ON FREE ENERGY CALCULATIONS USING FLUCTUATION THEOREMS OF WORK

Doctoral Thesis by Aljoscha M. Hahn

Von der Fakultät für Mathematik und Naturwissenschaften der
CARL VON OSSIETZKY UNIVERSITÄT OLDENBURG
zur Erlangung des Grades und Titels eines Doctor rerum naturalium (Dr. rer. nat.)
angenommene Dissertation von Herrn Aljoscha Maria Hahn,
geboren am 20. Februar 1977 in Hagios Antonios (Kreta).

Gutachter: **Prof. Dr. Andreas Engel**
Zweitgutachter: **Prof. Dr. Holger Stark**
Tag der Disputation: 15. Oktober 2010

Summary

Free energy determination of thermodynamic systems which are analytical intractable is an intensively studied problem since at least 80 years. The basic methods are commonly traced back to the works of John Kirkwood in the 1930s and Robert Zwanzig in the 1950s, who developed the widely known thermodynamic integration and thermodynamic perturbation theory. Originally aiming analytic calculations of thermodynamic properties with perturbative methods, the full power of their methods was only revealed in conjunction with modern computer capabilities and Monte Carlo simulation techniques. In this alliance they allow for effective, nonperturbative treatment of model systems with high complexity, in specific calculations of free energy differences between thermodynamic states.

The recently established nonequilibrium work theorems, found by Christopher Jarzynski and Gavin Crooks in the late 1990s, revived traditional free energy methods in a quite unexpected form. Whilst formerly relying on computer simulations of microscopic distributions, in their new robe they are based on measurements of work of nonequilibrium processes. The nonequilibrium work theorems, i.e. the Jarzynski Equation and the Crooks Fluctuation Theorem meant a paradigmatic change with respect to theory, experiment, and simulation in admitting the extraction of equilibrium information from nonequilibrium trajectories.

The focus of the present thesis lies on three directions: first, on understanding and characterizing elementary methods for free energy calculations originating from the fluctuation theorem. Second, on analytic transformation of data in order to enhance the performance of the methods, and third, on the development of criteria which allow for judging the quality of free energy calculations. Calculation hereby actually means statistical estimation with data sampled or measured from random distributions. The main work of the present author is summarized as follows.

Inspired by the work of Charles Bennett on his acceptance ratio method for free energy calculations and its recent revival in the context of Crooks' Fluctuation Theorem, we studied this method in great detail to understand its overall observed superiority over related methods. The acceptance ratio method utilizes measurements of work in both directions of a process, and it was finally observed by Shirts and co-workers that it can

also be understood as a maximum likelihood estimator for a given amount of data, which greatly explains its exquisite properties from a totally different point of view than that of Bennett. Yet, a drawback of the maximum likelihood approach to the acceptance ratio method is the implicit switch to another process of data gathering via Bayes' Theorem, which no longer reflects the actual process of measurement. This drawback can be removed, as we have shown, by a slightly different ansatz, which reveals the acceptance ratio method to be a constrained maximum likelihood estimator. The great difference between the two approaches is that the latter permits more efficient estimators, whilst the former does not. Even more efficient estimators can be provided by some other means, but are always linked to the specific process and require knowledge on the functional dependence of the work distributions on the free energy. In contrast, the acceptance ratio method is always a valid method, and in fact the best method we can use with given measurements of work when having no further information on the work distributions – which is virtually always the case.

The performance of the acceptance ratio method depends on the partitioning of the number of work-measurements with respect to the direction of process. Bennett has already discussed this question in some detail and derived an equation whose solution specifies the optimal partitioning of measurements. Albeit, he could not gain relevance for it and suggested the problem to be untreatable in praxis. We have completed this issue, in first proving that the mean square error is a convex function of the fraction of measurements in one direction, which guarantees the existence of a unique optimal partitioning, and then demonstrating its practical relevance for the purpose of free energy calculations at maximum efficiency. In addition, the convexity of the mean square error explains analytically why the acceptance ratio method is generically superior to free energy calculations relying on the Jarzynski Equation.

Building up on Jarzynski's observation that traditional free energy perturbation can be markedly improved by inclusion of analytically defined phase space maps, we have put forward this new and promising direction and derived a fluctuation theorem for a generalized notion of work, defined with recourse to phase space maps. The generalized work fluctuation theorem has the same form as Crooks' Fluctuation Theorem, and can include it for specific choices of maps. This analogy allowed us to define the acceptance ratio method also for generalized work.

The high potential of the mapping methods can also be seen as its drawback: there is no general receipt for the construction of suitable maps. So the method seems to depend primarily on the extend of the user's insight into the problem at hand. However, we could demonstrate its applicability to the calculation of the chemical potential of a high-density Lennard-Jones fluid. Thereby we have constructed maps in two ways, by simulation and by an analytical approach. In the analytic case, the map was parametrized and the parameter numerically optimized. The maps in conjunction with the acceptance ratio method yielded high-accuracy results which outperformed those from traditional calculations by far, in particular with respect to the speed of convergence.

Convergence is critical to be achieved within statistical calculations for obtaining reliable results, but is in general not easy to verify - if possible at all. Because of their strong dependence on rarely observed events, free energy calculations with the Jarzynski Equation and the acceptance ratio method suffer from the tendency to seeming convergence. This means that a running calculation obeys the property to settle down on a stable value over long times – but without having reached the true value of the free energy. Moreover, seeming convergence is typically accompanied by a small and decreasing sample variance, which may harden the belief in that the calculation has converged. This is quite problematic, as then there is no reliance on the results of calculations. To resolve this, we have proposed a measure of convergence for the acceptance ratio method. The convergence measure relies on a simple-to-implement test of self-consistency of the calculations which implicitly monitors the sufficient observation of rare events. Our analytical and numerical studies validated its reliability as a measure of convergence.

Zusammenfassung

Die Bestimmung freier Energien analytisch unzugänglicher Systeme ist ein seit wenigstens 80 Jahren intensiv studiertes Problem. Die grundlegenden Methoden werden gemeinhin auf die Arbeiten von John Kirkwood in den 30er Jahren und Robert Zwanzig in den 50er Jahren zurückgeführt. Obgleich urpsrünglich zur störungstheoretischen Berechnung thermodynamischer Eigenschaften entwickelt, entfaltete sich die volle Reichweite ihrer Methoden erst in Verbindung mit der Leistungsfähigkeit moderner Computer und Monte-Carlo Simulationstechniken. In dieser Vereinigung erlauben sie die effektive, nichtstörungstheoretische Behandlung komplexer Modellsysteme, insbesondere die Berechnung von Differenzen der freien Energie.

Die in jüngster Zeit begründeten Fluktuationstheoreme der Arbeit im Nichtgleichgewicht, entdeckt von Christopher Jarzynski und Gavin Crooks in den späten 90ern, hatten eine Wiederbelebung traditioneller Methoden zur Berechnung der freien Energie in einer recht unerwarteten Form zur Folge. Ursprünglich auf Computersimulationen mikroskopischer Verteilungen gestützt, beruhen sie in ihrem neuen Gewand auf Messungen der Arbeit in Nichtgleichgewichtsprozessen. Die Fluktuationstheoreme der Arbeit, d.h. die Jarzynski Gleichung und das Crooks'sche Fluktuationstheorem, bedeuteten einen paradigmatischen Wechsel in Bezug auf Theorie, Experiment und Simulation, indem sie die Bestimmung von Gleichgewichtseigenschaften aus Nichtgleichgewichtstrajektorien erlauben.

Die vorliegende Dissertation hat drei Schwerpunkte: Zum ersten, die Charakterisierung derjenigen elementaren Methoden zur Berechnung von freien Energien, die auf dem Fluktuationstheorem gründen. Zum zweiten, die analytische Datentransformation mit dem Ziel, die Güte der Methoden zu verbessern; und drittens, die Entwicklung von Kriterien, die einen Rückschluß auf die Qualität der Berechnungen erlauben. Berechnung bedeutet hier genauer statistische Schätzung, denn die den Rechnungen zugrundeliegenden Daten gehorchen statistischen Verteilungen. Die wesentliche Arbeit des gegenwärtigen Autors lässt sich wie folgt zusammenfassen.

Inspiriert von Charles Bennett's Arbeit zu seiner "Acceptance-Ratio" Methode zur Berechnung freier Energien und deren aktueller Wiederbelebung durch das Crooks'sche Fluktuationstheorem, haben wir diese Methode im Detail untersucht, um ihre allgemein

beobachtete Überlegenheit über verwandte Methoden zu verstehen. Die Acceptance-Ratio-Methode nutzt Messungen der Arbeit in beiden Richtungen eines Prozesses und kann, wie von Shirts und Mitarbeitern gezeigt wurde, als Maximum-Likelihood Schätzer der freien Energie angesehen werden. Dies erklärt deren vorzügliche Eigenschaften von einem gänzlich anderen Gesichtspunkte aus als dem Bennett'schen. Ein Nachteil des Maximum-Likelihood Zugangs zur Acceptance-Ratio-Methode liegt jedoch in seinem impliziten Wechsel zu einem anderen Prozess der Datengewinnung, der demjenigen der Messung nicht mehr entspricht. Wie wir zeigen konnten, lässt sich dieser Nachteil durch einen leicht modifizierten Ansatz beseitigen, welcher zeigt, daß die Acceptance-Ratio-Methode ein Maximum-Likelihood-Schätzer unter Nebenbedingungen darstellt. Infolgedessen können effizientere Schätzer auf Grundlage derselben Daten existieren, was bei einem reinen Maximum-Likelihood Schätzer nicht der Fall ist. Es gibt Beispiele für effizientere Schätzer, allerdings sind sie immer an den speziellen Prozess gebunden und erfordern die Kenntnis der funktionellen Abhängigkeit der Arbeitsverteilungen von der freien Energie. Im Gegensatz dazu ist die Acceptance-Ratio-Methode immer zulässig, und tatsächlich ist sie die beste Methode, die auf Grundlage gegebener Arbeitsmessungen und des Fluktuationstheoremes genutzt werden kann, solange keine zusätzliche Information über die Arbeitsverteilungen vorliegt - was beinahe immer der Fall ist.

Die Güte der Acceptance-Ratio-Methode ist abhängig von der Aufteilung der Anzahl der Arbeitsmessungen auf die Vorwärts- und Rückwärtsrichtung des Prozesses. Dies wurde bereits von Bennett diskutiert, der auch eine Gleichung angeben konnte, deren Lösung die optimale Aufteilung bestimmt. Jedoch hielt er sie für unzureichend lösbar in der praktischen Anwendung. Wir haben die Erörterung dieser Problemstellung vervollständigt, indem wir zunächst gezeigt haben, daß der mittlere quadratische Fehler der Acceptance-Ratio-Methode eine konvexe Funktion des Anteils der Arbeitswerte in einer Richtung ist, womit die Existenz einer eindeutigen optimalen Aufteilung garantiert ist. Weiterhin konnten wir zeigen, daß die optimale Aufteilung der Messungen in der Praxis realisiert werden kann, und daß dies eine wesentliche Steigerung der Effizienz der Methode ermöglicht. Darüberhinaus erklärt die Konvexität des mittleren quadratischen Fehlers analytisch, warum die Acceptance-Ratio-Methode gewöhnlich von großem Vorteil gegenüber denjenigen Methoden ist, die die Jarzynski Gleichung ausnutzen.

Aufbauend auf Jarzynski's Beobachtung wonach sich die traditionelle "Free-Energy-

Perturbation" Methode durch bijektive Abbildungen des Phasenraumes wesentlich verbessern lässt, haben wir diese neue und vielversprechende Richtung fortgeführt und ein Fluktuationstheorem für einen verallgemeinerten Begriff der Arbeit aufgestellt, der unter Einbeziehung von Abbildungen definiert ist. Das Fluktuationstheorem der verallgemeinerten Arbeit hat dieselbe Form wie das Crooks'sche Fluktuationstheorem und kann dieses für spezielle Abbildungen enthalten. Diese Analogie erlaubte uns, die Acceptance-Ratio-Methode auch für verallgemeinerte Arbeit zu definieren.

Das hohe Potenzial der Abbildungsmethode kann auch als ihr Nachteil angesehen werden: Es gibt kein allgemeines Rezept für die Konstruktion geeigneter Abbildungen. Daher scheint die Methode in erster Linie von dem Einblick des Nutzers in die behandelten Probleme abzuhängen. Wir konnten jedoch zeigen, wie man diese Methode erfolgreich zur Berechnung des chemischen Potentials eines wechselwirkenden Fluides bei hoher Dichte einsetzen kann. Hierbei haben wir Abbildung auf zwei Wegen konstruiert, analytisch und durch Simulation. Im analytischen Falle wurde die Abbildung parametrisiert und der Parameter numerisch optimiert. Die Abbildungen in Kombination mit der Acceptance-Ratio-Methode zeitigten Ergebnisse von hoher Präzision, die denen vergleichbarer traditioneller Methoden weit überlegen sind, insbesondere hinsichtlich der Konvergenz der Berechnungen.

Das Erreichen von Konvergenz einer statistischen Schätzung ist von großer Bedeutung für die Zuverlässigkeit des Ergebnisses, aber im allgemeinen nicht einfach zu verifizieren - sofern dies überhaupt möglich ist. Wegen ihrer starken Abhängigkeit von seltenen Ereignissen leiden Berechnungen der freien Energie mithilfe der Jarzynski Gleichung und der Acceptance-Ratio-Methode unter der Tendenz zur scheinbaren Konvergenz. Unter letzterem verstehen wir die Eigenschaft einer laufenden Berechnung, sich über lange Zeit auf einem stabilen Plateau einzupendeln - ohne den wahren Wert der freien Energie erreicht zu haben. Darüberhinaus ist scheinbare Konvergenz typischerweise von einer kleinen, abnehmenden Stichproben-Varianz begeleitet, was den Eindruck von Konvergenz noch verstärken kann. Um dieses Problem zu lösen, haben wir ein Konvergenzmaß für die Acceptance-Ratio-Methode vorgeschlagen. Das Konvergenzmaß beruht auf einem einfach durchzuführenden Test der Selbstkonsistenz der Berechnungen, welcher implizit die hinreichende Beobachtung seltener Ereignisse prüft. Die Zuverlässigkeit des Konvergenzmasses konnten wir durch analytische und numerische Studien belegen.

Preface

This thesis is of cumulative character, i.e. the main work has already been published in peer-reviewed journals, consisting of the papers [1–3]. We will not repeat all details concerning that work here, but instead will try to give a comprehensive description of the scientific context and point to our own achievements at the appropriate place.

The literature on free energy calculation is vast and the methods are numerous. But concerning the principles, the different techniques are essentially understood by tracing them back to only a few elementary methods [4]. We will adopt this point of view here, and begin with an introduction to the basic methods of free energy calculation within their physical background in section I. The methods are divided into three classes, namely into equilibrium, nonequilibrium, and mapping methods. But instead of going into the details of free energy calculation, we merely state three fundamental "source-relations" from which free energy methods can be deduced. One of these relations is Crooks Fluctuation Theorem.

In section II elementary free energy estimators are introduced, namely free energy perturbation, the acceptance ratio method, umbrella sampling, thermodynamic integration, together with their generalizations to nonequilibrium and mapping methods. To do this in a compact way, we have chosen to deduce them all from a unified point of view, namely from a *formal* fluctuation theorem, which can be related to any of the three mentioned classes of methods. This has the benefit that the similarities of methods can be worked out clearly, based on quite simple notation. Moreover, all properties of free energy estimators observed within this formalized context hold equally well in their specialized versions. Yet, the drawback of such an approach is also clear: the effort of data gathering when using one and the same estimator in the different contexts is somewhat obscured. For example, it makes a great difference whether using free energy perturbation with Monte Carlo sampling of canonical distributions, or the formal identical Jarzynski estimator with simulations of nonequilibrium trajectories. Nevertheless, we hope our approach is helpful to clarify the affinities and interrelations of methods.

Contents

Summary 1

Zusammenfassung 4

Preface 7

I. Introduction to free energy methods 9

A. Equilibrium methods 10

B. Nonequilibrium work theorems 13

C. Nonequilibrium free energy calculations 18

D. Mapping methods 22

1. Fluctuation theorem of generalized work 23

2. Relation to Crook's Fluctuation Theorem 26

3. Construction of suitable maps 27

II. Elementary free energy estimators 29

A. One-sided estimation (free energy perturbation) 32

B. Two-sided estimation (acceptance ratio method) 36

1. The ansatz and the estimator 36

2. Rare events of two-sided estimation 41

3. Convexity of mean square error 43

4. Dynamic sampling strategy 44

5. Measure of convergence 46

C. Umbrella sampling 49

D. Thermodynamic integration 52

III. Final remarks 55

Acknowledgments 58

References 58

Published work 67

Paper [1] 67

Paper [2] 85

Paper [3] 97

I. INTRODUCTION TO FREE ENERGY METHODS

The central idea bridging the macroscopic laws of thermodynamics with the microscopic, Hamiltonian description of systems involving a large number of degrees of freedom can be cast in the notion of statistical ensembles. A statistical ensemble can be expressed as some probability density on the system's phase space and provides, if suitably chosen, a *model of thermodynamics* [5]. This means that ensemble averages of mechanical phase-functions associated to thermodynamic observables obey exactly the thermodynamic relations. Despite the long-standing and still highly topical problem of its justification from mechanical principles [6–14], this approach has proven to be one of the most fruitful in physics with regard to analytic and computational extraction of equilibrium properties and near-equilibrium fluctuations of thermodynamic systems. Prominent examples are the microcanonical, canonical, and grand-canonical ensemble, which can be viewed to be equivalent in the thermodynamic limit of a large number of particles (at least if no long-range interactions are present in the system [5]). Each of them is the appropriate ensemble for a class of experimental setups, defined by those macroscopic state-variables which are being controlled. The canonical ensemble, in specific, is the adequate ensemble for closed systems in (weak) contact with a heat bath [15].

In the latter case, the quantity of central interest is the (Helmholtz) *free energy*, as knowledge of free energy in dependence of externally controlled state variables allows for inference on the equilibrium properties of the system under study [16], but is also useful in the context of stability analysis through the minimum principle of free energy [17]. The canonical ensemble offers a direct route to free energy by identifying it with the logarithm of the *partition function*, which is the normalizing constant of the canonical density.

To provide the notions, let $H_\lambda(x)$ be the Hamiltonian of the system under study, which, in addition to the phase space variable x, shall explicitly depend on some externally controlled parameter λ coupled to the microscopic degrees of freedom x. For example, λ may be the strength of an applied field, the system's volume (if the spatial confinement is explicit in the Hamiltonian), or the value of some physical property of the systems constituents (e.g. charge of particles), but it may also be thought of as a "frozen" generalized coordinate, e.g. the distance of two bodies. The canconical probability density

$\rho_\lambda(x)$ associated with the thermodynamic equilibrium state (T, λ) reads

$$\rho_\lambda(x) = \frac{e^{-\beta H_\lambda(x)}}{Z_\lambda}, \tag{1}$$

with $\beta = 1/k_B T$ the inverse of the heat bath's temperature T times Boltzmann's constant k_B, and Z_λ the partition function,

$$Z_\lambda = \int e^{-\beta H_\lambda(x)} dx. \tag{2}$$

Finally, the free energy $F_\lambda = F_\lambda(T)$ is given by

$$F_\lambda = -\frac{1}{\beta} \ln Z_\lambda. \tag{3}$$

Due to its connection to free energy, calculation of the partition function is a major aim in physics, but in general the integral (2) cannot be carried out analytically for interacting systems, especially if the number of degrees of freedom is large. These difficulties are typically accompanied by another severe problem: sparse regions of phase space dominate the value of (2) if $H_\lambda(x)$ has numerous local minima which are distributed over phase space and separated by large energetic barriers. This also rules out successful numerical integration and "blind-shooting" Monte Carlo integration of the partition function; in the latter case one interprets the integral (2) as an average of the integrand with respect to a uniform distribution.

A. Equilibrium methods

By means of the Metropolis algorithm [18] or molecular simulation [19], it is possible to simulate random draws of phase space points x distributed according to the canonical density $\rho_\lambda(x)$ once the Hamiltonian is given, and *without* needing to know the value of the partition function. Hence, in principle one has access to the mentioned sparse regions of phase space. But as the desired partition function Z_λ can not be expressed as an average of a mechanical phase-function in the density ρ_λ (rather it's the normalizing constant), we cannot use this possibility for direct calculation of Z_λ.

However, *ratios* of partition functions, and thus free energy *differences*, can be expressed as averages in the canonical density. Assume we are interested in the free energy difference ΔF between two states "0" and "1" at the same temperature $1/\beta$, but with different values λ_0 and λ_1 of the parameter λ. Without loss of generality, we may assume $\lambda_0 = 0$ and $\lambda_1 = 1$, which can always be achieved by a suitable change of variables. Defining the free energy difference with

$$\Delta F := F_1 - F_0 = -\frac{1}{\beta} \ln \frac{Z_1}{Z_0}, \tag{4}$$

and the (microscopic) energy difference by

$$\Delta H(x) := H_1(x) - H_0(x), \tag{5}$$

the ratio of $\rho_0 = \rho_{\lambda_0}$ and $\rho_1 = \rho_{\lambda_1}$ reads

$$\frac{\rho_0(x)}{\rho_1(x)} = e^{\beta(\Delta H(x) - \Delta F)}. \tag{6}$$

This simple identity is the source for the most fundamental computational free energy (difference) methods, namely Zwanzig's free energy perturbation [20], Bennett's acceptance ratio method [21], Torrie and Valleau's umbrella sampling [22], and in some sense also of Kirkwood's thermodynamic integration [23] (by taking $\lambda_1 = \lambda_0 + d\lambda$). These methods again form the basis of a large variety of more specialized and generalized techniques, developed over the decades in adaption to the concrete problems treated [4].

To provide some insight into the very nature of these methods, we take a brief look on the popular *free energy perturbation* method. It relies on the identity

$$\int e^{-\beta \Delta H(x)} \rho_0(x)\, dx = e^{-\beta \Delta F}, \tag{7}$$

which is a simple integral consequence of Eq. (6). The exponential of ΔF is expressed here as *ensemble average* of $e^{-\beta \Delta H(x)}$ in the canonical density (ensemble) ρ_0. Accordingly, a statistical estimate of ΔF is obtained by calculating a *sample average* of $e^{-\beta \Delta H(x)}$ with a set of N randomly drawn phase space points x, distributed according to $\rho_0(x)$. Denoting

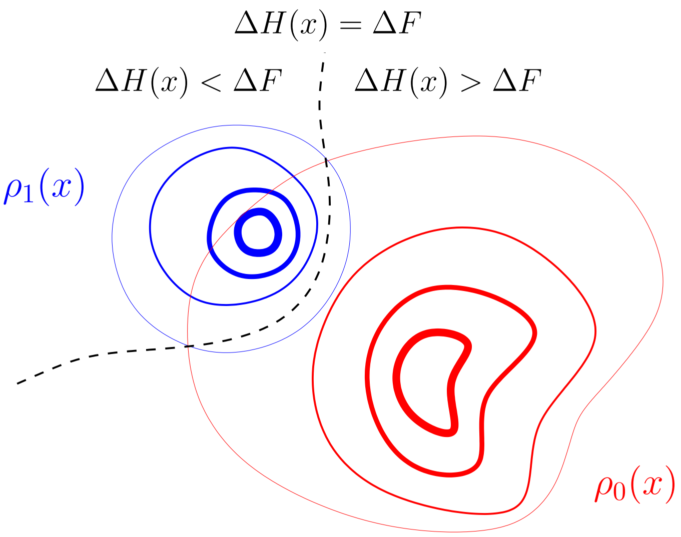

FIG. 1: Schematic contour plot of equilibrium densities $\rho_0(x)$ and $\rho_1(x)$, the drawing plane representing the phase space (thicker lines indicate larger probability density). The dashed line is determined by $\rho_0(x) = \rho_1(x)$, hence $\Delta H(x) = \Delta F$ (Eq. (6)). It divides the phase space into regions where $\Delta H(x) > \Delta F$ and $\Delta H(x) < \Delta F$. Precise free energy perturbation calculations with simulated draws from $\rho_0(x)$ require sampling the region where ρ_1 has its main probability mass.

the estimate with $\widehat{\Delta F}$, it is explicitly obtained by evaluating the estimator

$$\widehat{\Delta F} = -\frac{1}{\beta} \ln \frac{1}{N} \sum_{i=1}^{N} e^{-\beta \Delta H(x_i)} \tag{8}$$

with a set $\{x_i\}$ of N phase-space points, obtained e.g. from Metropolis Monte Carlo simulations of the distribution $\rho_0(x)$.

Hereby we have already indicated a common characteristic of computational free energy methods: the calculation of ΔF is no longer of bare analytic, but rather of stochastic nature. This means that the calculated value $\widehat{\Delta F}$ is a random variable with all its drawbacks: it spreads, i.e. repeated calculations in the same manner differ from each other, and it may even be biased, i.e. systematically off the desired value ΔF. The latter is actually the typical case in free energy calculations, albeit convenient methods (e.g. those mentioned above) possess the property of converging almost certainly to ΔF with unboundedly growing sample size N. Informally written,

$$\lim_{N\to\infty} \widehat{\Delta F} = \Delta F. \tag{9}$$

In praxis, of course, only finite sample sizes N are available, and therefore a critical question is whether the actual sample size is large enough to ensure convergence of the calculation (within a reasonable spread). Or, the other way round, how large needs the sample size N to be to obtain converging estimates $\widehat{\Delta F}$? Figure 1 illustrates this problem for free energy perturbation, which is discussed in greater detail in section II.

In essence, all traditional approaches can be called *equilibrium methods*, as they either use samples directly from equilibrium distributions, or from modified, so-called "biased" equilibrium distributions. For a long time, this seemed to be the only feasible way for properly obtaining free energy differences. The situation has changed dramatically with the discovery of nonequilibrium work relations by Jarzynski and Crooks. They have shown that it is indeed possible to extract *equilibrium information*, in particular free energy differences, from *nonequilibrium processes*, and thus from nonequilibrium distributions.

B. Nonequilibrium work theorems

The Jarzynski Equation [24–27] and Crooks' Fluctuation Theorem [28–30] are among the few exact results of nonequilibrium statistical physics which remain valid arbitrarily far from equilibrium. These closely connected *nonequilibrium work theorems* relate the statistics of work which is necessary for the realization of an externally driven, finite-time process with the equilibrium free energy difference between final and initial states of that process. Their great importance in view of fundamental theoretical issues results from the fact that they provide a basis for a new understanding of the second law of thermodynamics in a probabilistic sense. In specific, the Jarzynski Relation can be viewed as the second law in terms of an *equality* [31–33], as it implies the second law *inequality* [24]. Within the theoretical framework of the nonequilibrium work theorems, however, single realizations of nonequilibrium processes which violate the second law can, and even must occur, but the statistics of realizations in total is such, that it guarantees the validity of the second law *in average*.

The possibility of second law violations on the single-trajectory level, i.e. for single realizations of a process, has already been noticed by Boltzmann [10], and could be made evident by Evans, Cohen, Morris and Searles with numerical studies of small systems [34, 35]. Since then, strongly put forward by the nonequilibrium work theorems and experimental verifications of fluctuation theorems [36–43], it has come to attention that for small systems the laws of thermodynamics can still be expected to hold, but only in an averaged sense [31]. This has already been anticipated in the 1970s in the works of Bochkov and Kuzovlev [44], who derived relations similar to the nonequilibrium work theorems (involving no free energy difference), although with a somewhat different definition of work [45–47].

To prepare a quantitative formulation of the nonequilibrium work theorems, assume the following process. Initially in the equilibrium state $(T, \lambda = 0)$, our earlier defined system with Hamiltonian $H_\lambda(x)$ is being driven out of equilibrium by changing $\lambda = \lambda(t)$ in a predefined manner within some finite time $t = \tau$ from 0 to 1, while remaining coupled to the heat bath. The prescription $\lambda(\cdot) := \{\lambda(t)\}_0^\tau$ is commonly called the *protocol*. Schematically:

$$\lambda(0) = \ 0 \ \xrightarrow{\lambda(t)} \ 1 \ = \lambda(\tau). \tag{10}$$

The realization of that process requires an amount of work W to be done, whose value depends on the microscopic trajectory $x(\cdot) = \{x(t)\}_0^\tau$ on which the system evolves during the process. As the mechanical force acting "on the coordinate" λ is given by $-\frac{\partial}{\partial\lambda} H_\lambda(x)$ [15, 48], the work *applied to* the system along a single trajectory is given by the work functional $\mathcal{W}_{[\lambda]}[x(\cdot)]$ with [24, 31]

$$\mathcal{W}_{[\lambda]}[x(\cdot)] = \int_0^1 \frac{\partial}{\partial\lambda} H_{\lambda(t)}(x(t)) \ d\lambda(t) = \int_0^\tau \frac{\partial}{\partial\lambda} H_{\lambda(t)}(x(t)) \ \dot{\lambda}(t) \ dt. \tag{11}$$

Due to the random nature of trajectory, the value $W = \mathcal{W}_{[\lambda]}[x(\cdot)]$ of work will be a random variable, too, distributed according to some probability density $p_0(W)$ (0 indicates the initial equilibrium state and with this the direction $0 \to 1$ of process). In total, the process, which shall be called the *forward* process, drives the initial equilibrium distribution $\rho_0(x)$ to a final *nonequilibrium* distribution $\rho_0^{neq}(x, \tau)$, which will not equal $\rho_1(x)$, but rather "lag behind" [49, 50].

The forward process can be contrasted with its "time-reversed" counterpart, the *reverse* process, which also starts in equilibrium, but now in state $(T, \lambda = 1)$, with λ being traced back from 1 to 0 in exactly the opposite manner of the forward process, i.e. according to the time reversed protocol $\bar{\lambda}(\cdot) = \{\lambda(\tau - t)\}_{t=0}^{\tau}$. Again, some amount of work $\mathcal{W}_{[\bar{\lambda}]}[x(\cdot)]$ will be necessary, and we denote the probability density of work $W = -\mathcal{W}_{[\bar{\lambda}]}[x(\cdot)]$ *done by* the system in the reverse process with $p_1(W)$. (We compare the work supplied to the system in the forward process with the work gained from the system in the reverse process. To make contact with the common notation, we note that if $p_R(W)$ denotes the density of reverse work supplied to the system, then $p_1(W) = p_R(-W)$.)

The *Crooks Fluctuation Theorem* states the forward and reverse probability densities of work $p_0(W)$ and $p_1(W)$ to be related by [29]

$$\frac{p_0(W)}{p_1(W)} = e^{\beta(W - \Delta F)}. \tag{12}$$

This relation can be shown to hold under quite general assumptions on the underlying dynamics, including Hamiltonian dynamics [51], deterministic thermostatted dynamics [52], stochastic dynamics [30, 53–55], and quantum dynamics [56–59]. In the latter case, however, work cannot be defined with (11), but rather as difference of energy measurements at final and initial times of the process [60, 61]. In essence, the conditions for (12) to hold are invariance of the Hamiltonian under time-reversal, and a time-reversible dynamics [30, 31] which conserves the canonical distribution ρ_λ if λ is held fixed (i.e. , ρ_λ needs to be a stationary solution of a Liouville-type equation once $\lambda = const$).

A remarkable property of the fluctuation theorem is its generality with respect to the choice of protocol $\lambda(\cdot)$: it is valid for *any* protocol connecting $\lambda = 0$ and $\lambda = 1$ within *arbitrary* process duration τ. Nevertheless, the work densities are functionals of the protocol, and their shape depends strongly on its choice. But in each case are the forward and reverse work densities connected by Eq. (12). In specific, they will always and only intersect at $W = \Delta F$, which is schematically sketched in figure 2, along with an illustrative example of a forward and reverse process: the compression and expansion of a fluid, respectively [62–64].

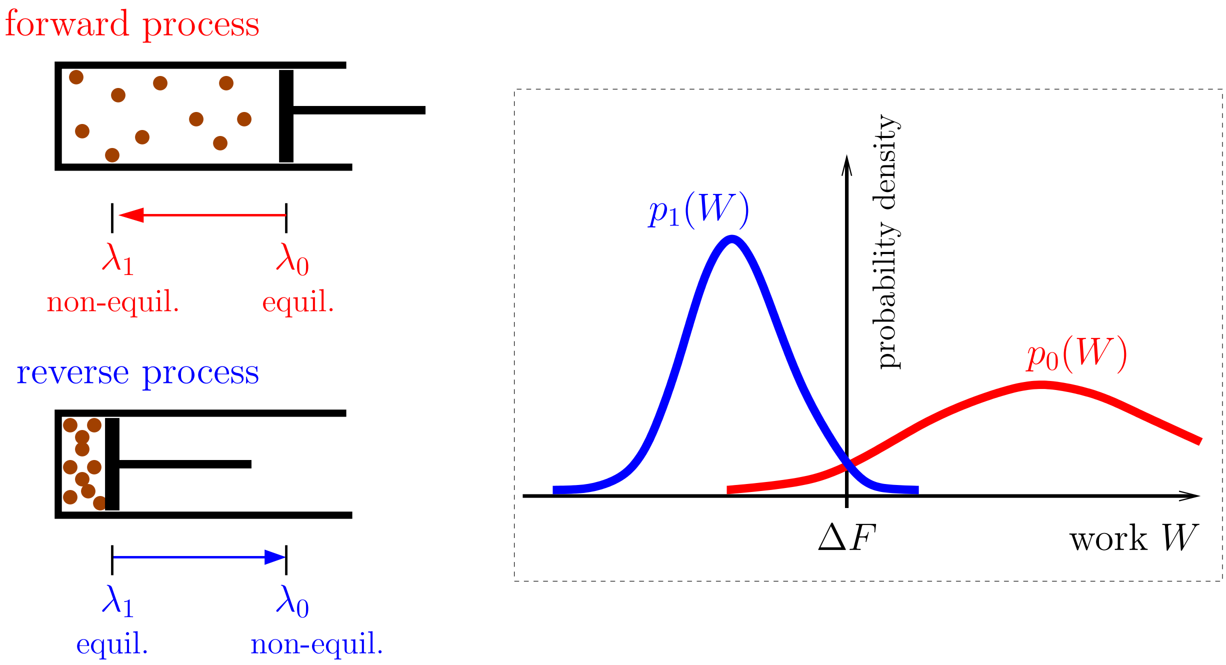

FIG. 2: Illustration of fluctuation theorem. *Left:* Example for a process: compression of fluid by moving a piston. The forward process starts with the fluid in thermal equilibrium with the surroundings and the piston at position λ_0. Then the piston is moved within finite time to the final position λ_1, resulting in a final non-equilibrium state. The reverse process starts likewise in thermal equilibrium, but with initial position of the piston at λ_1. Subsequently the piston is traced back to λ_0 in the opposite manner of the forward process. *Right:* Schematic forward and reverse work densities $p_0(W)$ and $p_1(W)$, respectively, which intersect at the equilibrium free energy difference ΔF. As a consequence of the fluctuation theorem, the average forward work $\langle W \rangle_0$ is larger than ΔF, while the average reverse work $\langle W \rangle_1$ is smaller than ΔF, in concord with the second law of thermodynamics.

Immediate consequence of the fluctuation theorem is the *Jarzynski Equation* [24]

$$\int e^{-\beta W} p_0(W)\, dW = e^{-\beta \Delta F}. \tag{13}$$

It has been derived even before the fluctuation theorem and states that the average of the exponentiated work equals the exponentiated free energy difference of the thermal equilibrium states *associated* with the final and initial value of λ at equal temperature $T = \frac{1}{k\beta}$. Thereby neither the final distribution of microscopic states x needs to be an equilibrium distribution (and will not be such for finite τ), nor the temperature needs to correspond with the system's temperature along the process (it may not be defined at all). The latter difficulty provoked early objections against the Jarzynski relation by Cohen and Mauzerall [65], but is resolved by accepting the temperature to belong to the heat bath, which is coupled to the system during the process [27]. Initially, the system is indeed assumed to be in equilibrium at that temperature, but is then driven out of equilibrium. Again, the identity (13) is independent of the choice of protocol, the speed of state transformation $\dot{\lambda}$ and process duration τ.

We shall close this section with commenting on some thermodynamic implications of the nonequilibrium work theorems. Jarzynski observed that from the equality (13) follows the inequality [24]

$$\langle W \rangle_0 \equiv \int W p_0(W)\, dW \geq \Delta F, \tag{14}$$

as a mere consequence of the convexity of the exponential function and Jensen's inequality. In words: the *average* work is larger than the free energy difference. Inequality (14) turns out to be an equality if and only if the work densities are indistinguishable, $p_0 \equiv p_1$. From the fluctuation theorem, however, it is evident that the latter is possible only if the work densities degenerate to delta-functions at $W = \Delta F$ where any randomness of work disappears. This can be expected to be case when the duration of process is stretched to infinity, $\tau \to \infty$, i.e. in the limit of quasi-static process [27, 66]. On the other hand, the second law of thermodynamics, applied to a process in an isothermal environment, asserts the thermodynamic work W_{td} to be larger than the free energy difference, $W_{td} \geq \Delta F$, with equality if and only if the state-transformation is carried out reversible [16].

That is, $\langle W \rangle_0$ and W_{td} behave in full analogy, which can be taken to justify their identification. A thorough thermodynamic analysis of Parrondo, Cleuren, and van den Broeck [12] has shown that the dissipated work

$$W_{diss} = \langle W \rangle_0 - \Delta F \geq 0 \tag{15}$$

can be understood as the entropy change of system plus reservoir, if relaxation of the system to the equilibrium state $(T, \lambda = 1)$ subsequent to the process is allowed (the analysis in [12] assumes that the process is carried out without heat exchange (weak coupling limit), followed by thermal relaxation; an analogous analysis with heat exchange during the process leads to the same result).

C. Nonequilibrium free energy calculations

The nonequilibrium work theorems allow for calculations of free energy differences by means of measuring a number of work-values of a nonequilibrium process, and evaluating suitable averages of them [67]. For example, the Jarzynski Equation (13) implies to take the sample average of the exponentiated work. This aspect of practical relevance has already been emphasized in the original works of Jarzynski [24] and Crooks [30]. As an essentially new tool, it overcomes former limitations of experimental determination of free energy, which was restricted to quasi-static processes (or near-equilibrium transformations [68]). By definition, such processes are long-lasting, and can even be difficult to realize for nano-scale systems [69], which are currently of strong interest (e.g. pulling a single molecule with an optical trap [70]). In fact, the nonequilibrium work theorems develop their full potential for small systems, only: for large systems, the work densities can be argued to be heavily peaked at $W \approx \langle W \rangle$, such that large fluctuations are "invisible", i.e. quite unlikely to be observed [24]. And exactly the large deviations are needed for nonequilibrium free energy calculations.

Besides in "real-world" experiments, work can also be simulated in computer experiments based on models of the physical process [19]. This requires simulation of phase space trajectories and accumulation of work along these trajectories according to definition (11). Therefore, the nonequilibrium work theorems considerably enriched the facilities of computational free energy determination, too. Thereby, the new nonequilibrium techniques appear as natural generalizations of traditional standard methods like free energy perturbation [20, 23] or the acceptance ratio method [21]. As major advances, the nonequilibrium methods need no equilibration routines (except for initial configurations), and the dynamic simulation of trajectory can in principle be carried out arbitrarily fast, by freedom of choice of τ.

However, certain inherent difficulties limit the advantage of nonequilibrium methods, whether experimental or computational: the convergence of the free energy calculation depends strongly on measuring a certain class of rarely observed work values, the so-called "rare-events" [71], which make the essential contributions to the free energy calculation. A common feature of the rare events is violation of the second law on the single-trajectory level, and the probability of observing such trajectories rapidly decreases with growing

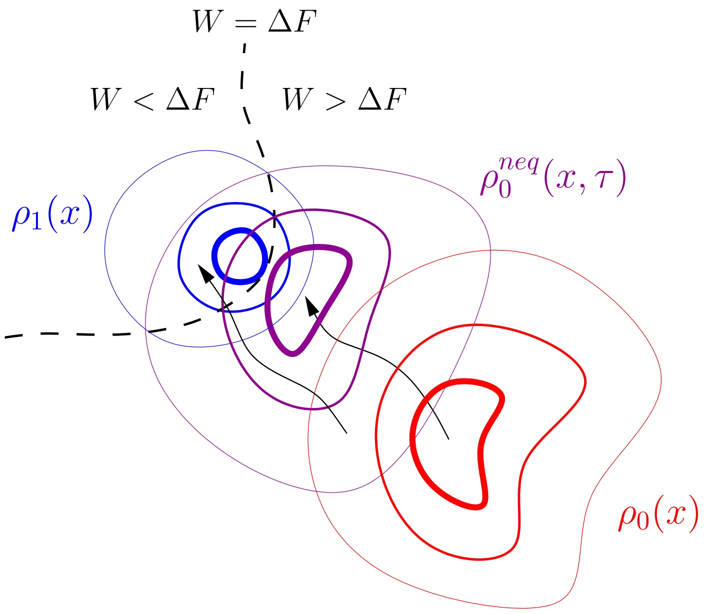

FIG. 3: Phase space picture of fluctuation theorem for Hamiltonian evolution (schematic, Eq. (19)). The forward process drives the initial equilibrium distribution ρ_0 to a final nonequilibrium distribution $\rho_0^{neq}(x,\tau)$ which overlaps better with the equilibrium density ρ_1. The dashed line is defined by the equation $\rho_0^{neq}(x,\tau) = \rho_1(x)$. Trajectories which *end* on it correspond to work values $W = \Delta F$. Two trajectories are indicated by lines with arrows. That on the right represents a typical event, as it starts in a region of large initial probability ρ_0. Its final point shows that it corresponds to a work value $W > \Delta F$. The other one represents an atypical event with $W < \Delta F$, starting in a region of low initial probability ρ_0. As this trajectory furthermore ends in the region where ρ_1 is large, it is one of the rare events needed for accurate free energy calculations with the Jarzynski Equation.

speed of state-transformations [66]. Therefore, a large number of work-values has to be measured in order to enclose the rare events if the process is carried out fast, i.e. far from equilibrium.

We can understand this situation with the following informal reflection. Assuming a forward process, we need, in order to calculate ΔF, information on the density ρ_1, as we aim to calculate the ratio of the normalizing constants of ρ_1 and ρ_0. The dynamic evolution drives the initial equilibrium density $\rho_0(x)$ to a final nonequilibrium density $\rho_0^{neq}(x,\tau)$ in the "region" of the equilibrium density ρ_1, so that we can nicely obtain the needed information, as illustrated in figure 3. This is at its best if the process is carried out very slowly, as then the final nonequilibrium density will almost match with ρ_1. On the other hand, if we carry out the process very fast, the phase space points cannot "follow" the dynamics properly, and $\rho_0^{neq}(x,\tau)$ essentially equals $\rho_0(x)$. Only very little trajectories will still explore the region of $\rho_1(x)$, namely those which *already started* near that region. These are the rare events, rare if the overlap of the densities $\rho_0(x)$ and $\rho_1(x)$ is small. (Note that the latter considerations need not always to apply fully. E.g., for a boundary switching process like that of figure 2, $\rho_0^{neq}(x,\tau)$ will deviate considerably from $\rho_0(x)$ for any $\tau > 0$.)

The named problem inspired the development of methods which bias the sampling of trajectories towards the rare events [72–74]. The main ideas and methods rely essentially on traditional importance or umbrella sampling techniques [22], transposed from phase space to trajectory space. In this context, also thermodynamic integration [23] found its generalization to nonequilibrium processes [72]. A brief description of these methods is given in section II.

To see where the similarity of traditional and nonequilibrium methods stems from, we note that Crook's Fluctuation Theorem originates from a relation in trajectory space which is reminiscent to the traditional source relation (6). Denoting with $\mathcal{P}_0[x(\cdot)]$ the probability density of observing a trajectory $x(\cdot)$ in the forward process, and with $\mathcal{P}_1[x(\cdot)]$ the analog for the reverse process, the general form of Crook's Fluctuation Theorem formulated in trajectory (or path) space reads [30, 53]

$$\frac{\mathcal{P}_0[x(\cdot)]}{\mathcal{P}_1[\bar{x}(\cdot)]} = e^{\beta\left(\mathcal{W}_{[\lambda]}[x(\cdot)]-\Delta F\right)}. \tag{16}$$

Thereby, the trajectory $\bar{x}(\cdot)$ denotes the "time-reversed" counterpart to $x(\cdot)$, obtained by tracing back $x(\cdot)$ in phase space with reversed momenta, cf. figure 4. Explicitly, the "conjugate" trajectory $\bar{x}(\cdot)$ is defined by

$$\bar{x}(t) = x^{\dagger}(\tau - t), \tag{17}$$

where the dagger operator means sign-reversal of momenta (more generally, sign-reversal of those generalized coordinates which are odd under time reversal). The reverse work along the trajectory $\bar{x}(\cdot)$ exactly equals the forward work along its conjugate $x(\cdot)$ [30]:

$$-\mathcal{W}_{[\bar{\lambda}]}[\bar{x}(\cdot)] = \mathcal{W}_{[\lambda]}[x(\cdot)] \tag{18}$$

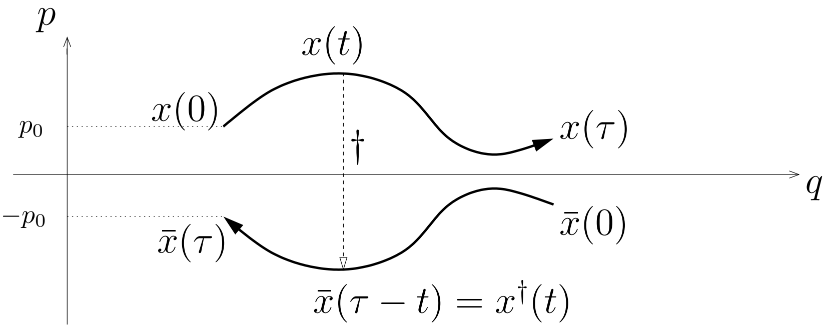

FIG. 4: A pair of conjugate trajectories in a two-dimensional phase space, $x = (q, p) = (\text{position}, \text{momentum})$. $\bar{x}(\cdot)$ is obtained from $x(\cdot)$ by mirroring the latter along the q-axis (dagger operator) and traversing it backwards in time.

(note our sign convention for the reverse work).

Equation (16) is adequate for continuous-time stochastic processes, but formally we can allow it to include also discrete-time stochastic evolution [30] and deterministic evolution [24, 52]. In the latter case, the path probability degenerates to a phase space distribution of initial configurations.

