhalskii and V. Spokoiny, "On large-deviation efficiency in statistical inference," Bernoulli **4**, 203–272 (1998).

[68] R. S. Ellis, K. Haven, and B. Turkington, "Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles," J. Stat. Phys. **101**, 999–1064 (2000).

[69] S. R. S. Varadhan, "Asymptotic probabilities and differential equations," Comm. Pure Appl. Math. **19**, 261–286 (1966).

[70] H. Touchette, "A basic introduction to large deviations: Theory, applications, simulations," in *Modern Computational Science 11: Lecture Notes from the 3rd International Oldenburg Summer School*, edited by R. Leidl and A. K. Hartmann (BIS-Verlag der Carl von Ossietzky Universität Oldenburg, Oldenburg, 2011).

[71] R. Chetrite and H. Touchette, "Nonequilibrium Markov processes conditioned on large deviations," Ann. Henri Poincaré **16**, 2005–2057 (2015).

[72] R. J. Harris and H. Touchette, "Current fluctuations in stochastic systems with long-range memory," J.

Phys. A: Math. Theor. **42**, 342001 (2009).
[73] U. Küchler and M. Sōrensen, "On exponential families of Markov processes," J. Stat. Planning and Inference **66**, 3–19 (1998).
[74] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes* (Springer, New York, 1979).
[75] R. Chetrite and H. Touchette, "Variational and optimal control representations of conditioned and driven processes," J. Stat. Mech. **2015**, P12001 (2015).
[76] R. T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970).
[77] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Vol. 317 (Springer, New York, 1988).
[78] J. Borwein and A. Lewis, *Convex Analysis and Nonlinear Optimization*, 2nd ed. (Springer, New York, 2006).