Let us briefly relate (16) to the important case of *Hamiltonian dynamics*. As Hamiltonian evolution is deterministic, we can substitute $\mathcal{P}_0[x(\cdot)]$ with with $\rho_0(x(0))$, and $\mathcal{P}_1[\bar{x}(\cdot)]$ with $\rho_1(\bar{x}(0))$. Further, by Liouville's theorem [7] we have $\rho_0(x(0)) = \rho_0^{neq}(x(\tau), \tau)$, and from the time-reversal invariance of the Hamiltonian together with definition (17) follows $\rho_1(\bar{x}(0)) = \rho_1(x^\dagger(\tau)) = \rho_1(x(\tau))$. Hence, the Fluctuation Theorem (16) for Hamiltonian dynamics reads

$$\frac{\rho_0^{neq}(x(\tau), \tau)}{\rho_1(x(\tau))} = e^{\beta\left(\mathcal{W}_{[\lambda]}[x(\cdot)] - \Delta F\right)}. \tag{19}$$

$x(\tau) = x(\tau; x(0))$ is understood here as the dynamical image of $x(0)$. Finally, the work functional can be simplified to $\mathcal{W}_{[\lambda]}[x(\cdot)] = H_1(x(\tau)) - H_0(x(0))$ [24]. Relation (19) is in perfect analogy to the traditional relation (6), but with the nonequilibrium distribution $\rho_0^{neq}(x, \tau)$ taking the place of the equilibrium distribution $\rho_0(x)$. Figure 3 gives an account of some implications of the Hamiltonian fluctuation theorem (19). Consequences of Eq. (19) with regard to dissipation and coarse graining are discussed in [50] and [49].

In the limiting case of an infinitely fast process, $\tau \to 0$, the "source-relation" of nonequilibrium methods (16) can formally be viewed to go over to the "source-relation" of equilibrium methods (6), and with it the free energy calculation methods relying on it. In detail, this limit means that the trajectory degenerates to a single point in phase space, $x(t) = const. = x$, hence $\mathcal{P}_0[x(\cdot)] \to \rho_0(x(0)) = \rho_0(x)$ (as initial configurations are drawn from ρ_0), $\mathcal{P}_1[\bar{x}(\cdot)] \to \rho_1(\bar{x}(0)) = \rho_1(x(\tau)) = \rho_1(x)$ (as initial configurations are drawn from $\rho_1(x)$ and by time-reversal invariance of the Hamiltonian), and finally $\mathcal{W}_{[\lambda]}[x(\cdot)] \to \frac{H_1(x(\tau))-H_0(x(0))}{\Delta\lambda}\,\Delta\lambda = \Delta H(x)$ (Eq.(5)).

It may seem paradoxical that nonequilibrium methods go over to equilibrium methods in the limit of highest nonequilibrium, $\tau \to 0$. In fact, from this point of view one should call the traditional methods "super-nonequilibrium" methods.

D. Mapping methods

A further conceptual new and promising method for free energy calculations has been introduced by Jarzynski with *targeted free energy perturbation* [75], which extends free energy perturbation by incorporation of bijective phase space maps. Astonishingly, the same idea has been established simultaneously and independently within mathematical statistics by Meng and Schilling [76], who called it "warp sampling". There it is used for the equivalent problem of estimating ratios of normalizing constants (or likelihood ratios). Similar to umbrella sampling, the idea behind targeted free energy perturbation is to use modified sampling-distributions in order to get access to the rare events. But instead of sampling directly from biased distributions, the targeted approach samples from the unbiased equilibrium distributions, and maps ordinary events to rare events.

This has far-reaching consequences: targeted free energy perturbation can, in principle, achieve immediate convergence of the calculation ("one throw of coin"). Immediate convergence is achieved if the map transforms the equilibrium densities $\rho_0(x)$ and $\rho_1(x)$ into each another, just like a reversible isothermal process would do. However, from this it becomes also clear that the arrangement of such an ideal map can be expected to be as difficult as equating the partition function itself. Maybe for this reason, targeted free energy perturbation found only little application up to now [77, 78]. Yet, it's successful use does *not* premise the map to be ideal, rather it suffices to go some way into this direction. Nevertheless, this also requires some degree of insight into the phase space landscape of the physical problem at hand and can hardly be automated in a "black-box" fashion. But also any other free energy method needs a portion of insight to guarantee an appropriate design of simulations, e.g. to achieve ergodic sampling. Targeted free energy calculations allow for analytic inclusion of all we know on the system *a priori*, but also of what we learn *a posteriori* within the simulations in a highly effective manner.

1. Fluctuation theorem of generalized work

Originally formulated for free energy perturbation, we have extended the method to the "targeted acceptance ratio method" by deriving a fluctuation theorem for a generalized notion of work, from which the acceptance ratio method can easily be invoked. The analysis behind shall briefly be introduced, for the details we refer to [1].

A (piecewise) differentiable, bijective map $\phi(x)$ of phase space,

$$x \longrightarrow \phi(x), \tag{20}$$

induces a mapped image $\widetilde{\rho_0}(x)$ of the canonical density $\rho_0(x)$,

$$\rho_0 \overset{\phi}{\rightarrow} \widetilde{\rho_0}, \tag{21}$$

which is related to its preimage by

$$\rho_0(x) = \widetilde{\rho_0}(\phi(x)) \; J(x). \tag{22}$$

Thereby $J(x)$ denotes the absolute of the map's Jacobian determinant,

$$J(x) = | \left| \frac{\partial \phi(x)}{\partial x} \right| |. \tag{23}$$

Then the following identity holds [1]:

$$\frac{\widetilde{\rho_0}(\phi(x))}{\rho_1(\phi(x))} = e^{\beta \left(\widetilde{\Delta H}(x) - \Delta F \right)}. \tag{24}$$

Here, the function of "generalized work" $\widetilde{\Delta H}(x)$ is introduced,

$$\widetilde{\Delta H}(x) := H_1(\phi(x)) - H_0(x) - \frac{1}{\beta} \ln J(x). \tag{25}$$

Relation (24) is in formal analogy with the traditional source relation (6), the latter being a special case for $\phi(x) = x$, but also with the trajectory formulation of Crook's Fluctuation Theorem (16). The similarity can even be enhanced, by noting that

$$\frac{\widetilde{\rho_0}(\phi(x))}{\rho_1(\phi(x))} \equiv \frac{\rho_0(x)}{\widetilde{\rho_1}(x)}, \tag{26}$$

where $\widetilde{\rho_1}$ is the mapped image of ρ_1 under the inverse map $\phi^{-1}(x)$:

$$\rho_1(x) = \widetilde{\rho_1}(\phi^{-1}(x)) \; J(\phi^{-1}(x))^{-1}. \tag{27}$$

Finally, from (24) we arrive at the generalized work fluctuation theorem [1]

$$\frac{\widetilde{p_0}(W)}{\widetilde{p_1}(W)} = e^{\beta(W - \Delta F)} \tag{28}$$

by adequately defining the "forward" and "reverse" densities of generalized work $\widetilde{p_0}(W)$ and $\widetilde{p_1}(W)$ through

$$\widetilde{p_0}(W) := \int \delta\left(\widetilde{\Delta H}(x) - W\right) \rho_0(x) \; dx, \tag{29}$$

$$\widetilde{p_1}(W) := \int \delta\left(\widetilde{\Delta H}(\phi^{-1}(x)) - W\right) \rho_1(x) \; dx. \tag{30}$$

From the generalized work fluctuation theorem follow analogies of the well known free energy estimators in new versions with maps, in specific the acceptance ratio method. The definitions (29) and (30) of "work"-densities implicitly also show how *samples thereof* are obtained: simply by drawing phase space points x from the equilibrium densities $\rho_0(x)$ and $\rho_1(x)$, and evaluating $\widetilde{\Delta H}(x)$ and $\widetilde{\Delta H}(\phi^{-1}(x))$, respectively. Figure 5 illustrates the

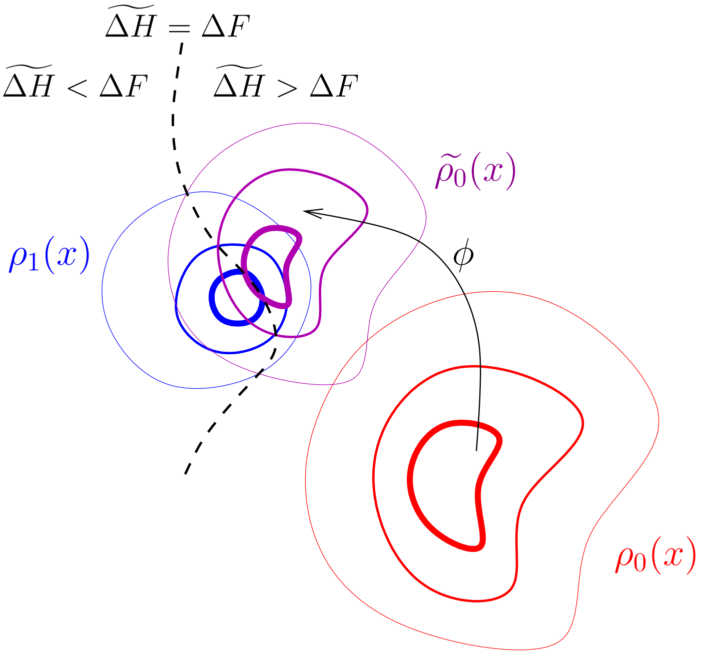
$\widetilde{\Delta H} = \Delta F$
$\widetilde{\Delta H} < \Delta F$
$\widetilde{\Delta H} > \Delta F$
$\widetilde{\rho}_0(x)$
$\rho_1(x)$
ϕ
$\rho_0(x)$

FIG. 5: Phase space picture of generalized work fluctuation theorem (schematic, Eq. (24)). A differentiable one-to-one map $\phi(x)$ of phase space onto itself causes a mapping of the distribution ρ_0 to a distribution $\widetilde{\rho}_0$. $\widetilde{\rho}_0$ and ρ_1 obey a fluctuation theorem for generalized work $\widetilde{\Delta H}$. Targeted free energy perturbation indirectly draws from $\widetilde{\rho}_0$, by drawing from the equilibrium distribution $\rho_0(x)$ and applying the map $\phi(x)$ to the phase space points drawn (arrow). The dashed line is defined by $\widetilde{\rho}_0(x) = \rho_1(x)$. Points x which are mapped onto the dashed line correspond to (forward) work values $\widetilde{\Delta H}(x) = \Delta F$. An ideal map is such, that $\widetilde{\rho}_0(x) = \rho_1(x)$ holds for all x.

generalized work fluctuation theorem from the phase space perspective.

We note that the mentioned work of Meng and Schilling [76], which came to our attention only after publishing [1], has already developed the acceptance ratio method including maps (but without formulating the fluctuation theorem (28) or (24)). Further, Zuckerman [79] called to our attention that he and Ytreberg had already used the acceptance ratio method including a constant translation in phase space [80], based on an old idea of Voter [81]. In our notation, this corresponds to the specific choice $\phi(x) = x + const.$ for the map.

2. Relation to Crook's Fluctuation Theorem

The generalized work fluctuation theorem (28) includes the Crooks Fluctuation Theorem (12) for certain types of deterministic dynamics, which is the reason for the notion "generalized work". A prerequisite for this is time reversal invariance of the Hamiltonian $H_\lambda(x)$. To see in which sense Crook's Fluctuation Theorem is contained, we note that e.g. Hamiltonian dynamics generates a phase space flow which establishes for each instance of time a one-to-one correspondence between initial configurations and their images under the dynamics. If we identify ϕ with the dynamical image of the initial configuration x_0 at time τ, $\phi(x_0) \equiv x(\tau; x_0)$, the generalized work $\widetilde{\Delta H}(x_0)$ equals the physical work $\mathcal{W}_{[\lambda]}[x(\cdot)] = H_1(x(\tau; x_0)) - H(x_0) = H_1(\phi(x_0)) - H_0(x_0) = \widetilde{\Delta H}(x_0)$, as the Jacobian $J(x)$ equals unity for Hamiltonian dynamics. Further, the mapped density equals the final nonequilibrium density of the Hamiltonian forward process, $\widetilde{\rho}_0(x) \equiv \rho_0^{neq}(x, \tau)$. Thus, Eq. (24) becomes the phase space representation of the fluctuation theorem (19) for Hamiltonian evolution, and therefore too, the generalized work fluctuation theorem (28) the Crooks Fluctuation Theorem (12).

In a similar manner it can also be shown that fluctuation theorems for deterministic thermostatted dynamics are included (e.g. Noisé-Hoover dynamics). In this case, the logarithm of the Jacobian appearing in the definition of the generalized work (25) can be interpreted as the heat-supply along the trajectory. The latter is worked out in [52]. See also [31].

Concerning stochastic evolution, Lechner *et. al.* [77] have pointed out that the Langevin equation with *fixed history of noise* can be regarded as deterministic, resulting in a bijective map for each realization of noise history. However, in this case the logarithm of the Jacobian can not be identified with heat [77], and consequently the generalized work does not equal the physical work.

3. Construction of suitable maps

We now come to the question of how to choose a map for the purpose of free energy calculations. Equation (24) shows that if $\widetilde{\rho}_0(x) = \rho_1(x)$, i.e. if the map is such, that ρ_1 is the mapped image of ρ_0, then $\widetilde{\Delta H}(x) = const. = \Delta F$. This is the *ideal* case, as then each measured "work"-value already yields the free energy difference.

As a simple example, think of an n-dimensional harmonic oscillator with $H_0(x) = \frac{1}{2}k_0^2x^2$ and $H_1(x) = \frac{1}{2}k_1^2(x-c)^2$, where x is, for now, an n-dimensional spatial coordinate (without momenta), c, k_0 and k_1 constants. Choosing the map $\phi(x) = \frac{k_0}{k_1}x + c$, we have $\widetilde{\Delta H}(x) = -\frac{1}{\beta}\frac{k_0}{k_1} = const.$ Therefore, this map is ideal for the present example, and we know $\Delta F = -\frac{1}{\beta}\ln\frac{k_0}{k_1}$, without needing to evaluate n-dimensional integrals (yet the same is concluded by carrying out the variable transform $x \to \phi(x)$ in $Z_1 = \int e^{-\beta H_1(x)}dx$).

For most problems, however, it will not be possible to find an ideal map. But their property of mapping ρ_0 to ρ_1 is the general guideline for construction of appropriate maps. What this actually means will depend to a large extend on the specific systems treated. We have studied the construction of maps for the purpose of calculating the chemical potential of a high-density Lennard-Jones fluid. Figure 6 gives a comprehensive account of the essence of our approach (details can be found in [1] on pp. 10-12). The type of map which proved to be useful for this problem will also be applicable in similar problems. For other problems, however, prototypes of maps have to be developed first.

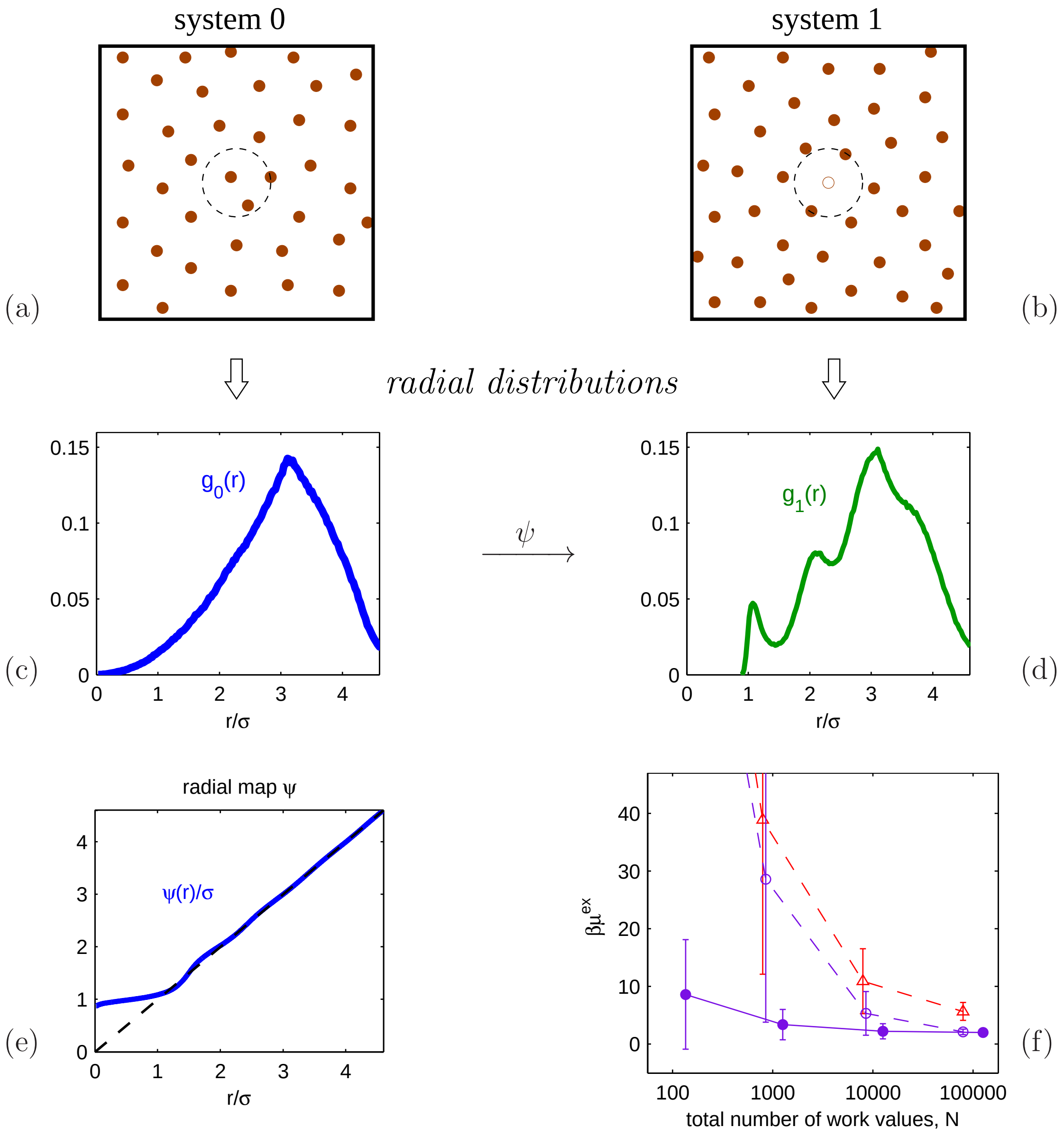

FIG. 6: Calculation of the chemical potential of a high density fluid with mapping methods [1]. *(a):* System 0, a homogeneous Lennard-Jones fluid of N_p particles (brown circles) confined in a box. *(b):* System 1, like (a), but with an additional interaction potential, due to a “ghost” particle (hollow brown circle) fixed in the center of the box. Thus the fluid is inhomogeneous. The chemical potential μ approximatively equals the free energy difference of these systems [82] and is difficult to calculate with traditional methods, as the overlap of the densities ρ_0 and ρ_1 is very small. This is because in system 1, there is nearly never met a particle within the dashed sphere of radius $\approx \sigma$, due to the strong repulsive part of interaction with the “ghost” particle for small distances (σ denotes the spatial parameter of the Lennard-Jones potential). In contrast, in system 0 we will nearly always find a particle within the same sphere. This fact is visualized in *(c)* and *(d)*, which show simulated probability distributions $g_0(r)$ and $g_1(r)$ of finding a particle in distance r from the center of the box in the respective systems. Consequently, a simple but effective mapping scheme employs a suitable radial map, which shifts each particle separately away from the center of the box. *(e):* A simulated radial map $\psi(r)$ which maps $g_0(r)$ to $g_1(r)$. I.e., ψ is an ideal map of the *radial* distributions of the systems. It’s application to the calculation of the chemical potential is shown in *(f):* Comparison of statistical properties of the targeted acceptance ratio method (solid line, full circles) with traditional equilibrium methods (dashed lines), namely the particle insertion (triangles), and the acceptance ratio method (hollow circles). The application of the radial map results in markedly improved convergence properties of the calculations.

II. ELEMENTARY FREE ENERGY ESTIMATORS

We shall now come to the heart of free energy calculations, the setting up of appropriate free energy estimators. In the last section we have introduced three main classes of free energy methods, namely the equilibrium, nonequilibrium, and mapping methods, which are condensed in the respective three "source-relations", Eqs. (6), (16), and (24). Depending on which class we choose, the methods quite differ with respect to data-gathering. It means sampling from canonical distributions for equilibrium and mapping methods, e.g. by Metropolis Monte Carlo simulations, and sampling from trajectory space for nonequilibrium methods, which is done by numerically solving the equations of motion for random initial conditions (again canonical), possibly with additional random forces in the course of time evolution (e.g. when the underlying dynamics is described by the Langevin equation). But with regard to *how* free energy is finally calculated, all classes yield the same types of (optimal) elementary estimators - in terms of work or generalized work. The reason for this lies in the similarity of the named source-relations, and the fact that they implicitly show that optimal estimators will be in terms of "work", whether using the acceptance ratio method or importance sampling (thermodynamic integration, however, is of other nature). The essence of reduction to the random variable "work" are the fluctuation theorems (12) and (28), respectively (the latter including the fluctuation theorem for equilibrium methods), which are of the same form.

We will not prove these statements in detail (there is probably nothing new in it), but rather use this point of view to discuss the basic free energy estimators in a unified way, based on a *formal* fluctuation theorem

$$\frac{p_0(w)}{p_1(w)} = e^{w-\Delta f}. \tag{31}$$

Its specific meaning is only obtained by referring it to the process of raw-data gathering and the work function. Table I summarizes how these connections are established.

TABLE I: Overview of source relations of free energy methods which lead to the fluctuation theorem (31), together with prescriptions of how to calculate work values in forward and reverse direction with the respective raw data. "$\sim$" means here "distributed according to". We note that the nonequilibrium reverse work can also be written $w = +\beta\mathcal{W}_{[\lambda]}[\bar{x}(\cdot)]$.

	equilibrium methods	nonequilibrium methods	mapping methods
source relation	$\frac{\rho_0(x)}{\rho_1(x)} = e^{\beta(\Delta H(x)-\Delta F)}$	$\frac{\mathcal{P}_0[x(\cdot)]}{\mathcal{P}_1[\bar{x}(\cdot)]} = e^{\beta\left(\mathcal{W}_{[\lambda]}[x(\cdot)]-\Delta F\right)}$	$\frac{\widetilde{\rho_0}(\phi(x))}{\rho_1(\phi(x))} = e^{\beta\left(\widetilde{\Delta H}(x)-\Delta F\right)}$
work function	$\Delta H(x) =$ $H_1(x) - H_0(x)$	$\mathcal{W}_{[\lambda]}[x(\cdot)] =$ $\int\limits_0^\tau \frac{\partial}{\partial\lambda} H_{\lambda(t)}(x(t))\dot{\lambda}(t)dt$	$\widetilde{\Delta H}(x) =$ $H_1(\phi(x)) - H_0(x) - \frac{1}{\beta}\ln\lvert\frac{\partial\phi}{\partial x}\rvert$
to choose	-	$\tau,\ \lambda(\cdot)$	$\phi(x)$
forward work	$w = \beta\Delta H(x)$	$w = \beta\mathcal{W}_{[\lambda]}[x(\cdot)]$	$w = \beta\widetilde{\Delta H}(x)$
raw data	$x \sim \rho_0(x)$	$x(\cdot) \sim \mathcal{P}_0[x(\cdot)]$	$x \sim \rho_0(x)$
reverse work	$w = \beta\Delta H(x)$	$w = -\beta\mathcal{W}_{[\bar{\lambda}]}[x(\cdot)]$	$w = \beta\widetilde{\Delta H}(\phi^{-1}(x))$
raw data	$x \sim \rho_1(x)$	$x(\cdot) \sim \mathcal{P}_1[x(\cdot)]$	$x \sim \rho_1(x)$
$p_0(w) =$	$\int\delta(\beta\Delta H(x) - w)\cdot$ $\cdot\rho_0(x)dx$	$\int\delta(\beta\mathcal{W}_{[\lambda]}[x(\cdot)] - w)\cdot$ $\cdot\mathcal{P}_0[x(\cdot)]\mathcal{D}x(\cdot)$	$\int\delta(\beta\widetilde{\Delta H}(x) - w)\cdot$ $\cdot\rho_0(x)dx$
$p_1(w) =$	$\int\delta(\beta\Delta H(x) - w)\cdot$ $\cdot\rho_1(x)dx$	$\int\delta(\beta\mathcal{W}_{[\bar{\lambda}]}[x(\cdot)] + w)\cdot$ $\cdot\mathcal{P}_1[x(\cdot)]\mathcal{D}x(\cdot)$	$\int\delta(\beta\widetilde{\Delta H}(\phi^{-1}(x)) - w)\cdot$ $\cdot\rho_1(x)dx$

To lighten the notation, we go over to express work and free energy in units of the thermal energy $1/\beta$, denoted by lowercase letters:

$$\begin{aligned} w &:= \beta W, \\ \Delta f &:= \beta \Delta F, \\ f_\lambda &:= \beta F_\lambda \end{aligned} \tag{32}$$

The work densities $p_i(w)$, $i = 0, 1$, are now understood to be the densities for the dimensionless work (i.e. $p_i^{\mathrm{new}}(w) \equiv p_i^{\mathrm{old}}(w/\beta)/\beta$). In advance, we also summarize some further definitions.

The *ensemble average* of an arbitrary function $f(w)$ in the work density $p_i(w)$ is abbreviated by angular brackets with subscript i:

$$\langle f(w) \rangle_i := \int f(w) p_i(w) \; dw. \tag{33}$$

In contrast, its *sample average* with a sample $\{w_k\} = \{w_1, \ldots, w_N\}$ of N work values drawn "from" the density $p_i(w)$ is written

$$\overline{f(w)}^{(i)} := \frac{1}{N} \sum_{k=1}^{N} f(w_k). \tag{34}$$

Finally, we define the variance operator

$$\mathrm{Var}_i\big(f(w)\big) := \left\langle f(w)^2 \right\rangle_i - \langle f(w) \rangle_i^2 \,. \tag{35}$$

A. One-sided estimation (free energy perturbation)

As simplest integral consequence of the fluctuation theorem (31), Δf can be expressed through an average in p_0,

$$\Delta f = -\ln \left\langle e^{-w} \right\rangle_0 , \tag{36}$$

which implies the definition of a free energy estimator $\widehat{\Delta f}_0$ by

$$\widehat{\Delta f}_0 = -\ln \overline{e^{-w}}^{(0)} . \tag{37}$$

Depending on how work is actually obtained, this estimator is equivalent with free energy perturbation [20], see Eq. (8), the Jarzynski estimator [24], or targeted free energy perturbation [75]. We will refer to $\widehat{\Delta f}_0$ with *one-sided* (forward) free energy estimator.

The practical applicability of one-sided estimation is considerably limited by the amount of *overlap* (or the *distance*) of the densities $p_0(w)$ and $p_1(w)$. This has been worked out in some detail by Lu, Wu, and Kofke in a series of papers [83–88]. The limitations come from the nonlinear, exponential average involved in (37), which is highly sensitive to the lowest observed work values. Considerable efforts have been done by Zuckerman and Woolf to understand the behavior of one-sided estimation on a firm analytic basis [89–92].

A precise estimate of Δf according to (37) requires that the sample size N is large enough to ensure that the left tail of $p_0(w)$ is sampled *accurately*, i.e. properly according to the statistical weights prescribed by $p_0(w)$, and further *stable* with respect to repetitions of drawing samples of the same size N. This applies fortunately not to the total of the left tail, but up to a certain region within it. The characteristic of this region is that $p_1(w)$ has its main probability mass therein [86, 93], cf. figure 7. Qualitatively, this can be seen by noting that according to the fluctuation theorem, the properly weighted exponential of work is proportional to the reverse work density: $e^{-w}p_0(w)dw \sim p_1(w)dw$. We will call work-values from that region "rare-events" [71]. Therefore, one-sided estimation is viable only, if we are capable to sample the main region of $p_1(w)$, *whilst drawing from*

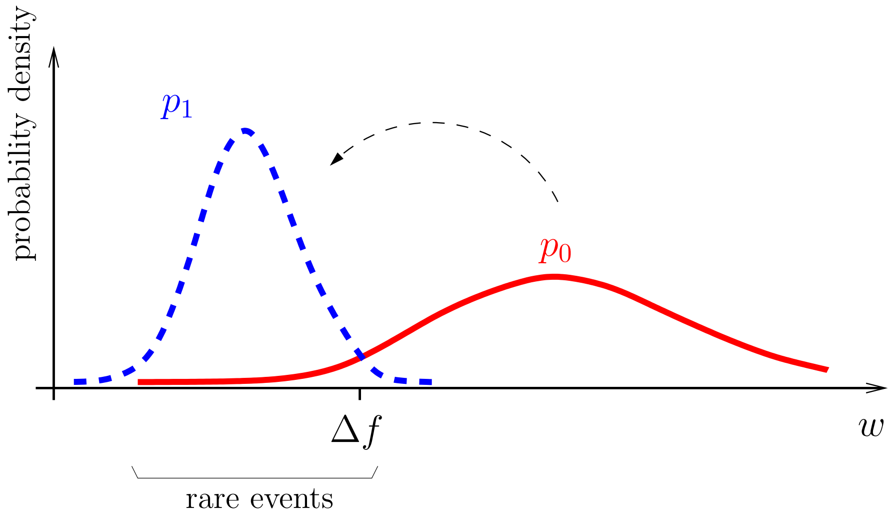

FIG. 7: Illustration of one-sided (forward) free energy estimation. Work is drawn from the forward density $p_0(w)$, but for a precise free energy estimate we need information on the reverse density $p_1(w)$ by sampling work values from the region where p_1 has its main probability mass. This region defines the “rare-events” of one-sided (forward) estimation.

$p_0(w)$. If the overlap of the work densities is too small, unattainable large sample sizes N are needed to sample the named region, resulting in strongly biased free energy estimates $\widehat{\Delta f}_0$. This problem inspired the recent development of methods which determine the asymptotic tails of work distributions [94].

Quantitatively, the performance of one-sided estimation is regulated by a certain measure of distance between the work densities, namely a chi-square distance. It naturally appears as proportionality factor of the asymptotic mean square error $X_0(N)$ of the estimator $\widehat{\Delta f}_0$. The latter reads [90, 92]

$$X_0(N) := \lim_{N\to\infty} \left\langle \left(\widehat{\Delta f}_0 - \Delta f \right)^2 \right\rangle = \frac{1}{N} \operatorname{Var}_0 \left(e^{-w+\Delta f} \right). \tag{38}$$

(this is the leading behavior of a power series in $\frac{1}{N}$, provided $\operatorname{Var}_0 \left(e^{-w+\Delta f} \right)$ exists [90]). By application of the fluctuation theorem (31), we see that the variance on the right-hand side is just the chi-square distance $\chi^2[p_1|p_0]$ of the work-densities, defined by

$$\chi^2[p_1|p_0] := \int \frac{\left(p_1(w) - p_0(w)\right)^2}{p_0(w)} \, dw \geq 0. \tag{39}$$

Thus

$$X_0(N) = \frac{1}{N}\chi^2[p_1|p_0]. \tag{40}$$

The dependence of the mean square error on the chi-square distance shows clearly that the performance of one-sided estimation is highly sensitive to the extend to which the left tail of p_0 reaches into p_1: because if $p_0(w)$ is small whenever $p_1(w)$ is large, then $\chi^2[p_1|p_0]$ attains very large values. It can even diverge – for example, this happens when the forward work is defined by a process which consists of decreasing instantaneously the frequency of a 2-dimensional harmonic oscillator, see [2] and [95]. Finally, we note an important relation to the Kullback-Leibler divergence $KL[p_1|p_0]$ [96], defined by

$$KL[p_1|p_0] := \int p_1(w) \ln \frac{p_1(w)}{p_0(w)} dw \geq 0. \tag{41}$$

In [1] we have shown the following inequality to hold:

$$\chi^2[p_1|p_0] \geq e^{KL[p_1|p_0]} - 1, \tag{42}$$

with equality if and only if $p_1 \equiv p_0$ ((42) holds for any pair of densities and does not rely on the fluctuation theorem). Thus, the mean square error X_0 is bounded from below with the exponentiated Kullback-Leibler divergence of the work densities. This has an interesting physical aspect by means of the latter's well-known relation to the dissipated work *in reverse direction*:

$$KL[p_1|p_0] = \Delta f - \langle w \rangle_1 . \tag{43}$$

(Note our sign convention for the reverse work.) This connection may seem puzzling, as $\widehat{\Delta f}_0$ uses work values of the forward process, only. But formally, the appearance of the reverse instead of the forward dissipation $\langle w \rangle_0 - \Delta f = KL[p_0|p_1]$ is evident from the properties of the Kullback-Leibler divergence. $KL[p_1|p_0]$, and not $KL[p_0|p_1]$, behaves qualitatively like $\chi^2[p_1|p_0]$: from (41) we see that $KL[p_1|p_0]$ is likewise sensitive to the extend to which p_0 reaches into p_1. In specific, it attains large values if $p_0(w)$ is small whenever $p_1(w)$ is large.

The strong dependence of the performance of one-sided estimation from the dissipated work in reverse direction was already noted e.g. by Jarzynski [71], who estimated the number N^* of work-measurements needed for a converging estimate $\widehat{\Delta f}_0$ with $N^* \approx e^{\Delta f - \langle w \rangle_1}$. With aid of inequality (42), we arrived from another direction at the sharpened statement [1]

$$N^* > e^{\Delta f - \langle w \rangle_1} - 1. \tag{44}$$

Interest on such relations comes not only from theoretical grounds, but also from practical questions, e.g. for having criteria on the preferential direction of process (or "perturbation") [83]. Or when numerically searching for the optimal protocol $\lambda(\cdot)$ which minimizes the error of one-sided estimation, as was done in [97]: minimization of the mean square error (38) was found there to be numerically quite costly. To resolve this, one could think of minimizing the reverse dissipation, instead, which is an active field of current research [98–101].

To accomplish the notions, we note that a second one-sided estimator $\widehat{\Delta f}_1$ in reverse direction exists. It relies on the identity

$$\int e^{+w} p_1(w)\, dw = e^{+\Delta f} \tag{45}$$

and is defined by

$$\widehat{\Delta f}_1 = +\ln \overline{\, e^{+w} \,}^{(1)}. \tag{46}$$

In contrast to the forward estimator $\widehat{\Delta f}_0$, the reverse estimator uses samples of work from the reverse density $p_1(w)$. The formal properties of the reverse estimator are essentially obtained from those of the forward estimator by interchanging the indices 0 and 1.

B. Two-sided estimation (acceptance ratio method)

Instead of estimating Δf with work-values from only one direction, one can also use work values from both directions of process. This leads to *two-sided* estimation. The simplest extension of one-sided estimation to a two-sided method would consist in drawing a pair of work-samples from the densities $p_0(w)$ and $p_1(w)$, then to calculate the corresponding one-sided estimates in each direction separately, and finally to take some average of them. Yet, any such procedure which combines independent free energy estimates will yield a suboptimal result, regardless of how they are averaged (arithmetic, harmonic, exponential, ...). Figuratively spoken, this is because we then neglect that information on the interrelation of p_0 and p_1 which could be obtained from *first* combining the samples and *then* calculating an overall estimate.

The mutual information of the densities is encoded in the fluctuation theorem (31): if we know, e.g., the value of $p_0(w)$ for some w, we can precisely tell which value $p_1(w)$ attains at the same point, given we know Δf. The other way round, given information on the work densities via independent samples from both, we can adjust the value of Δf such, that the fluctuation theorem is empirically optimally satisfied. This is what the acceptance ratio method essentially does, which was derived by Bennett in 1976 [21], and independently once again by Meng and Wong in 1996 in the context of estimation of ratios of normalizing constants [102]. Finally it was observed by Crooks that this method can also be used for free energy estimation with nonequilibrium work data [30].

In order to get a clear notion of the acceptance ratio method, we need to go somewhat into the details of its derivation in the following.

1. *The ansatz and the estimator*

With the fluctuation theorem (31), the identity [21, 30]

$$\left\langle t(w)e^{-w+\Delta f}\right\rangle_0 = \langle t(w)\rangle_1 \tag{47}$$

holds, for any choice of function $t(w)$. Given a pair of samples from the work densities, of size $n_0 > 0$ from $p_0(w)$ and of size $n_1 > 0$ from $p_1(w)$, one can define an estimate $\widehat{\Delta f}^*$

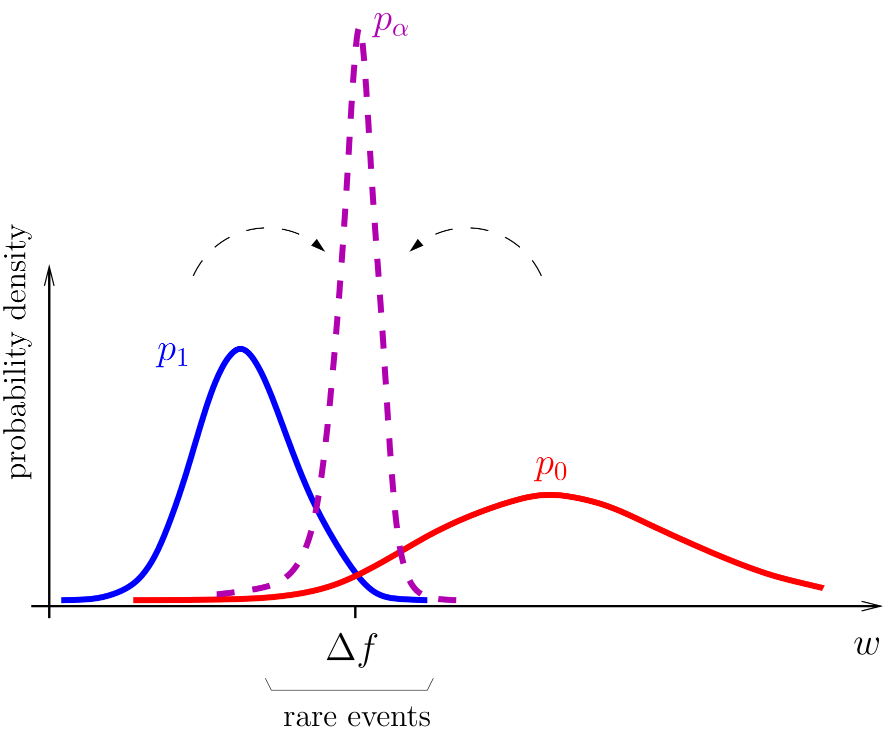

FIG. 8: Two-sided free energy estimation draws work values from both densities, $p_0(w)$ and $p_1(w)$, and then calculates an overall estimate of Δf. A precise estimate requires that the main region of the harmonic overlap density $p_\alpha(w)$ (here schematically for $\alpha \approx \frac{1}{2}$) is sampled well by the forward as well as the reverse draws. The value of α is given as the fraction of the number of forward draws. For $\alpha \to 1$ the overlap density p_α converges to the reverse work density p_1, which is the limiting case of one-sided (forward) estimation.

for any choice of $t(w)$ with

$$\overline{t(w)e^{-w+\widehat{\Delta f}^*}}^{(0)} = \overline{t(w)}^{(1)}, \tag{48}$$

or equivalently

$$\widehat{\Delta f}^* = \ln \frac{\overline{t(w)}^{(1)}}{\overline{t(w)e^{-w}}^{(0)}}. \tag{49}$$

If we choose, e.g., $t(w) = 1$, then this estimator coincides with the one-sided forward estimator $\widehat{\Delta f}_0$ based on n_0 draws, Eq. (37), whilst the information on Δf contained in the reverse sample is left unused. Therefore, the quality of the estimator (49) will depend strongly on the choice of $t(w)$, and we may ask (with Bennett) for the *optimal* choice of $t(w)$.

This requires some measure of performance. A common choice is the mean square error $\left\langle\left(\widehat{\Delta f}^{*}-\Delta f\right)^{2}\right\rangle$, which can be calculated explicitly as a functional of $t(w)$ for *large* sample sizes, $n_0, n_1 \to \infty$. Variational minimization of the asymptotic mean square error with respect to the function $t(w)$ results in a Fermi function [21, 102]:

$$t(w) = \frac{1}{\alpha + \widetilde{\alpha} e^{-w+\Delta f}}. \tag{50}$$

α and $\widetilde{\alpha}$ denote the fraction of forward and reverse number of work values, respectively,

$$\alpha = \frac{n_0}{N}, \quad \widetilde{\alpha} = \frac{n_1}{N} = 1 - \alpha, \tag{51}$$

with N the total sample size,

$$N = n_0 + n_1. \tag{52}$$

The optimal $t(w)$ is not of direct use, as it depends itself on the unknown quantity Δf. However, by freedom of choice on $t(w)$, one could instead use the function

$$t_c(w) := \frac{1}{\alpha + \widetilde{\alpha} e^{-w+c}} \tag{53}$$

for any choice of parameter c. Insertion into Eq. (49) results in a family of estimators $\widehat{\Delta f}^{*} = \widehat{\Delta f}_C$ [21],

$$\widehat{\Delta f}_C := c + \ln \frac{\overline{t_c(w)}^{(1)}}{\overline{t_c(w) e^{-w+c}}^{(0)}}. \tag{54}$$

Yet, as the optimal choice on c is $c = \Delta f$, Bennett proposed to choose c such, that $c = \widehat{\Delta f}_C$ holds. This is tantamount to solving

$$\overline{t_{\widehat{\Delta f}}(w) e^{-w+\widehat{\Delta f}}}^{(0)} = \overline{t_{\widehat{\Delta f}}(w)}^{(1)} \tag{55}$$

for $\widehat{\Delta f}$. Writing the function t_c explicitly, the latter equation reads

$$\overline{\frac{1}{\alpha e^{w-\widehat{\Delta f}} + \widetilde{\alpha}}}^{(0)} = \overline{\frac{1}{\alpha + \widetilde{\alpha} e^{-w+\widehat{\Delta f}}}}^{(1)} \tag{56}$$

Notably, a solution of Eq. (56) always exists and is unique, essentially because $t_c(w)$ is monotonically decreasing, whilst $t_c(w)e^{-w+c}$ is monotonically increasing in c. We will refer to this solution with *two-sided estimate*, and denote it with $\widehat{\Delta f}$. This implicit estimator is commonly referred to as *Bennett's acceptance ratio method* (although usually written in a slightly different form).

The solution of (56) can be obtained iteratively from Eq. (54) by starting with an arbitrary value of c, then calculating $\widehat{\Delta f}_C$ and using it as the value of c in the next step of iteration. This iteration can be shown to converge to the solution $\widehat{\Delta f}$ of (56) [102]. In other words, $\widehat{\Delta f} \equiv \widehat{\Delta f}_{\widehat{\Delta f}}$ is global fixed point of an iterative sequence defined via $\widehat{\Delta f}_C$.

Interestingly, the two-sided estimator $\widehat{\Delta f}$ is *not* within the ansatz (49), as there exists no function $t(w)$ which would lead to Eq. (56) for *any* drawn samples. Or differently expressed: for a given pair of samples, there is such a function, namely $t_c(w)$ with the *value* of c equal to $\widehat{\Delta f}$, and $\widehat{\Delta f}$ again the solution of (56) for that samples (we could by chance have chosen $c = \widehat{\Delta f}$). But if we want this to hold for any pair of samples, we must make c a function of *all* work values, $c = \widehat{\Delta f}\left(\{w_i^{(0)}\}, \{w_k^{(1)}\}; \alpha\right)$, and therefore $t_c(w)$ is no longer a function of a single w, as assumed in the ansatz (49).

Not only for this reason, it is interesting that the two-sided estimator can also be understood as a maximum likelihood estimator, as demonstrated by Shirts *et.al.* [103]. Maximum likelihood estimators for parameters of distributions are highly valuable, as they can be shown to be (asymptotically) optimal in the sense that they reach the Cramér-Rao lower bound on variance, which is a lower bound *for all* asymptotically unbiased estimators relying on the same data [104]. But this does not fully apply to the acceptance ratio method, as we have shown with its re-derivation by constrained maximum likelihood methods (see the Appendix of [1] and remark [105]) under reference to [106, 107]. The conclusion of this re-derivation can be formulated as follows: given we have no further information on the work densities p_0 and p_1 *despite* their relation via the fluctuation theorem (31), the acceptance ratio method is the optimal method for free energy calculations based on a given amount of work data. Maragakis *et.al.* came to essentially the same conclusions with Bayesian considerations [108].

To give a specific example for the difference between the acceptance ratio method and a "real" maximum likelihood estimator, assume the work densities are Gaussian. Then they have to read

$$p_0(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(w-\Delta f-\sigma^2/2\right)^2},$$
$$p_1(w) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(w-\Delta f+\sigma^2/2\right)^2}, \tag{57}$$

to be compatible with the fluctuation theorem. Assuming we have drawn samples from the Gaussian work densities of equal size $n_0 = n_1 = N/2$, the maximum likelihood estimator for Δf is given by

$$\widehat{\Delta f}_{ml} = \frac{1}{2}\left(\overline{w}^{(0)} + \overline{w}^{(1)}\right), \tag{58}$$

which is obtained by maximizing the likelihood of the observed data with respect to Δf, with explicit recourse to the Gaussian form of the work densities. This estimator makes full sense: as the forward and reverse work densities (57) are symmetric to each other with respect to Δf, the average of their mean values equals exactly Δf: $\frac{1}{2}\left(\langle w\rangle_0 + \langle w\rangle_1\right) = \Delta f$. The variance of the estimator (58) is readily equated to $\frac{1}{N}\sigma^2$. A somewhat more complex form for $\widehat{\Delta f}_{ml}$ is found if $n_0 \neq n_1$, involving estimation of the variance σ^2.

We note that Gaussian work densities are not a merely academic example. Rather, Speck and Seifert have proven that for stochastic dynamics the work densities converge to Gaussians in the limit of slow processes [66], $\tau \to \infty$. However, in general we have no information on the detailed dependence of the work densities on Δf. Then, the best we can do with given amount of forward and reverse work data is applying the acceptance ratio method.

2. Rare events of two-sided estimation

The distribution of two-sided estimates $\widehat{\Delta f}$ is asymptotically normal with mean Δf and variance $X(\alpha, N)$, given by [21, 102]

$$X(\alpha, N) = \frac{1}{N} M(\alpha), \tag{59}$$

with proportionality factor

$$M(\alpha) = \frac{1}{\alpha\widetilde{\alpha}} \left(\frac{1}{U_\alpha} - 1 \right). \tag{60}$$

U_α denotes the (harmonic-mean) overlap of densities:

$$U_\alpha = \int \frac{p_0(w) p_1(w)}{\alpha p_0(w) + \widetilde{\alpha} p_1(w)} \, dw \leq 1. \tag{61}$$

From this we see that the magnitude of the overlap U_α regulates the performance of two-sided estimation. Moreover, two-sided estimation is intimately related to the (harmonic mean) overlap density $p_\alpha(w)$ [2, 3, 102],

$$p_\alpha(w) := \frac{1}{U_\alpha} \frac{p_0(w) p_1(w)}{\alpha p_0(w) + \widetilde{\alpha} p_1(w)}. \tag{62}$$

The region where p_α has its main probability mass is the region which defines the rare events of two-sided estimation [3], i.e. the most important but rarely observed contributions to $\widehat{\Delta f}$. These rare events have to be sampled by both, the forward and reverse draws. To see this, we note that U_α can alternatively be written as

$$U_\alpha = \left\langle \frac{1}{\alpha e^{w - \Delta f} + \widetilde{\alpha}} \right\rangle_0 = \left\langle \frac{1}{\alpha + \widetilde{\alpha} e^{-w + \Delta f}} \right\rangle_1. \tag{63}$$

Comparison with Eq. (56) shows that two-sided estimation can essentially be understood as estimation of U_α - by adjusting $\widehat{\Delta f}$ such, that the equality (63) is *empirically* satisfied [3]:

$$\widehat{U}_\alpha := \overline{\frac{1}{\alpha e^{w - \widehat{\Delta f}} + \widetilde{\alpha}}}^{(0)} \overset{!}{=} \overline{\frac{1}{\alpha + \widetilde{\alpha} e^{-w + \widehat{\Delta f}}}}^{(1)}. \tag{64}$$

Therefore, a precise estimate $\widehat{\Delta f}$ (with small mean square error) is equivalent to a precise estimate $\widehat{U}_\alpha$ of the overlap. The latter requires that the region where p_α has its main mass is sampled accurately and stable with the samples of work in each direction, cf. figure 8. This point of view allows for the characterization of the convergence properties of two-sided estimation [3].

In contrast to the rare events of one-sided estimation, those of two-sided' will in general not lie so far in the tails of work distributions, but are placed somewhere "between" $p_0(w)$ and $p_1(w)$. If, for example, the work densities are Gaussian according to (57), then they are symmetric to each other with respect to Δf, $p_0(\Delta f + c) = p_1(\Delta f - c)$, and $p_\alpha(w)$ with $\alpha \approx \frac{1}{2}$ is sharply peaked at $w \approx \Delta f$ ("sharply" compared to p_0 and p_1). A quantitative example is included in figure 9.

Jarzynski [71] pointed out an interesting physical peculiarity of the rare events of one-sided estimation: namely that they correspond to typical trajectories of the reverse process, but observed in time-reversed manner within the forward process – which explains the unlikeliness of observing them out of equilibrium. Concerning the two-sided estimator, we may say that here the rare events correspond to trajectories which are equilibrium-like, in the sense that for these $w \approx \Delta f$ holds, and that they occur with approximate equal probability in either direction of process – as consequence of the fluctuation theorem: $p_0(\Delta f) = p_1(\Delta f)$. To confirm the relation to trajectories, we note that the overlap density $p_\alpha(w)$ is "induced" by its analog $\mathcal{P}_\alpha[x(\cdot)]$ in path space, i.e.

$$p_\alpha(w) = \int \delta\left(\beta \mathcal{W}_{[\lambda]}[x(\cdot)] - w\right) \mathcal{P}_\alpha[x(\cdot)] \; \mathcal{D}x(\cdot), \tag{65}$$

with

$$\mathcal{P}_\alpha[x(\cdot)] := \frac{1}{U_\alpha} \frac{\mathcal{P}_0[x(\cdot)]\mathcal{P}_1[\bar{x}(\cdot)]}{\alpha \mathcal{P}_0[x(\cdot)] + \widetilde{\alpha} \mathcal{P}_1[\bar{x}(\cdot)]}. \tag{66}$$

This is shown by applying the Fluctuation Theorem (16) in the right-hand side of (65).

Hence, two-sided estimation will typically be dominated by trajectories which occur with approximate the same probability in forward and (time-reversed) in backward direction of process, as for these $\mathcal{W}_{[\lambda]}[x(\cdot)] \approx \Delta f$ holds, see Eq. (16). Other interrelations of the acceptance ratio method with the distinguishability of forward and reverse trajectories have been worked out by Feng and Crooks with different methods and focus [109].

3. Convexity of mean square error

The two-sided estimator (56) is optimal for a given amount of forward and reverse work data. On the other hand, its performance depends on the fraction of forward draws $\alpha = n_0/N$, and thus on the partitioning of the total amount of work values $N = n_0 + n_1$ into forward and reverse "measurements". Consequently, supposed we aim to calculate Δf with a limited total number of work values N, a critical question is how to choose α optimally. Is it always $\alpha = \frac{1}{2}$? Or may it also be, e.g., $\alpha = 1$ or $\alpha = 0$? In the latter cases, we would perform one-sided estimation, to which two-sided estimation formally converges in the limits $\alpha \to 0, 1$.

If performance is measured with the asymptotic mean square error $X(\alpha, N) = \frac{1}{N}M(\alpha)$, Eq. (59), the optimal choice of α is such, that it minimizes X for fixed N, and thus the proportionality factor $M(\alpha)$. Figure 9 provides a quantitative example for the function $M(\alpha)$ for Gaussian work densities ($\Delta f = 0,\ \sigma = 4$, Eq. (57)), which shows a very strong dependence of M on α near the boundaries $\alpha = 0, 1$, and a weak one for a wide range of intermediate values of α, with minimum at $\alpha = \frac{1}{2}$ The symmetry of M with respect to $\alpha = \frac{1}{2}$ results from the symmetry of the Gaussian work densities with respect to Δf.

We have investigated the general dependence of two-sided estimation on α and could prove a certain characteristic, namely the strict convexity of the mean square error $X = \frac{1}{N}M(\alpha)$ with respect to α:

$$\frac{\partial^2 M(\alpha)}{\partial \alpha^2} > 0, \tag{67}$$

see theorem and proof in sec. IV of [2]. From this property follow some important statements for the practical application of two-sided estimation [2]:

(i) the optimal value α_o of α is unique;

(ii) for symmetric work-distributions, α_o always equals $\frac{1}{2}$;

(iii) for near symmetric work-distributions, α_o equals approximatively $\frac{1}{2}$;

(iv) two-sided estimation generically outperforms one-sided estimation.

This suggests $\alpha = \frac{1}{2}$ as suitable *a priori* choice, as was already highlighted by Bennett [21]. Nevertheless, it is also possible to simulate the optimal fraction α_o "on the fly"

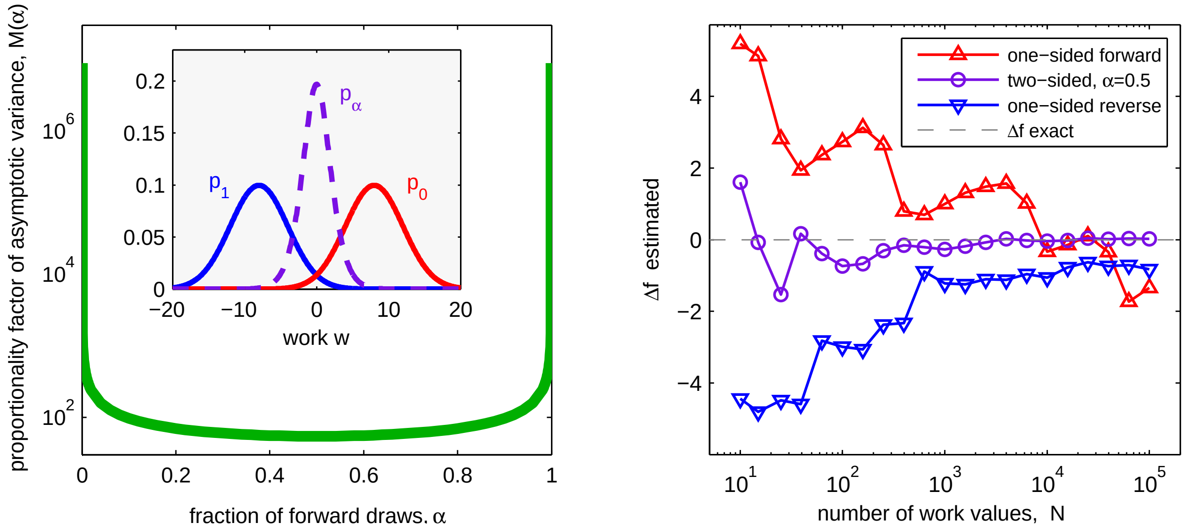

FIG. 9: *Left:* Proportionality factor $M(\alpha)$ of the mean square error of two-sided estimation in dependence of the fraction of forward draws α for Gaussian work densities. The inset displays the corresponding work densities, together with the overlap density $p_{\alpha=\frac{1}{2}}$. *Right:* Examples for running free energy calculations with increasing number of the total sample size N. The one-sided estimations correspond to $M(\alpha=1)$ (up triangles) and $M(\alpha=0)$ (down triangles). Compared with two-sided estimation ($\alpha=\frac{1}{2}$), the one-sided' converge very slowly.

within a simulation run, as will be discussed next.

4. Dynamic sampling strategy

The general equation for the optimal fraction α_o, which follows from the requirement $\frac{\partial M(\alpha)}{\partial \alpha}=0$, is known since Bennett [21]. In our notation it reads

$$\widetilde{\alpha}^2 \,\mathrm{Var}_0\left(\frac{1}{\alpha e^{w-\Delta f}+\widetilde{\alpha}}\right)=\alpha^2\,\mathrm{Var}_1\left(\frac{1}{\alpha+\widetilde{\alpha}e^{-w+\Delta f}}\right), \tag{68}$$

and has to be solved for α (with $\widetilde{\alpha}=1-\alpha$) (Eq. (34) in [2]). The solution can, in principle, be estimated with preliminary samples of work. As this involves estimating the second moments of the Fermi functions, the convergence of this estimate can be expected to be too slow to be of practical relevance [21]. Clearly: estimating α_o within a simulation run of finite time makes sense only if we can obtain reasonable estimates with relatively small sample sizes already, as only then it is possible to adjust the value of α towards α_o through additional draws of work before the simulation time has run out.

However, α_o can also be estimated with first moments of the fermi functions, only [2].

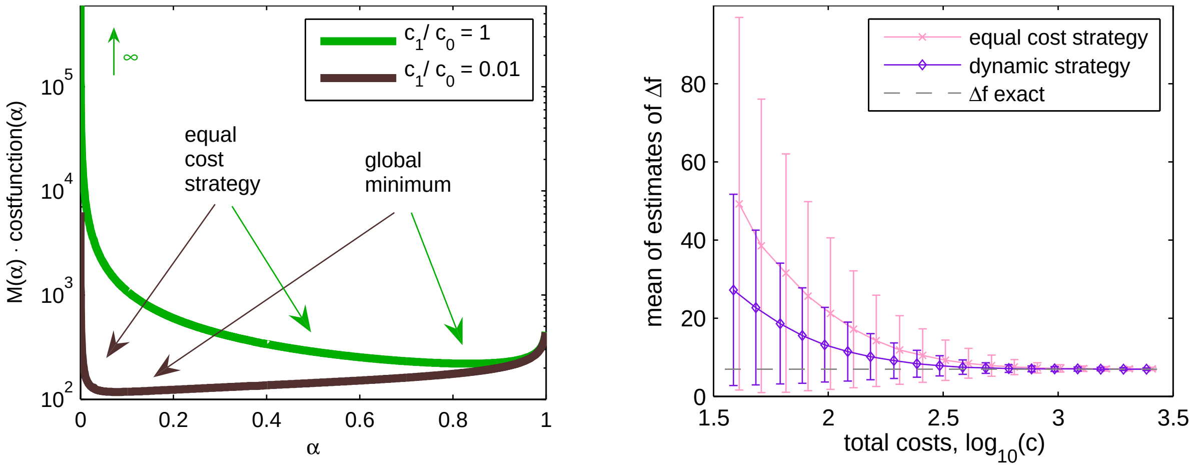

FIG. 10: Optimal fraction of forward draws and dynamic sampling strategy (exponential work densities). *Left:* The global minimum of $M(\alpha)\cdot\mathrm{costfunction}(\alpha)$ determines the optimal fraction α_o of forward draws used in two-sided free energy estimation, taking into account different costs c_0 and c_1 for obtaining a single forward respectively reverse work value. The two graphs shown are obtained with the same function $M(\alpha)$, but with different assumed cost ratios $\frac{c_1}{c_0}$. As the optimal fraction α_o is *a priori* unknown, Bennett [21] suggested to use the equal cost strategy, which chooses α such, that $n_0 c_0 = n_1 c_1$ holds. *Right:* Comparison of free energy estimation with the equal cost strategy and the dynamic sampling strategy, which approaches the optimal fraction α_o iteratively with growing sample size N. Shown is the average $\left\langle \widehat{\Delta f} \right\rangle$ in dependence of the total costs $c = n_0 c_0 + n_1 c_1$ spent, with cost ratio $\frac{c_1}{c_0} = 0.01$. The error bars reflect the standard deviation of $\widehat{\Delta f}$.

The key is an appropriate estimator of the function U_α for the entire range $\alpha \in [0,1]$ (see Eq. (46) in [2]), from which follows an estimate of the function $M(\alpha) = \frac{1}{\alpha\bar{\alpha}}(\frac{1}{U_\alpha} - 1)$ for $\alpha \in [0,1]$. Finally, α_o can then be estimated via the minimum of the estimated function $M(\alpha)$.

Based on this scheme for estimating α_o, we have developed a “dynamic” strategy of sampling work values [2]. This strategy iteratively estimates α_o and draws additional samples such, that the actual fraction α of forward draws approaches the estimated optimal fraction α_o. (To avoid misleading estimates of α_o, the convexity of $M(\alpha)$ enters as reliability criterion.)

More general, we can also take into account different costs c_0 and c_1 of single forward and reverse work values, respectively [2] (the costs can, e.g., be determined by the CPU-time needed for the simulation of a single work value). In this case, the optimal fraction α_o is determined by minimizing the mean square error $X(\alpha, N)$ with respect to α, subject to the constraint of fixed total costs $c = n_0 c_0 + n_1 c_1$. The total number $N = n_0 + n_1$ of draws becomes then a function of α and c,

$$N(c, \alpha) = \frac{c}{\alpha c_0 + \widetilde{\alpha} c_1}. \tag{69}$$

Consequently, $M(\alpha) \cdot \; costfunction(\alpha)$ has to be minimized instead of $M(\alpha)$, where $costfunction(\alpha) = \alpha c_0 + \widetilde{\alpha} c_1 = \frac{c}{N}$ accounts for the average costs per work-measurement. Without loss of generality, we may assume the normalization $c_0 + c_1 = 2$, which has the effect that $costfunction(\alpha) = 1$ and $c = N$ if $\frac{c_1}{c_0} = 1$.

An example for the shift of α_o due to different cost ratios $\frac{c_1}{c_0}$ is given in figure 10 (left). The right panel shows an example for the effect of the dynamic sampling strategy in comparison with Bennett's equal cost strategy [21] (see also figs. 8 and 9 in [2]). The equal cost strategy is the best available "static" sampling strategy. It draws according to a fixed fraction $\alpha = \alpha_{ec}$, where α_{ec} is determined by the requirement of equal total costs for forward and reverse measurements, hence $n_0 c_0 = n_1 c_1$ or $\alpha = \frac{c_1/c_0}{1 + c_1/c_0}$.

5. *Measure of convergence*

For an estimator to be of use, we need to have some quantitative notions of its bias and spread, as otherwise any estimate obtained would be without guarantee, not even within statistical bounds. Analytic statements on the properties of an estimator can in general only be made for its asymptotic behavior in the limit of infinitely large sample sizes N. This is of worth if we can expect to reach this limit at least approximatively for some finite and available N. But the crux is, we then also need some criterion for *whether* we actually have reached this limit in a particular calculation. Typically, this cannot be judged from the data immediately. Especially with regard to the nonlinear one-sided and two-sided free energy estimators, which tend to be plagued by large biases for low N while *seemingly* converging, this question is of great concern [91, 110]. Figure 11 (which is fig. 1 of [3]) illustrates this problem with an example of a running two-sided free energy calculation.

Based on the identification of the rare events of two-sided estimation, we have formulated a measure of convergence for two-sided free energy calculation [3]. In essence, the convergence measure is a test of consistency of the work-data with the assumption of being in the limit of large N. The motivation of the convergence measure from the

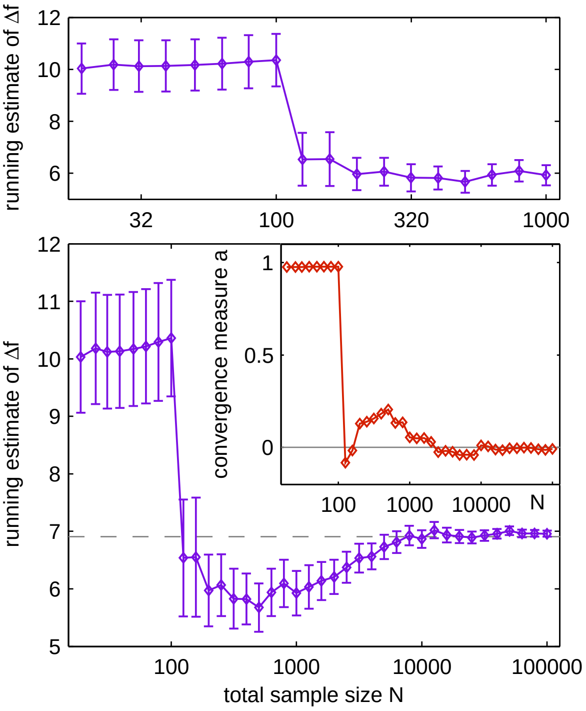

FIG. 11: *Top:* A running free energy calculation with increasing total sample size N up to $N = 10^3$. The free energy estimates $\widehat{\Delta f}$ settle down on a stable plateau with decreasing estimated standard deviation (shown as error bars) over about one order of magnitude of N-values. Has the calculation converged? *Bottom:* Same as before, but with additional samples of work up to $N = 10^5$. The prolonged calculation shows that we had not reached convergence in the calculations before. Now again, the calculation finally settles down on a stable plateau, raising the question of convergence once more. The answer is given by the behavior of the convergence measure a, which is shown in the inset. The fact that a is observed to converge to zero allows us to be (almost) sure that the corresponding free energy calculation has converged, too. As a result, we obtain a free energy estimate with trustable confidence interval (error bar). The exact value of Δf is indicated by the dashed horizontal line in the lower panel.

question of whether the region of rare events has been sampled sufficiently is reported in detail in [3], along with analytic investigations of its properties and numerical tests.

Its definition is as follows. Given samples from both work densities, the two-sided estimate $\widehat{\Delta f}$ according to Eq. (56) is calculated, along with two *dependent* estimates $\widehat{U}_\alpha$ and $\widehat{U}_\alpha^{(II)}$ of the harmonic overlap U_α. The estimate $\widehat{U}_\alpha$ has already been introduced with Eq. (64), for the definition of $\widehat{U}_\alpha^{(II)}$ we refer to [3] (Eq. (19) there).

Finally, the convergence measure a is given as the relative difference

$$a = \frac{\widehat{U}_\alpha - \widehat{U}_\alpha^{(II)}}{\widehat{U}_\alpha}. \tag{70}$$

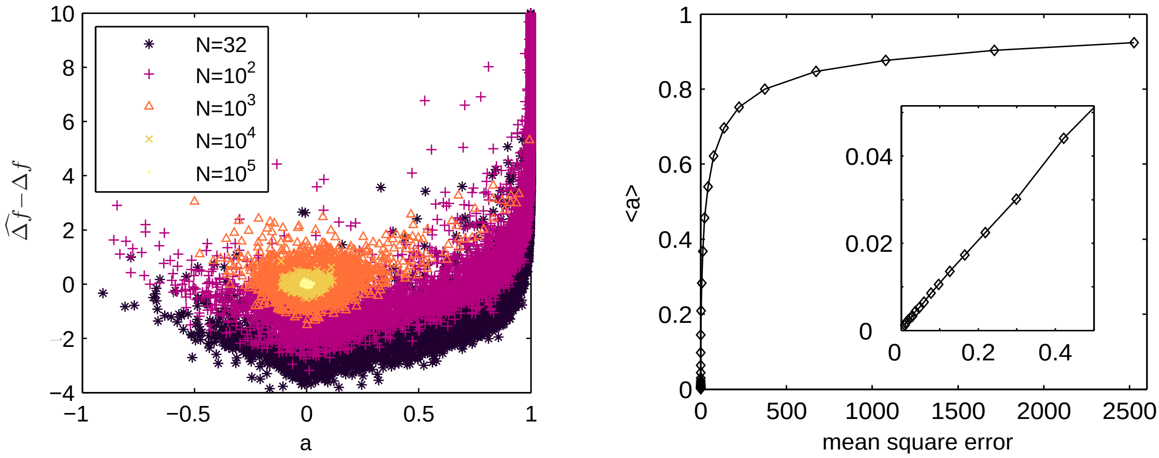

FIG. 12: Relation of free energy estimates $\widehat{\Delta f}$ to convergence measure a (example with exponential work densities). *Left:* Scatter plot of many individual observed pairs $(\widehat{\Delta f}, a)$ for different sample sizes N, showing no correlation, but uniform convergence of $\widehat{\Delta f}$ and a for large sample sizes $N \gtrsim 10^3$ ("onset" of large N limit for the present example). *Right:* Average convergence measure $\langle a \rangle$ plotted against the mean square error $\left\langle (\widehat{\Delta f} - \Delta f)^2 \right\rangle$ (each data point belongs to one value of N). The inset shows an enlargement for small values of a (corresponding to $N \geq 10^3$).

The measure a is bounded from above and below,

$$-1 < a < 1, \tag{71}$$

and converges to zero:

$$a \xrightarrow{N \to \infty} 0. \tag{72}$$

It is relatively easy to judge on the convergence of a, as we explicitly know its asymptotic value. Since we actually want to measure the convergence of $\widehat{\Delta f}$, however, the task is to show *when* the convergence of a to zero takes place with respect to the convergence of $\widehat{\Delta f}$ to Δf. We could argue and demonstrate that a and $\widehat{\Delta f}$ converge *almost simultaneously*, the measure a slightly later than $\widehat{\Delta f}$ (which is better than the other way round). These properties validate a as measure for the convergence of two-sided free energy calculations.

Yet, it must be emphasized that a is *not* directly correlated with the deviation $\widehat{\Delta f} - \Delta f$ of the free energy estimate from the true value. Instead, it merely converges together with $\widehat{\Delta f}$, with average value $\langle a \rangle$ proportional to the mean square error of two-sided estimation (in the limit of large N). This is demonstrated in figure 12, which shows numerical results for exponential work densities (see figs. 8 and 9 in [3] for details).

Hereby we close the discussion of the one-sided and two-sided estimators and introduce umbrella sampling and thermodynamic integration next. In contrast to the former, the latter estimators can not be applied to experimental data. We have not studied them in great detail, and hence will merely state some of their general properties which can also be found in the named literature. Maybe new and interesting, however, is their unified derivation from the fluctuation theorem, although somewhat artificial for thermodynamic integration.

C. Umbrella sampling

Umbrella sampling [22] can be an effective way of solving the rare-event problems with which one- and two-sided estimations are faced. It means sampling not from the original, but from modified, so-called "biased" densities of work and is a special case of importance sampling when applied to estimation of free energy. The biasing, however, is at the cost of loosing information on the equilibrium states and nonequilibrium processes. A unified description of umbrella sampling including modern variants of biased path sampling [72–74] can again be given by starting with the fluctuation theorem (31).

Assuming an arbitrary normalized "umbrella" density $q(w)$ with same support as the work densities $p_0(w)$ and $p_1(w)$, we can write Eq. (36) in the modified form

$$e^{-\Delta f} = \int \frac{e^{-w} p_0(w)}{q(w)} \, q(w) dw \equiv \left\langle \frac{e^{-w} p_0(w)}{q(w)} \right\rangle_q . \tag{73}$$

The average is now in the density $q(w)$, indicated by the subscript q at the angular brackets. We may think of q being proportional to p_0, in order to get rid of the unknown p_0. As, however, we must also assume not to know the normalizing constant of q, which appears explicitly in the averaged function in (73), this expression is not of use for free energy estimation. But we can complete it to a useful expression, taking into account

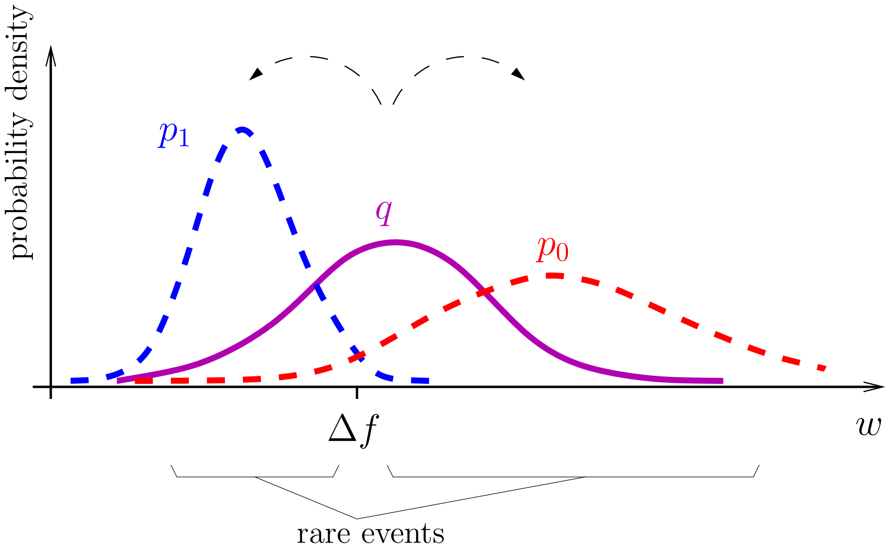

FIG. 13: Umbrella estimation draws work-values from a "biased" density $q(w)$, which can, in principle, be chosen freely. Precise free energy calculations require accurately sampling the regions where $p_0(w)$ and $p_1(w)$ have most of its mass ("rare events"). Thus, $q(w)$ should be chosen such, that it has significant overlap with both, $p_0(w)$ and $p_1(w)$. The suboptimal choice $q(w) = p_0(w)$ results in one-sided (forward) free-energy estimation.

that $1 = \int \frac{p_0(w)}{q(w)}\, q(w) dw$:

$$e^{-\Delta f} = \frac{\left\langle \dfrac{e^{-w} p_0(w)}{q(w)} \right\rangle_q}{\left\langle \dfrac{p_0(w)}{q(w)} \right\rangle_q}. \tag{74}$$

Now the unknown constant cancels. The corresponding estimator $\widehat{\Delta f}_{(q)}$ reads

$$\widehat{\Delta f}_{(q)} = -\ln \frac{\overline{\frac{e^{-w} p_0(w)}{q(w)}}^{(q)}}{\overline{\frac{p_0(w)}{q(w)}}^{(q)}}. \tag{75}$$

The overbar with index q indicates a sample average with a sample drawn from the umbrella density $q(w)$. Similar as before with the t-function in the context of two-sided estimation, we have freedom of choice on the density q. But here this "auxiliary" density does not only define the explicit form of the free energy algorithm, but also the density from which is being drawn.

What is a good choice for $q(w)$? The important contributions to the numerator of (75) originate from w-values for which p_1 is large, whilst the denominator is dominated by values for which p_0 is large. Therefore, a convenient umbrella density has good overlap with both, $p_0(w)$ and $p_1(w)$, which gave rise for the attribute "umbrella" [22]. See figure 13 for an illustration.

For example, a possible choice is the geometric mean $q(w) \sim \sqrt{p_0(w)p_1(w)}$, which yields the estimator

$$\widehat{\Delta f}_{(q)} = -\ln \frac{\overline{e^{-\frac{1}{2}w}}^{(q)}}{\overline{e^{+\frac{1}{2}w}}^{(q)}}. \tag{76}$$

This estimator has been studied within the nonequilibrium methods by Ytreberg and Zuckerman, with which they obtained impressive results [73]. It means sampling paths (trajectories) $x(\cdot)$ from the path density $\sim \sqrt{\mathcal{P}_0[x(\cdot)]\mathcal{P}_1[\bar{x}(\cdot)]} \sim e^{-\frac{1}{2}\beta\mathcal{W}_{[\lambda]}[x(\cdot)]}\mathcal{P}_0[x(\cdot)]$, and calculating work according to $w = \beta\mathcal{W}_{[\lambda]}[x(\cdot)]$, which yields a value distribution of work $q(w) = \sqrt{p_0(w)p_1(w)}/\int\sqrt{p_0(w')p_1(w')}dw'$. Algorithms for biased path sampling were developed by Sun [72] and are essentially based on a Metropolis Monte Carlo-type run for whole trajectories, starting with "unbiased" initial trajectories which are generated by the underlying dynamics (and distributed either according to $\mathcal{P}_0[x(\cdot)]$ or $\mathcal{P}_1[x(\cdot)]$).

Yet, the "geometric" choice $q \sim \sqrt{p_0p_1}$ is still suboptimal. Minimization of the asymptotic mean square error

$$X_{[q]}(N) := \lim_{N\to\infty}\left\langle\left(\widehat{\Delta f}_{(q)} - \Delta f\right)^2\right\rangle = \frac{1}{N}\int\frac{(p_1(w)-p_0(w))^2}{q(w)}\,dw, \tag{77}$$

with respect to $q(w)$ yields the optimal umbrella density [74, 112]

$$q_{opt}(w) = \frac{|p_1(w)-p_0(w)|}{\int|p_1(w')-p_0(w')|\,dw'}. \tag{78}$$

With this choice, the mean square error becomes

$$X_{[q_{opt}]}(N) = \frac{1}{N}\left[\int|p_1(w)-p_0(w)|\ dw\right]^2. \tag{79}$$

The integral under the square is known as the variational distance. It attains at maximum the value 2, namely in the limit of vanishing overlap of the densities ($p_0(w)p_1(w) \to 0\ \forall w$). Note that all methods discussed before become useless in this limit. Further, it can explicitly be shown that the optimal umbrella estimator outperforms one-sided estimation, two-sided estimation, and thermodynamic integration [112] (with respect to the means square error). However, we cannot directly use it, as the optimal umbrella density depends itself on Δf. Investigations of near-optimal choices of $q(w)$ can be found in [112] and [74].

D. Thermodynamic integration

Finally, we come to one of the oldest methods, thermodynamic integration [23]. Actually it is not directly related to the fluctuation theorem, as it calculates free energy from thermodynamic forces instead of work. To clarify the notions, we will begin with stating it from the traditional perspective, and afterwards show that it can also be recovered from the fluctuation theorem, together with possible extensions.

Taking the derivative of the free energy $f_\lambda = \beta F_\lambda$ with respect to λ (temperature $\frac{1}{\beta}$ fixed), we have from Eqs. (2) and (3)

$$\frac{d}{d\lambda} f_\lambda = -\frac{d}{d\lambda} \ln Z_\lambda = \int \rho_\lambda(x) \frac{\partial}{\partial\lambda} \beta H_\lambda(x)\; dx. \tag{80}$$

Finally, integration with respect to λ yields the *thermodynamic integration* identity [23]

$$\Delta f = \int\limits_{\lambda_0=0}^{\lambda_1=1} \left\langle \frac{\partial}{\partial\lambda} \beta H_\lambda(x) \right\rangle_{\rho_\lambda} d\lambda. \tag{81}$$

It is usually used for free energy calculations in the following manner. For a dense sequence $\{\lambda_i\}$ of λ values between λ_0 and λ_1, samples from the canonical densities $\rho_{\lambda_i}(x)$ are drawn, and for each λ_i the average force $\left\langle \frac{\partial}{\partial\lambda} H_\lambda(x) \right\rangle_{\rho_\lambda}$ is estimated. Finally, the integral over λ according to (81) is carried out approximatively. For an overview of optimization issues we refer to [113].

In order to derive thermodynamic integration from the fluctuation theorem, we need to consider it not only between initial and final values $\lambda_0 = 0$ and $\lambda_1 = 1$, but also between $\lambda = 0$ and $\lambda = \lambda(s)$ for all $s \in (0, \tau)$, with $\lambda(0) = 0$ and $\lambda(\tau) = 1$. Thus, we define the forward work density $p_0(w|s)$ and the reverse work density $p_1(w|s)$ belonging to the "process" between $\lambda = 0$ and $\lambda = \lambda(s)$. For each s, the fluctuation theorem

$$\frac{p_0(w|s)}{p_1(w|s)} = e^{w-\Delta f_s} \tag{82}$$

holds, with free energy difference $\Delta f_s = f_{\lambda(s)} - f_0$. Therefore we can write

$$\frac{d}{ds}\Delta f_s = -\frac{\partial}{\partial s} \ln \frac{p_0(w|s)}{p_1(w|s)}, \tag{83}$$

which yields by multiplication with $p_1(w|s)$ and subsequent integration with respect to w

$$\frac{d}{ds}\Delta f_s = -\int p_1(w|s)\frac{\partial}{\partial s} \ln p_0(w|s)\ dw. \tag{84}$$

Up to integration with respect to s, this can be called the work density analog of thermodynamic integration, although in this form it is not of practical use for free energy calculations. To proceed, we need to give an explicit description of the work density $p_0(w|s)$ in terms of the underlying microscopic (path or phase space) density. As an example, we relate the work densities to the equilibrium methods (cf. table I), i.e.

$$p_0(w|s) = \int \delta\left(\beta\Delta H(x,s) - w\right) \rho_0(x) dx, \tag{85}$$

with

$$\Delta H(x,s) = H_{\lambda(s)}(x) - H_0(x), \tag{86}$$

we have

$$\frac{\partial}{\partial s} p_0(w|s) = -\int \frac{\partial}{\partial s}\beta\Delta H(x,s) \frac{d}{dw}\delta\left(\beta\Delta H(x,s) - w\right)\rho_0(x) dx. \tag{87}$$

Using this together with (82) in (84) gives

$$\begin{aligned}\frac{d}{ds}\Delta f_s &= \int dx\ \frac{\partial}{\partial s}\beta\Delta H(x,s)\rho_0(x) \int dw\ e^{-w+\Delta f_s}\frac{d}{dw}\delta\left(\beta\Delta H(x,s) - w\right) \\ &= \int dx\ \frac{\partial}{\partial s}\beta\Delta H(x,s)\rho_0(x) e^{-\beta\Delta H(x,s)+\Delta f_s} = \int \frac{\partial}{\partial s}\beta\Delta H(x,s)\rho_{\lambda(s)}(x)\ dx.\end{aligned} \tag{88}$$

Thus

$$\frac{d}{ds}\Delta f_s = \left\langle \frac{\partial}{\partial s}\beta\Delta H(x,s)\right\rangle_{\rho_{\lambda(s)}}, \tag{89}$$

which is equivalent to (80). Of course, this is a rather complicated way for arriving at this result, but may be of interest with respect to nonequilibrium methods.

Let us finally mention that Sun derived an interesting free energy formula related to thermodynamic integration via the moment generating function of the work density. In our notation, his formula reads [72]

$$\Delta f = \int_0^1 \langle w\rangle_{q_s}\ ds, \tag{90}$$

where the average work $\langle w\rangle_{q_s}$ is calculated with the “moment-generating-density”

$$q_s(w) = \frac{e^{-sw}p_0(w)}{\int e^{-sw'}p_0(w')dw'}. \tag{91}$$

See [72] for details.

We now close the introduction of elementary estimators and come to some final remarks.

III. FINAL REMARKS

The previous sections have shown that three general classes of free energy methods, named here equilibrium, nonequilibrium, and mapping methods, can be considered in a unified way through reduction to the random variable "work" and its fluctuation theorem (31). From the latter follow elementary free energy estimators, which consequently can be applied within any class of methods by measuring or simulating work-values according to their respective definitions (table I).

Yet with respect to application, a specific estimator is not equivalent in the distinct classes. This has two main reasons. First, the way of obtaining the raw-data for the calculation of work quite differs, ranging from simple Monte Carlo sampling of phase space distributions to much more involved and time-consuming simulations of nonequilibrium trajectories. In addition, nonequilibrium and mapping methods require the definition a suitable protocol $\lambda(\cdot)$ and phase space map $\phi(x)$, respectively. Altogether, these measures determine the expenses for the simulation of work. Second, the work densities $p_0(w)$ and $p_1(w)$ obtained within the different classes will be far distinct. This greatly affects the convergence properties of the estimators, which depend crucially on the extend of overlap of the work densities. For example, we have seen that the performance of two-sided estimation depends on the harmonic overlap $U_\alpha = \int \frac{p_0(w)p_1(w)}{\alpha p_0(w)+\widetilde{\alpha}p_1(w)}dw$. Relating the work densities to the equilibrium methods (table I), the overlap can be equivalently written

$$U_\alpha = U_\alpha^{eq} = \int \frac{\rho_0(x)\rho_1(x)}{\alpha\rho_0(x)+\widetilde{\alpha}\rho_1(x)}dx; \tag{92}$$

for mapping methods, it equals

$$U_\alpha = U_\alpha^{map} = \int \frac{\widetilde{\rho_0}(x)\rho_1(x)}{\alpha\widetilde{\rho_0}(x)+\widetilde{\alpha}\rho_1(x)}dx; \tag{93}$$

and for nonequilibrium methods

$$U_\alpha = U_\alpha^{neq} = \int \frac{\mathcal{P}_0[x(\cdot)]\mathcal{P}_1[\bar{x}(\cdot)]}{\alpha\mathcal{P}_0[x(\cdot)]+\widetilde{\alpha}\mathcal{P}_1[\bar{x}(\cdot)]}\mathcal{D}x(\cdot). \tag{94}$$

Hence, the overlap of the microscopic distributions is "conserved" when going over to the corresponding work distributions. But we can expect $U_\alpha^{neq} > U_\alpha^{eq}$ and $U_\alpha^{map} > U_\alpha^{eq}$

(for an appropriate map ϕ). More general, this "conservation" applies for any overlap and distance measure which can be written in the generic form $\int f(\frac{p_0}{p_1})p_0 dw$. Examples include the harmonic overlap, the chi-square distance, the Kullback-Leibler distance and the variational distance, i.e. measures which we have seen to regulate the performance of free energy estimators which are based on the fluctuation theorem.

Therefore, the advantage of nonequilibrium and mapping methods in comparison with equilibrium methods relies on enhancement of overlap (or reduction of distance) of the work densities, but is accompanied with increased expenses for the simulation of work. This touches a long-standing problem, namely the comparison of free energy methods with respect to efficiency, which needs to take into account both, precision as well as time and effort. Some studies in that direction have been done recently [114–117], and some of them indicate that nonequilibrium methods could be less efficient than traditional equilibrium methods.

The elementary free energy estimators can be taken as the starting point for more advanced techniques. For example, if the distance of the distributions $\rho_0(x)$ and $\rho_1(x)$ is quite large, such that applying one- or two-sided estimation in their basic form is in vain, a possible extension could be to use some intermediate state ρ_{λ_i} with $\lambda_i \in (\lambda_0, \lambda_1)$, and to carry out free energy calculations once between λ_0 and λ_i, and once between λ_i and λ_1, thereby having reduced the effective distance between ρ_0 and ρ_1. Such procedures are known under the name "staging" or "stratification", and can be extended to many intermediates. In the limit of infinitely many intermediates, stratification of two-sided estimation converges to thermodynamic integration [113]. Another example, referring to umbrella sampling, is to use adaptive biasing techniques [4] in order to improve the umbrella density $q(w)$ stepwise within the simulation, based on the growth of our information on the system with increasing sample size. Further, so far we have only focused on calculating a single free energy difference ΔF. But in many applications, one is interested in knowing the free energy function in dependence of λ, or likewise, the free energy difference $\Delta F_\lambda = F_\lambda - F_0$ for all values of $\lambda \in [0, 1]$. This can conveniently be achieved with an extended version of two-sided estimation, as worked out in [118–120].

Nevertheless, the study of elementary estimators is still important, insofar as advanced techniques do rely on them and share many of their properties. Among the elementary estimators, the one- and two-sided' are of special importance for physics because of their applicability to laboratory measurements of work.

Concerning the present thesis, our main contributions to the problem of free energy calculations are summarized as follows. We have extended targeted free energy perturbation [75] to a targeted two-sided method by deriving a fluctuation theorem for generalized work [1]. The utility of this new method could be demonstrated with its application to the calculation of the chemical potential of a high-density Lennard-Jones fluid, which required a deepened study of construction and performance of phase space maps. Further, we have studied the general properties of two-sided estimation and could show that: (a) two-sided estimation is a constrained maximum likelihood estimator [1], (b) its asymptotic mean square error $X(\alpha, N)$ is convex in the fraction $\alpha = \frac{n_0}{N}$ of forward draws [2], and (c) its convergence properties are characterized by the ability of sampling a certain overlap region of the work distributions [1, 3]. Based on these observations, we have developed and tested the dynamic sampling strategy [2], which optimizes the efficiency of two-sided estimation. And finally, we have proposed [1], investigated and tested a measure for the convergence of two-sided estimation [3].

We belief that in specific the convergence measure could prove to be a valuable standard instrument for reliable free energy calculations, which closes the lack of appreciation [121] of the convergence properties of two-sided estimation. But the convergence measure a can also be used to test something quite different: namely whether the simulated (or measured) data are distributed according to densities which obey the fluctuation theorem. If this is not the case, then a will converge to some value $\neq 0$ [122]. This can be useful, e.g., to check the consistency of setup of simulations or experiments.

Acknowledgments

I cordially thank Andreas Engel for the supervision of this theses, and Holger Stark for assuming the secondary report. I'm greatly indebted to Holger Then, with whom to work was a real pleasure for me. Finally, I'm most obliged to Thembi for her selfless support of this work.

---

[1] A. M. Hahn and H. Then, *Using bijective maps to improve free energy estimates*, Phys. Rev. E **79**, 011113 (2009).

[2] A. M. Hahn and H. Then, *Characteristic of Bennett's acceptance ratio method*, Phys. Rev. E **80**, 031111 (2009).

[3] A. M. Hahn and H. Then, *Measuring the convergence of Monte-Carlo free energy calculations*, Phys. Rev. E **81**, 041117 (2010).

[4] Ch. Chipot and A. Pohorille (eds.), *Free Energy Calculations*, Springer Series in Chem. Phys. 86 (Springer, Berlin, 2007).

[5] N. Gallavotti, *Statistical mechanics: a short treatise*, (Springer, Berlin, 1999).

[6] J. von Neumann, *Beweis des Ergodensatzes und des H-Theorems in der neuen Mechanik*, Zeitschrift fuer Physik **57**, 30 (1929).

[7] A. J. Chintschin, *Mathematische Grundlagen der statistischen Mechanik*, German translation of the Russian original from 1941, (Bibliographisches Institut, Mannheim, 1964).

[8] J. L. Lebowitz, *Macroscopic laws, microscopic dynamics, time's arrow and Boltzmann's entropy*, Physica A **194**, 1 (1993).

[9] M. Campisi, *On the mechanical foundations of thermodynamics: The generalized Helmholtz theorem*, Stud. Hist. Phil. Mod. Phys. **36**, 275 (2005).

[10] J. Uffink, *Compendium of the foundations of classical statistical physics*, in J. Butterfield and J. Earman (eds), *Handbook for the Philosophy of Physics* (Elsevier, Amsterdam, 2007) pp. 924-1074.

[11] P. Reimann, *Foundation of Statistical Mechanics under Experimentally Realistic Conditions*, Phys. Rev. Lett. **101**, 190403 (2008)

[12] J. M. R. Parrondo, C. Van den Broeck and R. Kawai, *Entropy production and the arrow*

of time, New Journal of Physics **11**, 073008 (2009).

[13] S. Goldstein, J. L. Lebowitz, R. Tumulka and N. Zanghi, *Long-Time Behavior of Macroscopic Quantum Systems: Commentary Accompanying the English Translation of John von Neumann's 1929 Article on the Quantum Ergodic Theorem*, arXiv:1003.2129v1 (2010).

[14] H. Tasaki, *The approach to thermal equilibrium and "thermodynamic normality" — An observation based on the works by Goldstein, Lebowitz, Mastrodonato, Tumulka, and Zanghi in 2009, and by von Neumann in 1929*, arXiv:1003.5424v4 (2010).

[15] R. Becker, *Theorie der Wärme*, 3rd ext. ed. (Springer, Berlin, 1985).

[16] M. Planck, *Vorlesungen über Thermodynamik*, 11th ext. ed. (Walter de Gruyter, Berlin, 1964).

[17] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, 1987).

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of State Calculations by Fast Computing Machines*, J. Chem. Phys. **21**, 1087 (1953).

[19] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic Press, London, 2002).

[20] R. W. Zwanzig, *High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases*, J. Chem. Phys. **22**, 1420 (1954).

[21] C. H. Bennett, *Efficient Estimation of Free Energy Differences from Monte Carlo Data*, J. Comput. Phys. **22**, 245 (1976).

[22] G. M. Torrie and J. P. Valleau, *Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling* J. Comput. Phys. **23**, 187 (1977).

[23] J. G. Kirkwood, *Statistical Mechanics of Fluid Mixtures*, J. Chem. Phys. **3**, 300 (1935).

[24] C. Jarzynski, *Nonequilibrium Equality for Free Energy Differences*, Phys. Rev. Lett. **78**, 2690 (1997).

[25] C. Jarzynski, *Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach*, Phys. Rev. E **56**, 5018 (1997).

[26] C. Jarzynski, *Equilibrium free-energy differences from nonequilibrium Processes*, Acta Physica Polonica B **29**, 1609 (1997).

[27] C. Jarzynski, *Nonequilibrium work theorem for a system strongly coupled to a thermal environment* J. Stat. Mech.: Theor. Exp. P09005 (2004).

[28] G. E. Crooks, *Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems*, J. Stat. Phys. **90**, 1481 (1998).

[29] G. E. Crooks, *Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences*, Phys. Rev. E **60**, 2721 (1999).

[30] G. E. Crooks, *Path-ensemble averages in systems driven far from equilibrium*, Phys. Rev. E **61**, 2361 (2000).

[31] E. M. Sevick, R. Prabhakar, S. R. Williams and D. J. Searles, *Fluctuation Theorems*, Annu. Rev. Phys. Chem. **59**, 603 (2008).

[32] A. Engel, *The second law as identity: Exact relations for non-equilibrium processes*, in T. Pöschel, H. Malchow und L. Schimansky-Geier (eds.), *Irreversible Prozesse und Selbstorganisation* (Logos-Verlag, Berlin, 2006).

[33] A. Gomez-Marin, J. M. R. Parrondo and C. Van den Broeck, *The "footprints" of irreversibility*, EPL **82**, 50002 (2008).

[34] D. J. Evans and D. J. Searles, *Equilibrium microstates which generate second law violating steady states*, Phys. Rev. E **50**, 1645 (1994).

[35] D. J. Evans, E. G. D. Cohen and G.P. Morriss, *Probability of Second Law Violations in Shearing Steady States*, Phys. Rev. Lett. **71**, 2401 (1993).

[36] J. Liphardt, S. Dummont, S. B. Smith, I. Tinoco, and C. Bustamante, *Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski's Equality* Science **296**, 1832 (2002).

[37] G. M. Wang, E. M. Sevick, E. Mittag, D. J. Searles and D. J. Evans, *Experimental Demonstration of Violations of the Second Law of Thermodynamics for Small Systems and Short Time Scales*, Phys. Rev. Lett. **89**, 050601 (2002).

[38] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco and C. Bustamante, *Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies*, Nature *437*, 231 (2005).

[39] C. Bustamante, J. Liphardt and F. Ritort, *The Nonequilibrium Thermodynamics of Small Systems*, Phys. Today **58**, issue 7, p. 43 (2005).

[40] S. Schuler, T. Speck, C. Tierz, J. Wrachtrup and U. Seifert, *Experimental Test of the Fluctuation Theorem for a Driven Two-Level System with Time-Dependent Rates*, Phys. Rev. Lett. **94**, 180602 (2005).

[41] I. Junier, A. Mossa, M. Manosas and F. Ritort, *Recovery of Free Energy Branches in Single Molecule Experiments*, Phys. Rev. Lett. **102**, 070602 (2009).

[42] A. Mossa, M. Manosas, N. Forns, J. M. Huguet and F. Ritort, *Dynamic force spectroscopy of DNA hairpins: I. Force kinetics and free energy landscapes*, J. Stat. Mech.: Theor. Exp., P02060 (2009).

[43] Y. Utsumi, D.S. Golubev, M. Marthaler, K. Saito, T. Fujisawa and Gerd Schön, *Bidirectional single-electron counting and the fluctuation theorem*, Phys. Rev. B **81**, 125331 (2010).

[44] G. N. Bochkov and I. E. Kuzovlev, *General theory of thermal fluctuations in nonlinear systems*, Sov. Phys. JETP **45**, 125 (1977).

[45] C. Jarzynski, *Comparison of far-from-equilibrium work relations*, Comptes Rendus Physique **8**, 495 (2007).

[46] J. Horowitz and C. Jarzynski, *Comparison of work fluctuation relations*, J. Stat. Mech. P11002 (2007).

[47] M. Campisi, P. Talkner and P. Hänggi, *Quantum Bochkov-Kuzovlev Work Fluctuation Theorem*, arXiv:1003.1052v1 (2010); Philos. Trans. Roy. Soc. A **369**, 291 (2011).

[48] A. Münster, *Statistische Thermodynamik* (Springer, Berlin, 1956).

[49] S. Vaikuntanathan and C. Jarzynski, *Dissipation and lag in irreversible processes*, EPL **87**, 60005 (2009).

[50] R. Kawai, J. Parrondo and C. Van den Broeck, *Dissipation: The Phase-Space Perspective*, Phys. Rev. Lett. **98** 080602 (2007).

[51] C. Jarzynski, *Hamiltonian Derivation of a Detailed Fluctuation Theorem*, J. Stat. Phys. **98**, 77 (2000).

[52] E. Schöll-Paschinger and C. Dellago, *A proof of Jarzynski's nonequilibrium work theorem for dynamical systems that conserve the canonical distribution*, J. Chem. Phys. **125**, 054105 (2006).

[53] V. Y. Chernyak, M. Chertkov and C. Jarzynski, *Path-integral analysis of fluctuation theorems for general Langevin processes*, J. Stat. Mech.: Theor. Exp. P08001 (2006).

[54] U. Seifert, *Entropy Production along a Stochastic Trajectory and an Integral Fluctuation Theorem*, Phys. Rev. Lett. **95**, 040602 (2005).

[55] T. Speck and U. Seifert, *The Jarzynski relation, fluctuation theorems, and stochastic ther-*

modynamics for non-Markovian processes, J. Stat. Mech.: Theor. Exp. L09002 (2007).

[56] H. Tasaki, *Jarzynski Relations for Quantum Systems and Some Applications*, arXiv:cond-mat/0009244v2 (2000).

[57] S. Mukamel, *Quantum Extension of the Jarzynski Relation: Analogy with Stochastic Dephasing*, Phys. Rev. Lett. 90, 170604 (2003).

[58] M. Campisi, P. Talkner, and P. Hänggi, *Fluctuation Theorem for Arbitrary Open Quantum Systems*, Phys. Rev. Lett. **102**, 210401 (2009).

[59] M. Campisi, P. Talkner, and P. Hänggi, *Fluctuation Theorems for Continuously Monitored Quantum Fluxes*, arXiv:1006.1542v1 (2010); Phys. Rev. Lett. **105**, 140601 (2010).

[60] A. Engel and R. Nolte, *Jarzynski equation for a simple quantum system: Comparing two definitions of work*, EPL 79, 10003 (2007).

[61] P. Talkner, E. Lutz and P. Hänggi, *Fluctuation theorems: Work is not an observable*, Phys. Rev. E **69**, 026115 (2007).

[62] R. C. Lua, A. Y. Grosberg, *Practical Applicability of the Jarzynski Relation in Statistical Mechanics: A Pedagogical Example*, J. Phys. Chem. B **109**, 6805 (2005).

[63] R. Nolte and A. Engel, *Jarzynski equation for the expansion of a relativistic gas and black-body radiation*, Physica A **388**, 3752 (2009).

[64] B. Cleuren, K. Willaert, A. Engel and C. Van den Broeck, *Fluctuation theorem for entropy production during effusion of a relativistic ideal gas*, Phys. Rev. E **77**, 022103 (2008).

[65] E. G. D. Cohen and D. Mauzerall, *A note on the Jarzynski equality*, J. Stat. Mech.: Theor. Exp. P07006 (2004).

[66] T. Speck and U. Seifert, *Distribution of work in isothermal nonequilibrium processes*, Phys. Rev. E **70**, 066112 (2004).

[67] G. Hummer and A. Szabo, *Free energy reconstruction from nonequilibrium single-molecule pulling experiments*, PNAS **98**, 3658 (2001).

[68] J. E. Hunter III, W. P. Reinhardt and T. F. Davis, *A finite-time variational method for determining optimal paths and obtaining bounds on free energy changes from computer simulations*, J. Chem. Phys. **99**, 6856 (1993).

[69] G. Hummer, *Nonequilibrium Methods for Equilibrium Free Energy Calculations*, in Ch. Chipot and A. Pohorille (eds.), *Free Energy Calculations*, Springer Series in Chem. Phys. 86 (Springer, Berlin, 2007).

[70] F. Ritort, *Single-molecule experiments in biological physics: methods and applications*, J. Phys.: Condens. Matter **18**, R531 (2006).

[71] C. Jarzynski, *Rare events and the convergence of exponentially averaged work values*, Phys. Rev E **73**, 046105 (2006).

[72] S. X. Sun, *Equilibrium free energies from path sampling of nonequilibrium trajectories*, J. Chem. Phys. **118**, 5769 (2003).

[73] F. M. Ytreberg and D. M. Zuckerman, *Single-ensemble nonequilibrium path-sampling estimates of free energy differences*, J. Chem. Phys. **120**, 10876 (2004).

[74] H. Oberhofer and C. Dellago, *Optimum bias for fast-switching free energy calculations*, Comput. Phys. Comm. **179** 41 (2008).

[75] C. Jarzynski, *Targeted free energy perturbation*, Phys. Rev. E **65**, 046122 (2002).

[76] X.-L. Meng and S. Schilling, *Warp Bridge Sampling*, J. Comput. Graph. Stat. **11**, 552 (2002).

[77] W. Lechner, H. Oberhofer, C. Dellago, and P. L. Geissler, *Equilibrium free energies from fast-switching trajectories with large time steps*, J. Chem. Phys. **124**, 044113 (2006).

[78] H. Oberhofer, C. Dellago, and S. Boresch, *Single molecule pulling with large time steps*, Phys. Rev. E **75**, 061106 (2007).

[79] D. M. Zuckerman, private communication (2009).

[80] F. M. Ytreberg and D. M. Zuckerman, *Peptide Conformational Equilibria Computed via a Single-Stage Shifting Protocol*, J. Phys. Chem. B. **109**, 9096 (2005).

[81] A. F. Voter, *A Monte Carlo method for determining free-energy differences and transition state theory rate constants*, J. Chem. Phys. **82**, 1890 (1985).

[82] B. Widom, *Some Topics in the Theory of Fluids*, J. Chem. Phys. **39**, 2808 (1963).

[83] N. Lu and D. A. Kofke, *Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling*, J. Chem. Phys. **114**, 7303 (2001).

[84] N. Lu and D. A. Kofke, *Accuracy of free-energy perturbation calculations in molecular simulation. II. Heuristics*, J. Chem. Phys. **115**, 6866 (2001).

[85] D. Wu and D. A. Kofke, *Model for small-sample bias of free-energy calculations applied to Gaussian-distributed nonequilibrium work measurements*, J. Chem. Phys. **121**, 8742 (2004).

[86] D. Wu and D. A. Kofke, *Asymmetric bias in free-energy perturbation measurements using*

two Hamiltonian-based models, Phys. Rev. E **70**, 066702 (2004).

[87] D. Wu and D. A. Kofke, *Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation*, J. Chem. Phys. **123**, 054103 (2005).

[88] D. Wu and D. A. Kofke, *Phase-space overlap measures. II. Design and implementation of staging methods for free-energy calculations*, J. Chem. Phys. **123**, 084109 (2005).

[89] D. M. Zuckerman and T. B. Woolf, *Theory of a Systematic Computational Error in Free Energy Differences*, Phys. Rev. Lett. **89**, 180602 (2002).

[90] D. M. Zuckerman and T. B. Woolf, *Overcoming finite-sampling errors in fast-switching free-energy estimates: extrapolative analysis of a molecular system*, Chem. Phys. Lett. **351**, 445 (2002).

[91] D. M. Zuckerman and T. B. Woolf, *Systematic Finite-Sampling Inaccuracy in Free Energy Differences and Other Nonlinear Quantities*, J. Stat. Phys. **114**, 1303 (2004).

[92] J. Gore, F. Ritort and C. Bustamante, *Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements*, Proc. Natl. Acad. Sci. **100**, 12564 (2003).

[93] F. Ritort, *Work and heat fluctuations in two-state systems: a trajectory thermodynamics formalism*, J. Stat. Phys.: Theor. Exp. P10016 (2004).

[94] A. Engel, *Asymptotics of work distributions in nonequilibrium systems*, Phys. Rev. E **80**, 021120 (2009).

[95] M. R. Shirts and V. S. Pande, *Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration*, J. Chem. Phys **122**, 144107 (2005).

[96] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, Ann. Math. Statist. **22**, 79 (1951).

[97] P. Geiger and C. Dellago, *Optimum protocol for fast-switching free-energy calculations*, Phys. Rev. E **81**, 021127 (2010).

[98] T. Schmiedl and U. Seifert, *Optimal Finite-Time Processes In Stochastic Thermodynamics*, Phys. Rev. Lett. **98**, 108301 (2007).

[99] H. Then and A. Engel, *Computing the optimal protocol for finite-time processes in stochastic thermodynamics*, Phys. Rev. E **77**, 041105 (2008).

[100] A. Gomez-Marin, T. Schmiedl and U. Seifert, *Optimal protocols for minimal work processes*

in underdamped stochastic thermodynamics, J. Chem. Phys. **129**, 024114 (2008).

[101] T. Schmiedl, E. Dieterich, P.-S. Dieterich and U. Seifert, *Optimal protocols for Hamiltonian and Schrödinger dynamics*, J. Stat. Mech.: Theor. Exp. P07013 (2009).

[102] X.-L. Meng and W. H. Wong, *Simulating ratios of normalizing constants via a simple identity: a theoretical exploration*, Stat. Sin. **6**, 831 (1996).

[103] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, *Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods*, Phys. Rev. Lett. **91**, 140601 (2003).

[104] R. H. Norden, *A Survey of Maximum Likelihood Estimation*, Intern. Statist. Rev. **40**, 329 (1972).

[105] *We note that the constrained maximum likelihood ansatz used in [1] can be extended as follows: write without loss of generality* $p_0(w) = g_0(w)h(w)$ *and* $p_1(w) = g_1(w)h(w)$*, with* g_0 *and* g_1 *unspecified functions which obey the fluctuation theorem Eq.* (31)*. In maximizing the likelihood for the observed data under the assumption that* $h(w)$ *is independent from* Δf*, one arrives at the acceptance ratio estimator Eq.* (56) *(taking into account normalization constraints).*

[106] J. A. Anderson, *Separate sample logistic discrimination*, Biometrika **59**, 19 (1972).

[107] R. L. Prentice and R. Pyke, *Logistic disease incidence models and case-control studies*, Biometrika **66**, 403 (1979).

[108] P. Maragakis, F. Ritort, C. Bustamante, M. Karplus, and G. E. Crooks, *Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise*, J. Chem. Phys. **129**, 024102 (2004).

[109] E. H. Feng and G. E. Crooks, *Length of Time's Arrow*, Phys. Rev. Lett. **101**, 090602 (2008).

[110] N. Lu and T. B. Woolf, *Understanding and Improving Free Energy Calculations in Molecular Simulations: Error Analyses and Reduction Methods*, in Ch. Chipot and A. Pohorille (eds.), *Free Energy Calculations*, Springer Series in Chem. Phys. 86 (Springer, Berlin, 2007) pp. 199–247.

[111] N. Lu, J. K. Singh and D. A. Kofke, *Appropriate methods to combine forward and reverse free-energy perturbation averages*, J. Chem. Phys. *118*, 2977 (2003).

[112] M.-H. Chen and Q.-M. Shao, *On Monte Carlo methods for estimating ratios of normalizing*

constants, Annals of Stat. **25**, 1563 (1997).

[113] A. Gelman and X.-L. Meng, *Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling*, Stat. Science. **13**, 163 (1998).

[114] G. Hummer, *Fast-growth thermodynamic integration: Error and efficiency analysis*, J. Chem. Phys. **114** 7330 (2001).

[115] D. Rodriguez-Gomez, E. Darve and A. Pohorille, *Assessing the efficiency of free energy calculation methods*, J. Chem. Phys. **120**, 3563 (2004).

[116] H. Oberhofer, C. Dellago and P. L. Geissler, *Biased Sampling of Nonequilibrium Trajectories: Can Fast Switching Simulations Outperform Conventional Free Energy Calculation Methods?*, J. Phys. Chem. B **109**, 6902 (2005).

[117] W. Lechner and C. Dellago, *On the efficiency of path sampling methods for the calculation of free energies from non-equilibrium simulations*, J. Stat. Mech.: Theor. Exp. P04001 (2007).

[118] D. D. L. Minh and A. B. Adib, *Optimized Free Energies from Bidirectional Single-Molecule Force Spectroscopy*, Phys. Rev. Lett. **100**, 180602 (2008).

[119] D. D. L. Minh and J. D. Chodera, *Optimal estimators and asymptotic variances for nonequilibrium path-ensemble averages*, J. Chem. Phys. **131**, 134110 (2009).

[120] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan, *A theory of statistical models for Monte Carlo integration*, J. R. Stat. Soc. B **65**, 585 (2003).

[121] A. Pohorille, C. Jarzynski and C. Chipot, *Good Practices in Free-Energy Calculations*, J. Phys. Chem. B **114**, 10235 (2010).

[122] A. M. Hahn and H. Then, unpublished.

Published work

Paper [1]

PHYSICAL REVIEW E **79**, 011113 (2009)

Using bijective maps to improve free-energy estimates

A. M. Hahn and H. Then
Institut für Physik, Carl von Ossietzky Universität, 26111 Oldenburg, Germany

(Received 12 September 2008; published 13 January 2009)

We derive a fluctuation theorem for generalized work distributions, related to bijective mappings of the phase spaces of two physical systems, and use it to derive a two-sided constraint maximum likelihood estimator of their free-energy difference which uses samples from the equilibrium configurations of both systems. As an application, we evaluate the chemical potential of a dense Lennard-Jones fluid and study the construction and performance of suitable maps.

DOI: 10.1103/PhysRevE.79.011113 PACS number(s): 05.40.−a, 05.70.Ln

I. INTRODUCTION

Extracting free-energy differences from a suitable set of computer simulation data is an active field of research and of interest, e.g., for drug design [1] or nonperturbative quantum chromodynamics [2]. Concerning estimators for the free-energy difference, an extensive literature can be found. Probably the most elementary estimator is the traditional free-energy perturbation [3], which is briefly introduced in the following. Assume we have given two systems, arbitrarily labeled as system 0 and system 1, that are characterized by Hamiltonians $H_0(x)$ and $H_1(x)$, respectively, depending on the point x in phase space. Further, let $\rho_i(x)$ denote the thermal equilibrium phase space density of system i,

$$\rho_i(x) = \frac{e^{-\beta H_i(x)}}{Z_i}, \quad i = 0,1, \tag{1}$$

where $Z_i = \int e^{-\beta H_i(x)}dx$ denotes the partition function and $\beta = \frac{1}{kT}$ the inverse temperature. We are interested in the Helmholtz free-energy difference ΔF of the systems, defined as $\Delta F = -\frac{1}{\beta}\ln\frac{Z_1}{Z_0}$. Traditional free-energy perturbation [3] originates from the equality

$$\frac{\rho_0(x)}{\rho_1(x)} = e^{\beta[\Delta H(x) - \Delta F]}, \tag{2}$$

with $\Delta H(x) := H_1(x) - H_0(x)$. The latter quantity may be interpreted as the work performed during an infinitely fast switching process transforming system 0 to system 1, with initial configuration x [4]. A direct consequence of Eq. (2) is the perturbation identity

$$e^{-\beta\Delta F} = \int e^{-\beta\Delta H(x)}\rho_0(x)dx, \tag{3}$$

which is frequently used to obtain an estimate of ΔF in drawing a sample $\{x_1, \ldots, x_N\}$ from $\rho_0(x)$ (e.g., by Monte Carlo simulations) and evaluating the estimator

$$\widehat{\Delta F}_0^{\text{trad}} = -\frac{1}{\beta}\ln \overline{e^{-\beta\Delta H(x)}}. \tag{4}$$

The overbar denotes a sample average [i.e., $\overline{f(x)} = \frac{1}{N}\sum_{k=1}^{N} f(x_k)$ where f stands for an arbitrary function]. As can be seen by comparison with Eq. (2), the integrand appearing in Eq. (3) is proportional to ρ_1, and thus the main contributions to an accurate estimate of ΔF with Eq. (4) come from realizations x (drawn from ρ_0) that are typical for the density ρ_1. This means that the performance of such an estimate depends strongly on the degree of overlap of ρ_0 with ρ_1. If the overlap is small, the traditional free-energy perturbation is plagued with a slow convergence and a large bias. This can be overcome by using methods that bridge the gap between the densities ρ_0 and ρ_1, for instance the thermodynamic integration. Since thermodynamic integration samples a sequence of many equilibrium distributions, it soon becomes computationally expensive. Another method is umbrella sampling [5] which distorts the original distribution in order to sample regions that are important for the average. Because of the distortion, the latter method is in general restricted to answer only one given question, e.g., the value of the free-energy difference, but fails to give further answers. This is of particular concern, if in addition the values of some other thermodynamic variables are sought, for example, pressure or internal energy. There are dynamical methods [6] that make use of the Jarzynski work theorem [4]. They allow to base the estimator on work values of fast, finite time, nonequilibrium processes connecting system 0 with system 1. However, the dynamic simulation of the trajectories is typically very expensive.

Six years ago, the targeted free-energy perturbation method [7] was introduced; a promising method which is based on mapping equilibrium distributions close to each other in order to overcome the problem of insufficient overlap, without the need to draw from biased distributions. However, this method is rarely used in the literature [8,9]. An obstacle might be that there is no general description of how to construct a suitable map. A related idea [10] was applied in [11]. A recent improvement is the escorted free-energy simulation [12] which is a dynamical generalization of the targeted free-energy perturbation.

Any free-energy difference refers to two equilibrium ensembles. The above mentioned methods draw only from one of the two ensembles and propagate the system in direction of the other. Insofar, they are "one-sided" methods. However, it is of advantage to draw from both equilibrium distributions and combine the obtained "two-sided" information. Optimizing the elementary two-sided estimator for free-energy differences results in the acceptance ratio method [13–15]. The next step of improvement is to implement a two-sided targeted free-energy method that optimally employs the information of drawings from both equilibrium distributions. Our aim is to combine the advantages of the acceptance ratio

©2009 The American Physical Society

method with the advantages of the targeted free-energy perturbation.

The central result of this paper is a fluctuation theorem for the distributions of generalized work values that is derived and presented in Sec. III. From this fluctuation theorem, the desired optimal two-sided targeted free-energy estimator follows in Sec. IV. In Sec. V, appropriate measures are introduced which relate the overlap of $\tilde{\rho}_0$ with ρ_1 to the mean square errors of the one- and two-sided free-energy estimators. In Sec. VI, a convergence criterion for the two-sided estimator is proposed. From Sec. VII on, numerics plays an important part. In particular, Sec. VII A deals with explicit numerical applications. Based on the two-sided targeted free-energy estimator, in Sec. VII B, an estimator for the chemical potential of a high-density homogeneous fluid is established and applied to a dense Lennard-Jones fluid. Finally, the construction and performance of suitable maps is studied. In order to set some notation straight, we start by recalling the targeted free-energy perturbation method.

II. TARGETED FREE-ENERGY PERTURBATION

Let Γ_0 and Γ_1 denote the phase spaces of the systems 0 and 1, respectively. We require that Γ_i contains only those points x for which $\rho_i(x)$ is nonzero.

Mapping the phase space points of system 0, $x \rightarrow \phi(x)$, such that the mapped phase space $\tilde{\Gamma}_0 = \phi(\Gamma_0)$ coincides with the phase space Γ_1 and such that the mapped distribution $\tilde{\rho}_0$ overlaps better with the canonical distribution ρ_1 results in the *targeted free-energy perturbation* [7] where the samples are drawn *effectively* from $\tilde{\rho}_0$ instead.

Following the idea of Jarzynski [7], we introduce such a phase space map. If Γ_0 and Γ_1 are diffeomorph, there exists a bijective and differentiable map $\mathcal{M}$ from Γ_0 to Γ_1,

$$\mathcal{M}:\Gamma_0 \rightarrow \Gamma_1, \quad \mathcal{M}:x \rightarrow \phi(x), \tag{5}$$

where the absolute value of the Jacobian is

$$K(x) = \left| \left| \frac{\partial \phi}{\partial x} \right| \right|. \tag{6}$$

The inverse map reads

$$\mathcal{M}^{-1}:\Gamma_1 \rightarrow \Gamma_0, \quad \mathcal{M}^{-1}:y \rightarrow \phi^{-1}(y). \tag{7}$$

According to $\mathcal{M}$, the phase space density ρ_0 is mapped to the density $\tilde{\rho}_0$,

$$\tilde{\rho}_0(y) = \int_{\Gamma_0} \delta[y - \phi(x)]\rho_0(x)dx, \tag{8}$$

which can be written as

$$\tilde{\rho}_0(\phi(x)) = \frac{\rho_0(x)}{K(x)} \tag{9}$$

or

$$\int_{\phi(\Gamma)} \tilde{\rho}_0(y)dy = \int_{\Gamma} \rho_0(x)dx, \quad \forall \ \Gamma \subset \Gamma_0. \tag{10}$$

In analogy to Eq. (2), the targeted free-energy perturbation is based on the identity

$$\frac{\tilde{\rho}_0(\phi(x))}{\rho_1(\phi(x))} = e^{\beta[\widetilde{\Delta H}(x) - \Delta F]} \quad \forall \ x \in \Gamma_0, \tag{11}$$

which follows from the densities (1) and (9) with $\widetilde{\Delta H}$ being defined by

$$\widetilde{\Delta H}(x) := H_1(\phi(x)) - H_0(x) - \frac{1}{\beta} \ln K(x). \tag{12}$$

Multiplying Eq. (11) by $e^{-\beta\widetilde{\Delta H}(x)}\rho_1(\phi(x))K(x)$ and integrating over Γ_0 yields the targeted free-energy perturbation formula,

$$e^{-\beta\Delta F} = \int_{\Gamma_0} e^{-\beta\widetilde{\Delta H}(x)}\rho_0(x)dx. \tag{13}$$

An alternative derivation is given in [7]. The traditional free-energy perturbation formula (3) can be viewed as a special case of Eq. (13). The latter reduces to the former if $\mathcal{M}$ is chosen to be the identity map, $\phi(x) = x$. [This requires that $\Gamma_1 = \Gamma_0$ holds.]

Now an obvious estimator for ΔF, given a sample $\{x_k\}$ drawn from $\rho_0(x)$, is

$$\widehat{\Delta F}_0 = -\frac{1}{\beta} \ln \overline{e^{-\beta\widetilde{\Delta H}(x)}}, \tag{14}$$

which we refer to as the targeted *forward* estimator for ΔF. The convergence problem of the traditional forward estimator, Eq. (4), in the case of insufficient overlap of ρ_0 with ρ_1 is overcome in the targeted approach by choosing a suitable map $\mathcal{M}$ for which the image $\tilde{\rho}_0$ of ρ_0 overlaps better with ρ_1. Indeed, suppose for the moment that the map is chosen to be ideal, namely such that $\tilde{\rho}_0(x)$ coincides with $\rho_1(x)$. Then, as a consequence of Eq. (11), the quantity $\widetilde{\Delta H}(x)$ is constant and equals ΔF, and the convergence of the targeted estimator (14) is immediate. Although the construction of such an ideal map is impossible in general, the goal of approaching an ideal map guides the design of suitably good maps.

To complement the *one-sided* targeted estimator, a second perturbation formula in the "reverse" direction is derived from Eq. (11),

$$e^{+\beta\Delta F} = \int_{\Gamma_1} e^{+\beta\widetilde{\Delta H}(\phi^{-1}(y))}\rho_1(y)dy, \tag{15}$$

leading to the definition of the targeted *reverse* estimator $\widehat{\Delta F}_1$ of ΔF,

$$\widehat{\Delta F}_1 = +\frac{1}{\beta} \ln \overline{e^{+\beta\widetilde{\Delta H}(\phi^{-1}(y))}}. \tag{16}$$

The index 1 indicates that the set $\{y_k\}$ is drawn from ρ_1. Using the identity map $\phi(x) = x$ in Eq. (16) gives the traditional reverse estimator, which is valid if $\Gamma_0 = \Gamma_1$ holds.

It will prove to be beneficial to switch from phase space densities to one-dimensional densities which describe the value distributions of $\widetilde{\Delta H}(x)$ and $\widetilde{\Delta H}(\phi^{-1}(y))$, cf. Eqs. (13) and (15). This is done next and results in the fluctuation theorem for generalized work distributions.

III. FLUCTUATION THEOREM FOR GENERALIZED WORK DISTRIBUTIONS

We call $\widetilde{\Delta H}(x)$, $x\in\Gamma_0$, function of the *generalized work* in *forward* direction and $\widetilde{\Delta H}(\phi^{-1}(y))$, $y\in\Gamma_1$, function of the *generalized work* in *reverse* direction, having in mind that these quantities are the functions of the actual physical work for special choices of the map $\mathcal{M}$ [16].

The probability density $p(W|0;\mathcal{M})$ for the outcome of a specific value W of the generalized work in forward direction subject to the map $\mathcal{M}$ when sampled from ρ_0 is given by

$$p(W|0;\mathcal{M}) = \int_{\Gamma_0} \delta[W-\widetilde{\Delta H}(x)]\rho_0(x)dx. \tag{17}$$

Conversely, the probability density $p(W|1;\mathcal{M})$ for the observation of a specific value W of the generalized work in reverse direction when sampled from ρ_1 reads

$$p(W|1;\mathcal{M}) = \int_{\Gamma_1} \delta[W-\widetilde{\Delta H}(\phi^{-1}(y))]\rho_1(y)dy. \tag{18}$$

Relating the forward and reverse "work" probability densities to each other results in the fluctuation theorem

$$\frac{p(W|0;\mathcal{M})}{p(W|1;\mathcal{M})} = e^{\beta(W-\Delta F)}. \tag{19}$$

This identity provides the main basis for our further results. It is established by multiplying Eq. (11) with $\delta[W-\widetilde{\Delta H}(x)]\rho_1(\phi(x))$ and integrating with respect to $\phi(x)$. The left-hand side yields

$$\int_{\phi(\Gamma_0)} \delta[W-\widetilde{\Delta H}(x)]\tilde{\rho}_0(\phi(x))d\phi(x)$$
$$= \int_{\Gamma_0} \delta[W-\widetilde{\Delta H}(x)]\rho_0(x)dx = p(W|0;\mathcal{M}), \tag{20}$$

and the right-hand side gives

$$\int_{\phi(\Gamma_0)} e^{\beta[\widetilde{\Delta H}(x)-\Delta F]}\delta[W-\widetilde{\Delta H}(x)]\rho_1(\phi(x))d\phi(x)$$
$$= e^{\beta(W-\Delta F)}\int_{\Gamma_1} \delta[W-\widetilde{\Delta H}(\phi^{-1}(y))]\rho_1(y)dy$$
$$= e^{\beta(W-\Delta F)}p(W|1;\mathcal{M}). \tag{21}$$

It is worthwhile to emphasize that the fluctuation theorem (19) is an exact identity for *any* differentiable, bijective map $\mathcal{M}$ from Γ_0 to Γ_1. Especially, it covers known fluctuation theorems [17–20] related to the *physical* work applied to a system that is driven externally and evolves in time according to some deterministic equations of motion, e.g., those of Hamiltonian dynamics, Nosé-Hoover dynamics, or Gaussian isokinetic dynamics [16].

As an example, consider the time-reversible adiabatic evolution of a conservative system with Hamiltonian $H_\lambda(x)$, depending on an externally controlled parameter λ (e.g., the strength of an external field). Let $x(t)=\phi(x_0,t;\lambda(\cdot))$ with $x(0)=x_0$ be the flow of the Hamiltonian system which is a functional of the parameter $\lambda(t)$ that is varied from $\lambda(0)=0$ to $\lambda(\tau)=1$ according to some prescribed protocol that constitutes the forward process. The Hamiltonian flow can be used to define a map, $\mathcal{M}: x\to\phi(x) \coloneqq \phi(x,\tau;\lambda(\cdot))$. Since the evolution is adiabatic and Hamiltonian, no heat is exchanged, $Q=0$, and the Jacobian is identical to one, $|\frac{\partial\phi}{\partial x}|=1$. Consequently, the generalized work in the forward direction reduces to the physical work applied to the system, $W^0 \coloneqq \widetilde{\Delta H}(x)=H_1(\phi(x))-H_0(x)=W$. For each forward path $\{x(t),\lambda(t);0\leqslant t\leqslant\tau\}$ we have a reverse path $\{x^T(\tau-t),\lambda^T(\tau-t);0\leqslant t\leqslant\tau\}$, where the superscript T indicates that quantities that are odd under time reversal (such as momenta) have changed their sign. The generalized work in reverse direction reduces to the physical work done *by* the system

$$W^1 \coloneqq \widetilde{\Delta H}(\phi^{-1}(y)) = H_1(y) - H_0(\phi^{-1}(y)) - \tfrac{1}{\beta}\ln K(\phi^{-1}(y))$$
$$= -[H_0(\phi^{-1}(y)^T) - H_1(y^T)] = -W.$$

Starting the forward process with an initial canonical distribution, $\rho_0(x)$, some probability distribution for the physical work in forward direction follows, $p(W|0;\mathcal{M}) \eqqcolon p^F(W)$. Starting the reverse process with an initial canonical distribution, $\rho_1(y)$, some probability distribution for the physical work in reverse direction follows, $p(W|1;\mathcal{M}) \eqqcolon p^R(-W)$. The distributions $p^F(W)$ and $p^R(-W)$ are related to each other by the identity (19) which coincides with the fluctuation theorem of Crooks [17].

From the fluctuation theorem (19) some important inequalities follow that are valid for any map $\mathcal{M}$. First of all we state that the targeted free-energy perturbation formulas (13) and (15) can be regarded as a simple consequence of the fluctuation theorem (19) and can be rewritten in terms of the generalized work distributions, $e^{-\beta\Delta F}=\langle e^{-\beta W}\rangle_0$ and $e^{+\beta\Delta F}=\langle e^{+\beta W}\rangle_1$, where the angular brackets with subscript i denote an ensemble average with respect to the density $p(W|i;\mathcal{M})$, $i=0,1$. The monotonicity and convexity of the exponential function appearing in the above averages allows the application of Jensen's inequality, $\langle e^{\mp\beta W}\rangle \geqslant e^{\mp\beta\langle W\rangle}$. From this follows the fundamental inequality

$$\langle W\rangle_1 \leqslant \Delta F \leqslant \langle W\rangle_0, \tag{22}$$

which shows that the values of the average work in forward and reverse direction constitute an upper and a lower bound on ΔF, respectively.

Concerning one-sided estimates of ΔF, the targeted forward and reverse estimators (14) and (16) can be written $\widehat{\Delta F}_0=-\frac{1}{\beta}\overline{e^{-\beta W^0}}$ and $\widehat{\Delta F}_1=\frac{1}{\beta}\overline{e^{\beta W^1}}$, where the overbar denotes a sample average according to a sample $\{W_k^0\}$ and $\{W_k^1\}$ of forward and reverse work values, respectively. Similarly to Eq. (22) one finds the inequalities $\widehat{\Delta F}_0\leqslant\overline{W^0}$ and $\widehat{\Delta F}_1\geqslant\overline{W^1}$.

Taking the ensemble averages $\langle\widehat{\Delta F_i}\rangle_i = \mp\frac{1}{\beta}\langle\ln\overline{e^{\mp\beta W^i}}\rangle_i$, $i=0,1$, of the one-sided estimates and applying Jensen's inequality to the averages of the logarithms, $\langle\ln\overline{e^{\mp\beta W^i}}\rangle_i \leqslant \ln\langle e^{\mp\beta W^i}\rangle_i = \mp\beta\Delta F$, one obtains

$$\langle W\rangle_1 \leqslant \langle\widehat{\Delta F_1}\rangle_1 \leqslant \Delta F \leqslant \langle\widehat{\Delta F_0}\rangle_0 \leqslant \langle W\rangle_0. \tag{23}$$

In other words, the forward and reverse estimators are biased in opposite directions for any finite size N of the work samples, but their mean values form closer upper and lower bounds on ΔF than the values of the mean work do.

So far, we were concerned with one-sided estimates of ΔF only. However, the full power of the fluctuation theorem (19) will develop when dealing with a two-sided targeted free-energy estimator where a sample of forward *and* reverse work values is used simultaneously, since the fluctuation theorem relates the forward and reverse work probability densities to each other in dependence of the free-energy difference.

In the next section, we will not mention the target map $\mathcal{M}$ explicitly in order to simplify the notation. For instance, we will write $p(W|i)$, but mean $p(W|i;\mathcal{M})$ instead.

IV. TWO-SIDED TARGETED FREE-ENERGY ESTIMATOR

An important feature of the fluctuation theorem (19) is that it provides a way to answer the following question: Given a sample of n_0 work values $\{W_i^0\}=\{W_1^0,\ldots,W_{n_0}^0\}$ in the forward direction and a second sample of n_1 work values $\{W_j^1\}=\{W_1^1,\ldots,W_{n_1}^1\}$ in the reverse direction, what would be the best estimator of ΔF that utilizes the entire two samples?

If drawn from an ensemble that consists of forward and reverse work values, the elements are given by a pair of values (W,Y) of work and direction, where $Y=0$ indicates the forward and $Y=1$ the reverse direction. The probability density of the pairs (W,Y) is $p(W,Y)$. The probability density for the work is $p(W) := p(W,0)+p(W,1)$, and that for the direction is $p_Y := \int p(W,Y)dW$.

Bayes theorem,

$$p(W|Y)p_Y = p(Y|W)p(W), \tag{24}$$

implies the "balance" equation

$$p_1\int p(0|W)p(W|1)dW = p_0\int p(1|W)p(W|0)dW. \tag{25}$$

From the fluctuation theorem (19) and Bayes theorem (24) follows

$$\frac{p(0|W)}{p(1|W)} = e^{\beta(W-C)} \tag{26}$$

with

$$C = \Delta F + \frac{1}{\beta}\ln\frac{p_1}{p_0}. \tag{27}$$

Together with the normalization $p(0|W)+p(1|W)=1$, Eq. (26) determines the explicit form of the conditional direction probabilities [15],

$$p(Y|W) = \frac{e^{Y\beta(C-W)}}{1+e^{\beta(C-W)}}, \quad Y=0,1. \tag{28}$$

Replacing both, the ensemble averages by sample averages and the ratio $\frac{p_1}{p_0}$ by $\frac{n_1}{n_0}$, the balance equation, $p_1\langle p(0|W)\rangle_1 = p_0\langle p(1|W)\rangle_0$, results in the two-sided targeted free-energy estimator, $n_1\overline{p(0|W^1)} = n_0\overline{p(1|W^0)}$, which reads

$$\sum_{j=1}^{n_1}\frac{1}{1+e^{\beta\left(\widehat{\Delta F}_{01}+\frac{1}{\beta}\ln\frac{n_1}{n_0}-W_j^1\right)}} = \sum_{i=1}^{n_0}\frac{1}{1+e^{-\beta\left(\widehat{\Delta F}_{01}+\frac{1}{\beta}\ln\frac{n_1}{n_0}-W_i^0\right)}}. \tag{29}$$

It is worth it to emphasize that this estimator is *the* optimal two-sided estimator, a result that is shown with a constraint maximum likelihood approach in the Appendix. A derivation of this estimator is also given by Shirts *et al.* [15] in the framework of a maximum likelihood approach.

If samples of n_0 forward and n_1 reverse work values $\{W_i^0\}$ and $\{W_j^1\}$ are given, but no further information is present, it is the two-sided estimator (29) that yields the best estimate of the free-energy difference with respect to the mean square error. If needed, the samples $\{W_i^0\}$ and $\{W_j^1\}$ can be obtained indirectly by drawing samples $\{x_i\}$ and $\{y_j\}$ of ρ_0 and ρ_1 and setting $W_i^0=\widehat{\Delta H}(x_i)$ and $W_j^1=\widehat{\Delta H}(\phi^{-1}(y_j))$.

Opposed to the one-sided estimators (14) and (16), the two-sided targeted free-energy estimator (29) is an implicit equation that needs to be solved for $\widehat{\Delta F}_{01}$. Note however that the solution $\widehat{\Delta F}_{01}$ is unique.

Let us mention a subtlety concerning the choice of the ratio $\frac{p_1}{p_0}$. The mixed ensemble $\{(W,Y)\}$ is specified by the mixing ratio $\frac{p_1}{p_0}$, and by the conditional work probability densities $p(W|Y)$. With the mixed ensemble we are free to choose the mixing ratio. For instance, replacing the ensemble averages in the balance equation (25) by sample averages results in an estimator $p_1\overline{p(0|W^1)} = p_0\overline{p(1|W^0)}$ for ΔF that depends on the value of the mixing ratio. This raises the question of the optimal choice for $\frac{p_1}{p_0}$. As shown in the Appendix, it is optimal to choose the mixing ratio equal to the sample ratio, $\frac{p_1}{p_0}=\frac{n_1}{n_0}$. A result that may be clear intuitively, since then the mixed ensemble reflects the actual samples best.

Other free-energy estimators follow, if the explicit expressions (28) and the definition of the constant C, Eq. (27), are inserted in the balance equation (25). The latter can then be expressed as

$$e^{\beta\Delta F} = e^{\beta C}\frac{\displaystyle\int\frac{1}{1+e^{\beta(C-W)}}p(W|1)dW}{\displaystyle\int\frac{1}{1+e^{-\beta(C-W)}}p(W|0)dW}, \tag{30}$$

and results in the estimator

$$\widehat{\Delta F}_{\mathrm{B}}(C) = C + \frac{1}{\beta}\ln\frac{\dfrac{1}{n_1}\displaystyle\sum_{j=1}^{n_1}\frac{1}{1+e^{\beta(C-W_j^1)}}}{\dfrac{1}{n_0}\displaystyle\sum_{i=1}^{n_0}\frac{1}{1+e^{-\beta(C-W_i^0)}}}. \tag{31}$$

The nontargeted version of this estimator, i.e., for $\mathcal{M}=id$, is due to Bennett [13] who used a variational principle in order to find the estimator for the free-energy difference that minimizes the mean square error.

Equation (30) is an identity for any value of C, since with the ratio $\frac{p_1}{p_0}$ the value of $C=\Delta F+\frac{1}{\beta}\ln\frac{p_1}{p_0}$ can be chosen arbitrarily. However, concerning the estimator (31), different values of C yield different estimates. Bennett's choice is

$$C_{\mathrm{B}} = \Delta F + \frac{1}{\beta}\ln\frac{n_1}{n_0}, \tag{32}$$

i.e., $\frac{p_1}{p_0}=\frac{n_1}{n_0}$, which results from minimizing the mean square error $\langle(\widehat{\Delta F}_{\mathrm{B}}-\Delta F)^2\rangle$, where the angular brackets denote an average over infinitely many repetitions of the estimation process (31) with n_0 and n_1 being fixed. According to the Appendix, Bennett's choice is also optimal for any target map $\mathcal{M}$.

With $C=C_{\mathrm{B}}$, Eq. (31) has to be solved in a self-consistent manner which is tantamount to solve the two-sided targeted estimator (29). In other words, $\widehat{\Delta F}_{\mathrm{B}}(C_{\mathrm{B}})$ is the unique root $\widehat{\Delta F}_{01}$ of Eq. (29).

V. OVERLAP MEASURES AND MEAN SQUARE ERRORS

In this section we introduce measures for the overlap of $\tilde{\rho}_0$ with ρ_1, or, equivalently, of $p(W|0;\mathcal{M})$ with $p(W|1;\mathcal{M})$ and relate them to the mean square error of one- and two-sided estimators.

The estimators (14), (16), and (29) are subject to both, bias and variance. Taking both errors into account results in the mean square error. Let us consider the mean square errors of the one-sided targeted estimators first. They read $X_0 := \langle(\widehat{\Delta F}_0-\Delta F)^2\rangle_0 = \langle(\ln \overline{e^{-\beta(W^0-\Delta F)}})^2\rangle_0$ in forward direction, and analogously in backward direction. In the forward direction, it can be quantified by expanding the logarithm into a power series about the mean value of its argument, $\langle\overline{e^{-\beta(W^0-\Delta F)}}\rangle_0=1$, and neglecting terms of higher order in $\frac{1}{N}$, which gives

$$\beta^2 X_0 \approx \frac{1}{N}\langle(e^{-\beta(W-\Delta F)}-1)^2\rangle_0. \tag{33}$$

Equation (33) is valid for a sufficiently large sample size N (large N limit) [21]. With the use of the fluctuation theorem (19), the variance appearing on the right-hand side of Eq. (33) can be written $\langle(e^{-\beta(W-\Delta F)}-1)^2\rangle_0=\langle e^{-\beta(W-\Delta F)}\rangle_1-1 \geqslant e^{-\beta(\langle W\rangle_1-\Delta F)}-1$. This yields the inequality

$$\beta^2 X_0 \geqslant \frac{1}{N}(e^{\beta(\Delta F-\langle W\rangle_1)}-1). \tag{34}$$

In the same manner as above the inequality

$$\beta^2 X_1 \geqslant \frac{1}{N}(e^{\beta(\langle W\rangle_0-\Delta F)}-1) \tag{35}$$

is obtained for the mean square error X_1 of the reverse estimator $\widehat{\Delta F}_1$.

The inequalities (34) and (35) specify the minimum sample size N that is required to obtain a forward and reverse estimate $\widehat{\Delta F}$, respectively, whose root mean square error $\sqrt{X}$ is not larger than kT. Namely, $N\geqslant e^{\beta(\Delta F-\langle W\rangle_1)}$ is required for a forward, and $N\geqslant e^{\beta(\langle W\rangle_0-\Delta F)}$ for a reverse estimate. Similar expressions are found in Ref. [22]. Since the required sample size N depends exponentially on the dissipation, it is good to choose a target map $\mathcal{M}$ which reduces the dissipation in the opposite direction.

The dissipation is related to the overlap of $\tilde{\rho}_0$ with ρ_1. The overlap of two probability densities $\pi_a(z)$ and $\pi_b(z)$ of a random variable z can be quantified with the Kullback-Leibler divergence

$$D(\pi_a\|\pi_b) := \int \pi_a(z)\ln\frac{\pi_a(z)}{\pi_b(z)}dz, \tag{36}$$

a positive semidefinite measure that yields zero if and only if π_a is identical to π_b. Applied to the densities ρ_1 and $\tilde{\rho}_0$, the Kullback-Leibler divergence turns out to be identical with the Kullback-Leibler divergence of $p(W|1;\mathcal{M})$ with $p(W|0;\mathcal{M})$ and results in the generalized dissipated work in reverse direction,

$$D(\rho_1\|\tilde{\rho}_0) = D(p(W|1;\mathcal{M})\|p(W|0;\mathcal{M})) = \beta(\Delta F-\langle W\rangle_1), \tag{37}$$

which is established with the use of Eqs. (11) and (18), and the fluctuation theorem (19). Similarly, we have

$$D(\tilde{\rho}_0\|\rho_1) = D(p(W|0;\mathcal{M})\|p(W|1;\mathcal{M})) = \beta(\langle W\rangle_0-\Delta F). \tag{38}$$

For the one-sided targeted free-energy estimators this means that choosing a target map which reduces the dissipation in the opposite direction is the same as choosing a target map which enhances the overlap of $\tilde{\rho}_0$ with ρ_1.

Now, we proceed with the overlap measure and the mean square error of the two-sided free-energy estimator (29). In order to keep the notation simple, we assume that the samples of forward and reverse work values are of equal size, $n_0=n_1=\frac{N}{2}$, i.e., $n_0+n_1=N$. (A generalization to $n_0\neq n_1$ is straightforward possible, but not given in this paper.)

Consider the overlap density $p_{\mathrm{ol}}(W|\mathcal{M})$,

$$p_{\mathrm{ol}}(W|\mathcal{M}) := \frac{1}{A_{\mathrm{ol}}}\frac{p(W|0;\mathcal{M})p(W|1;\mathcal{M})}{\frac{1}{2}[p(W|0;\mathcal{M})+p(W|1;\mathcal{M})]}, \tag{39}$$

where the normalization constant A_{ol} reads

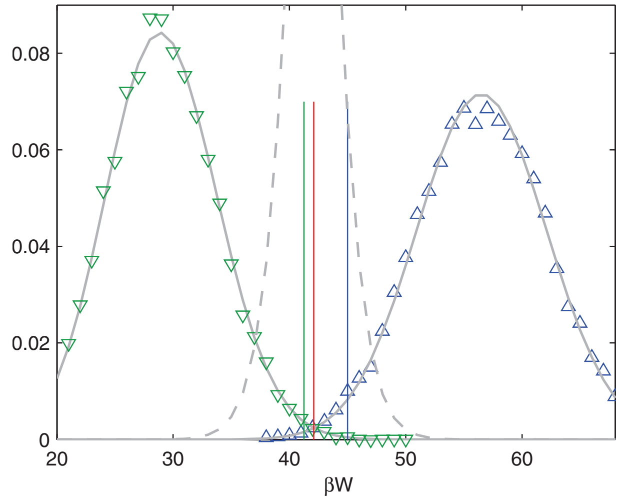

FIG. 1. (Color online) Targeted work probability distributions for the expansion of a cavity in an ideal gas and the associated overlap distribution. The up (down) triangles display the normalized histogram of a sample of forward (reverse) work values. The smooth solid curves are the exact analytic work distributions $p(W|0;\mathcal{M})$ (right) and $p(W|1;\mathcal{M})$ (left), and the dashed curve shows their overlap distribution $p_{ol}(W|\mathcal{M})$. The straight vertical lines show the values of the targeted estimates of ΔF on the abscissa. From left to right: the reverse, the two-sided (which is indistinguishable from the exact analytic value), and the forward estimate.

$$A_{ol} = \int \frac{p(W|0;\mathcal{M})p(W|1;\mathcal{M})}{\frac{1}{2}[p(W|0;\mathcal{M})+p(W|1;\mathcal{M})]}dW$$

$$= \int \frac{\tilde{\rho}_0(y)\rho_1(y)}{\frac{1}{2}[\tilde{\rho}_0(y)+\rho_1(y)]}dy. \tag{40}$$

A_{ol} is a measure for the overlap area of the distributions and takes its maximum value 1 in case of coincidence. Using the fluctuation theorem (19), the two-sided overlap measure can be written

$$\frac{1}{2}A_{ol} = \left\langle \frac{1}{1+e^{\beta(\Delta F-W)}} \right\rangle_1 = \left\langle \frac{1}{1+e^{-\beta(\Delta F-W)}} \right\rangle_0. \tag{41}$$

Comparing Eq. (41) with the two-sided targeted free-energy estimator (29), one sees that the two-sided targeted free-energy estimation method readily estimates the two-sided overlap measure. The accuracy of the estimate depends on how good the sampled work values reach into the main part of the overlap distribution $p_{ol}(W|\mathcal{M})$. By construction, the overlap region is sampled far earlier than the further distant tail that lies in the peak of the other distribution, cf. Fig. 1. This is the reason why the two-sided estimator is superior if compared to the one-sided estimators.

In the large N limit the mean square error $X_{01}(N) = \langle(\widehat{\Delta F}_{01}-\Delta F)^2\rangle$ of the two-sided estimator can be expressed in terms of the overlap measure and reads

$$X_{01}(N) = \frac{4}{N}\left(\frac{1}{A_{ol}}-1\right), \tag{42}$$

cf. [13,15]. Note that if an estimated value $\hat{A}_{ol}$ is plugged in, this formula is valid in the limit of large N only, but it is not clear *a priori* when this limit is reached. Therefore, we develop a simple convergence criterion for the two-sided estimate.

VI. CONVERGENCE

In this section, a measure for the convergence of the two-sided estimate is developed, again for the special case $n_0 = n_1 = \frac{N}{2}$. First, we define the estimate $\hat{A}_{ol}$ of the overlap measure A_{ol} with

$$\frac{1}{2}\hat{A}_{ol}(N) = \frac{1}{n_1}\sum_{j=1}^{n_1}\frac{1}{1+e^{\beta(\widehat{\Delta F}_{01}-W_j^1)}}, \tag{43}$$

which is equal to $\frac{1}{n_0}\Sigma_{i=1}^{n_0}\frac{1}{1+e^{-\beta(\widehat{\Delta F}_{01}-W_i^0)}}$, as we understand the estimate $\widehat{\Delta F}_{01}$ to be obtained according to Eq. (29) with the same samples of forward and reverse work values. Since the accuracy of the estimated value $\hat{A}_{ol}$ is unknown, we need an additional quantity to compare with.

Another expression for the overlap measure is

$$\frac{1}{2}A_{ol} = \left\langle \left(\frac{1}{1+e^{\beta(\Delta F-W)}}\right)^2 \right\rangle_1 + \left\langle \left(\frac{1}{1+e^{-\beta(\Delta F-W)}}\right)^2 \right\rangle_0, \tag{44}$$

which can be verified with the fluctuation theorem (19). Based on Eq. (44), we define the overlap estimator of second order,

$$\frac{1}{2}\hat{A}_{ol}^{(II)}(N) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{1}{1+e^{\beta(\widehat{\Delta F}_{01}-W_j^1)}}\right)^2 + \frac{1}{n_0}\sum_{i=1}^{n_0}\left(\frac{1}{1+e^{-\beta(\widehat{\Delta F}_{01}-W_i^0)}}\right)^2. \tag{45}$$

Because $\widehat{\Delta F}_{01}$ converges to ΔF, both $\hat{A}_{ol}$ and $\hat{A}_{ol}^{(II)}$ converge to A_{ol} in the limit $N\to\infty$. However, the second-order estimator $\hat{A}_{ol}^{(II)}$ converges slower and is for small N typically much smaller than $\hat{A}_{ol}$, since the main contributions to the averages appearing in Eq. (44) result from work values that lie somewhat further in the tails of the work distributions.

We use the relative difference

$$a(N) = \frac{\hat{A}_{ol}-\hat{A}_{ol}^{(II)}}{\hat{A}_{ol}} \tag{46}$$

to quantify the convergence of the two-sided estimate $\widehat{\Delta F}_{01}$, where $\hat{A}_{ol}$, $\hat{A}_{ol}^{(II)}$, and $\widehat{\Delta F}_{01}$ are understood to be calculated with the same two samples of forward and reverse work values.

From Eqs. (45), (43), and (29) follows that $0 \leqslant \hat{A}_{ol}^{(II)} \leqslant 2\hat{A}_{ol}$ holds. Hence, the convergence measure $a(N)$ is bounded by

$$-1 \leqslant a(N) \leqslant 1 \tag{47}$$

for any N. A necessary convergence condition is $a(N) \rightarrow 0$. This means that only if $a(N)$ is close to zero, the two-sided overlap estimators can have converged. Typically, $a(N)$ being close to zero is also a sufficient convergence condition. Hence, if $a(N)$ is close to zero, the mean square error of $\widehat{\Delta F}_{01}$ is given by Eq. (42) with $A_{\mathrm{ol}} \approx \hat{A}_{\mathrm{ol}}$. As can be seen from Eq. (42), the mean square error and in turn the variance and the bias are reduced by both, by taking a larger sample size N and by choosing a map $\mathcal{M}$ that enhances the overlap of $\tilde{\rho}_0$ with ρ_1.

With the targeted free-energy estimators at hand, together with their mean square errors, we are now ready to compute free-energy differences numerically.

VII. NUMERICAL EXAMPLES

We investigate two numerical applications. One is the free-energy difference of a fluid subject to the expansion of a cavity which allows the comparison with published results [7]. The other is the chemical potential of a fluid in the high density regime.

Beneath an ideal gas, the fluid is chosen to be a Lennard-Jones fluid with pairwise interaction

$$V(r_{kl}) = 4\epsilon\left[\left(\frac{\sigma}{r_{kl}}\right)^{12} - \left(\frac{\sigma}{r_{kl}}\right)^{6}\right], \tag{48}$$

where r_{kl} is the distance between the kth and lth particle, $r_{kl} = |\mathbf{r}_k - \mathbf{r}_l|$. The parameters used are those of argon, $\sigma = 3.542$ Å, and $\epsilon/k = 93.3$ K [24].

In all applications, the samples from the densities ρ_0 and ρ_1 are simulated with the Metropolis algorithm [23]. In order to simulate macroscopic behavior with a small number N_p of particles, periodic boundary conditions and the minimum image convention [6] are used. Pairwise interactions are truncated at half of the box length $R_{\mathrm{box}} = L/2$, but are not shifted, and the appropriate cutoff corrections are applied [6].

A. Expansion of a cavity in a fluid

The expansion of a cavity in a fluid is given by the following setup: Consider a fluid of N_p point molecules with pairwise interaction $V(r_{kl})$ confined in a cubic box of side length $2R_{\mathrm{box}}$, but excluded from a sphere of radius $R \leqslant R_{\mathrm{box}}$, compare with Fig. 2. Both the box and the sphere are centered at the origin $\mathbf{r} = 0$. A configurational microstate of the system is given by a set $x = (\mathbf{r}_1, \ldots, \mathbf{r}_{N_p})$ of particle positions $\mathbf{r}_k$. Growing the sphere from $R = R_0$ to $R = R_1$ decreases the volume accessible to the particles and the fluid is compressed. We are interested in the increase of free-energy ΔF subject to the compression of the fluid. Since the kinetic contribution to the free-energy is additive and independent of R, the difference ΔF depends only on the configurational part of the Hamiltonian. The latter reads

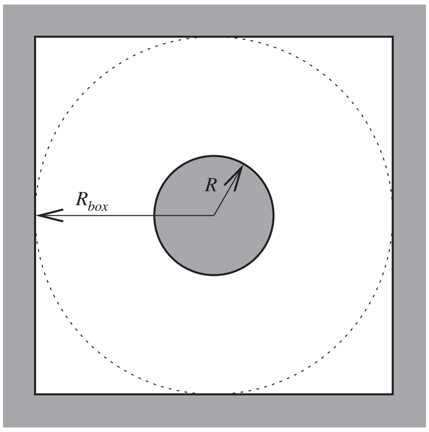

FIG. 2. The geometric setup.

$$H_i(x) = \begin{cases} \sum\limits_{k<l} V(r_{kl}) & \text{if } x \in \Gamma_i, \\ \infty & \text{if } x \notin \Gamma_i, \end{cases} \tag{49}$$

with $i = 0, 1$. Γ_0 and Γ_1 denote the accessible parts of configuration space of the system 0 ($R = R_0$) and 1 ($R = R_1$), respectively. We assume that $R_0 < R_1$ holds which implies $\Gamma_1 \subset \Gamma_0$.

Drawing a sample $\{x_k\}$ from ρ_0 and applying the traditional forward estimator (4) results in the following: $e^{-\beta\Delta H(x_k)}$ takes the values one and zero depending on whether $x_k \in \Gamma_1$ or not, i.e., whether the region between the two spheres of radius R_0 and R_1 is found vacant of particles or not. A comparison with Eq. (3) reveals that $e^{-\beta\Delta F}$ is the probability for the spherical shell being observed devoid of particles [7]. Hence, the rate of convergence of $\overline{e^{-\beta\Delta H}}$ decreases with the latter probability and will in general be poor.

Conversely, drawing a sample y_k from ρ_1 and applying the traditional reverse estimator $\widehat{\Delta F}_1^{\mathrm{trad}} = \frac{1}{\beta}\overline{e^{\beta\Delta H(y)}}$ [Eq. (16) with $\phi(x) = x$] figures out to be invalid, because the term $e^{\beta\Delta H(y_k)}$ takes always the value one. In consequence, the traditional reverse estimator is inconsistent. The deeper reason for this is that $\Gamma_1 \subset \Gamma_0$ holds: Eq. (2) is valid only for $x \in \Gamma_1$. By the same reason, the traditional two-sided estimator is inconsistent, too.

The mentioned shortcomings are avoided with a well chosen target map. Consider mapping each particle separately according to

$$\phi(x) = \left(\psi(r_1)\frac{\mathbf{r}_1}{r_1}, \ldots, \psi(r_{N_p})\frac{\mathbf{r}_{N_p}}{r_{N_p}}\right), \tag{50}$$

where $r_k = |\mathbf{r}_k|$ is the distance of the kth particle with respect to the origin, and $\psi: (R_0, R_{\max}] \rightarrow (R_1, R_{\max}]$ is a bijective and piecewise smooth radial mapping function. In order not to map particles out of the confining box, it is required that $\psi(r) = r$ holds for $r > R_{\mathrm{box}}$. The Jacobian for the radial map (50) reads

$$\left|\frac{\partial\phi}{\partial x}\right| = \prod_{j=1}^{N_p} \frac{\psi(r_j)^2}{r_j^2}\frac{\partial\psi(r_j)}{\partial r_j}. \tag{51}$$

(This formula is immediately clear when changing to polar coordinates.) We use the map of Ref. [7] which is designed to uniformly compress the volume of the shell $R_0 < r \leq R_{\text{box}}$ to the volume of the shell $R_1 < r \leq R_{\text{box}}$. Thus, for $r \in (R_0, R_{\text{box}}]$ the radial mapping function $\psi(r)$ is defined by

$$\psi(r)^3 - R_1^3 = c(r^3 - R_0^3), \tag{52}$$

with the compression factor $c = (R_{\text{box}}^3 - R_1^3)/(R_{\text{box}}^3 - R_0^3)$. According to Eq. (51), we have $\ln K(x) = \nu(x) \ln c$, where $\nu(x)$ is the number of particles in the shell $R_0 < r \leq R_{\text{box}}$.

1. Ideal gas

As a first illustrative and exact solvable example we choose the fluid to be an ideal gas, $V(r_{kl}) = 0$. In this case the free-energy difference is solely determined by the ratio of the confined volume $V_i = 8R_{\text{box}}^3 - \frac{4}{3}\pi R_i^3$, $i = 0, 1$, and is given by $\beta\Delta F = -N_p \ln(V_1/V_0)$. Using the radial map (52), the work in the forward direction as a function of x reads $\widehat{\Delta H}(x) = -\frac{1}{\beta}\nu(x) \ln c$ and takes discrete values only, as $\nu(x) = n$ holds with $n \in \{0, 1, \ldots, N_p\}$. Consequently, the probability $p(W_n | 0; \mathcal{M})$ of observing the work $W_n = -\frac{n}{\beta} \ln c$ in forward direction is binomial,

$$p(W_n|0;\mathcal{M}) = \binom{N_p}{n} q_0^n (1 - q_0)^{N-n}, \tag{53}$$

where $q_0 = \frac{4}{3}\pi(R_{\text{box}}^3 - R_0^3)/V_0$ is the probability of any fixed particle to be found in the shell $R_0 < r \leq R_{\text{box}}$. In analogy, the probability distribution $p(W_n | 1; \mathcal{M})$ for observing the work $W = W_n$ in reverse direction is given by replacing the index 0 with 1 in Eq. (53). Finally, the work probability distributions (rather then the densities) obey the fluctuation theorem (19) for any $n = 0, 1, \ldots, N_p$,

$$\frac{p(W_n|0;\mathcal{M})}{p(W_n|1;\mathcal{M})} = \frac{1}{c^n}\left(\frac{V_1}{V_0}\right)^{N_p} = e^{\beta(W_n - \Delta F)}. \tag{54}$$

A simple numerical evaluation highlights the convergence properties. Choosing the parameters to be $2R_{\text{box}} = 22.28$ Å, $R_0 = 7$ Å, $R_1 = 10$ Å, and $N_p = 125$ (β arbitrary), the free-energy difference takes the value $\beta\Delta F = 42.1064$. Because $e^{-\beta\widehat{\Delta H}(x)}$ can take only the numbers zero and one, the probability of observing a configuration x with nonvanishing contribution in the traditional forward estimator of ΔF is $e^{-\beta\Delta F} \approx 10^{-19}$. Hence, in practice it is impossible to use the traditional method successfully, since it would require at least $N_p \times 10^{19}$ Monte Carlo trial moves. However, the targeted approach already gives reasonable estimates with a sample size of just a few thousands. Figure 1 shows estimates of the targeted work probability distributions for samples of size $n_i = 10^4$ $(i = 0, 1)$ from ρ_0 and ρ_1 each. While the forward distribution $p(W | 0; \mathcal{M})$ is obviously well sampled in the central region, the sampling size is too small in order to reach the small values of βW where the reverse distribution $p(W | 1; \mathcal{M})$ is peaked. Exactly the latter values would be required for an *accurate* exponential average in the targeted forward estimator, Eq. (14). Therefore, the targeted forward estimate of ΔF is still inaccurate; it yields $\beta\widehat{\Delta F}_0 = 45.0 \pm 0.3$. The same is true for the targeted reverse estimate (16) which gives $\beta\widehat{\Delta F}_1 = 41.3 \pm 0.5$. The errors are calculated using root mean squares and propagation of uncertainty. A more accurate estimate follows from the targeted two-sided estimator (29) which yields $\beta\widehat{\Delta F}_{01} = 42.1 \pm 0.1$ ($n_0 = n_1 = 10^4$). This is clear, as for the two-sided estimate it is sufficient yet that the forward and reverse work values sample the region where the overlap distribution $p_{\text{ol}}(W | \mathcal{M})$, Eq. (39), is peaked, which is obviously the case, cf. Fig. 1.

The ideal gas is an exactly solvable model. This raises the question of whether a "perfect" or an ideal map can be constructed. The answer is yes, however such an ideal map would not be in the set of radial maps as defined with Eq. (50). Instead, the ideal map would also depend on the angles and would have a more complicated structure. The reason for this is the geometry of the simulation box: An ideal map needs to compress the fluid of uniform density with $R = R_0$ to the fluid of uniform density with $R = R_1$. The radial mapping function $\psi(r)$, Eq. (52), can be viewed as a good approximation to the ideal map within the set of radial maps.

2. Lennard-Jones fluid

We now focus on particles with Lennard-Jones interaction (48). The parameters are chosen to coincide with those of Ref. [7], i.e., $2R_{\text{box}} = 22.28$ Å, $R_0 = 9.209$ Å, $R_1 = 9.386$ Å, $N_p = 125$, and $T = 300$ K. In Lennard-Jones units, the reduced densities $\rho_i^* = \sigma^3 N_p / V_i$ of the systems 0 ($R = R_0$) and 1 ($R = R_1$) are $\rho_0^* = 0.713$ and $\rho_1^* = 0.731$, respectively, and $T^* = 1/(\beta\epsilon) = 3.215$ holds for both. If we had an ideal gas, the probability of observing the space between the spheres of radius R_0 and R_1 to be vacant of particles would be $(V_1/V_0)^{N_p} = 0.044$. Because of the strong repulsive part of interaction, this probability is much smaller in case of a dense Lennard-Jones fluid.

We generate samples of $\rho_0(x)$ and $\rho_1(x)$ with a Metropolis Monte Carlo simulation. Each run starts with 1000 equilibration sweeps, followed by the production run. In the production run the configurational microstate x is being sampled every fourth sweep only in order to reduce correlations between successive samples. The use of decorrelated data is of particular importance for the self-consistent two-sided estimate $\widehat{\Delta F}_{01}$ because it depends intrinsically on the ratio $\frac{n_1}{n_0}$ of the numbers of *uncorrelated* samples, cf. Eq. (29).

Figure 3 gives an overview of independent runs with different sample sizes N, where the one- and two-sided targeted estimators can be compared with each other and with the traditional forward estimator. Displayed is the estimated mean $\overline{\widehat{\Delta F}(N)}$ in dependence of the sample size N. The error bars reflect the estimated standard deviation

$$\overline{[\widehat{\Delta F}(N) - \overline{\widehat{\Delta F}(N)}]^2}^{1/2}.$$

Each mean and each standard deviation is estimated using $z(N)$ independent estimates $\widehat{\Delta F}(N)$. In ascending order of N,

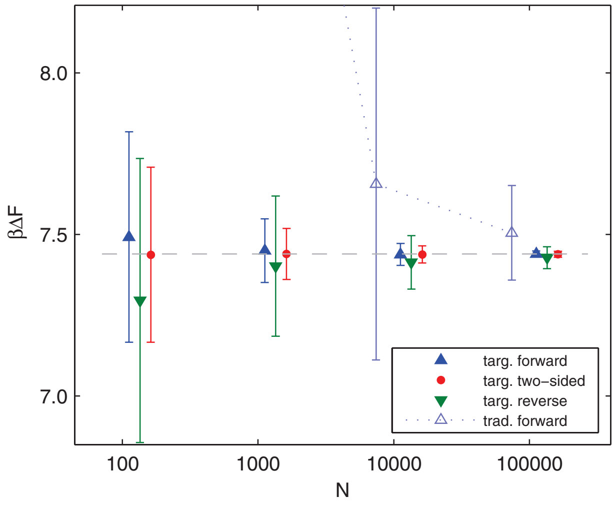

FIG. 3. (Color online) Free-energy estimates for the expansion of a cavity in a Lennard-Jones fluid. Shown are the average values of traditional and targeted estimates of ΔF in dependence of the sample size N, with an errorbar of one standard deviation. In order to distinguish the data points those corresponding to targeted estimates are shifted to the right and are spread, whereas those corresponding to traditional estimates are shifted to the left. For example, all four data points in the vicinity of $N=10\,000$ refer to $N=10\,000$. The dashed horizontal line represents a targeted two-sided estimate with $N=10^6$, see Table I.

$z(N)$ reads 440, 170, 40, 13. For the two-sided estimates, $n_0=n_1=\frac{N}{2}$ is used and Eq. (29) is solved.

Note that the theoretical mean of *traditional* forward estimates of ΔF is infinite *for any finite* N, because of the finite probability of observing a sequence of length N of solely vanishing contributions to the exponential average $\overline{e^{-\beta\Delta H}}$. Strictly spoken, the estimator $\beta\widehat{\Delta F}_0^{\rm trad}=-\ln \overline{e^{-\beta\Delta H}}$ is not well defined, because $\Gamma_1 \subset \Gamma_0$. Nevertheless, in Fig. 3 there are two finite *observed* mean values of traditional forward estimates displayed, what by no means is a contradiction. Infinite values are observed in the cases where $N<10^4$ holds. This is symbolized by the rising dotted line. The mentioned ill definiteness of the traditional estimator is removed by using the map (52). Figure 3 shows that all three targeted estimators are consistent even for small N in the sense that the error bars overlap. Whereas the targeted forward and reverse estimators show to be decreasingly biased with increasing N, the targeted two-sided estimator does not show any noticeable bias at all. This example demonstrates how worth it can be to take all three estimators, forward, reverse, and two-sided, into account. The one-sided estimators are biased in opposite directions and may serve as upper and lower bounds for ΔF, Eq. (23), whereas the two sided is typically placed in between the one sided.

We conclude this example with explicit estimates obtained from a single run with $N=10^6$, which are summarized in Table I. The errors are derived using block averages [25] and propagation of uncertainty.

B. Chemical potential of a homogeneous fluid

Consider a fluid of N_p particles confined within a cubic box of volume $V_c=(2R_{\rm box})^3$ with pairwise interaction $V(r_{ij})$. The configurational Hamiltonian for the N_p-particle system at $x=(\mathbf{r}_1,\ldots,\mathbf{r}_{N_p})$ reads

TABLE I. Cavity in a Lennard-Jones fluid. Estimated free-energy differences $\beta\widehat{\Delta F}$ for the expansion of a cavity, using targeted and traditional estimators. $N=10^6$.

Method	$\beta\widehat{\Delta F}$
Traditional forward	7.500 ± 0.043
Targeted forward	7.439 ± 0.003
Targeted two sided	7.440 ± 0.002
Targeted reverse	7.420 ± 0.009

$$H_{N_p}(x) = \sum_{i<j}^{N_p} V(r_{ij}). \tag{55}$$

The configurational density for the N_p system is given by

$$\rho_{N_p}(x) = e^{-\beta H_{N_p}(x)}/Z_{N_p}, \tag{56}$$

with the partition function $Z_{N_p}=\int e^{-\beta H_{N_p}(x)}dx$. Now consider one particle is added: the position of this new particle may be $\mathbf{r}_{N_p+1}$. The equilibrium density of the (N_p+1)-particle system reads

$$\rho_{N_p+1}(x) = e^{-\beta H_{N_p+1}(x,\mathbf{r}_{N_p+1})}/Z_{N_p+1}. \tag{57}$$

Taking the ratio of the densities (56) and (57) leads to Widom's particle insertion method [26] for estimating the excess chemical potential $\mu^{\rm ex}$ of the N_p system, defined as the excess of the chemical potential μ to that of an ideal gas at the same temperature and density. For sufficiently large N_p, $\mu^{\rm ex}$ can be approximated with

$$\mu^{\rm ex} = -\frac{1}{\beta}\ln\frac{Z_{N_p+1}}{Z_{N_p}V_c}. \tag{58}$$

Turning the tables, we use Eq. (58) to be the definition of the quantity $\mu^{\rm ex}$. The particle insertion method inserts at a random position an extra particle to the N_p system and measures the increase of energy that results from this particle. Since we consider a homogeneous fluid, we may as well fix the position of insertion arbitrarily, for instance at the origin, what is done in the following. We define system 1 through the configuration-space density $\rho_1(x)$ as follows:

$$\rho_1(x) = V_c\int \delta(\mathbf{r}_{N_p+1})\rho_{N_p+1}(x,\mathbf{r}_{N_p+1})d\mathbf{r}_{N_p+1}. \tag{59}$$

The factor V_c ensures normalization. Written in the usual form $\rho_1(x)=e^{-\beta H_1(x)}/Z_1$, we have

$$H_1(x) = H_{N_p}(x) + \sum_{k=1}^{N_p} V(r_k) \tag{60}$$

and $Z_1=Z_{N_p+1}/V_c$. System 1 can be understood as an equilibrium system of N_p interacting particles in the external potential $\Sigma_{k=1}^{N_p}V(r_k)$, due to one extra particle fixed at the origin $\mathbf{r}$

$=0$. Further, we identify system 0 with the N_p-particle system and rewrite

$$\rho_0(x) = \rho_{N_p}(x), \quad H_0(x) = H_{N_p}(x) \tag{61}$$

and $Z_0 = Z_{N_p}$. The ratio of ρ_0 and ρ_1 has the familiar form of Eq. (2), with ΔF being identical to μ^{ex},

$$\frac{\rho_0(x)}{\rho_1(x)} = e^{\beta[\Delta H(x) - \mu^{\text{ex}}]}. \tag{62}$$

The energy difference $\Delta H(x) = H_1(x) - H_0(x)$ is the increase of energy due to an added particle at the origin $\mathbf{r}=0$,

$$\Delta H(x) = \sum_{k=1}^{N_p} V(r_k). \tag{63}$$

Assume a finite potential $V(r)$ for nonvanishing r (i.e., no hard-core potential), but with a strong repulsive part for $r \to 0$ (a so-called soft-core potential), e.g., a Lennard-Jones potential. In this case, the configuration spaces of system 0 and 1 coincide, i.e., $\Gamma_0 = \Gamma_1$. Thus a traditional estimate of μ^{ex} is in principle valid in both directions, forward and reverse. In the forward direction we have the equivalent to the particle insertion method [26], $\widehat{\beta\mu}_0^{\text{extrad}} = -\ln \overline{e^{-\beta\Delta H(x)}}$, but with the fixed position of insertion $\mathbf{r}=0$. Here x is drawn from ρ_0 and we will typically find a particle in a sphere of radius $\bar{r}$ centered at the origin. $\bar{r}$ can roughly be estimated by the mean next-neighbor distance $(V_c/N_p)^{1/3}$ of an ideal gas. The dominant contributions to the exponential average come from realizations x that resemble typical realizations of system 1 [22]. However, typical realizations x of system 1 do not contain any particle within a sphere of some radius r_{hc} centered at the origin, because of the extra particle fixed at the origin and the strong repulsive part of the interaction. r_{hc} may be regarded as a temperature-dependent effective hard-core radius of the interaction $\beta V(r)$. We conclude that the insertion method is accurate and fast convergent only if $r_{hc}^3 \ll \bar{r}^3$, i.e., for low densities. Concerning the reverse traditional estimator $\widehat{\beta\mu}_1^{\text{extrad}} = \ln \overline{e^{\beta\Delta H(y)}}$, where y is drawn from ρ_1, the same argumentation reveals the impossibility of obtaining an accurate estimate in this way. Effectively, the particles of system 1 cannot access the vicinity of the origin, no matter how large the sample size will be. In this sense, Γ_1 can be substituted with an effective $\Gamma_1^{\text{eff}} \subset \Gamma_1 = \Gamma_0$, implying that the traditional reverse estimator tends to be inconsistent.

1. Constructing a map

Again, we use a map that changes each particle's distance to the origin separately, $\phi(x) = (\mathbf{R}_1, \ldots, \mathbf{R}_{N_p})$, with $\mathbf{R}_k = \psi(r_k)\frac{\mathbf{r}_k}{r_k}$. In searching a suitable radial mapping function $\psi(r)$, we are guided by the mean radial properties of the systems themselves. The radial probability density $g_0(r)$ of finding a particle in distance r from origin in system 0 is

$$g_0(r) = \frac{1}{N_p}\sum_{k=1}^{N_p} \int \delta(r_k - r)\rho_0(x)dx, \tag{64}$$

and that for system 1 is

$$g_1(r) = \frac{1}{N_p}\sum_{k=1}^{N_p} \int \delta(r_k - r)\rho_1(x)dx. \tag{65}$$

Due to the interaction with the extra particle fixed at the origin in system 1, $g_1(r)$ will in general be quite different from $g_0(r)$. The latter is related to a homogeneous fluid and is proportional to r^2 (for $r < R_{\text{box}}$), whereas the former refers to an inhomogeneous one and is proportional to $r^2 e^{-\beta V(r)}$ in the limit $r \to 0$ [26]. For large r, however, the influence of the extra particle vanishes and $g_1(r) \to g_0(r)$. Evaluation of the definition (64) of g_0 yields

$$g_0(r) = \frac{r^2}{V_c} h_0(r), \tag{66}$$

where $h_0(r)$ accounts for the decay of volume in the corners of the confining box and is given by $h_0(r) = \iint_{A(r)} \sin\theta d\phi d\theta$. The integration extends over the fraction of surface $A(r)$ of a sphere with radius r that lies inside the confining box. Note that $h_0(r) = 4\pi$ for $r < R_{\text{box}}$. In contrast to g_0, g_1 depends on the interaction $V(r)$. After some transformations of the right-hand side of Eq. (65), g_1 can be written

$$g_1(r) = \frac{r^2 e^{-\beta V(r)}}{V_c} h_1(r). \tag{67}$$

The function $h_1(r)$ can be written (cf. [26])

$$h_1(r) = e^{2\beta\mu^{\text{ex}}} h_0(r) \left\langle \exp\left(-\beta \sum_{k=1}^{N_p - 1} [V(r_k) + V(|\mathbf{r}_k - \mathbf{r}_{N_p}|)] \right) \right\rangle_{(N_p - 1)}, \tag{68}$$

where the angular brackets denote an average with a $N_p - 1$ particle density according to Eq. (56) and the vector $\mathbf{r}_{N_p}$ is arbitrarily fixed, but of magnitude r. Further, the approximation $V_c^2 Z_{N_p - 1}/Z_{N_p + 1} \approx e^{2\mu^{\text{ex}}}$ is used.

We note that the ratio of $g_1(r)/g_0(r)$ equals the well-known radial distribution function of the $N_p + 1$-particle fluid,

$$\Pi(r) = \frac{g_1(r)}{g_0(r)}. \tag{69}$$

Figure 4 shows estimates of g_0 and g_1 for a dense Lennard-Jones fluid with parameter values of argon [see below Eq. (48)], obtained from Monte Carlo simulations.

Now define a function $\psi^*(r)$ by requiring that it maps the mean radial behavior of system 0 to that of system 1. This is done by demanding

$$\int_0^{\psi^*(r)} g_1(t)dt = \int_0^r g_0(t)dt, \tag{70}$$

which yields

$$\frac{\partial\psi^*}{\partial r} = \frac{g_0(r)}{g_1(\psi^*(r))}. \tag{71}$$

In the limiting case of an ideal gas, $g_1 = g_0$ holds and the map becomes an identity, $\psi^*(r) = r$. Of practical interest are the

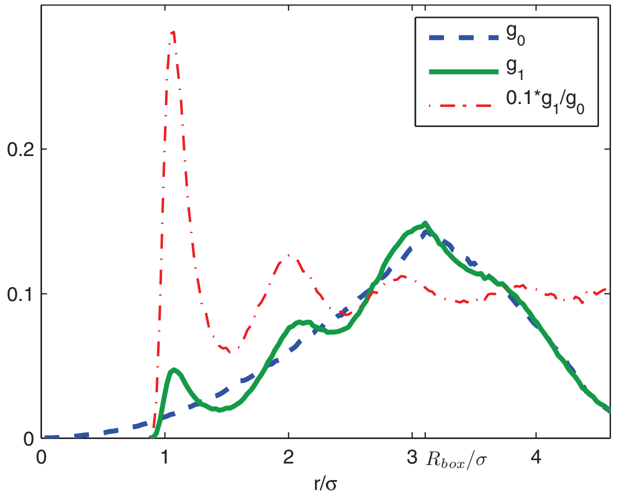

FIG. 4. (Color online) The radial densities $g_1(r)$ and $g_0(r)$ for a dense Lennard-Jones fluid ($\rho^*=0.9$ and $T^*=1.2$), estimated from simulated data. The ratio $g_1(r)/g_0(r)$ equals the radial distribution function.

cases where g_1 is unknown and thus Eq. (70) cannot be used to derive $\psi^*(r)$. However, the function ψ^* can be estimated with Monte Carlo simulations without knowledge of g_1 and g_0 as follows.

Take a sufficiently large amount n of samples $x_j=(\mathbf{r}_{1j},\ldots,\mathbf{r}_{N_pj})$, $j=1,\ldots,n$, drawn from $\rho_0(x)$ together with the same number of samples $y_j=(\mathbf{R}_{1j},\ldots,\mathbf{R}_{N_pj})$ drawn from $\rho_1(y)$. Calculate the distances to the origin $r_{ij}=|\mathbf{r}_{ij}|$ and $R_{ij}=|\mathbf{R}_{ij}|$ and combine all r_{ij} to the set $(r_a,r_b,r_c\ldots)$, as well as all R_{ij} to the set $(R_a,R_b,R_c,\ldots)$. Provided in both sets the elements are ordered ascending, $r_a\leqslant r_b\leqslant r_c\leqslant\ldots$ and $R_a\leqslant R_b\leqslant R_c\leqslant\ldots$, ψ^* is simulated by constructing a one-to-one correspondence $r_a\to R_a$, $r_b\to R_b,\ldots$ and estimating $\psi^*(r_\alpha)$ to be R_α, $\alpha=a,b,c,\ldots$. In effect, we have drawn the r_α and R_α from the densities $g_0(r)$ and $g_1(r)$, respectively, and have established a one-to-one correspondence between the ordered samples. We refer to this scheme as the simulation of the map of g_0 to g_1.

The solid curve shown in Fig. 5 is the result of a simulation of the function ψ^* for a Lennard-Jones fluid (parameters of argon, $\rho^*=0.9$, $T^*=1.2$). The corresponding densities g_0 and g_1 are plotted in Fig. 4. Noticeable is the sudden "start" of ψ^* with a value of roughly σ. This is due to the strong repulsive part of the interaction that keeps particles in system 1 approximately a distance σ away from the origin. Therefore, the behavior of $\psi^*(r)$ for $r\to0$ is not obtainable from finite-time simulations. However, the definition of ψ^* implies that for any soft-core potential $\psi^*(0)=0$ holds. To model ψ^* for small r, the limit $g_1(r)\xrightarrow{r\to0} ar^2e^{-\beta V(r)}4\pi/V_c$ can be used, where a is a constant. Thus, Eq. (70) can be written

$$[\psi^{*-1}(r)]^3 = 3a\int_0^r r'^2 e^{-\beta V(r')}dr' \tag{72}$$

in the limit $r\to0$, with ψ^{*-1} being the inverse of ψ^*. The constant a is in general unknown, but here it can be chosen

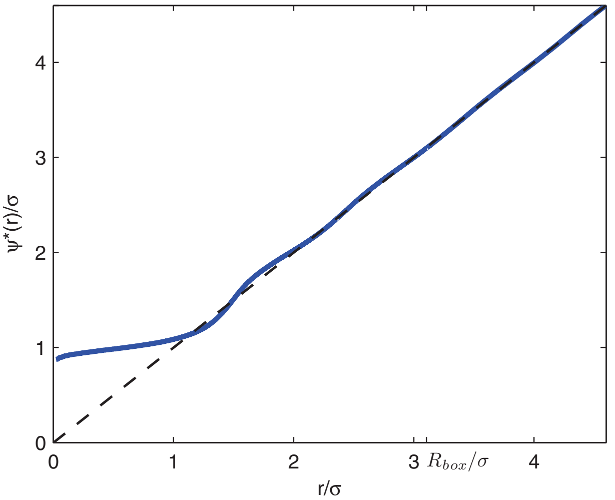

FIG. 5. (Color online) Simulated radial mapping function $\psi^*(r)$ for a dense Lennard-Jones fluid (solid). ψ^* maps the radial density $g_0(r)$ to $g_1(r)$, cf. Fig. 4. For the ideal gas, ψ^* is the identity map (dashed).

such that a continuous fit to the simulated part of ψ^{*-1} is obtained.

When the function ψ^* is used in the configuration space map ϕ according to Eq. (50), then, by definition of ψ^*, the radial density $\tilde{g}_0(r)$ of the mapped distribution $\tilde{\rho}_0(x)$, Eq. (8), is identical to the one of $\rho_1(x)$,

$$\begin{aligned}\tilde{g}_0(R) &:= \frac{1}{N_p}\sum_k \int \delta(|\mathbf{R}_k|-R)\tilde{\rho}_0(\phi)d\phi \\ &= \int \delta[\psi(r_1)-R]\rho_0(x)dx \\ &= \int\int \delta[\psi(r)-R]\delta(r_1-r)\rho_0(x)dxdr \\ &= \int \delta[\psi(r)-R]g_0(r)dr = g_1(R).\end{aligned} \tag{73}$$

Therefore we expect that the overlap of the mapped distribution $\tilde{\rho}_0$ with ρ_1 is larger than the overlap of the unmapped distribution ρ_0 with ρ_1. However, it must be noted that the use of ψ^* in the map ϕ is in general valid only in the limit of an infinite large system ($N,V_c\to\infty$; $N/V_c=$const), since we have not yet taken into account the requirement that particles may not be mapped out of the confining box. If R_{box} is chosen large enough, this might not be a serious problem, cf. Fig. 5.

2. Application of the radial map ψ^*

We now apply ψ^* and estimate the chemical potential of a dense Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$, parameters of argon) with $R_{\text{box}}=3.1056\sigma$ and $N_p=216$ particles. Configurations are drawn from ρ_0 and ρ_1 using a Metropolis algorithm with seven decorrelation sweeps between successive drawings. From every drawn configuration there results one value for the traditional work and one for the work related to

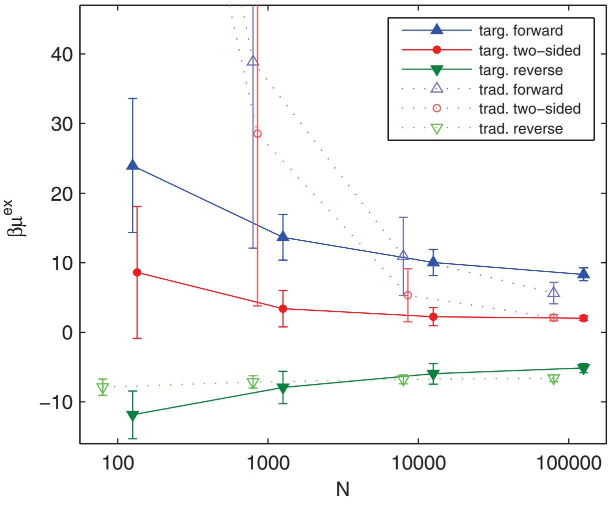

FIG. 6. (Color online) Targeted estimates of the excess chemical potential μ^{ex} of a dense Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$) compared to traditional estimates.

the map. The usual cutoff corrections [6] are applied. To avoid mapping particles out of the confining box, we simulate the map on the interval $0 \leqslant r \leqslant R_{\text{box}}$ subject to the condition $\psi^*(R_{\text{box}})=R_{\text{box}}$ and use $\psi^*(r)=r$ for $r>R_{\text{box}}$. The derivatives of ψ^* and ψ^{*-1} are obtained numerically. For the calculation of the work values in the simulation, the functions $\psi^*(r)$ and $\psi^{*-1}(r)$ as well as their derivatives are discretized in steps Δr with $R_{\text{box}}/\Delta r = 11\times 10^4$.

A comparison of the behavior of the targeted and traditional forward, reverse, and two-sided estimators in dependence of the sample size N is given in Fig. 6 (for the two-sided estimators $n_0=n_1=\frac{N}{2}$ is used). Each data point represents the average value of $z(N)$ independent estimates $\widehat{\mu^{\text{ex}}}(N)$. The error bars display one standard deviation. $z(N)$ reads $z(N)=450,250,45,5$ for $N=100,1000,10\,000,100\,000$, respectively.

As can be seen from Fig. 6, the traditional one-sided estimators behave quite different. The reverse estimator converges extremely slow in comparison to the forward estimator. This can be understood by comparing the average work values $\overline{W^i}$ in forward ($i=0$) and reverse ($i=1$) direction, see Table II. Since the absolute value of $\beta\Delta F=\beta\mu^{\text{ex}}$ is small, the traditional reverse estimator practically never converges, whereas for an accurate traditional forward estimate we need some 10^5 work values, cf. Eqs. (34) and (35). In contrast, the targeted one-sided estimators both show a similar convergence behavior if compared with each other. However, the convergence is slow.

The two-sided estimators converge much faster, in particular, the targeted two-sided estimator converges fastest, see Fig. 6. The convergence of the latter was checked with the convergence measure $a(N)$, Eq. (46). For instance, the convergence measure $a(N)$ takes the values 0.08 and 0.01 for the traditional and the targeted estimator, respectively, if $n_0 = n_1 = \frac{N}{2} = 10^5$ configurations per direction are sampled.

TABLE II. Estimated values of the mean forward and reverse work, obtained from $N=10^5$ sampled work values each.

	$\beta\overline{W^0}$	$\beta\overline{W^1}$
Traditional	10^{20}	-9.8
Targeted	10^5	-10^6

Investigating the histograms of the generalized work distributions in the traditional and the targeted case visualizes the effectiveness of the mapping. The histograms are similar to those displayed in Fig. 8 for $m=0$ (traditional) and $m=0.005$ (targeted).

A moderate gain in precision for the two-sided targeted estimator is found if compared to the precision of the two-sided traditional estimator which can be quantified with the overlap measure $\hat{A}_{\text{ol}}$ (43). Namely, $\hat{A}_{\text{ol}}=3.0\times 10^{-4}$ for the targeted case, and $\hat{A}_{\text{ol}}=2.2\times 10^{-4}$ for the traditional case.

We also studied other radial mapping functions ψ. Some of them turned out to give much better results and are easier to deal with.

3. Other radial mapping functions

The radial mapping function ψ^* was obtained from simulations, because the distribution $g_1(r)$ is analytically unknown. However, we are free to use any radial mapping function $\psi(r)$ and can thus in turn fix the function g_1 appearing in Eq. (70). To do this, we introduce the normalized, positive definite function $g_1'(r)$,

$$g_1'(r) = \frac{r^2}{c_1} e^{-\beta[V(r)+Q(r)]}, \quad r \in [0,R_{\text{box}}]. \tag{74}$$

$Q(r)$ is an arbitrary finite function over $(0,R_{\text{box}}]$ and $c_1 = \int_0^{R_{\text{box}}} r^2 e^{-\beta[V(r)+Q(r)]} dr$ a normalization constant. Further, let $g_0'(r)$ be a normalized quadratic density,

$$g_0'(r) = \frac{r^2}{c_0}, \quad r \in [0,R_{\text{box}}], \tag{75}$$

with $c_0 = R_{\text{box}}^3/3$.

The general (monotonically increasing) radial mapping function $\psi(r)$ can be expressed in terms of the equation

$$\int_0^{\psi(r)} g_1'(t)dt = \int_0^r g_0'(t)dt \tag{76}$$

for $r\in[0,R_{\text{box}}]$. For $r>R_{\text{box}}$ it shall be understood that $\psi(r)=r$. Given the function $Q(r)$, ψ and ψ^{-1} are determined uniquely by Eq. (76). An advantage of defining ψ with Eq. (76) is that the derivative $\partial\psi/\partial r$ is given in terms of V and Q,

$$\frac{\partial\psi(r)}{\partial r} = \frac{r^2}{\psi(r)^2} e^{\beta\{V(\psi(r))+Q(\psi(r))-f\}}, \tag{77}$$

with $f=-\frac{1}{\beta}\ln\frac{c_1}{c_0}$. Using ψ in the configuration space map $\phi(x)$ according to Eq. (50) yields the work function

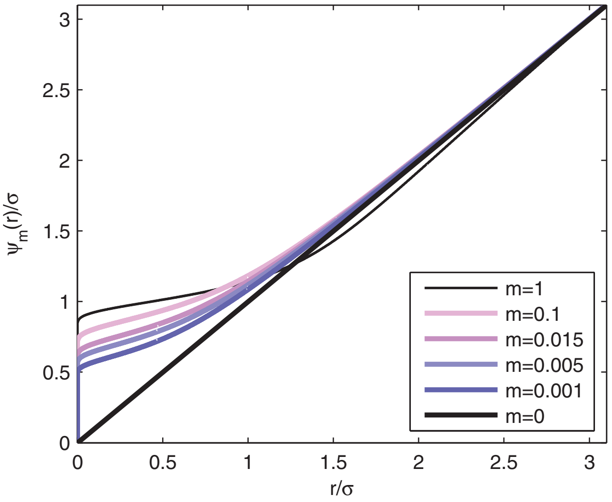

FIG. 7. (Color online) Members of the family of radial mapping functions ψ_m for the Lennard-Jones potential. For $m \rightarrow 0$, ψ_m converges to the identity map $\psi_0(r)=r$.

$$\widetilde{\Delta H}(x) = \sum_{i<j}^{N_p} \{V(|\mathbf{R}_i - \mathbf{R}_j|) - V(|\mathbf{r}_i - \mathbf{r}_j|)\} - \sum_{r_i \leqslant R_{\text{box}}} \{Q(\psi(r_i)) - f\}. \tag{78}$$

Here $\mathbf{R}_i$ is understood to be $\mathbf{R}_i = \psi(r_i)\frac{\mathbf{r}_i}{r_i}$, and the sum in the second line extends only over those particles for which $r \leqslant R_{\text{box}}$ holds. Note that the potential-energy contribution of the extra particle fixed at the origin is eliminated in the work function, due to the definition of ψ. However, in Eq. (78) we have already assumed $V(r)$ to be cut off at $r=R_{\text{box}}$, i.e., $V(r)=0$ for $r \geqslant R_{\text{box}}$. Otherwise we had to add $\Sigma_{r_i > R_{\text{box}}} V(\psi(r_i)) = \Sigma_{r_i > R_{\text{box}}} V(r_i)$ to the right-hand side of Eq. (78).

4. A family of maps

We now introduce a family $\{\psi_m\}$ of radial mapping functions, where each member ψ_m is defined by Eq. (76) with the choice

$$Q(r) = (m-1)V(r) \tag{79}$$

in the expression (74). This choice is motivated by the following: Consider a one particle system, $N_p=1$. In this case the optimal radial map can be computed analytically and results in $g_1'(r) = \frac{r^2}{c_1} e^{-\beta V(r)}$. We formally use this map for a system of $N_p \gg 1$ particles, but weaken the potential $V(r)$ by multiplying it with a small parameter m, i.e., $g_1'(r) = \frac{r^2}{c_1} e^{-\beta m V(r)}$, since the potential is screened by the $N_p - 1$ other particles. This results in Eq. (79). Useful maps are obtained for $m \in [0,1]$. Since we have no *a priori* knowledge on the optimal value of m, we determine the best value of m numerically.

Figure 7 depicts some members of the family $\{\psi_m\}$ for Lennard-Jones interaction (with parameters of argon). Again, we apply these functions discretized (in steps Δr with $R_{\text{box}}/\Delta r = 11 \times 10^4$) in the calculation of the targeted forward and reverse work $\widetilde{\Delta H}(x)$ and $\widetilde{\Delta H}(\phi^{-1}(x))$. Any pair of forward and reverse targeted work distributions belonging to the same value of m obeys the fluctuation theorem (19). In particular they cross at $W=\mu^{\text{ex}}$ ($\Delta F = \mu^{\text{ex}}$ here), see Fig. 9. Nevertheless, the shape of these distributions is sensitive to the value of m. This is demonstrated in Fig. 8. There, normalized histograms of βW are shown. They result from 10^4 work values per m and per direction. We emphasize that all of the targeted forward (reverse) work values were obtained with *one* sample of $N=10^4$ configurations x from ρ_0 (ρ_1). Figure 9

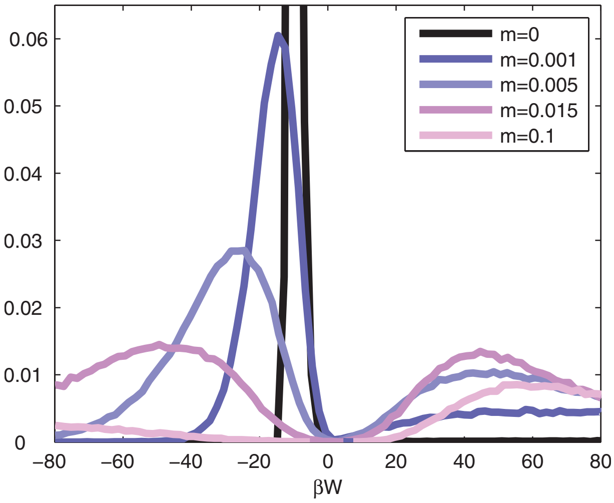

FIG. 8. (Color online) Forward (right) and reverse (left) work distributions of a Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$) for different radial mapping functions ψ_m. ($m=0$ results in the traditional work distributions.)

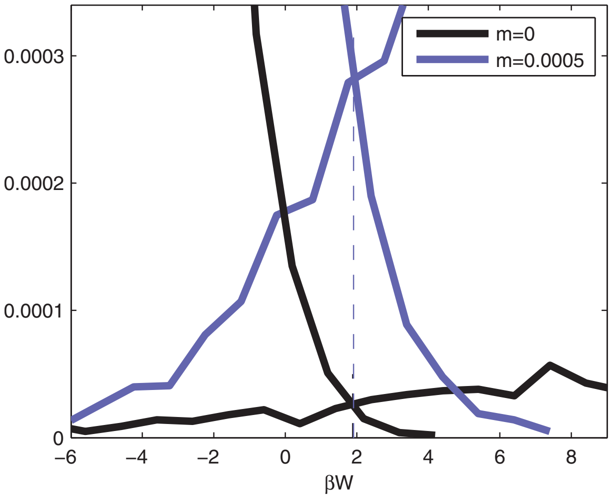

FIG. 9. (Color online) A detailed enlargement of forward (ascending lines) and reverse (descending lines) work distributions of the Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$) for two different radial mapping functions ψ_m. Notice the enhancement of overlap for $m=0.0005$. The vertical dashed line displays the estimated two-sided targeted value, $\widehat{\beta\mu^{\text{ex}}}_{01}=1.91$.

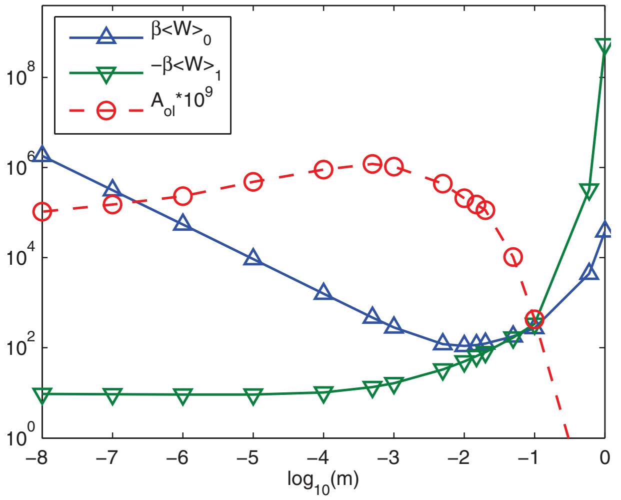

FIG. 10. (Color online) The average generalized work $\langle W\rangle_0$ and $\langle W\rangle_1$ in forward and reverse direction, respectively, and the two-sided overlap measure $A_{\rm ol}$ in dependence of the mapping parameter m. The forward dissipation is reduced up to 18 orders of magnitude if compared with the traditional dissipation, cf. Table II. Among the one-sided estimators the best is found for $m=0$ in forward direction. The optimal two-sided estimator results from using the m that maximizes $A_{\rm ol}$.

is a detailed enlargement where a sample of $n_0=n_1=10^6$ forward (reverse) configurations is used.

Instructive is the comparison of the mean work $\langle W\rangle$ related to different values of m. In Fig. 10 estimated values of mean work are shown in dependence of m. From these values one sees that the dissipation is minimal for $m=0$ in the reverse direction. Therefore, the best one-sided targeted estimate of $\mu^{\rm ex}$ among the family $\{\psi_m\}$ is obtained with $m=0$ in forward direction, i.e., with the traditional particle insertion. However, the same is not true for two-sided estimates. Using the same data as before and performing two-sided estimates with $n_0=n_1=\frac{N}{2}=10^4$ work values per direction, we obtain the displayed values $\widehat{\mu^{\rm ex}}_{01}$ of Fig. 11. In order to compare the performance of two-sided estimators for different maps, we estimate the overlap measures $A_{\rm ol}$. The latter are shown in Fig. 10. The maximum value for $A_{\rm ol}$ is found with m being 0.0005. This indicates that $m\approx 0.0005$ is the optimal choice for m. The estimates $\hat{A}_{\rm ol}$ are used to calculate the mean square errors X_{01} of the estimates $\widehat{\mu^{\rm ex}}_{01}$. The square roots of the X_{01} enter in Fig. 11 as error bars.

We are left to check the convergence properties of two-sided estimators. Figure 12 displays the convergence measure $a(N)$ for some parameter values m. Best convergence is found for $m=0.0005$ (not shown in Fig. 12, but very similar to $m=0.001$). The same value of the mapping parameter m was found to maximize the overlap $A_{\rm ol}$.

Employing the optimal value 0.0005 for the mapping parameter and using $n_0=n_1=\frac{N}{2}=10^6$ forward and reverse samples, we have computed the chemical potential. The results are given in Table III. The listed error is the square root of the X_{01} according to Eq. (42) with $A_{\rm ol}=\hat{A}_{\rm ol}$. This is justified with the observed values of the convergence measure a which are listed in the table, too.

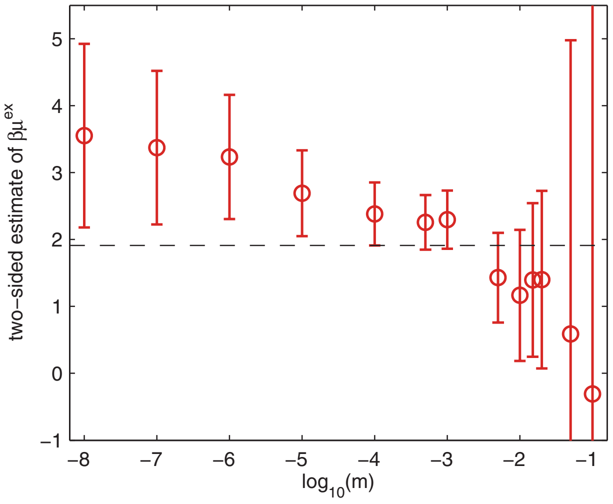

FIG. 11. (Color online) Two-sided estimates of $\mu^{\rm ex}$ as function of the mapping parameter m out of $n_0=n_1=\frac{N}{2}=10^4$ work values per direction for each m. The value of the traditional estimate ($m=0$) is $\widehat{\mu^{\rm ex}}_{01}=4.0\pm 2.0$. The error bars show the square root of the estimated mean square errors X_{01}. For comparison, the dashed line represents a two-sided estimate with $\frac{N}{2}=10^6$ and $m=0.0005$ (standard deviation 0.03).

It should be mentioned that the optimal value of m found here is not universal, but depends on the density ρ^*. If another value is chosen for ρ^*, the optimal m can again be found from numerical simulations. Note that the maps used here can be applied to simulations where particles are inserted and deleted at random [26], too. One simply has to use the point of insertion (deletion) as temporary origin of the coordinate system and apply the map there. This might enhance the efficiency of the simulation.

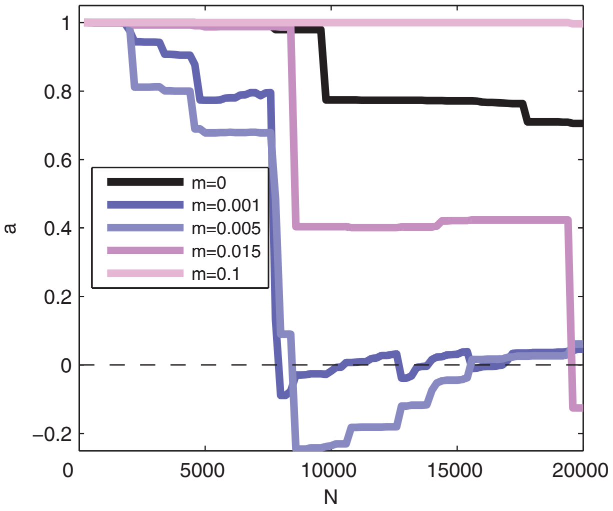

FIG. 12. (Color online) Convergence measure $a(N)$ of two-sided estimates for some parameter values m, depending on the sample size N. A faster decrease of a towards the value 0 indicates a faster convergence of the two-sided estimator.

TABLE III. Two-sided estimates $\widehat{\mu^{\mathrm{ex}}}_{01}$ of the excess chemical potential of a Lennard-Jones fluid ($\rho^*=0.9$, $T^*=1.2$). Also listed is the two-sided overlap measure A_{ol} and the convergence measure a. For the targeted estimate the radial mapping function ψ_m with $m=0.0005$ is used. The number of work values in each direction is 10^6 and the number of particles in the simulation is $N_p=216$.

	$\beta\widehat{\mu^{\mathrm{ex}}}_{01}$	$10^4\hat{A}_{\mathrm{ol}}$	a
Traditional	1.88 ± 0.08	2.4	0.05
Targeted	1.91 ± 0.03	19	−0.02

VIII. CONCLUSION

The central result of this paper, a fluctuation theorem for generalized work distributions, allowed us to establish an optimal targeted two-sided estimator of the free-energy difference ΔF. We have numerically tested this estimator and found it to be superior with respect to one-sided and nontargeted estimators. In addition we have demonstrated that this estimator can be applied successfully to estimate the chemical potential of a Lennard-Jones fluid in the high density regime.

In order to use the targeted two-sided estimator it is however crucial to use a suitable map. We have investigated the construction of maps and developed appropriate measures which enabled a quantitative comparison of the performance of different maps. Especially, a measure for the convergence of the two-sided estimate was designed. This points the way for better results when free-energy differences or chemical potentials need to be estimated numerically.

ACKNOWLEDGMENT

We thank Andreas Engel for helpful hints and discussions.

APPENDIX: CONSTRAINT MAXIMUM LIKELIHOOD DERIVATION OF THE TWO-SIDED ESTIMATOR

Deriving the optimal estimator of ΔF, given a collection of n_0 forward $\{W_i^0\}$ and n_1 reverse $\{W_j^1\}$ work values drawn from $p(W|0)$ and $p(W|1)$, respectively, leads to Bennett's acceptance ratio method [13] with the target map included.

In Sec. IV, the mixed ensemble is introduced, where the elements are given by pairs of values (W,Y) of work and direction, and which is specified by the probabilities of direction p_Y and the densities $p(W|Y)$. With the mixture ensemble, the mixing ratio $\frac{p_1}{p_0}$ can be chosen arbitrarily. Crucial about the mixture ensemble is that, according to the fluctuation theorem (19), the analytic form of the conditional probabilities $p(Y|W)$ can be derived explicitly, regardless of whether $p(W|Y)$ is known, see Sec. IV. This provides a natural way to construct a constraint maximum likelihood estimator [27–29] for ΔF.

Since it is only possible to draw from the ensembles $p(W|Y)$, but not from $p(Y|W)$, $Y=0,1$, the proper log-likelihood is

$$\ln\mathcal{L}=\sum_{i=1}^{n_0}\ln p(W_i^0|0)+\sum_{j=1}^{n_1}\ln p(W_j^1|1). \tag{A1}$$

A *direct* maximization of Eq. (A1) with respect to ΔF is impossible without knowledge of the analytic form of the probability densities $p(W|Y)$. However, according to Bayes theorem (24) the likelihood can be split into

$$\ln\mathcal{L}=\ln\mathcal{L}_{\mathrm{post}}(\Delta F)+\ln\mathcal{L}_{\mathrm{prior}}+\ln\mathcal{L}_{p_Y} \tag{A2}$$

with

$$\ln\mathcal{L}_{\mathrm{post}}(\Delta F)=\sum_{i=1}^{n_0}\ln p(0|W_i^0)+\sum_{j=1}^{n_1}\ln p(1|W_j^1), \tag{A3}$$

$$\ln\mathcal{L}_{\mathrm{prior}}=\sum_{k=1}^{n_0+n_1}\ln p(W_k), \tag{A4}$$

and

$$\ln\mathcal{L}_{p_Y}=n_0\ln\frac{1}{p_0}+n_1\ln\frac{1}{p_1}, \tag{A5}$$

where the sum in the prior likelihood (A4) runs over all n observed forward and reverse work values.

Since the definite form of $p(W)$ is unknown, we treat it in the manner of an unstructured prior distribution and maximize (A2) with respect to the constant ΔF *and* to the function $p(W)$ [29]. Thereby,

$$1=\int p(W)dW \tag{A6}$$

and

$$p_1=\int p(1|W)p(W)dW \tag{A7}$$

enter as constraints. Using Lagrange parameters λ and μ, the constrained likelihood reads

$$\ln\mathcal{L}^{\mathrm{c}}=\ln\mathcal{L}+\lambda\left(p_1-\int p(1|W)p(W)dW\right)+\mu\left(1-\int p(W)dW\right). \tag{A8}$$

The conditional direction probabilities $p(Y|W)$ are known explicitly in dependence of ΔF, Eq. (28), and their partial derivatives read $\frac{1}{\beta}\frac{\partial}{\partial\Delta F}\ln p(0|W)=-p(1|W)$ and $\frac{1}{\beta}\frac{\partial}{\partial\Delta F}\ln p(1|W)=p(0|W)=1-p(1|W)$. This allows one to extremize the constraint likelihood (A8) with respect to ΔF,

$$0=\frac{1}{\beta}\frac{\partial}{\partial\Delta F}\ln\mathcal{L}^{\mathrm{c}}=n_1-\sum_{k=1}^{n_0+n_1}p(1|W_k)-\lambda\int[1-p(1|W)]p(1|W)p(W)dW. \tag{A9}$$

Extremizing the conditional likelihood (A8) with respect to the function $p(W)$ gives

$$0 = \frac{\delta}{\delta p(W)} \ln \mathcal{L}^{c} = \frac{1}{p(W)} \sum_{k=1}^{n} \delta(W - W_k) - \lambda p(1|W) - \mu, \tag{A10}$$

which can be solved in $p(W)$,

$$p(W) = \frac{\Sigma_k \delta(W - W_k)}{\lambda p(1|W) + \mu}, \tag{A11}$$

or written as

$$\lambda p(1|W) p(W) = - \mu p(W) + \sum_k \delta(W - W_k). \tag{A12}$$

If interested in the values of the Lagrange multipliers λ and μ, one multiplies Eq. (A10) with $p(W)$ and integrates. This yields

$$0 = n - \lambda p_1 - \mu. \tag{A13}$$

A second independent equation follows from inserting Eq. (A12) into Eq. (A9) which results in

$$0 = n_1 + \mu - \mu p_1 - n, \tag{A14}$$

and the Lagrange multipliers take the values

$$\mu = \frac{n_0}{p_0} \quad \text{and} \quad \lambda = \frac{n p_0 - n_0}{p_0 p_1}. \tag{A15}$$

With the distribution (A11) the constraints (A6) and (A7) read

$$1 = \sum_k \frac{1}{\lambda p(1|W_k) + \mu} \tag{A16}$$

and

$$p_1 = \sum_k \frac{p(1|W_k)}{\lambda p(1|W_k) + \mu} = \frac{p_1}{n_1} \sum_k p^{\mathrm{B}}(1|W_k), \tag{A17}$$

where $p^{\mathrm{B}}(1|W)$ denotes $p(1|W)$ with $C = \Delta F + \frac{1}{\beta} \ln \frac{n_1}{n_0}$. Whenever the constraint (A17) is fulfilled, the constraint (A16) and the variational equations (A9) and (A10) are automatically satisfied. In consequence, Eq. (A17) defines the constrained maximum likelihood estimate of ΔF. Note that the estimator (A17) is *independent* of the choice of $\frac{p_1}{p_0}$. Moreover, Eq. (A17) is equivalent to Eq. (29) regardless of the choice of $\frac{p_1}{p_0}$.

An alternative derivation of the estimator (A17) was presented by Shirts *et al.* [15]. There, the specific choice $\frac{p_1}{p_0} = \frac{n_1}{n_0}$ was *necessary*. With this choice, the Lagrange parameter λ is identical to zero. Hence, there is no need to take any constraint into consideration and the posterior likelihood (A3) results directly in the estimator of ΔF.

---

[1] M. R. Reddy and M. D. Erion, *Free Energy Calculations in Rational Drug Design* (Kluwer Academic, New York, 2001).

[2] T. Schäfer and E. V. Shuryak, Rev. Mod. Phys. **70**, 323 (1998).

[3] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).

[4] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[5] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).

[6] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic Press, London, 2002).

[7] C. Jarzynski, Phys. Rev. E **65**, 046122 (2002).

[8] W. Lechner, H. Oberhofer, C. Dellago, and P. L. Geissler, J. Chem. Phys. **124**, 044113 (2006).

[9] H. Oberhofer, C. Dellago, and S. Boresch, Phys. Rev. E **75**, 061106 (2007).

[10] A. F. Voter, J. Chem. Phys. **82**, 1890 (1985).

[11] F. M. Ytreberg and D. M. Zuckerman, J. Phys. Chem. B **109**, 9096 (2005).

[12] S. Vaikuntanathan and C. Jarzynski, Phys. Rev. Lett. **100**, 190601 (2008).

[13] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).

[14] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).

[15] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).

[16] E. Schöll-Paschinger and C. Dellago, J. Chem. Phys. **125**, 054105 (2006).

[17] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[18] C. Jarzynski, J. Stat. Phys. **98**, 77 (2000).

[19] D. J. Evans, Mol. Phys. **101**, 1551 (2003).

[20] M. A. Cuendet, Phys. Rev. Lett. **96**, 120602 (2006).

[21] J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **100**, 12564 (2003).

[22] C. Jarzynski, Phys. Rev. E **73**, 046105 (2006).

[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[24] R. C. Reid, J. M. Prausnitz, and T. K. Sherwood, *The Properties of Gases and Liquids*, 3rd ed. (McGraw-Hill, New York, 1977), Appendixes A and C.

[25] H. Flyvbjerg and H. G. Petersen, J. Chem. Phys. **91**, 461 (1989).

[26] B. Widom, J. Chem. Phys. **39**, 2808 (1963).

[27] J. Aitchison and S. D. Silvey, Ann. Math. Stat. **29**, 813 (1958).

[28] J. A. Anderson, Biometrika **59**, 19 (1972).

[29] R. L. Prentice and R. Pyke, Biometrika **66**, 403 (1979).

Paper [2]

PHYSICAL REVIEW E **80**, 031111 (2009)

Characteristic of Bennett's acceptance ratio method

Aljoscha M. Hahn* and Holger Then
Institut für Physik, Carl von Ossietzky Universität, 26111 Oldenburg, Germany

(Received 12 June 2009; published 10 September 2009)

A powerful and well-established tool for free-energy estimation is Bennett's acceptance ratio method. Central properties of this estimator, which employs samples of work values of a forward and its time-reversed process, are known: for given sets of measured work values, it results in the best estimate of the free-energy difference in the large sample limit. Here we state and prove a further characteristic of the acceptance ratio method: the convexity of its mean-square error. As a two-sided estimator, it depends on the ratio of the numbers of forward and reverse work values used. Convexity of its mean-square error immediately implies that there exists a unique optimal ratio for which the error becomes minimal. Further, it yields insight into the relation of the acceptance ratio method and estimators based on the Jarzynski equation. As an application, we study the performance of a dynamic strategy of sampling forward and reverse work values.

DOI: 10.1103/PhysRevE.80.031111 PACS number(s): 02.50.Fz, 05.40.−a, 05.70.Ln

I. INTRODUCTION

A quantity of central interest in thermodynamics and statistical physics is the (Helmholtz) free energy, as it determines the equilibrium properties of the system under consideration. In practical applications, e.g., drug design, molecular association, thermodynamic stability, and binding affinity, it is usually sufficient to know free-energy differences. As recent progress in statistical physics has shown, free-energy differences, which refer to equilibrium, can be determined via nonequilibrium processes [1,2].

Typically, free-energy differences are beyond the scope of analytic computations and one needs to measure them experimentally or compute them numerically. Highly efficient methods have been developed in order to estimate free-energy differences precisely, including thermodynamic integration [3,4], free-energy perturbation [5], umbrella sampling [6–8], adiabatic switching [9], dynamic methods [10–12], asymptotics of work distributions [13], optimal protocols [14], targeted, and escorted free-energy perturbation [15–19].

A powerful [20–22] and frequently [23–25] used method for free-energy determination is two-sided estimation, i.e., Bennett's acceptance ratio method [26], which employs a sample of work values of a driven nonequilibrium process together with a sample of work values of the time-reversed process [27].

The performance of two-sided free-energy estimation depends on the ratio

$$r = \frac{n_1}{n_0} \tag{1}$$

of the number of forward and reverse work values used. Think of an experimenter who wishes to estimate the free-energy difference with Bennett's acceptance ratio method and has the possibility to generate forward as well as reverse work values. The capabilities of the experiment give rise to an obvious question: if the total amount of draws is intended to be $N=n_0+n_1$, which is the optimal choice of partitioning N into the numbers n_0 of forward and n_1 of reverse work values or, equivalently, what is the optimal choice r_o of the ratio r? The problem is to determine the value of r that minimizes the (asymptotic) mean-square error of Bennett's estimator when $N=n_0+n_1$ is held constant.

While known since Bennett [26], the optimal ratio is underutilized in the literature. Bennett himself proposed to use a suboptimal equal-time strategy, instead, because his estimator for the optimal ratio converges too slowly in order to be practicable. Even questions as fundamental as the existence and uniqueness are unanswered in the literature. Moreover, it is not always clear *a priori* whether two-sided free-energy estimation is better than one-sided exponential work averaging. For instance, Shirts and Pande presented a physical example where it is optimal to draw work values from only one direction [28].

The paper is organized as follows. In Secs. II and III we rederive two-sided free-energy estimation and the optimal ratio. We also remind that two-sided estimation comprises one-sided exponential work averaging as limiting cases for $\ln r \to \pm\infty$, a result that is also true for the mean-square errors of the corresponding estimators.

The central result is stated in Sec. IV: the asymptotic mean-square error of two-sided estimation is convex in the fraction $\frac{n_0}{N}$ of forward work values used. This fundamental characteristic immediately implies that the optimal ratio r_o exists and is unique. Moreover, it explains the generic superiority of two-sided estimation if compared with one sided, as found in many applications.

To overcome the slow convergence of Bennett's estimator of the optimal ratio, which is based on estimating second moments, in Sec. V we transform the problem into another form such that the corresponding estimator is entirely based on first moments, which enhances the convergence enormously.

As an application, in Sec. VII we present a dynamic strategy of sampling forward and reverse work values that maximizes the efficiency of two-sided free-energy estimation.

---

[*] Present address: Technische Universität Berlin, Institut für Theoretische Physik, 10623 Berlin, Germany.

©2009 The American Physical Society

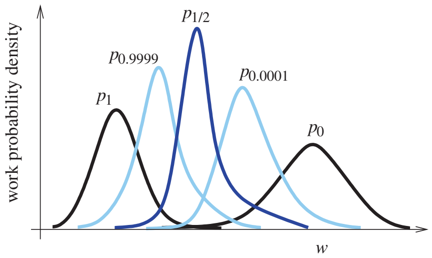

FIG. 1. (Color online) The overlap density $p_\alpha(w)$ bridges the densities $p_0(w)$ and $p_1(w)$ of forward and reverse work values, respectively. α is the fraction $\frac{n_0}{n_0+n_1}$ of forward work values, here schematically shown for $\alpha=0.0001$, $\alpha=0.5$, and $\alpha=0.9999$. The accuracy of two-sided free-energy estimates depends on how good $p_\alpha(w)$ is sampled when drawing from $p_0(w)$ and $p_1(w)$.

II. TWO-SIDED FREE-ENERGY ESTIMATION

Given a pair of samples of n_0 forward and n_1 reverse work values drawn from the probability densities $p_0(w)$ and $p_1(w)$ of forward and reverse work values and provided the latter are related to each other via the fluctuation theorem [2],

$$\frac{p_0(w)}{p_1(w)} = e^{w-\Delta f}, \tag{2}$$

Bennett's acceptance ratio method [20,26,27,29] is known to give the optimal estimate of the free-energy difference Δf in the limit of large sample sizes. Throughout the paper, $\Delta f = \Delta F/kT$ and $w=W/kT$ are understood to be measured in units of the thermal energy kT. The normalized probability densities $p_0(w)$ and $p_1(w)$ are assumed to have the same support Ω, and we choose the following sign convention: $p_0(w) := p_{\text{forward}}(+w)$ and $p_1(w) := p_{\text{reverse}}(-w)$.

Now define a normalized density $p_\alpha(w)$ with

$$p_\alpha(w) = \frac{1}{U_\alpha}\frac{p_0(w)p_1(w)}{\alpha p_0(w) + \beta p_1(w)}, \tag{3}$$

$w \in \Omega$, where $\alpha \in [0,1]$ is a real number and

$$\alpha + \beta = 1. \tag{4}$$

The normalization constant U_α is given by

$$U_\alpha = \int_\Omega \frac{p_0 p_1}{\alpha p_0 + \beta p_1} dw. \tag{5}$$

The density $p_\alpha(w)$ is a normalized harmonic mean of p_0 and p_1, $\frac{p_0p_1}{\alpha p_0+\beta p_1} = [\alpha\frac{1}{p_1}+\beta\frac{1}{p_0}]^{-1}$, and thus bridges between p_0 and p_1 (see Fig. 1). In the limit $\alpha \to 0$, $p_\alpha(w)$ converges to the forward work density $p_0(w)$ and, conversely, for $\alpha \to 1$ it converges to the reverse density $p_1(w)$. As a consequence of the inequality of the harmonic and arithmetic mean $[\alpha\frac{1}{p_1} + \beta\frac{1}{p_0}]^{-1} \leq \alpha p_1 + \beta p_0$, U_α is bounded from above by unity,

$$U_\alpha \leq 1, \tag{6}$$

$\forall \alpha \in [0,1]$. Except for $\alpha=0$ and $\alpha=1$, the equality holds if and only if $p_0 \equiv p_1$. Using the fluctuation theorem (2), U_α can be written as an average in p_0 and p_1,

$$U_\alpha = \left\langle \frac{1}{\alpha + \beta e^{-w+\Delta f}} \right\rangle_1 = \left\langle \frac{1}{\beta + \alpha e^{w-\Delta f}} \right\rangle_0, \tag{7}$$

where the angular brackets with subscript $\gamma \in [0,1]$ denote an ensemble average with respect to p_γ, i.e.,

$$\langle g \rangle_\gamma = \int_\Omega g(w) p_\gamma(w) dw, \tag{8}$$

for an arbitrary function $g(w)$.

In setting $\alpha=1$, Eq. (7) reduces to the nonequilibrium work relation [1],

$$1 = \langle e^{-w+\Delta f} \rangle_0, \tag{9}$$

in the *forward* direction and, conversely, with $\alpha=0$ we obtain the nonequilibrium work relation in the *reverse* direction,

$$1 = \langle e^{w-\Delta f} \rangle_1. \tag{10}$$

The last two relations can, of course, be obtained more directly from the fluctuation theorem (2). An important application of these relations is the *one-sided* free-energy estimation. Given a sample $\{w_1^0 \ldots w_N^0\}$ of N forward work values drawn from p_0, Eq. (9) is commonly used to define the *forward* estimate $\widehat{\Delta f}_0$ of Δf with

$$\widehat{\Delta f}_0 = -\ln \frac{1}{N}\sum_{k=1}^{N} e^{-w_k^0}. \tag{11}$$

Conversely, given a sample $\{w_1^1 \ldots w_N^1\}$ of N reverse work values drawn from p_1, Eq. (10) suggests the definition of the *reverse* estimate $\widehat{\Delta f}_1$ of Δf,

$$\widehat{\Delta f}_1 = \ln \frac{1}{N}\sum_{l=1}^{N} e^{w_l^1}. \tag{12}$$

If we have drawn both, a sample of n_0 forward *and* a sample of n_1 reverse work values then Eq. (7) can serve us to define a two-sided estimate $\widehat{\Delta f}$ of Δf by replacing the ensemble averages with sample averages,

$$\frac{1}{n_1}\sum_{l=1}^{n_1} \frac{1}{\alpha + \beta e^{-w_l^1 + \widehat{\Delta f}}} = \frac{1}{n_0}\sum_{k=1}^{n_0} \frac{1}{\beta + \alpha e^{w_k^0 - \widehat{\Delta f}}}. \tag{13}$$

$\widehat{\Delta f}$ is understood to be the unique root of Eq. (13), which exists for any $\alpha \in [0,1]$. Different values of α result in different estimates for Δf. Choosing

$$\alpha = \frac{n_0}{N}, \quad \beta = \frac{n_1}{N}, \tag{14}$$

$N=n_0+n_1$, the estimate (13) coincides with Bennett's optimal estimate, which defines the two-sided estimate with a least asymptotic mean-square error for a given value $\alpha=\frac{n_0}{N}$ or, equivalently, *for a given ratio* $r=\frac{\beta}{\alpha}=\frac{n_1}{n_0}$ [20,26]. We denote the optimal two-sided estimate, i.e., the solution of Eq. (13) under the constraint (14), by $\widehat{\Delta f}_{1-\alpha}$ and simply refer to it as the two-sided estimate. Note that the optimal estimator can be written in the familiar form

$$\sum_{l=1}^{n_1} \frac{1}{1+e^{-w_l^1+\widehat{\Delta f}+\ln n_1/n_0}} = \sum_{k=1}^{n_0} \frac{1}{1+e^{w_k^0-\widehat{\Delta f}-\ln n_1/n_0}}. \tag{15}$$

In the limit $\alpha=\frac{n_0}{N}\to 1$, the two-sided estimate reduces to the one-sided forward estimate (11), $\widehat{\Delta f}_{1-\alpha} \xrightarrow{\alpha\to 1} \widehat{\Delta f}_0$, and, conversely, $\widehat{\Delta f}_{1-\alpha} \xrightarrow{\alpha\to 0} \widehat{\Delta f}_1$. Thus, the one-sided estimates are the optimal estimates if we have given draws from only one of the densities p_0 or p_1.

A characteristic quantity to express the performance of the estimate $\widehat{\Delta f}_{1-\alpha}$ is the mean-square error,

$$\langle(\widehat{\Delta f}_{1-\alpha} - \Delta f)^2\rangle, \tag{16}$$

which depends on the total sample size $N=n_0+n_1$ and the fraction $\alpha=\frac{n_0}{N}$. Here, the average is understood to be an ensemble average in the value distribution of the estimate $\widehat{\Delta f}_{1-\alpha}$ for fixed N and α. In the limit of large n_0 and n_1, the asymptotic mean-square error X (which then equals the variance) can be written as [20,26]

$$X(N,\alpha) = \frac{1}{N}\frac{1}{\alpha\beta}\left(\frac{1}{U_\alpha} - 1\right). \tag{17}$$

Provided the right-hand side of Eq. (17) exists, which is guaranteed for any $\alpha \in (0,1)$, the N dependence of X is simply given by the usual $\frac{1}{N}$ factor, whereas the α dependence is determined by the function U_α given in Eq. (5). Note that if a two-sided estimate $\widehat{\Delta f}_{1-\alpha}$ is calculated then essentially the normalizing constant U_α is estimated from two sides 0 and 1 [cf. Eqs. (7) and (13)]. With an estimate $\widehat{\Delta f}_{1-\alpha}$, we therefore always have an estimate of the mean-square error at hand. However, the reliability of the latter naturally depends on the degree of convergence of the estimate $\widehat{\Delta f}_{1-\alpha}$. The convergence of the two-sided estimate can be checked with the convergence measure introduced in Ref. [19].

In the limits $\alpha=\frac{n_0}{N}\to 1$ and $\alpha\to 0$, respectively, the asymptotic mean-square error X of the two-sided estimator converges to the asymptotic mean-square error of the appropriate one-sided estimator [30],

$$\lim_{\alpha\to 1} X(N,\alpha) = \frac{1}{N}\mathrm{Var}_0\left(\frac{p_1}{p_0}\right) = \frac{1}{N}\mathrm{Var}_0(e^{-w+\Delta f}), \tag{18}$$

and

$$\lim_{\alpha\to 0} X(N,\alpha) = \frac{1}{N}\mathrm{Var}_1\left(\frac{p_0}{p_1}\right) = \frac{1}{N}\mathrm{Var}_1(e^{w-\Delta f}), \tag{19}$$

where Var_γ denotes the variance operator with respect to the density p_γ, i.e.,

$$\mathrm{Var}_\gamma(g) = \langle(g-\langle g\rangle_\gamma)^2\rangle_\gamma, \tag{20}$$

for an arbitrary function $g(w)$ and $\gamma \in [0,1]$.

III. THE OPTIMAL RATIO

Now we focus on the question raised in the introduction: which value α_o of α in the range [0,1] minimizes the mean-square error (17) when the total sample size $N=n_0+n_1$ is held fixed?

Let M be the rescaled asymptotic mean-square error given by

$$M(\alpha) = NX(N,\alpha), \tag{21}$$

which is a function of α only. Assuming $\alpha_o \in (0,1)$, a necessary condition for a minimum of M is that the derivative $M'(\alpha)=\frac{dM}{d\alpha}$ of M vanishes at α_o. Before calculating M' explicitly, it is beneficial to rewrite M by using the identity

$$\begin{aligned} U_\alpha &= \int_\Omega \frac{p_0 p_1(\alpha p_0+\beta p_1)}{(\alpha p_0+\beta p_1)^2}dw \\ &= \alpha\left\langle \frac{p_0^2}{(\alpha p_0+\beta p_1)^2}\right\rangle_1 + \beta\left\langle \frac{p_1^2}{(\alpha p_0+\beta p_1)^2}\right\rangle_0. \end{aligned} \tag{22}$$

Subtracting $(\alpha+\beta)U_\alpha^2=U_\alpha^2$ from Eq. (22) and recalling the definition (3) of p_α, one obtains

$$U_\alpha(1-U_\alpha) = [\alpha\theta_1(\alpha)+\beta\theta_0(\alpha)]U_\alpha^2, \tag{23}$$

where the functions θ_i are defined as

$$\theta_1(\alpha) = \mathrm{Var}_1\left(\frac{p_\alpha}{p_1}\right) = \frac{1}{U_\alpha^2}\mathrm{Var}_1\left(\frac{1}{\alpha+\beta e^{-w+\Delta f}}\right),$$

$$\theta_0(\alpha) = \mathrm{Var}_0\left(\frac{p_\alpha}{p_0}\right) = \frac{1}{U_\alpha^2}\mathrm{Var}_0\left(\frac{1}{\beta+\alpha e^{w-\Delta f}}\right). \tag{24}$$

θ_0 and θ_1 describe the relative fluctuations of the quantities that are averaged in the two-sided estimation of Δf [cf. Eq. (13)].

With the use of formula (23), M can be written as

$$M(\alpha) = \frac{\theta_0(\alpha)}{\alpha} + \frac{\theta_1(\alpha)}{\beta}, \tag{25}$$

and the derivative yields

$$M'(\alpha) = \frac{\theta_1(\alpha)}{\beta^2} - \frac{\theta_0(\alpha)}{\alpha^2} + \frac{\beta\theta_0'(\alpha)+\alpha\theta_1'(\alpha)}{\alpha\beta}. \tag{26}$$

The derivatives of the θ functions involve the first two derivatives of U_α, which will thus be computed first,

$$U_\alpha' := \frac{d}{d\alpha}U_\alpha = \int_\Omega \frac{p_0 p_1(p_1-p_0)}{(\alpha p_0+\beta p_1)^2}dw, \tag{27}$$

and

$$U_\alpha'' := \frac{d^2}{d\alpha^2}U_\alpha = 2\int_\Omega \frac{p_0 p_1(p_1-p_0)^2}{(\alpha p_0+\beta p_1)^3}dw. \tag{28}$$

From this equation, it is clear that U_α is convex in α, $U_\alpha'' \geq 0$, with a unique minimum in (0,1) (as $U_0=U_1=1$). We can rewrite the θ functions with U_α and U_α' as follows:

$$\theta_1(\alpha) = \frac{U_\alpha - \beta U_\alpha'}{U_\alpha^2} - 1,$$

$$\theta_0(\alpha) = \frac{U_\alpha + \alpha U'_\alpha}{U_\alpha^2} - 1. \tag{29}$$

Differentiating these expressions gives

$$\theta'_1(\alpha) = -\frac{\beta}{U_\alpha^3}(U''_\alpha U_\alpha - 2U'^2_\alpha),$$

$$\theta'_0(\alpha) = \frac{\alpha}{U_\alpha^3}(U''_\alpha U_\alpha - 2U'^2_\alpha). \tag{30}$$

θ_0 and θ_1 are monotonically increasing and decreasing, respectively. This immediately follows from writing the term occurring in the brackets of Eqs. (30) as a variance in the density p_α,

$$U''_\alpha U_\alpha - 2U'^2_\alpha = 2\ \mathrm{Var}_\alpha\left(\frac{p_1 - p_0}{\alpha p_0 + \beta p_1}\right) U_\alpha^2, \tag{31}$$

which is thus positive.

As a consequence of Eq. (30), the relation

$$\beta\theta'_0(\alpha) + \alpha\theta'_1(\alpha) = 0 \qquad \forall\ \alpha \in [0,1] \tag{32}$$

holds and M' reduces to

$$M'(\alpha) = \frac{\theta_1(\alpha)}{\beta^2} - \frac{\theta_0(\alpha)}{\alpha^2}. \tag{33}$$

The derivatives of the θ functions do not contribute to M' due to the fact that the specific form of the two-sided estimator (13) originates from minimizing the asymptotic mean-square error (cf. [26]). The necessary condition for a local minimum of M at α_o, $M'(\alpha_o)=0$, now reads as

$$\frac{\beta_o^2}{\alpha_o^2} = \frac{\theta_1(\alpha_o)}{\theta_0(\alpha_o)}, \tag{34}$$

where $\beta_o = 1 - \alpha_o$ is introduced. Using Eqs. (24) and (2), the condition (34) results in

$$\mathrm{Var}_1\left(\frac{1}{1 + e^{-w+\Delta f+\ln r_o}}\right) = \mathrm{Var}_0\left(\frac{1}{1 + e^{w-\Delta f-\ln r_o}}\right). \tag{35}$$

This means, the optimal ratio r_o is such that the variances of the random functions, which are averaged in the two-sided estimation (15), are equal. However, the existence of a solution of $M'(\alpha)=0$ is not guaranteed in general.

Writing Eq. (35) in the form

$$\mathrm{Var}_1\left(\frac{p_1 - p_0}{\alpha p_0 + \beta p_1}\right) = \mathrm{Var}_0\left(\frac{p_1 - p_0}{\alpha p_0 + \beta p_1}\right) \tag{36}$$

prevents the equation from becoming a tautology.

IV. CONVEXITY OF THE MEAN-SQUARE ERROR

Theorem. The asymptotic mean-square error $M(\alpha)$ is convex in α.

In order to prove the convexity, we introduce the operator $\Gamma_\alpha(f)$, which is defined for an arbitrary function $f(w)$ by

$$\Gamma_\alpha(f) = \beta\ \mathrm{Var}_0(f) + \alpha\ \mathrm{Var}_1(f) - U_\alpha\ \mathrm{Var}_\alpha(f). \tag{37}$$

Lemma. Γ_α is positive semidefinite, i.e.,

$$\Gamma_\alpha(f) \geq 0 \qquad \forall\, f(w). \tag{38}$$

For $\alpha \in (0,1)$ and $f(w) \neq \mathrm{const}$, the equality holds if and only if $p_0 \equiv p_1$.

Proof of the Lemma. Let $\delta f_\gamma = f(w) - \langle f\rangle_\gamma$, $\gamma \in [0,1]$. Then

$$\begin{aligned}\Gamma_\alpha(f) &= \int_\Omega \left(\beta\delta f_0^2 p_0 + \alpha\delta f_1^2 p_1 - \delta f_\alpha^2 \frac{p_0 p_1}{\alpha p_0 + \beta p_1}\right) dw \\ &= \int_\Omega \frac{(\beta\delta f_0^2 p_0 + \alpha\delta f_1^2 p_1)(\alpha p_0 + \beta p_1) - \delta f_\alpha^2 p_0 p_1}{\alpha p_0 + \beta p_1} dw \\ &= \alpha\beta \int_\Omega \frac{(\delta f_1 p_1 - \delta f_0 p_0)^2}{\alpha p_0 + \beta p_1} dw + U_\alpha(\beta\langle f\rangle_0 + \alpha\langle f\rangle_1 \\ &\quad - \langle f\rangle_\alpha)^2,\end{aligned} \tag{39}$$

which is clearly positive. Provided $f \neq \mathrm{const}$ and $\alpha \neq 0, 1$, the integrand in the last line is zero $\forall w$ if and only if $p_0 \equiv p_1$. This completes the proof of the lemma. □

Proof of the Theorem. Consulting Eqs. (33) and (32), the second derivative of M reads as

$$M''(\alpha) = 2\left(\frac{\theta_1(\alpha)}{\beta^3} + \frac{\theta_0(\alpha)}{\alpha^3}\right) - \frac{1}{\alpha^2\beta}\theta'_0(\alpha). \tag{40}$$

Expressing $p_0 = p - \beta d$ and $p_1 = p + \alpha d$ in center and relative "coordinates" $p = \alpha p_0 + \beta p_1$ and $d = p_1 - p_0$, respectively, gives

$$\theta_1(\alpha) = \frac{1}{U_\alpha^2}\mathrm{Var}_1\left(\frac{p_0}{p}\right) = \frac{\beta^2}{U_\alpha^2}\mathrm{Var}_1\left(\frac{d}{p}\right),$$

$$\theta_0(\alpha) = \frac{1}{U_\alpha^2}\mathrm{Var}_0\left(\frac{p_1}{p}\right) = \frac{\alpha^2}{U_\alpha^2}\mathrm{Var}_0\left(\frac{d}{p}\right),$$

$$\theta'_0(\alpha) = \frac{2\alpha}{U_\alpha}\mathrm{Var}_\alpha\left(\frac{d}{p}\right). \tag{41}$$

Therefore, $\frac{1}{2}\alpha\beta U_\alpha^2 M'' = \Gamma_\alpha(\frac{d}{p})$, which is positive according to the lemma. □

The convexity of the mean-square error is a fundamental characteristic of Bennett's acceptance ratio method. This characteristic allows us to state a simple criterion for the existence of a *local* minimum of the mean-square error in terms of its derivatives at the boundaries. Namely, if

$$M'(0) = \mathrm{Var}_1(e^{w-\Delta f}) - \mathrm{Var}_0(e^{w-\Delta f}) \tag{42}$$

is negative and

$$M'(1) = \mathrm{Var}_1(e^{-w+\Delta f}) - \mathrm{Var}_0(e^{-w+\Delta f}) \tag{43}$$

is positive there exists a local minimum of $M(\alpha)$ for $\alpha \in (0,1)$. Otherwise, no local minimum exists and the global minimum is found on the boundaries of α: if $M'(0) > 0$, the global minimum is found for $\alpha = 0$; thus, it is optimal to measure work values in the reverse direction only and to use the one-sided reverse estimator (12). Else, if $M'(1) < 0$, the global minimum is found for $\alpha = 1$, implying the one-sided forward estimator (11) to be optimal.

In addition, the convexity of the mean-square error proves the existence and uniqueness of the optimal ratio since a convex function has a global minimum on a closed interval.

Corollary. If a solution of $M'(\alpha)=0$ exists, it is unique and $M(\alpha)$ attains its global minimum ($\alpha \in [0,1]$) there.

V. ESTIMATING THE OPTIMAL RATIO WITH FIRST MOMENTS

In situations of practical interest, the optimal ratio is not available *a priori*. Thus, we are going to estimate the optimal ratio. There exist estimators of the optimal ratio since Bennett. In addition, we have just proven that the optimal ratio exists and is unique. However, there is still one obstacle to overcome. Yet, all expressions for estimating the optimal ratio are based on second moments [see, e.g., Eq. (35)]. Due to convergence issues, it is not practicable to base any estimator on expressions that involve second moments. The estimator would converge far too slowly. For this reason, we transform the problem into a form that employs first moments only.

Assume we have given n_0 and n_1 work values in forward and reverse direction, respectively, and want to estimate U_a, with $0 \leq a \leq 1$. According to Eq. (7), we can estimate the overlap measure U_a by using draws from the forward direction,

$$\hat{U}_a^{(0)} = \frac{1}{n_0}\sum_{k=1}^{n_0} \frac{1}{b + a e^{w_k^0 - \widehat{\Delta f}}}, \tag{44}$$

where b equals $1-a$ and for $\widehat{\Delta f}$ the best available estimate of Δf is inserted, i.e., the two-sided estimate based on the $n_0 + n_1$ work values. Similarly, we can estimate the overlap measure by using draws from the reverse direction,

$$\hat{U}_a^{(1)} = \frac{1}{n_1}\sum_{l=1}^{n_1} \frac{1}{a + b e^{-w_l^1 + \widehat{\Delta f}}}. \tag{45}$$

Since in general draws from both directions are available, it is reasonable to take an arithmetic mean of both estimates

$$\hat{U}_a = a\hat{U}_a^{(1)} + b\hat{U}_a^{(0)}, \tag{46}$$

where the weighting is chosen such that the better estimate $\hat{U}_a^{(0)}$ or $\hat{U}_a^{(1)}$ contributes stronger: with increasing a the estimate $\hat{U}_a^{(1)}$ becomes more reliable, as U_a is the normalizing constant of the bridging density p_a [Eq. (3)] and $p_a \xrightarrow{a\to 1} p_1$, and conversely for decreasing a.

From the estimate of the overlap measure, we can estimate the rescaled mean-square error by

$$\hat{M}(a) = \frac{1}{ab}\left(\frac{1}{\hat{U}_a} - 1\right) \tag{47}$$

for all $a \in (0,1)$, a result that is entirely based on first moments. The infimum of $\hat{M}(a)$ finally results in an estimate $\hat{\alpha}_o$ of the optimal choice α_o of $\frac{n_0}{N}$,

$$\hat{\alpha}_o : \Leftrightarrow \hat{M}(\hat{\alpha}_o) = \inf_a \hat{M}(a). \tag{48}$$

When searching for the infimum, we also take

$$\hat{M}(0) = \frac{1}{n_0}\sum_{k=1}^{n_0} e^{w_k^{(0)} - \widehat{\Delta f}} - \frac{1}{n_1}\sum_{l=1}^{n_1} e^{w_l^{(1)} - \widehat{\Delta f}},$$

$$\hat{M}(1) = \frac{1}{n_1}\sum_{l=1}^{n_1} e^{-w_l^{(1)} + \widehat{\Delta f}} - \frac{1}{n_0}\sum_{k=1}^{n_0} e^{-w_k^{(0)} + \widehat{\Delta f}}, \tag{49}$$

into account which follow from a series expansion of Eq. (47) in a at $a=0$ and $a=1$, respectively.

VI. INCORPORATING COSTS

The costs of measuring a work value in forward direction may differ from the costs of measuring a work value in reverse direction. The influence of costs on the optimal ratio of sample sizes is investigated here.

Different costs can be due to a direction dependent effort of experimental or computational measurement of work (unfolding a RNA may be much easier than folding it). We assume the work values to be uncorrelated, which is essential for the validity of the theory presented in this paper. Thus, a source of nonequal costs, which arises especially when work values are obtained via computer simulations, is the difference in the strength of correlations of consecutive Monte Carlo steps in forward and reverse direction. To achieve uncorrelated draws, the "correlation lengths" or "correlation times" have to be determined within the simulation too. However, this is advisable in any case of two-sided estimation, independent of the sampling strategy.

Let c_0 and c_1 be the costs of drawing a single forward and reverse work value, respectively. Our goal is to minimize the mean-square error $X = \frac{1}{N}M$ while keeping the total costs $c = n_0 c_0 + n_1 c_1$ constant. Keeping c constant results in

$$N(c,\alpha) = \frac{c}{\alpha c_0 + \beta c_1}, \tag{50}$$

which in turn yields

$$X(c,\alpha) = \frac{1}{N(c,\alpha)} M(\alpha). \tag{51}$$

If a local minimum exists, it results from $\frac{\partial}{\partial\alpha}X(c,\alpha)=0$, which leads to

$$\frac{\beta_o^2}{\alpha_o^2} = \frac{c_0 \theta_1(\alpha_o)}{c_1 \theta_0(\alpha_o)}, \tag{52}$$

a result Bennett was already aware of [26]. However, based on second moments, it was not possible to estimate the optimal ratio r_o accurately and reliably. Hence, Bennett proposed to use a suboptimal *equal-time strategy* or *equal cost strategy*, which spends an equal amount of expenses to both directions, i.e., $n_0 c_0 = n_1 c_1 = \frac{c}{2}$ or

$$\frac{\beta_{ec}}{\alpha_{ec}} = \frac{c_0}{c_1}, \tag{53}$$

where $\alpha_{ec}=1-\beta_{ec}$ is the equal cost choice for $\alpha=\frac{n_0}{N}$. This choice is motivated by the following result:

$$X(c,\alpha) \geq \frac{1}{2}X(c,\alpha_{ec}) \quad \forall\ \alpha \in [0,1], \tag{54}$$

which states that the asymptotic mean-square error of the equal cost strategy is at most suboptimal by a factor of 2 [26]. Note, however, that the equal cost strategy can be far more suboptimal if the asymptotic limit of large sample sizes is not reached.

Since we can base the estimator for the optimal ratio r_o on first moments (see Sec. V), we propose a *dynamic strategy* that performs better than the equal cost strategy. The infimum of

$$\hat{X}(c,a) = \frac{ac_0 + bc_1}{c}\hat{M}(a) \tag{55}$$

results in the estimate $\hat{\alpha}_o$ of the optimal choice α_o of $\frac{n_0}{N}$,

$$\hat{\alpha}_o: \Leftrightarrow \hat{X}(c,\hat{\alpha}_o) = \inf_a \hat{X}(c,a). \tag{56}$$

We remark that opposed to $M(\alpha)$, $X(c,\alpha)$ is not necessarily convex. However, a global minimum clearly exists and can be estimated.

VII. A DYNAMIC SAMPLING STRATEGY

Suppose we want to estimate the free-energy difference with the acceptance ratio method but have a limit on the total amount of expenses c that can be spend for measurements of work. In order to maximize the efficiency, the measurements are to be performed such that $\frac{n_0}{N}$ finally equals the optimal fraction α_o of forward measurements.

The dynamic strategy is as follows:

(1) In absence of preknowledge on α_o, we start with Bennett's equal cost strategy (53) as an initial guess of α_o.

(2) After drawing a small number of work values, we make preliminary estimates of the free-energy difference, the mean-square error, and the optimal fraction α_o.

(3) Depending on whether the estimated rescaled mean-square error $\hat{M}(a)$ is convex, which is a necessary condition for convergence, our algorithm updates the estimate $\hat{\alpha}_o$ of α_o.

(4) Further work values are drawn such that $\frac{n_0}{N}$ dynamically follows $\hat{\alpha}_o$, while $\hat{\alpha}_o$ is updated repeatedly.

There is no need to update $\hat{\alpha}_o$ after each individual draw. Splitting the total costs into a sequence $0<c^{(1)}<\ldots<c^{(p)}=c$, not necessarily equidistant, we can predefine when and how often an update in $\hat{\alpha}_o$ is made. Namely, this is done whenever the actually spent costs reach the next value $c^{(\nu)}$ of the sequence.

The dynamic strategy can be cast into an algorithm.

Algorithm. Set the initial values $n_0^{(0)}=n_1^{(0)}=0$, $\hat{\alpha}_o^{(1)}=\alpha_{ec}$. In the νth step of the iteration $\nu=1,\ldots,p$ determine

$$n_0^{(\nu)} = \lfloor \hat{\alpha}_o^{(\nu)} N^{(\nu)} \rfloor,$$

$$n_1^{(\nu)} = \lfloor \hat{\beta}_o^{(\nu)} N^{(\nu)} \rfloor, \tag{57}$$

with

$$N^{(\nu)} = \frac{c^{(\nu)}}{\hat{\alpha}_o^{(\nu)} c_0 + \hat{\beta}_o^{(\nu)} c_1}, \tag{58}$$

where $\lfloor\ \rfloor$ means rounding to the next lower integer. Then, $\Delta n_0^{(\nu)}=n_0^{(\nu)}-n_0^{(\nu-1)}$ additional forward and $\Delta n_1^{(\nu)}=n_1^{(\nu)}-n_1^{(\nu-1)}$ additional reverse work values are drawn. Using the entire present samples, an estimate $\widehat{\Delta f}^{(\nu)}$ of Δf is calculated according to Eq. (13). With the free-energy estimate at hand, $\hat{M}^{(\nu)}(a)$ is calculated for all values of $a \in [0,1]$ via Eqs. (44)–(47) and (49) discretized, say in steps $\Delta a=0.01$. If $\hat{M}^{(\nu)}(a)$ is convex, we update the recent estimate $\hat{\alpha}_o^{(\nu)}$ of α_o to $\hat{\alpha}_o^{(\nu+1)}$ via Eqs. (55) and (56). Otherwise, if $\hat{M}^{(\nu)}(a)$ is not convex, the corresponding estimate of α_o is not yet reliable and we keep the recent value, $\hat{\alpha}_o^{(\nu+1)}=\hat{\alpha}_o^{(\nu)}$. Increasing ν by one, we iteratively continue with Eq. (57) until we finally obtain $\widehat{\Delta f}^{(p)}$, which is the optimal estimate of the free-energy difference after having spend all costs c.

Note that an update in $\hat{\alpha}_o^{(\nu)}$ may result in negative values of $\Delta n_0^{(\nu)}$ or $\Delta n_1^{(\nu)}$. Should $\Delta n_0^{(\nu)}$ happen to be negative, we set $n_0^{(\nu)}=n_0^{(\nu-1)}$ and

$$n_1^{(\nu)} = \left\lfloor \frac{c^{(\nu)} - c_0 n_0^{(\nu-1)}}{c_1} \right\rfloor. \tag{59}$$

We proceed analogously, if $\Delta n_1^{(\nu)}$ happens to be negative.

The optimal fraction α_o depends on the cost ratio c_1/c_0, i.e., the algorithm needs to know the costs c_0 and c_1. However, the costs are not always known in advance and may also vary over time. Think of a long-time experiment which is subject to currency changes, inflation, terms of trade, innovations, and so on. Of advantage is that the dynamic sampling strategy is capable of incorporating varying costs. In each iteration step of the algorithm, one just inserts the actual costs. If desired, the breakpoints $c^{(\nu)}$ may also be adapted to the actual costs. Should the costs initially be unknown (e.g., the "correlation length" of a Monte Carlo simulation needs to be determined within the simulation first) one may use any reasonable guess until the costs are known.

VIII. EXAMPLE

For illustration of results, we choose exponential work distributions

$$p_i(w) = \frac{1}{\mu_i}e^{-w/\mu_i}, \quad w \in \Omega = \mathbb{R}^+, \tag{60}$$

$\mu_i>0$, $i=0,1$. According to the fluctuation theorem (2), we have $\mu_1=\frac{\mu_0}{1+\mu_0}$ and $\Delta f=\ln(1+\mu_0)$.

Exponential work densities arise in a natural way in the context of a two-dimensional harmonic oscillator with Boltzmann distribution $\rho(x,y)=e^{-(1/2)\omega^2(x^2+y^2)}/Z$, where Z

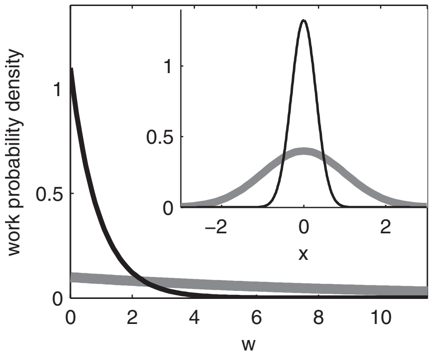

FIG. 2. The main figure displays the exponential work densities p_0 (thick line) and p_1 (thin line) for the choice of $\mu_0=10$ and, according to the fluctuation theorem, $\mu_1=10/11$. The inset displays the corresponding Boltzmann distributions $\rho_0(x,y)$ (thick) and $\rho_1(x,y)$ (thin) both for $y=0$. Here, ω_0 is set equal to 1 arbitrarily, hence, $\omega_1^2=(1+\mu_0)\omega_0^2=11$. The free-energy difference is $\Delta f=\ln(1+\mu_0)=\ln(\omega_1^2/\omega_0^2)\approx 2.38$.

$=2\pi/\omega^2$ is a normalizing constant (partition function) and $(x,y)\in\mathbb{R}^2$ [28]. Drawing a point (x,y) from the initial density $\rho=\rho_0$ defined by setting $\omega=\omega_0$, and switching the frequency to $\omega_1>\omega_0$ instantaneously amounts in the work $\frac{1}{2}(\omega_1^2-\omega_0^2)(x^2+y^2)$. The probability density of observing a specific work value w is given by the exponential density p_0 with $\mu_0=\frac{\omega_1^2-\omega_0^2}{\omega_0^2}$. Switching the frequency in the reverse direction $\omega_1\to\omega_0$, with the point (x,y) drawn from $\rho=\rho_1$ with $\omega=\omega_1$, the density of work (with interchanged sign) is given by p_1 with $\mu_1=\frac{\omega_1^2-\omega_0^2}{\omega_1^2}=\frac{\mu_0}{1+\mu_0}$. The free-energy difference of the states characterized by ρ_0 and ρ_1 is the log ratio of their normalizing constants $\Delta f=-\ln\frac{Z_1}{Z_0}=\ln(1+\mu_0)$. A plot of the work densities for $\mu_0=10$ is enclosed in Fig. 2.

Now, with regard to free-energy estimation, is it better to use one- or two-sided estimators? In other words, we want to know whether the global minimum of $M(\alpha)$ is on the boundaries $\{0,1\}$ of α or not. By the convexity of M, the answer is determined by the signs of the derivatives $M'(0)$ and $M'(1)$ at the boundaries. The asymptotic mean-square errors (18) and (19) of the one-sided estimators are calculated to be

$$M(1)=\mathrm{Var}_0(e^{-w+\Delta f})=\frac{\mu_0^2}{1+2\mu_0}, \tag{61}$$

for the forward direction and

$$M(0)=\mathrm{Var}_1(e^{w-\Delta f})=\frac{\mu_0^2}{1-\mu_0^2},\quad \mu_0<1, \tag{62}$$

for the reverse direction. For $\mu_0\geq 1$, the variance of the reverse estimator diverges. Note that $M(0)>M(1)$ holds for all $\mu_0>0$, i.e., forward estimation of Δf is always superior if compared to reverse estimation. Furthermore, a straightforward calculation gives

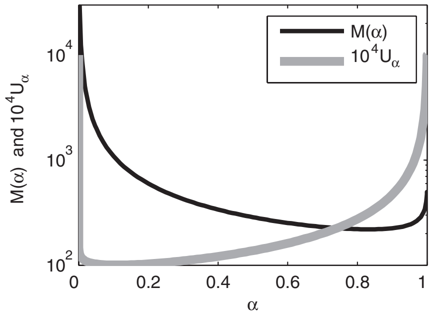

FIG. 3. The overlap function U_α and the rescaled asymptotic mean-square error M for $\mu_0=1000$. Note that $M(\alpha)$ diverges for $\alpha\to 0$.

$$M'(1)=\frac{\mu_0^3(\mu_0+\xi_-)(\mu_0-\xi_+)}{(1+2\mu_0)^2(1+3\mu_0)}, \tag{63}$$

where $\xi_\pm=\frac{1}{2}(\sqrt{17}\pm 3)$, and

$$M'(0)=-\frac{\mu_0^3(2+(1-2\mu_0)\mu_0)}{(1-\mu_0^2)^2(1-2\mu_0)},\quad \mu_0<\frac{1}{2}, \tag{64}$$

and $M'(0)=-\infty$ for $\mu_0\geq\frac{1}{2}$. Thus, for the range $\mu_0\in(0,\xi_+)$ we have $M'(0)<0$ as well as $M'(1)<0$ and therefore $\alpha_o=1$, i.e., the forward estimator is superior to any two-sided estimator (13) in this range. For $\mu_0\in(\xi_+,\infty)$, we have $M'(0)<0$ and $M'(1)>0$, specifying that $\alpha_o\in(0,1)$, i.e., two-sided estimation with an appropriate choice of α is optimal.

Numerical calculation of the function U_α and subsequent evaluation of $M(\alpha)$ allows to find the "exact" optimal fraction α_o. Examples for U_α and M are plotted in Fig. 3.

The behavior of α_o as a function of μ_0 is quite interesting (see Fig. 4). We can interpret this behavior in terms of the Boltzmann distributions as follows. Without loss of generality, assume $\omega_0=1$ is fixed. Increasing μ_0 then means increasing ω_1. The density ρ_1 is fully nested in ρ_0 (cf. the inset of Fig. 2) (remember that $\omega_1>\omega_0$) and converges to a delta peak at the origin with increasing ω_1. This means that by

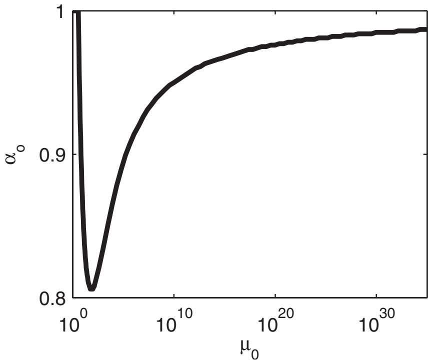

FIG. 4. The optimal fraction $\alpha_o=\frac{n_0}{N}$ of forward work values for the two-sided estimation in dependence of the average forward work μ_0. For $\mu_0\leq\xi_+\approx 3.56$, the one-sided forward estimator is optimal, i.e., $\alpha_o=1$.

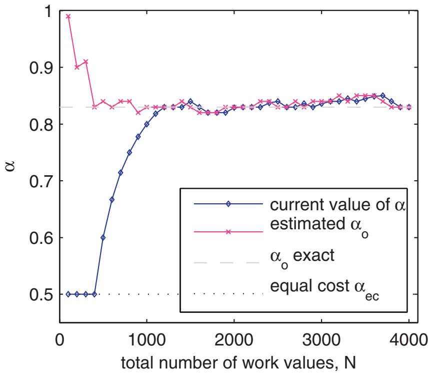

FIG. 5. (Color online) Example of a single run using the dynamic strategy: the optimal fraction α_o of forward measurements for the two-sided free-energy estimation is estimated at predetermined values of total sample sizes $N=n_0+n_1$ of forward and reverse work values. Subsequently, taking into account the current actual fraction $\alpha=\frac{n_0}{N}$, additional work values are drawn such that we come closer to the estimated $\hat{\alpha}_o$.

sampling from ρ_0 we can obtain information about the full density ρ_1 quite easily, whereas sampling from ρ_1 provides only poor information about ρ_0. This explains why $\alpha_o=1$ holds for small values of μ_0. However, with increasing ω_1 the density ρ_1 becomes so narrow that it becomes difficult to obtain draws from ρ_0 that fall into the main part of ρ_1. Therefore, it is better to add some information from ρ_1, hence, α_o decreases. Increasing ω_1 further, the relative number of draws needed from ρ_1 will decrease, as the density converges toward the delta distribution. Finally, it will become sufficient to make only *one* draw from ρ_1 in order to obtain the full information available. Therefore, α_o converges toward 1 in the limit $\mu_0 \to \infty$.

In the following, the dynamic strategy proposed in Sec. VII is applied. We choose $\mu_0=1000$ and $c_0=c_1$. The equal cost strategy draws according to $\alpha_{ec}=0.5$, which is used as initial value in the dynamic strategy. The results of a single run are presented in Figs. 5–7. Starting with $N=100$, the estimate of α_o is updated in steps of $\Delta N=100$. The actual forward fractions α together with the estimated values of the optimal fraction α_o are shown in Fig. 5. The first three estimates of α_o are rejected because the estimated function $\hat{M}(\alpha)$ is not yet convex. Therefore, α remains unchanged at the beginning. Afterward, α follows the estimates of α_o and starts to fluctuate about the exact value of α_o. Some estimates of the function M corresponding to this run are depicted in Fig. 6. For these estimates, α is discretized in steps $\Delta\alpha=0.01$. Remarkably, the estimates of α_o that result from these curves are quite accurate even for relatively small N. Finally, Fig. 7 shows the free-energy estimates of the run (not for all values of N) compared with those of a single run where the equal cost strategy is used. We find some increase in accuracy when using the dynamic strategy.

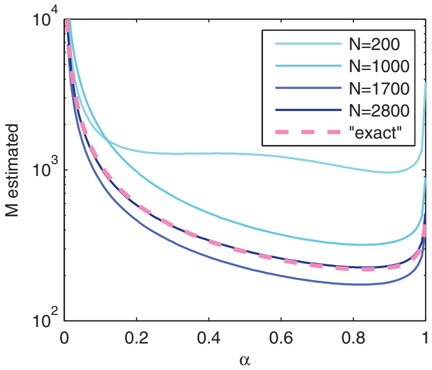

FIG. 6. (Color online) Displayed are estimated mean-square errors $\hat{M}$ in dependence of α for different sample sizes. The global minimum of the estimated function $\hat{M}$ determines the estimate of the optimal fraction α_o of forward work measurements.

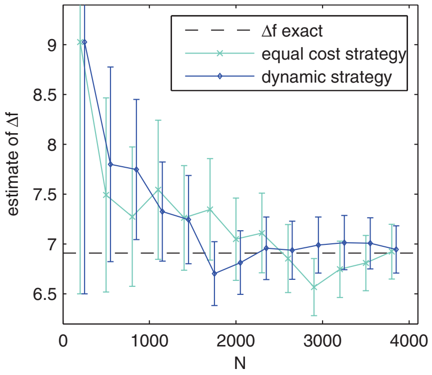

FIG. 7. (Color online) Comparison of a single run of free-energy estimation using the equal cost strategy versus a single run using the dynamic strategy. The error bars are the square roots of the estimated mean-square error X.

In combination with a good *a priori* choice of the initial value of α, the use of the dynamic strategy enables a superior convergence and precision of free-energy estimation (see Figs. 8 and 9). Due to insight into some particular system under consideration, it is not unusual that one has *a priori*

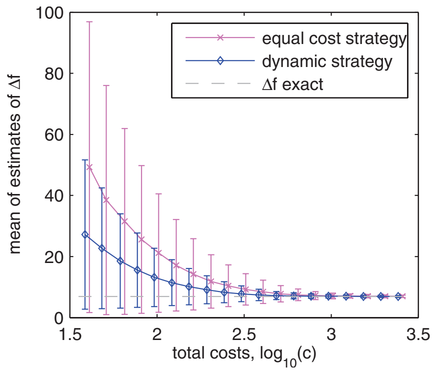

FIG. 8. (Color online) Averaged estimates from 10 000 independent runs with dynamic strategy versus 10 000 runs with equal cost strategy in dependence of the total cost $c=n_0c_0+n_1c_1$ spend. The cost ratio is $c_1/c_0=0.01$, $c_0+c_1=2$, and $\mu_0=1000$. The error bars represent one standard deviation. Here, the initial value of α in the dynamic strategy is 0.5, while the equal cost strategy draws with $\alpha_{ec}\approx 0.01$. We note that $\alpha_o\approx 0.08$.

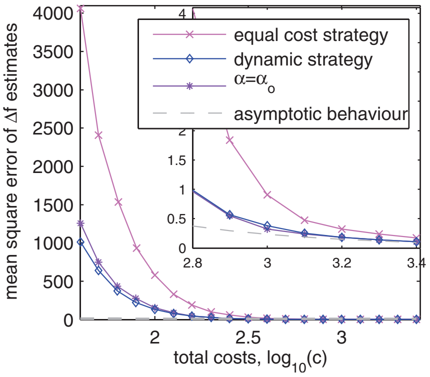

FIG. 9. (Color online) Displayed are mean-square errors of free-energy estimates using the same data as in Fig. 8. In addition, the mean-square errors of estimates with constant $\alpha=\alpha_o$ are included, as well as the asymptotic behavior (51). The inset shows that the mean-square error of the dynamic strategy approaches the asymptotic optimum, whereas the equal cost strategy is suboptimal. Note that for small sample sizes, the asymptotic behavior does not represent the actual mean-square error.

knowledge which results in a better guess for the initial choice of α in the dynamic strategy than starting with $\alpha=\alpha_{ec}$. For instance, a good initial choice is known when estimating the chemical potential via Widom's particle insertion and deletion [31]. Namely, it is *a priori* clear that inserting particles yields much more information than deleting particles since the phase space which is accessible to particles in the "deletion system" is effectively contained in the phase space accessible to the particles in the "insertion system" (cf., e.g., [19]). A good *a priori* initial choice for α may be $\alpha=0.9$ with which the dynamic strategy outperforms any other strategy that the authors are aware of.

Once reaching the limit of large sample sizes, the dynamic strategy is insensitive to the initial choice of α since the strategy is robust and finds the optimal fraction α_o of forward measurements itself.

IX. CONCLUSION

Two-sided free-energy estimation, i.e., the acceptance ratio method [26], employs samples of n_0 forward and n_1 reverse work measurements in the determination of free-energy differences in a statistically optimal manner. However, its statistical properties depend strongly on the ratio $\frac{n_1}{n_0}$ of work values used. As a central result, we have proven the convexity of the asymptotic mean-square error of two-sided free-energy estimation as a function of the fraction $\alpha=\frac{n_0}{N}$ of forward work values used. From here follows immediately the existence and uniqueness of the optimal fraction α_o, which minimizes the asymptotic mean-square error. This is of particular interest if we can control the value of α, i.e., can make additional measurements of work in either direction. Drawing such that we finally reach $\frac{n_0}{N}=\alpha_o$, the efficiency of two-sided estimation can be enhanced considerably. Consequently, we have developed a dynamic sampling strategy which iteratively estimates α_o and makes additional draws or measurements of work. Thereby, the convexity of the mean-square error enters as a key criterion for the reliability of the estimates. For a simple example, which allows to compare with analytic calculations, the dynamic strategy has shown to work perfectly.

In the asymptotic limit of large sample sizes, the dynamic strategy is optimal and outperforms any other strategy. Nevertheless, in this limit it has to compete with the near optimal equal cost strategy of Bennett, which also performs very good. It is worth mentioning that even if the latter comes close to the performance of ours, it is worthwhile the effort of using the dynamic strategy since the underlying algorithm can be easily implemented and does cost quite anything if compared to the effort required for drawing additional work values.

Most important for experimental and numerical estimation of free-energy differences is the range of small and moderate sample sizes. For this relevant range, it is found that the dynamic strategy performs very good too. It converges significantly better than the equal cost strategy. In particular, for small and moderate sample sizes it can improve the accuracy of free-energy estimates by half an order of magnitude.

We close our considerations by mentioning that the two-sided estimator is typically far superior with respect to one-sided estimators: assume the support of p_0 and p_1 is symmetric about Δf [32]; then, if the densities are symmetric to each other, $p_0(\Delta f+w)=p_1(\Delta f-w)$, the optimal fraction of forward draws is $\frac{n_0}{N}=\frac{1}{2}$ by symmetry. Therefore, if the symmetry is violated not too strongly, the optimum will remain near 0.5. Continuous deformations of the densities change the optimal fraction α_o continuously. Thus, α_o does not reach 0 and 1, respectively, for some certain strength of asymmetry. It is exceptionally hard to violate the symmetry such that α_o hits the boundary 0 or 1. In consequence, in almost all situations, the two-sided estimator is superior.

ACKNOWLEDGMENTS

We thank Andreas Engel for a critical reading of the manuscript.

---

[1] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[2] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[3] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[4] A. Gelman and X.-L. Meng, Stat. Sci. **13**, 163 (1998).
[5] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[6] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[7] M.-H. Chen and Q.-M. Shao, Ann. Stat. **25**, 1563 (1997).

[8] H. Oberhofer and C. Dellago, Comput. Phys. Commun. **179**, 41 (2008).

[9] M. Watanabe and W. P. Reinhardt, Phys. Rev. Lett. **65**, 3301 (1990).

[10] S. X. Sun, J. Chem. Phys. **118**, 5769 (2003).

[11] F. M. Ytreberg and D. M. Zuckerman, J. Chem. Phys. **120**, 10876 (2004).

[12] C. Jarzynski, Phys. Rev. E **73**, 046105 (2006).

[13] A. Engel, Phys. Rev. E **80**, 021120 (2009).

[14] H. Then and A. Engel, Phys. Rev. E **77**, 041105 (2008).

[15] X.-L. Meng and S. Schilling, J. Comput. Graph. Stat. **11**, 552 (2002).

[16] C. Jarzynski, Phys. Rev. E **65**, 046122 (2002).

[17] H. Oberhofer, C. Dellago, and S. Boresch, Phys. Rev. E **75**, 061106 (2007).

[18] S. Vaikuntanathan and C. Jarzynski, Phys. Rev. Lett. **100**, 190601 (2008).

[19] A. M. Hahn and H. Then, Phys. Rev. E **79**, 011113 (2009).

[20] X.-L. Meng and W. H. Wong, Stat. Sin. **6**, 831 (1996).

[21] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan, J. R. Stat. Soc. Ser. B (Stat. Methodol.) **65**, 585 (2003).

[22] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).

[23] D. M. Ceperley, Rev. Mod. Phys. **67**, 279 (1995).

[24] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic Press, London, 2002).

[25] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, Jr., and C. Bustamante, Nature (London) **437**, 231 (2005).

[26] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).

[27] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).

[28] M. R. Shirts and V. S. Pande, J. Chem. Phys. **122**, 144107 (2005).

[29] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).

[30] J. Gore, F. Ritort, and C. Bustamante, Proc. Natl. Acad. Sci. U.S.A. **100**, 12564 (2003).

[31] B. Widom, J. Chem. Phys. **39**, 2808 (1963).

[32] Which is not the case for the densities studied in Sec. VIII.

Paper [3]

PHYSICAL REVIEW E **81**, 041117 (2010)

Measuring the convergence of Monte Carlo free-energy calculations

Aljoscha M. Hahn* and Holger Then†
Institut für Physik, Carl von Ossietzky Universität, 26111 Oldenburg, Germany

(Received 15 October 2009; revised manuscript received 24 February 2010; published 16 April 2010)

The nonequilibrium work fluctuation theorem provides the way for calculations of (equilibrium) free-energy based on work measurements of nonequilibrium, finite-time processes, and their reversed counterparts by applying Bennett's acceptance ratio method. A nice property of this method is that each free-energy estimate readily yields an estimate of the asymptotic mean square error. Assuming convergence, it is easy to specify the uncertainty of the results. However, sample sizes have often to be balanced with respect to experimental or computational limitations and the question arises whether available samples of work values are sufficiently large in order to ensure convergence. Here, we propose a convergence measure for the two-sided free-energy estimator and characterize some of its properties, explain how it works, and test its statistical behavior. In total, we derive a convergence criterion for Bennett's acceptance ratio method.

DOI: 10.1103/PhysRevE.81.041117 PACS number(s): 02.50.Fz, 05.40.−a, 05.70.Ln, 05.10.−a

I. INTRODUCTION

Many methods have been developed in order to estimate free-energy differences, ranging from thermodynamic integration [1,2], path sampling [3], free-energy perturbation [4], umbrella sampling [5–7], adiabatic switching [8], dynamic methods [9–12], optimal protocols [13,14], asymptotic tails [15], to targeted and escorted free-energy perturbation [16–20]. Yet, the reliability and efficiency of the approaches have not been considered in full depth. Fundamental questions remain unanswered [21], e.g., what method is best for evaluating the free-energy? Is the free-energy estimate reliable and what is the error in it? How can one assess the quality of the free-energy result when the true answer is unknown? Generically, free-energy estimators are strongly biased for finite sample sizes such that the bias constitutes the main source of error of the estimates. Moreover, the bias can manifest itself in a seemingly convergence of the calculation by reaching a stable value, although far apart from the desired true value. Therefore, it is of considerable interest to have reliable criteria for the convergence of free-energy calculations.

Here we focus on the convergence of Bennett's acceptance ratio method. Thereby, we will only be concerned with the intrinsic statistical errors of the method and assume uncorrelated and unbiased samples from the work densities. For incorporation of instrument noise, see Ref. [22].

With emerging results from nonequilibrium stochastic thermodynamics, Bennett's acceptance ratio method [23–26] has revived actual interest.

Recent research has shown that the isothermal free-energy difference $\Delta f = f_1 - f_0$ of two thermal equilibrium states 0 and 1, both at the same temperature T, can be determined by externally driven nonequilibrium processes connecting these two states. In particular, if we start the process with the initial thermal equilibrium state 0 and perturb it towards 1 by varying the control parameter according to a predefined protocol, the work w applied to the system will be a fluctuating random variable distributed according to a probability density $p_0(w)$. This direction will be denoted with *forward*. Reversing the process by starting with the initial equilibrium state 1 and perturbing the system towards 0 by the time reversed protocol, the work w done *by* the system in the *reverse* process will be distributed according to a density $p_1(w)$. Under some quite general conditions, the forward and reverse work densities $p_0(w)$ and $p_1(w)$ are related to each other by Crooks fluctuation theorem [27,28]

$$\frac{p_0(w)}{p_1(w)} = e^{w-\Delta f}. \tag{1}$$

Throughout the paper, all energies are understood to be measured in units of the thermal energy kT, where k is Boltzmann's constant. The fluctuation theorem relates the equilibrium free-energy difference Δf to the nonequilibrium work fluctuations which permits calculation (estimation) of Δf using samples of work values measured either in only one direction (*one-sided* estimation) or in both directions (*two-sided* estimation). The one-sided estimators rely on the Jarzynski relation [29] $e^{-\Delta f} = \int e^{-w} p_0(w) dw$ which is a direct consequence of Eq. (1), and the free-energy is estimated by calculating the sample mean of the exponential work. In general, however, it is of great advantage to employ optimal two-sided estimation with Bennett's acceptance ratio method [23], although one has to measure work values in both directions.

The work fluctuations necessarily allow for events which "violate" the second law of thermodynamics such that $w < \Delta f$ holds in forward direction and $w > \Delta f$ in reverse direction, and the accuracy of any free-energy estimate solely based on knowledge of Eq. (1) will strongly depend on the extend to which these events are observed. The fluctuation theorem indicates that such events will in general be exponentially rare; at least, it yields the inequality $\langle w \rangle_1 \le \Delta f \le \langle w \rangle_0$ [29], which states the second law in terms of the average work $\langle w \rangle_0$ and $\langle w \rangle_1$ in forward and reverse direction,

---

*Present address: Technische Universität Berlin, Institut für Theoretische Physik, 10623 Berlin, Germany.

†Present address: University of Bristol, Department of Mathematics, University Walk, Bristol BS8 1TW, UK.

©2010 The American Physical Society

respectively. Reliable free-energy calculations will become harder the larger the dissipated work $\langle w\rangle_0 - \Delta f$ and $\Delta f - \langle w\rangle_1$ in the two directions is [20], i.e., the farther from equilibrium the process is carried out, resulting in an increasing number N of work values needed for a converging estimate of Δf. This difficulty can also be expressed in terms of the overlap area $\mathcal{A} = \int \min\{p_0(w), p_1(w)\} dw \leq 1$ of the work densities, which is just the sum of the probabilities $\int_{-\infty}^{\Delta f} p_0 dw$ and $\int_{\Delta f}^{\infty} p_1 dw$ of observing second law "violating" events in the two directions. Hence, N has to be larger than $1/\mathcal{A}$. However, an *a priori* determination of the number N of work values required will be impossible in situations of practical interest. Instead, it may be possible to determine *a posteriori* whether a given calculation of Δf has converged. The present paper develops a criterion for the convergence of two-sided estimation which relies on monitoring the value of a suitably bounded quantity a, the convergence measure. As a key feature, the convergence measure a checks if the relevant second law "violating" events are observed sufficiently and in the right proportion for obtaining an accurate and precise estimate of Δf.

Two-sided free-energy estimation, i.e., Bennett's acceptance ratio method, incorporates a pair of samples of both directions. Given a sample $\{w_k^0\}$ of n_0 forward work values, drawn independently from $p_0(w)$, together with a sample $\{w_l^1\}$ of n_1 reverse work values drawn from $p_1(w)$, the asymptotically optimal estimate $\widehat{\Delta f}$ of the free-energy difference Δf is the unique solution of [23–26]

$$\frac{1}{n_0}\sum_{k=1}^{n_0}\frac{1}{\beta + \alpha e^{w_k^0 - \widehat{\Delta f}}} = \frac{1}{n_1}\sum_{l=1}^{n_1}\frac{1}{\alpha + \beta e^{-w_l^1 + \widehat{\Delta f}}}, \tag{2}$$

where α and $\beta \in (0,1)$ are the fraction of forward and reverse work values used, respectively,

$$\alpha = \frac{n_0}{N} \quad \text{and} \quad \beta = \frac{n_1}{N}, \tag{3}$$

with the total sample size $N = n_0 + n_1$.

Originally found by Bennett [23] in the context of free-energy perturbation [4], with "work" being simply an energy difference, the two-sided estimator (2) was generalized by Crooks [30] to actual work of nonequilibrium finite-time processes. We note that the two-sided estimator has remarkably good properties [21,23,24,31]. Although in general biased for small sample sizes N, the bias

$$b = \langle \widehat{\Delta f} - \Delta f\rangle, \tag{4}$$

asymptotically vanishes for $N \to \infty$ and the estimator is the one with least mean-square error (viz. variance) in the limit of large sample sizes n_0 and n_1 within a wide class of estimators. In fact, it is the optimal estimator if no further knowledge on the work densities besides the fluctuation theorem is given [20,22]. It comprises one-sided Jarzynski estimators as limiting cases for $\alpha \to 0$ and $\alpha \to 1$, respectively. Recently [32], the asymptotic mean square error has been shown to be a convex function of α for fixed N, indicating that typically two-sided estimation is superior if compared to one-sided estimation.

In the limit of large N, the mean-square error

$$m = \langle (\widehat{\Delta f} - \Delta f)^2\rangle, \tag{5}$$

converges to its asymptotics

$$X(N,\alpha) = \frac{1}{N}\frac{1}{\alpha\beta}\left(\frac{1}{U_\alpha} - 1\right), \tag{6}$$

where the overlap (integral) U_α is given by

$$U_\alpha = \int \frac{p_0 p_1}{\alpha p_0 + \beta p_1} dw. \tag{7}$$

Likewise, in the large N limit the probability density of the estimates $\widehat{\Delta f}$ (for fixed N and α) converges to a Gaussian density with mean Δf and variance $X(N,\alpha)$ [24]. Thus, within this regime a reliable confidence interval for a particular estimate $\widehat{\Delta f}$ is obtained with an estimate $\hat{X}(N,\alpha)$ of the variance,

$$\hat{X}(N,\alpha) := \frac{1}{N\alpha\beta}\left(\frac{1}{\hat{U}_\alpha} - 1\right), \tag{8}$$

where the overlap estimate $\hat{U}_\alpha$ is given through

$$\hat{U}_\alpha := \frac{1}{n_0}\sum_{k=1}^{n_0}\frac{1}{\beta + \alpha e^{w_k^0 - \widehat{\Delta f}}} = \frac{1}{n_1}\sum_{l=1}^{n_1}\frac{1}{\alpha + \beta e^{-w_l^1 + \widehat{\Delta f}}}. \tag{9}$$

To get some feeling for when the large N limit "begins," we state a close connection between the asymptotic mean-square error and the overlap area $\mathcal{A}$ of the work densities as follows:

$$\frac{1 - 2\mathcal{A}}{N\mathcal{A}} < X(N,\alpha) \leq \frac{1 - \mathcal{A}}{\alpha\beta N\mathcal{A}}, \tag{10}$$

see Appendix A. Using $\alpha \approx 0.5$ and assuming that the estimator has converged once $X < 1$, we find the "onset" of the large N limit for $N > \frac{1}{\mathcal{A}}$. However, this onset may actually be one or more orders of magnitude larger.

If we do not know whether the large N limit is reached, we cannot state a reliable confidence interval of the free-energy estimate: a problem which encounters frequently within free-energy calculations is that the estimates "converge" towards a stable plateau. While the sample variance can become small, it remains unclear whether the reached plateau represents the correct value of Δf. Possibly, the found plateau is subject to some large bias, i.e., far off the correct value. A typical situation is displayed in Fig. 1 which shows successive two-sided free-energy estimates in dependence of the sample size N. The errorbars are obtained with an error-propagation formula for the variance of $\widehat{\Delta f}$ which reflects the sample variances, see Appendix C after reading Sec. III. If we take a look on the top panel of Fig. 1, we might have the impression that the free-energy estimate has converged at $N \approx 300$ already, while the bottom panel reaches out to larger sample sizes where it becomes visible that the "convergence" in the top panel was just pretended. Finally, we may ask if the estimates shown in the bottom panel have converged at $N \gtrsim 10000$? As we know the true

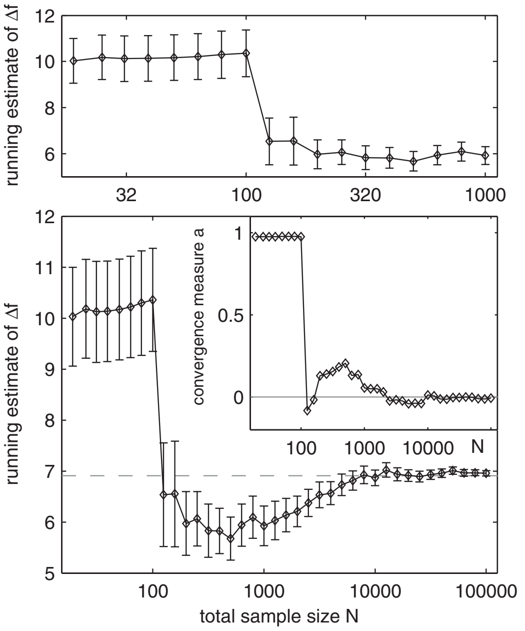

FIG. 1. Displayed are free-energy estimates $\widehat{\Delta f}$ in dependence of the sample size N, reaching a seemingly stable plateau if N is restricted to $N=1000$ (top panel). Another stable plateau is reached if the sample size is increased up to $N=100\,000$ (bottom panel). Has the estimate finally converged? The answer is given by the corresponding graph of the convergence measure a which is shown in the inset. The fluctuations around zero indicate convergence. The exact value of the free-energy difference is visualized by the dashed horizontal line.

value of Δf, which is depicted in the figure as a dashed line, we can conclude that convergence actually happened.

The main result of the present paper is the statement of a convergence criterion for two-sided free-energy estimation in terms of the *behavior* of the convergence measure a. As will be seen, a converges to zero. Moreover, this happens almost simultaneously with the convergence of $\widehat{\Delta f}$ to Δf. The procedure is as follows: while drawing an increasing number of work values in both directions (with fixed fraction α of forward draws), successive estimates $\widehat{\Delta f}$ and corresponding values of a, based on the present samples of work, are calculated. The values of a are displayed graphically in dependence of N, preferably on a log scale. Then the typical situation observed is that a is close to it's upper bound for small sample sizes $N<\frac{1}{A}$, which indicates lack of "rare events" which are required in the averages of Eq. (2) (i.e., those events which "violate" the second law). Once N becomes comparable to $\frac{1}{A}$, single observations of rare events happen and change the value of $\widehat{\Delta f}$ and a rapidly. In this regime of N, rare events are likely to be observed either disproportionally often or seldom, resulting in strong fluctuations of a around zero. This indicates the transition region to the large N limit. Finally, at some $N\gg\frac{1}{A}$, the large N limit is reached, and a typically fluctuates close around zero, cf. the inset of Fig. 1.

The paper is organized as follows. In Sec. II, we first consider a simple model for the source of bias of two-sided estimation which is intended to obtain some insight into the convergence properties of two-sided estimation. The convergence measure a, which is introduced in Sec. III, however, will not depend on this specific model. As the convergence measure is based on a sample of forward and reverse work values, it is itself a random variable, raising the question of reliability once again. Using numerically simulated data, the statistical properties of the convergence measure will be elaborated in Sec. IV. The convergence criterion is stated in Sec. V, and Sec. VI presents an application to the estimation of the chemical potential of a Lennard-Jones fluid.

II. NEGLECTED TAIL MODEL FOR TWO-SIDED ESTIMATION

To obtain some first qualitative insight into the relation between the convergence of Eq. (9) and the bias of the estimated free-energy difference, we adopt the neglected tails model [33] originally developed for one-sided free-energy estimation.

Two-sided estimation of Δf essentially means estimating the overlap U_α from two sides, however in a dependent manner, as $\widehat{\Delta f}$ is adjusted such that both estimates are equal in Eq. (9).

Consider the (normalized) overlap density $p_\alpha(w)$, defined as harmonic mean of p_0 and p_1

$$p_\alpha(w)=\frac{1}{U_\alpha}\frac{p_0(w)p_1(w)}{\alpha p_0(w)+\beta p_1(w)}. \tag{11}$$

For $\alpha\rightarrow 0$ and $\alpha\rightarrow 1$, p_α converges to p_0 and p_1, respectively. The dominant contributions to U_α come from the overlap region of p_0 and p_1 where p_α has its main probability mass, see Fig. 2 (top).

In order to obtain an accurate estimate of Δf with the two-sided estimator (2), the sample $\{w_k^0\}$ drawn from p_0 has to be representative for p_0 up to the *overlap region* in the left tail of p_0 and the sample $\{w_k^1\}$ drawn from p_1 has to be representative for p_1 up to the overlap region in the right tail of p_1. For small n_0 and n_1, however, we will have certain effective cut-off values w_c^0 and w_c^1 for the samples from p_0 and p_1, respectively, beyond which we typically will not find any work values, see Fig. 2 (bottom).

We introduce a model for the bias (4) of two-sided free-energy estimation as follows. Assuming a "semilarge" $N=n_0+n_1$, the *effective* behavior of the estimator for fixed n_0 and n_1 is modeled by substituting the sample averages appearing in the estimator (2) with ensemble averages, however truncated at w_c^0 and w_c^1, respectively,

$$\int_{w_c^0}^{\infty}\frac{p_0(w)}{\beta+\alpha e^{w-\langle\widehat{\Delta f}\rangle}}dw=\int_{-\infty}^{w_c^1}\frac{p_1(w)}{\alpha+\beta e^{-w+\langle\widehat{\Delta f}\rangle}}dw. \tag{12}$$

Thereby, the cutoff values w_c^i are thought fixed (only depending on n_0 and n_1) and the expectation $\langle\widehat{\Delta f}\rangle$ is understood to be the unique root of Eq. (12), thus being a function of the cut-off values w_c^i, $i=0,1$.

In order to elaborate the implications of this model, we rewrite Eq. (12) with the use of the fluctuation theorem (1) such that the integrands are equal,

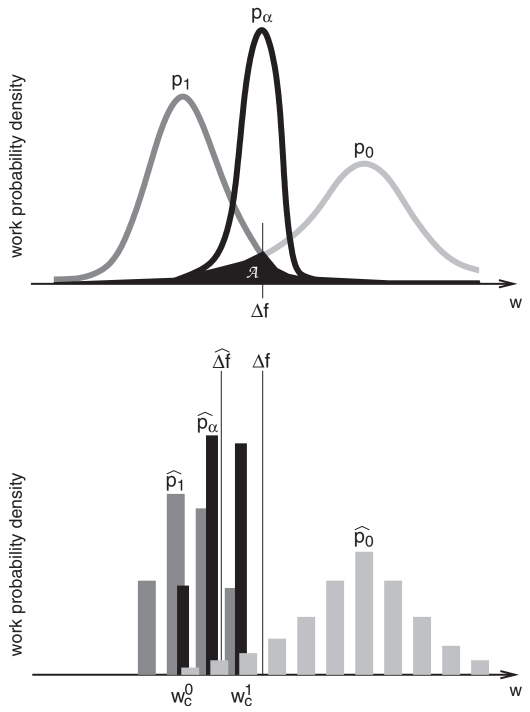

FIG. 2. Schematic diagram of reverse p_1, overlap p_α, and forward p_0 work densities (top). Schematic histograms of finite samples from p_0 and p_1, where in particular the latter is imperfectly sampled, resulting in a biased estimate $\widehat{\Delta f}$ of the free-energy difference (bottom).

$$e^{\langle\widehat{\Delta f}-\Delta f\rangle} = \frac{\displaystyle\int_{-\infty}^{w_c^1} \frac{p_0(w)}{\alpha e^{w-\langle\widehat{\Delta f}\rangle}+\beta}dw}{\displaystyle\int_{w_c^0}^{\infty} \frac{p_0(w)}{\alpha e^{w-\langle\widehat{\Delta f}\rangle}+\beta}dw}, \tag{13}$$

and consider two special cases:

(1) *Large* n_1 limit: assume the sample size n_1 is large enough to ensure that the overlap region is fully and accurately sampled (large n_1 limit). Thus, w_c^1 can be safely set equal to ∞ in Eq. (13), and the right-hand side becomes larger than unity. Accordingly, our model predicts a positive bias.

(2) *Large* n_0 limit: turning the tables and using $w_c^0=-\infty$ in Eq. (13), the model implies a negative bias.

In essence, $\langle\widehat{\Delta f}\rangle$ is shifted away from Δf towards the insufficiently sampled density. In general, when none of the densities is sampled sufficiently, the bias will be a trade off between the two cases.

Qualitatively, from the neglected tails model, we find the main source of bias resulting from a different convergence behavior of forward and reverse estimates (9) of U_α. The task of the next section will be to develop a quantitative measure of convergence.

III. CONVERGENCE MEASURE

In order to check convergence, we propose a measure which relies on a consistency check of estimates based on first and second moments of the Fermi functions that appear in the two-sided estimator (9). In a recent study [20], we already used this measure for the special case of $\alpha=\frac{1}{2}$. Here, we give a generalization to arbitrary α, study the convergence measure in greater detail, and justify its validity and usefulness. In the following we will assume that the densities p_0 and p_1 have the same support.

It was discussed in the preceding section that the large N limit is reached and hence the bias of two-sided estimation vanishes if the overlap U_α is (in average) correctly estimated from both sides, 0 and 1. Defining the complementary Fermi functions $t_c(w)$ and $b_c(w)$ (for given α) with

$$t_c(w) = \frac{1}{\alpha+\beta e^{-w+c}},$$

$$b_c(w) = \frac{1}{\alpha e^{w-c}+\beta}, \tag{14}$$

such that $\alpha t_c(w)+\beta b_c(w)=1$ and $t_c(w)=e^{w-c}b_c(w)$ holds. The overlap (7) can be expressed in terms of first moments,

$$U_\alpha = \int t_{\Delta f}(w)p_1(w)dw = \int b_{\Delta f}(w)p_0(w)dw, \tag{15}$$

and the overlap estimate $\hat{U}_\alpha$, Eq. (9), is simply obtained by replacing in Eq. (15) the ensemble averages by sample averages,

$$\hat{U}_\alpha = \overline{t_{\widehat{\Delta f}}}^{(1)} = \overline{b_{\widehat{\Delta f}}}^{(0)}. \tag{16}$$

According to Eq. (2), the value of $\widehat{\Delta f}$ is defined such that the above relation holds. Note that $\widehat{\Delta f}=\widehat{\Delta f}(w_1^0,\ldots,w_{n_1}^1)$ is a single-valued function depending on all work values used in both directions. The overbar with index (i) denotes an average with a sample $\{w_k^i\}$ drawn from p_i, $i=0,1$. For an arbitrary function $g(w)$ it explicitly reads

$$\overline{g}^{(i)} = \frac{1}{n_i}\sum_{k=1}^{n_i} g(w_k^i). \tag{17}$$

Interestingly, U_α can be expressed in terms of second moments of the Fermi functions such that it reads

$$U_\alpha = \alpha\int t_{\Delta f}^2 p_1 dw + \beta\int b_{\Delta f}^2 p_0 dw. \tag{18}$$

A useful test of self-consistency is to compare the first-order estimate $\hat{U}_\alpha$, with the second order estimate $\hat{U}_\alpha^{(II)}$, where the latter is defined by replacing the ensemble averages in Eq. (18) with sample averages

$$\hat{U}_\alpha^{(II)} = \alpha\overline{t_{\widehat{\Delta f}}^2}^{(1)} + \beta\overline{b_{\widehat{\Delta f}}^2}^{(0)}. \tag{19}$$

Thereby, the estimates $\widehat{\Delta f}$, $\hat{U}_\alpha$, and $\hat{U}_\alpha^{(II)}$, are understood to be calculated with the same pair of samples $\{w_k^0\}$ and $\{w_l^1\}$.

The relative difference of this comparison results in the definition of the convergence measure,

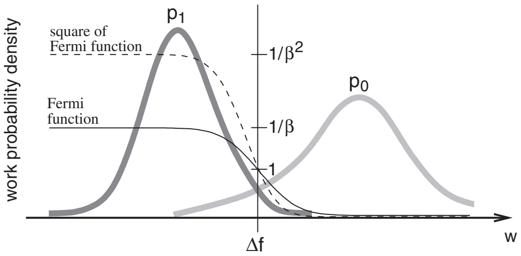

FIG. 3. Schematic plot which shows that the forward work density, $p_0(w)$, samples the Fermi function $b_{\Delta f}(w)=1/(\beta+\alpha e^{w-\Delta f})$ somewhat earlier than its square.

$$a=\frac{\hat{U}_\alpha-\hat{U}_\alpha^{(II)}}{\hat{U}_\alpha}, \tag{20}$$

for all $\alpha\in(0,1)$. Clearly, in the large N limit, a will converge to zero, as then $\widehat{\Delta f}$ converges to Δf and thus $\hat{U}_\alpha$ as well as $\hat{U}_\alpha^{(II)}$ converge to U_α. As argued below, it is the estimate $\hat{U}_\alpha^{(II)}$ that converges last, hence a converges somewhat later than $\widehat{\Delta f}$.

Below the large N limit, a will deviate from zero. From the general inequality

$$\hat{U}_\alpha^2\leq\hat{U}_\alpha^{(II)}<2\hat{U}_\alpha, \tag{21}$$

(see Appendix B) follow upper and lower bounds on a which read

$$-1<a\leq 1-\hat{U}_\alpha<1. \tag{22}$$

The behavior of a with increasing sample size $N=n_0+n_1$ (while keeping the fraction $\alpha=\frac{n_0}{N}$ constant) can roughly be characterized as follows: a "starts" close to its upper bound for small N and decreases towards zero with increasing N. Finally, a begins to fluctuate around zero when the large N limit is reached, i.e., when the estimate $\widehat{\Delta f}$ converges.

To see this qualitatively, we state that the second order estimate $\hat{U}_\alpha^{(II)}$ converges later than the first order estimate $\hat{U}_\alpha$, as the former requires sampling the tails of p_0 and p_1 to a somewhat wider extend than the latter, cf. Fig. 3. For small N, both, $\hat{U}_\alpha$ and $\hat{U}_\alpha^{(II)}$, will typically underestimate U_α, as the "rare events" which contribute substantially to the averages (16) and (19) are quite likely not to be observed sufficiently, if at all. For the same reason, generically $\hat{U}_\alpha^{(II)}<\hat{U}_\alpha$ will hold, since $b_{\widehat{\Delta f}}(w^0)^2\leq b_{\widehat{\Delta f}}(w^0)$ holds for $w^0\geq\widehat{\Delta f}$ and similar $t_{\widehat{\Delta f}}(w^1)^2\leq t_{\widehat{\Delta f}}(w^1)$ for $w^1\leq\widehat{\Delta f}$. Therefore, a is typically positive for small N. In particular, if N is so small that *all* work values of the forward sample are larger than $\widehat{\Delta f}$ and all work values of the reverse sample are smaller than $\widehat{\Delta f}$, then $\hat{U}_\alpha^{(II)}$ becomes much smaller than $\hat{U}_\alpha$, resulting in $a\approx 1$.

Analytic insight into the behavior of a for small N results from the fact that $n\bar{x}^2\geq\overline{x^2}$ for any set $\{x_1,\dots x_n\}$ of positive numbers x_k. Using this in Eq. (19) yields

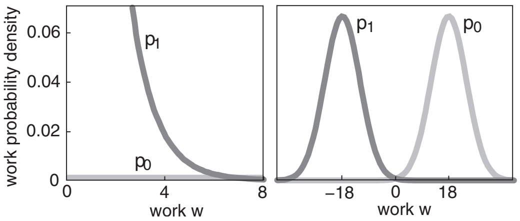

FIG. 4. Exponential (left panel) and Gaussian (right panel) work densities.

$$\hat{U}_\alpha^{(II)}\leq 2N\alpha\beta\hat{U}_\alpha^2, \tag{23}$$

and

$$1-2\alpha\beta N\hat{U}_\alpha\leq a\leq 1-\hat{U}_\alpha. \tag{24}$$

This shows that as long as $N\hat{U}_\alpha\ll 1$ holds, a is close to its upper bound $1-\hat{U}_\alpha\approx 1$. In particular, if $\alpha=\frac{1}{2}$ and $N=2$, then $a=1-\hat{U}_\alpha$ holds exactly.

Averaging the inequality for some N sufficiently large to ensure $\langle a\rangle\approx 0$ and $\langle\hat{U}_\alpha\rangle\approx U_\alpha$, we get a lower bound on N which reads $N\geq\frac{1}{2\alpha\beta U_\alpha}$. Again, this bound can be related to the overlap area $\mathcal{A}$ taking $\alpha=\frac{1}{2}$ and using $U_{1/2}\leq 2\mathcal{A}$ (see Appendix A), we obtain $N\geq\frac{1}{\mathcal{A}}$, in concordance with the lower bound for the large N limit stated in Sec. I.

Last we note that the convergence measure a can also be understood as a measure of the sensibility of relation (2) with respect to the value of $\widehat{\Delta f}$. In the low N regime, the relation is highly sensible to the value of $\widehat{\Delta f}$, resulting in large values of a, whereas in the limit of large N, relation (2) becomes insensible to small perturbations of $\widehat{\Delta f}$, corresponding to $a\approx 0$. The details are summarized in Appendix D.

IV. STUDY OF STATISTICAL PROPERTIES OF THE CONVERGENCE MEASURE

In order to demonstrate the validity of a as a measure of convergence of two-sided free-energy estimation, we apply it to two qualitatively different types of work densities, namely exponential and Gaussian, see Fig. 4. Samples from these densities are easily available by standard (pseudo)random generators. Statistical properties of a are obtained by means of independent repeated calculations of $\widehat{\Delta f}$ and a. While the two types of densities used are fairly simple, they are entirely different and general enough to reflect the statistical properties of the convergence measure.

A. Exponential work densities

The first example uses exponential work densities, i.e.,

$$p_i(w)=\frac{1}{\mu_i}e^{-w/\mu_i},\quad w\geq 0, \tag{25}$$

$\mu_i>0$, $i=0,1$. According to the fluctuation theorem (1), the mean values μ_i of p_0 and p_1 are related to each other, μ_1

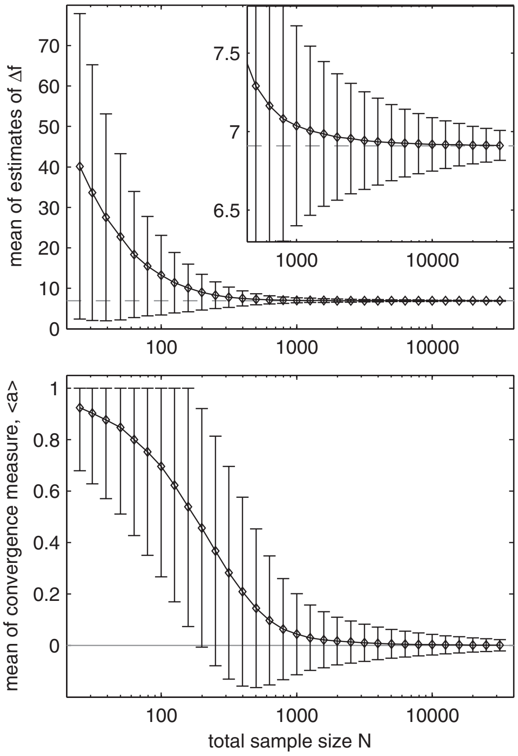

FIG. 5. Statistics of two-sided free-energy estimation (exponential work densities): shown are averaged estimates of Δf in dependence of the total sample size N. The error bars reflect the standard deviation. The dashed line shows the exact value of Δf and the inset the details for large N (top). Statistics of the convergence measure a corresponding to the estimates of the top panel: shown are the average values of a together with their standard deviation in dependence of the sample size N. Note the characteristic convergence of a towards zero in the large N limit (bottom).

$=\frac{\mu_0}{1+\mu_0}$, and the free-energy difference is known to be $\Delta f = \ln(1+\mu_0)$.

Choosing $\mu_0=1000$ and $\alpha=\frac{1}{2}$, i.e., $n_0=n_1$, we calculate free-energy estimates $\widehat{\Delta f}$ according to Eq. (2) together with the corresponding values of a according to Eq. (20) for different total sample sizes $N=n_0+n_1$. An example of a single running estimate and the corresponding values of the convergence measure are depicted in Fig. 1. Ten thousand repetitions for each value of N yield the results presented in Figs. 5–10. To begin with, the top panel of Fig. 5 shows the averaged free-energy estimates in dependence of N, where the errorbars show ± the estimated square root of the variance $\langle(\widehat{\Delta f}-\langle\widehat{\Delta f}\rangle)^2\rangle$. For small N, the bias $\langle\widehat{\Delta f}-\Delta f\rangle$ of free-energy estimates is large, but becomes negligible compared to the standard deviation for $N \gtrsim 5000$. This is a prerequisite of the large N limit, therefore we will view $N\approx 5000$ as the onset of the large N limit.

The bottom panel of Fig. 5 shows the averaged values of the convergence measure a corresponding to the free-energy estimates of the top panel. Again, the errorbars are ± one standard deviation $\sqrt{\langle a^2\rangle - \langle a\rangle^2}$, except that the upper limit is truncated for small N, as $a<1$ holds. The trend of the averaged convergence measure $\langle a\rangle$ is in full agreement with the general considerations given in the previous section. For small N, $\langle a\rangle$ starts close to its upper bound, decreases monotonically with increasing sample size, and converges towards zero in the large N limit. At the same time, its standard deviation converges to zero, too. This indicates that single values of a corresponding to single estimates $\widehat{\Delta f}$ will typically be found close to zero in the large N regime.

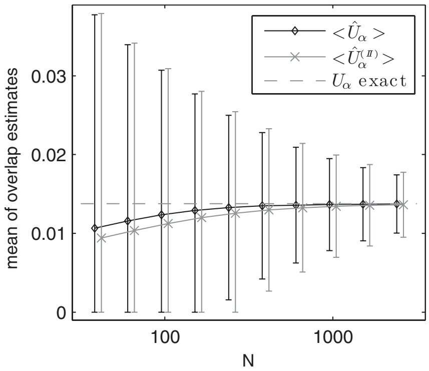

FIG. 6. Mean values of overlap estimates $\hat{U}_\alpha$ and $\hat{U}_\alpha^{(II)}$ of first and second order, respectively. The slightly slower convergence of $\hat{U}_\alpha^{(II)}$ towards U_α results in the characteristic properties of the convergence measure a. To enhance clarity, data points belonging to the same value of N are spread.

Noting that a is defined as relative difference of the overlap estimators: $\hat{U}_\alpha$ and $\hat{U}_\alpha^{(II)}$ of first and second order, respectively, we can understand the trend of the average convergence measure by taking into consideration the average values $\langle\hat{U}_\alpha\rangle$ and $\langle\hat{U}_\alpha^{(II)}\rangle$, which are shown in Fig. 6. For small sample sizes, U_α is *typically* underestimated by both, $\hat{U}_\alpha$ and $\hat{U}_\alpha^{(II)}$, with $\hat{U}_\alpha^{(II)} < \hat{U}_\alpha$.

The convergence measure takes advantage of the different convergence times of the overlap estimators: $\hat{U}_\alpha^{(II)}$ converges

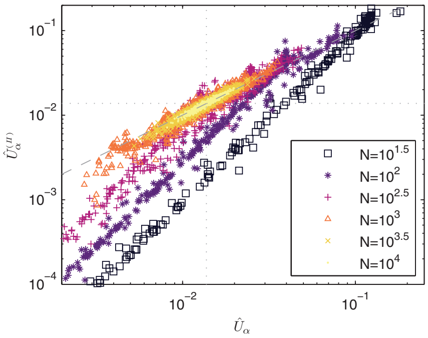

FIG. 7. (Color online) Double-logarithmic scatter plot of $\hat{U}_\alpha^{(II)}$ versus $\hat{U}_\alpha$ for many individual estimates in dependence of the sample size N. The dotted lines mark the exact value of U_α on the axes and the dashed line is the bisectrix $\hat{U}_\alpha^{(II)}=\hat{U}_\alpha$. The approximatively linear relation between the logarithms of $\hat{U}_\alpha^{(II)}$ and $\hat{U}_\alpha$ is continued up to the smallest observed values ($<10^{-100}$, not shown here).

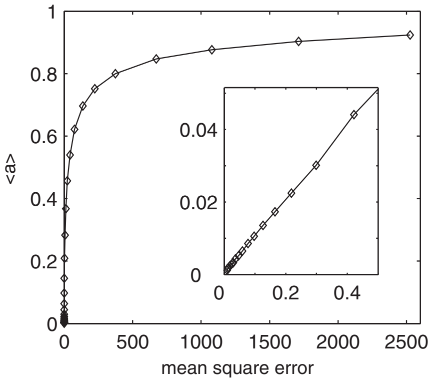

FIG. 8. The average convergence measure $\langle a\rangle$ plotted against the corresponding mean-square error $\langle(\widehat{\Delta f}-\Delta f)^2\rangle$ of the two-sided free-energy estimator. The inset shows an enlargement for small values of $\langle a\rangle$.

somewhat slower than $\hat U_\alpha$, ensuring that a approaches zero right after $\widehat{\Delta f}$ has converged. The large standard deviations shown as errorbars in Fig. 6 do not carry over to the standard deviation of a, because $\hat U_\alpha$ and $\hat U_\alpha^{(II)}$ are strongly correlated, as is impressively visible in Fig. 7. The estimated correlation coefficient

$$\frac{\langle(\hat U_\alpha^{(II)}-\langle\hat U_\alpha^{(II)}\rangle)(\hat U_\alpha-\langle\hat U_\alpha\rangle)\rangle}{\sqrt{\mathrm{Var}(\hat U_\alpha^{(II)})\mathrm{Var}(\hat U_\alpha)}}, \qquad (26)$$

is about 0.97 for the entire range of sample sizes N. In good approximation, $\hat U_\alpha$ and $\hat U_\alpha^{(II)}$ are related to each other according to a power law, $\hat U_\alpha^{(II)}\approx c_N\hat U_\alpha^{\gamma_N}$, where the exponent γ_N and the prefactor c_N depend on the sample size N (and α). We note that γ_N has a phase-transitionlike behavior: for small N, it stays approximately constant near two; right before the onset of the large N limit, it shows a sudden switch to a value close to one where it finally remains.

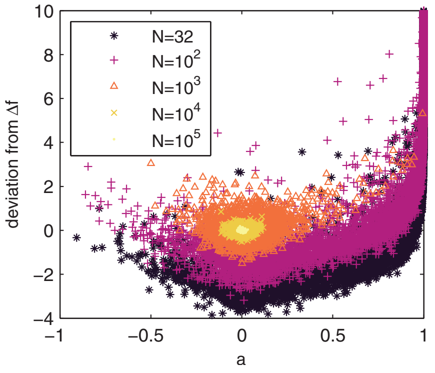

FIG. 9. (Color online) A scatter plot of the deviation $\widehat{\Delta f}-\Delta f$ versus the convergence measure a for many individual estimates in dependence of the sample size N. Note that the majority of estimates belonging to $N=32$ and 100 have large values of $\widehat{\Delta f}-\Delta f$ well outside the displayed range with a being close to one.

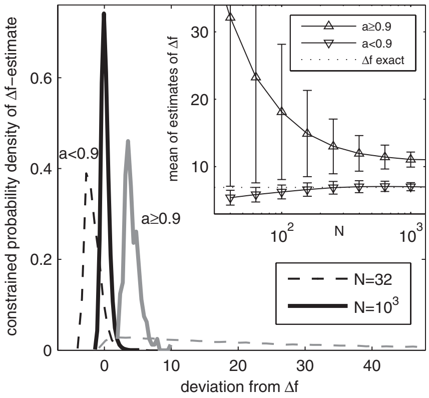

FIG. 10. Estimated constrained probability densities $p(\widehat{\Delta f}|a<0.9)$ (black) and $p(\widehat{\Delta f}|a\geq 0.9)$ (grayscale) for two different sample sizes N, plotted versus the deviation $\widehat{\Delta f}-\Delta f$. The inset shows averaged estimates of Δf over the total sample size N subject to the constraints $a\geq 0.9$ and $a<0.9$, respectively.

Figure 8 accents the decrease in the average $\langle a\rangle$ with decreasing mean square error (5) of two-sided estimation. The small N behavior is given by the upper right part of the graph, where $\langle a\rangle$ is close to its upper bound together with a large mean-square error of $\widehat{\Delta f}$. With increasing sample size, the mean-square error starts to drop somewhat sooner than $\langle a\rangle$, however, at the onset of the large N limit, they drop both and suggest a linear relation, as can be seen in the inset for small values of $\langle a\rangle$. The latter shows that $\langle a\rangle$ decreases to zero proportional to $\frac{1}{N}$ for large N (this is confirmed by a direct check, but not shown here).

The next point is to clarify the correlation of single values of the convergence measure with their corresponding free-energy estimates. For this issue, figure 9 is most informative, showing the deviations $\widehat{\Delta f}-\Delta f$ in dependence of the corresponding values of a for many individual observations. The figure makes clear that there is a *strong relation, but no one-to-one correspondence* between a and $\widehat{\Delta f}-\Delta f$: for large N, both a and $\widehat{\Delta f}-\Delta f$ approach zero with very weak correlations between them. However, the situation is different for small sample sizes N where the bias $\langle\widehat{\Delta f}-\Delta f\rangle$ is considerably large. There, the typically observed large deviations occur together with values of a close to the upper bound, whereas the atypical events with small (negative) deviations come together with values of a well below the upper limit. Therefore, small values of a detect exceptional events if N is well below the large N limit, and ordinary events if N is large.

To make this relation more visible, we split the estimates $\widehat{\Delta f}$ into the mutually exclusive events $a\geq 0.9$ and $a<0.9$. The statistics of the $\widehat{\Delta f}$ values within these cases are depicted in the inset of Fig. 10, where normalized histograms, i.e. estimates of the constrained probability densities $p(\widehat{\Delta f}|a\geq 0.9)$ and $p(\widehat{\Delta f}|a<0.9)$ are shown. The unconstrained probability density of $\widehat{\Delta f}$ can be reconstructed from a likelihood weighted sum of the constrained densities, $p(\widehat{\Delta f})$

$=p(\widehat{\Delta f}|a\geq 0.9)p_{a\geq 0.9}+p(\widehat{\Delta f}|a<0.9)p_{a<0.9}$. The likelihood ratios read $p_{a\geq 0.9}/p_{a<0.9}=6.2$ and 0.002 for $N=32$ and 1000, respectively. Finally, the inset of Fig. 10 shows the average values of constrained estimates $\widehat{\Delta f}$ over N with errorbars of ± one standard deviation, in dependence of the condition on a.

B. Gaussian work densities

For the second example the work densities are chosen to be Gaussian,

$$p_i(w)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(w-\mu_i)^2/2\sigma^2},\quad w\in\mathbb{R},\tag{27}$$

$i=0,1$. The fluctuation theorem (1) demands both densities to have the same variance σ^2 with mean values $\mu_0=\Delta f+\frac{1}{2}\sigma^2$ and $\mu_1=\Delta f-\frac{1}{2}\sigma^2$. Hence, p_0 and p_1 are symmetric to each other with respect to Δf, $p_0(\Delta f+w)=p_1(\Delta f-w)$. As a consequence of this symmetry, the two-sided estimator with equal sample sizes n_0 and n_1, i.e. $\alpha=0.5$, is unbiased for any N. However, this does not mean that the limit of large N is reached immediately.

In analogy to the previous example, we proceed in presenting the statistical properties of a. Choosing $\sigma=6$ and without loss of generality $\Delta f=0$, we carry out 10^4 estimations of Δf over a range of sample sizes N. The forward fraction is chosen to be equal to $\alpha=0.5$, and for comparison, $\alpha=0.999$, and $\alpha=0.99999$, respectively. In the latter two cases, the two-sided estimator is biased for small N. We note that $\alpha=0.5$ is always the optimal choice for symmetric work densities which minimizes the asymptotic mean-square error (6) with respect to α [32].

Comparing the top and the bottom panel of Fig. 11, which show the statistics (mean value and standard deviation as error bars) of the observed estimates $\widehat{\Delta f}$ and of the corresponding values of a, we find a coherent behavior for all three cases of α values. The trend of the average $\langle a\rangle$ shows in all cases the same features in agreement with the trend found for exponential work densities.

As before, the characteristics of a are understood by the slower convergence of $\hat{U}_\alpha^{(II)}$ compared to that of $\hat{U}_\alpha$, as can be seen in Fig. 12. A scatter plot of $\hat{U}_\alpha^{(II)}$ versus $\hat{U}_\alpha$ looks qualitatively such as Fig. 7, but is not shown here.

Figure 13 compares the average convergence measures as functions of the mean-square error of $\widehat{\Delta f}$ for the three values of α. For the range of small $\langle a\rangle$, all three curves agree and are linear. Again $\langle a\rangle$ decreases proportionally to $\frac{1}{N}$ for large N. Noticeable for small N is the shift of $\langle a\rangle$ towards smaller values with increasing α. This results from the definition of a: the upper bound $1-\hat{U}_\alpha$ of a tends to zero in the limits $\alpha\to 0,1$, as then $\hat{U}_\alpha\to 1$.

The relation of single free-energy estimates $\widehat{\Delta f}$ with the corresponding a values can be seen in the scatter plot of Fig. 14. The mirror symmetry of the plot originates from the symmetry of the work densities and the choice $\alpha=0.5$, i.e., of the unbiasedness of the two-sided estimator. Opposed to the foregoing example, the correlation between $\widehat{\Delta f}-\Delta f$ and a vanishes for any value of N. Despite the lack of any correlation, the figure reveals a strong relation between the deviation $\widehat{\Delta f}-\Delta f$ and the value of a: they converge equally to zero for large N.

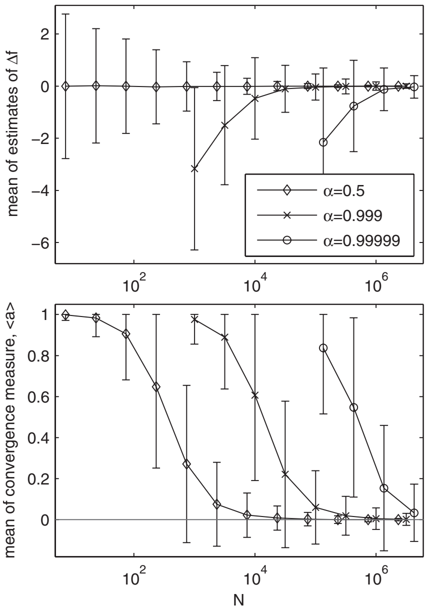

FIG. 11. Gaussian work densities result in the displayed averaged estimates of Δf. For comparison, three different fractions α of forward work values are used (top). Average values of the convergence measure a corresponding to the estimates of the top panel (bottom).

Last, Fig. 15 shows averages of constrained Δf estimates for the mutually exclusive conditions $a\geq 0.9$ and $a<0.9$, now with $\alpha=0.99999$ in order to incorporate some bias. We observe the same characteristics as before, cf. the inset of Fig. 10: the condition $a<0.9$ filters the estimates $\widehat{\Delta f}$ which are closer to the true value.

C. General case

The characteristics of the convergence measure are dominated by contributions of work densities inside and near the

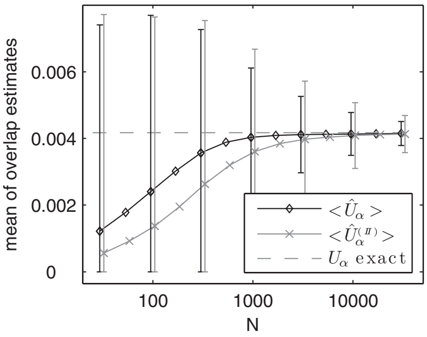

FIG. 12. Mean values of overlap estimates $\hat{U}_\alpha$ and $\hat{U}_\alpha^{(II)}$ of first and second order ($\alpha=0.5$).

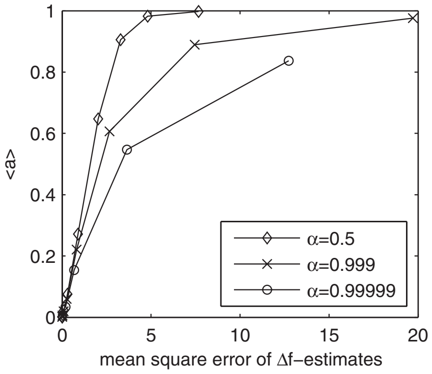

FIG. 13. The average convergence measure $\langle a \rangle$ plotted against the corresponding mean square error $\langle (\widehat{\Delta f} - \Delta f)^2 \rangle$ of free-energy estimates in dependence of N.

region where the overlap density $p_\alpha(w)$, Eq. (11), has most of its mass. We call this region the overlap region. In the overlap region, the work densities may have one of the following characteristic relation of shape:

(1) Having their maxima at larger and smaller values of work, respectively, the forward and reverse work densities both drop towards the overlap region. Hence, any of both densities sample the overlap region by rare events, only, which are responsible for the behavior of the convergence measure.

(2) Both densities decrease with increasing w and the overlap region is well sampled by the forward work density compared with the reverse density. Especially the "rare" events $w < \Delta f$ of forward direction are much more available than the rare events $w > \Delta f$ of reverse direction. Hence, more or less typical events of one direction together with atypical events of the other direction are responsible for the behavior of the convergence measure. Likewise if both densities increase with w.

(3) More generally, the work densities are some kind of interpolation between the above two cases.

(4) Finally, there remain some exceptional cases. For instance, if the forward and reverse work densities have different support or if they do not obey the fluctuation theorem at all.

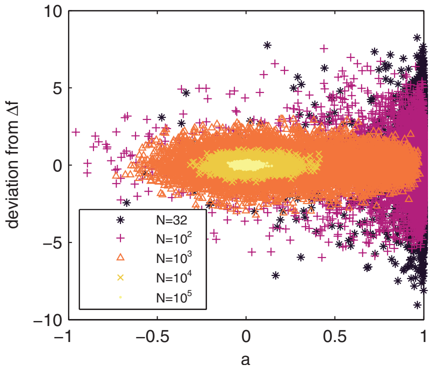

FIG. 14. (Color online) A scatter plot of the deviation $\widehat{\Delta f} - \Delta f$ versus the convergence measure a for many individual estimates in dependence of the sample size N (α=0.5).

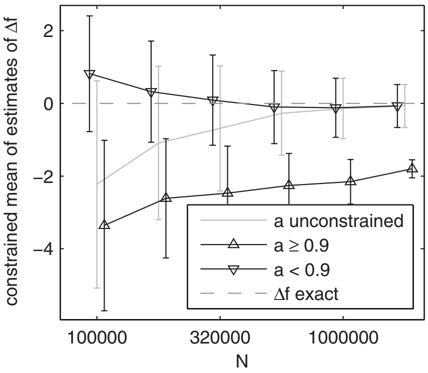

FIG. 15. Averaged two-sided estimates of Δf in dependence of the total sample size N for the constraints $a \geq 0.9$, a unconstrained, and $a < 0.9$ (α=0.99999).

With respect to the exceptional case, the convergence measure fails to work, since it requires that the forward and reverse work densities have the same support and that the densities are related to each other via the fluctuation theorem (1).

In all other cases, the convergence measure certainly will work and will show a similar behavior, regardless of the detailed nature of the densities. This can be explained as follows. In the preceding subsections, we have investigated exponential and Gaussian work densities, two examples that differ in their very nature. While exponential work densities cover case number two and Gaussians cover case number one, they show the same characteristics of a. This means that the characteristics of the convergence measure are insensitive to the individual nature of the work densities as long as they have the same support and obey the fluctuation theorem.

To this end, we want to point to some subtleties in the text of the actual paper. While the measure of convergence is robust with respect to the nature of work densities, some heuristic or pedagogic explanations in the text are written with regard to the typical case number one, where the overlap region is sampled by rare events, only. This concerns mainly Sec. II where we speak about effective cut-off values in the context of the neglected tail model. These effective cut-off values would become void if we would try to explain the bias of exponential work densities qualitatively via the neglected tail model. Also the explanations in the text of the next section are mainly focused on the typical case number one. This concerns the passages where we speak about rare events. Nevertheless, the main and essential statements are valid for all cases.

The most important property of a is its almost *simultaneous* convergence with the free-energy estimator $\widehat{\Delta f}$ to an *a priori known* value. This fact is used to develop a convergence criterion in the next section.

V. CONVERGENCE CRITERION

Elaborated the statistical properties of the convergence measure, we are finally interested in the convergence of a *single* free-energy estimate. In contrast to averages of many independent running estimates, estimates based on individual realization are not smooth in N, see e.g., Fig. 1.

For small N, typically $\hat{U}_\alpha^{(II)}$ underestimates U_α more than $\hat{U}_\alpha$ does, pushing a close to its upper bound. With increasing N, $\widehat{\Delta f}$ starts to "converge;" typically in a nonsmooth manner. The convergence of $\widehat{\Delta f}$ is triggered by the occurrence of rare events. Whenever such a rare event in the important tails of the work densities gets sampled, $\widehat{\Delta f}$ jumps, and between such jumps, $\widehat{\Delta f}$ stays rather on a stable plateau. The measure a is triggered by the same rare events, but the changes in a are smaller, unless convergence starts happening. Typically, the rare events that bring $\widehat{\Delta f}$ near to its true value are the rare events which change the value of a drastically. In the typical case, these rare events let a even undershoot below zero, before $\widehat{\Delta f}$ and a finally converge.

The features of the convergence measure,

(1) it is bounded, $a \in (-1, 1-\hat{U}_\alpha]$,

(2) it starts for small N at its upper bound,

(3) it converges to a known value, $a \to 0$,

(4) and typically it converges almost simultaneously with $\widehat{\Delta f}$,

simplify the task of monitoring the convergence significantly, since it is far easier to compare estimates of a with the known value zero than the task of monitoring convergence of $\widehat{\Delta f}$ to an unknown target value. The characteristics of the convergence measure enable us to state: typically, if it is close to zero, $\widehat{\Delta f}$ has converged.

Deviations from the typical situation are possible. For instance, $\widehat{\Delta f}$ may not show such clear jumps, neither may a. Occasionally, $\widehat{\Delta f}$ and a, may also fluctuate exceedingly strong. Thus, a single value of a close to zero does not guarantee convergence of the free-energy estimate as can be seen from some few individual events in the scatter plot of Fig. 14 that fail a correct estimate while a is close to zero. A single random realization may give rise to a fluctuation that brings a close to zero by chance, a fact that needs to be distinguished from a having converged to zero. The difference between random chance and convergence is revealed by increasing the sample size, since it is highly unlikely that a stays close to zero by random. It is the *behavior* of a with increasing N, that needs to be taken into account in order to establish an equivalence between $a \to 0$ and $\widehat{\Delta f} \to \Delta f$.

This allows us to state the convergence criterion: *if* a fluctuates close around zero, convergence is assured, implying that if a fluctuates around zero, $\widehat{\Delta f}$ fluctuates around its true value Δf, the bias vanishes, and the mean-square error reaches its asymptotics which can be estimated using Eq. (8). a fluctuating close around zero means that it does so over a suitable range of sample sizes, which extends over an order of magnitude or more.

VI. APPLICATION

As an example, we apply the convergence criterion to the calculation of the excess chemical potential μ^{ex} of a Lennard-Jones fluid. Using Metropolis Monte Carlo simulation [34] of a fluid of N_p particles, the forward work is defined as energy increase when inserting at random a particle into a given configuration [35], whereas the reverse work is defined as energy decrease when a random particle is deleted from a given N_p+1-particle configuration. The densities $p_0(w)$ and $p_1(w)$ of forward and reverse work obey the fluctuation theorem (1) with $\Delta f = \mu^{ex}$ [20]. Thus, Bennett's acceptance ratio method can be applied to the calculation of the chemical potential.

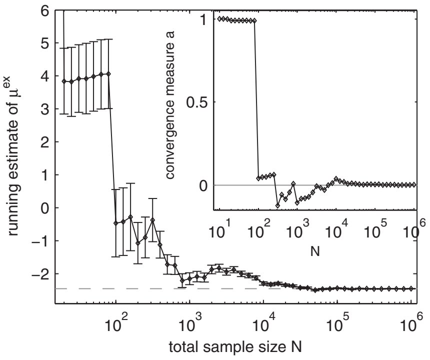

FIG. 16. Running estimates of the excess chemical potential μ^{ex} in dependence of the sample size N ($\alpha=0.9$). The inset displays the corresponding values of the convergence measure a.

Details of the simulation are reported in Ref. [20]. Here, the parameter values chosen read: $N_p=120$, reduced temperature $T^*=1.2$, and reduced density $\rho^*=0.5$.

Drawing work values up to a total sample size of 10^6 with fraction $\alpha=0.9$ of forward draws (which will be a near-optimal choice [32]), the successive estimates of the chemical potential together with the corresponding values of the convergence measure are shown in Fig. 16. The dashed horizontal line does not show the exact value of μ^{ex}, which is unknown, but rather the value of the last estimate with $N=10^6$. Taking a closer look on the behavior of the convergence measure with increasing N, we observe a near unity for $N \leq 10^2$, indicating the low N regime and the lack of observing rare events. Then, a sudden drop near to zero happens at $N=10^2$, which coincides with a large jump of the estimate of μ^{ex}, followed by large fluctuations of a with strong negative values in the regime $N=10^2$ to 10^4. This behavior indicates that the important but rare events which trigger the convergence of the μ^{ex} estimate are now sampled, but with strongly fluctuating relative frequency, which in specific cases causes the negative values of a (because of too many rare events). Finally, with $N>10^4$, a equilibrates and converges to zero. The latter is observed over two orders of magnitude, such that we can conclude that the latest estimate of μ^{ex} with $N=10^6$ has surely converged and yields a reliable value of the chemical potential. The confidence interval of the estimate can safely be calculated as the square root of Eq. (6) (one standard deviation) and we obtain explicitly $\widehat{\mu^{ex}} = -2.451 \pm 0.005$.

Interested in the statistical behavior of a for the present application, we carried out 270 simulation runs up to $N=10^4$ to obtain the average values and standard deviations of

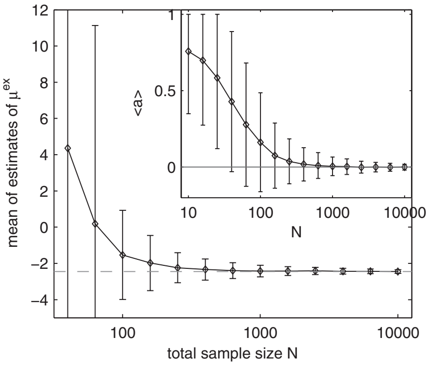

FIG. 17. Statistics of estimates of the excess chemical potential shown are the average value and the standard deviation (as error bars) in dependence of the sample size N. The statistics of the corresponding values of the convergence measure is shown in the inset.

$\widehat{\mu^{ex}}$ and a which are depicted in Fig. 17. The dashed line marks the same value as that in Fig. 16. Again, we observe the same qualitative behavior of a as in the foregoing examples of Sec. IV, especially a positive average value of $\langle a\rangle$ and a convergence to zero which occurs simultaneously with the convergence of Bennett's acceptance ratio method.

VII. CONCLUSIONS

Since its formulation a decade ago, the Jarzynski equation and the Crooks fluctuation theorem gave rise to enforced research of nonequilibrium techniques for free-energy calculations. Despite the variety of methods, in general little is known about their statistical properties. In particular, it is often unclear whether the methods actually converge to the desired value of the free-energy difference Δf, and if so, it remains in question whether convergence happened within a given calculation. This is of great concern, as usually the calculations are strongly biased before convergence starts happening. In consequence, it is impossible to state the result of a single calculation of Δf with a reliable confidence interval unless a convergence measure is evaluated.

In this paper, we presented and tested a quantitative measure of convergence for two-sided free-energy estimation, i.e., Bennett's acceptance ratio method, which is intimately related to the fluctuation theorem. From this follows a criterion for convergence relying on monitoring the convergence measure a within a running estimation of Δf. The heart of the convergence criterion is the nearly simultaneous convergence of the free-energy calculation and the convergence measure a. Whereas the former converges towards the unknown value Δf, which makes it difficult or even impossible to decide when convergence actually takes place, the latter converges to an *a priori known* value. If convergence is detected with the convergence criterion, the calculation results in a reliable estimate of the free-energy difference together with a precise confidence interval.

APPENDIX A

The derivation of inequality (10) relies on the close connection between the overlap U_α and the overlap area $\mathcal{A}$,

$$U_\alpha = \int \frac{p_0 p_1}{\alpha p_0 + \beta p_1} dw \geq \int \frac{p_0 p_1}{(\alpha+\beta)\max\{p_0,p_1\}} dw = \int \min\{p_0,p_1\} dw = \mathcal{A}, \tag{A1}$$

$$U_{1/2} = 2\int \frac{1}{1/p_1 + 1/p_0} dw < 2\int \min\{p_0,p_1\} dw = 2\mathcal{A}. \tag{A2}$$

Together with the inequality $\frac{1}{2}X(N,\frac{1}{2}) \leq X(N,\alpha)$ of Bennett [23], we obtain

$$\frac{1-2\mathcal{A}}{N\mathcal{A}} < \frac{1-U_{1/2}}{\frac{1}{2}NU_{1/2}} = \frac{1}{2}X\left(N,\frac{1}{2}\right) \leq X(N,\alpha) \leq \frac{1}{N}\frac{1}{\alpha\beta}\left(\frac{1}{\mathcal{A}}-1\right) \tag{A3}$$

which directly yields inequality (10).

APPENDIX B

Inequality (21) can be obtained as follows. Noting that $t_c(w) < \frac{1}{\alpha}$ and $b_c(w) < \frac{1}{\beta}$, cf. Eq. (14) we have

$$2\hat{U}_\alpha = \overline{t_{\widehat{\Delta f}}}^{(1)} + \overline{b_{\widehat{\Delta f}}}^{(0)} > \alpha\overline{t^2_{\widehat{\Delta f}}}^{(1)} + \beta\overline{b^2_{\widehat{\Delta f}}}^{(0)} = \hat{U}_\alpha^{(II)} \tag{B1}$$

and further,

$$\hat{U}_\alpha^{(II)} = \hat{U}_\alpha^2 + \alpha\overline{(t_{\widehat{\Delta f}} - \hat{U}_\alpha)^2}^{(1)} + \beta\overline{(b_{\widehat{\Delta f}} - \hat{U}_\alpha)^2}^{(0)} \geq \hat{U}_\alpha^2, \tag{B2}$$

which results in Eq. (21).

APPENDIX C

The error bars in Figs. 1 and 16 are obtained via the error-propagation formula for the variance of Bennett's acceptance ratio method.

A possible estimate $\hat{\sigma}^2_{ep}$ of the variance of the two-sided free-energy estimator obtained from error propagation reads

$$\hat{\sigma}^2_{ep} = \frac{1}{n_1}\frac{\overline{t^2_{\widehat{\Delta f}}}^{(1)} - \overline{t_{\widehat{\Delta f}}}^{(1)2}}{\overline{t_{\widehat{\Delta f}}}^{(1)2}} + \frac{1}{n_0}\frac{\overline{b^2_{\widehat{\Delta f}}}^{(0)} - \overline{b_{\widehat{\Delta f}}}^{(0)2}}{\overline{b_{\widehat{\Delta f}}}^{(0)2}}. \tag{C1}$$

Alternatively, $\hat{\sigma}^2_{ep}$ can be expressed through the overlap estimates $\hat{U}_\alpha$ and $\hat{U}_\alpha^{(II)}$ of first and second order, Eqs. (16) and (19),

$$\hat{\sigma}^2_{ep} = \frac{1}{\alpha\beta N}\frac{\hat{U}_\alpha^{(II)} - \hat{U}_\alpha^2}{\hat{U}_\alpha^2}. \tag{C2}$$

In the limit of large N, $\hat{\sigma}^2_{ep}$ converges to the asymptotic mean square error $X(N,\alpha)$, Eq. (6). An upper bound on $\hat{\sigma}^2_{ep}$ follows from inequality (23):

$$\hat{\sigma}^2_{ep} \le 2 - \frac{1}{\alpha\beta N}. \tag{C3}$$

Finally let us mention that the convergence measure a, Eq. (20), is closely related to the relative difference of the estimated asymptotic mean square error $\hat{X}$, Eq. (8), and $\hat{\sigma}^2_{ep}$

$$a = (1 - \hat{U}_\alpha)\frac{\hat{X} - \hat{\sigma}^2_{ep}}{\hat{X}}. \tag{C4}$$

APPENDIX D

Consider the family $\hat{\phi}(c)$ of Δf estimators, parameterized by the real number c [23]

$$\hat{\phi}(c) = c + \ln \frac{\overline{t_c}^{(1)}}{\overline{b_c}^{(0)}}. \tag{D1}$$

For any fixed value of c, $\hat{\phi}(c)$ defines a consistent estimator of Δf, $\hat{\phi}(c) \xrightarrow{N\to\infty} \Delta f \ \forall c$. For finite N, however, the performance of the estimator strongly depends on c. The (optimal) two-sided estimate (2) is obtained by the additional condition $\hat{\phi}(c)=c$ such that $\overline{t_c}^{(1)}=\overline{b_c}^{(0)}$ holds, and thus $c=\widehat{\Delta f}$. A possible measure for the sensibility of the estimate $\hat{\phi}(c)$ on c is it's derivative with respect to c. Using $\frac{\partial}{\partial c}t_c = -\beta t_c b_c$, $\frac{\partial}{\partial c}b_c = \alpha t_c b_c$, and $\alpha t_c + \beta b_c = 1$, we obtain

$$\frac{\partial}{\partial c}\hat{\phi}(c) = -1 + \alpha \frac{\overline{t_c^2}^{(1)}}{\overline{t_c}^{(1)}} + \beta \frac{\overline{b_c^2}^{(0)}}{\overline{b_c}^{(0)}}. \tag{D2}$$

Taking the derivative at $c=\widehat{\Delta f}$ directly results in the convergence measure a,

$$\frac{\partial}{\partial c}\hat{\phi}(c)\Big|_{\widehat{\Delta f}} = -a. \tag{D3}$$

[1] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[2] A. Gelman and X.-L. Meng, Stat. Sci. **13**, 163 (1998).
[3] D. D. L. Minh and J. D. Chodera, J. Chem. Phys. **131**, 134110 (2009).
[4] R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[5] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[6] M.-H. Chen and Q.-M. Shao, Ann. Stat. **25**, 1563 (1997).
[7] H. Oberhofer and C. Dellago, Comput. Phys. Commun. **179**, 41 (2008).
[8] M. Watanabe and W. P. Reinhardt, Phys. Rev. Lett. **65**, 3301 (1990).
[9] S. X. Sun, J. Chem. Phys. **118**, 5769 (2003).
[10] F. M. Ytreberg and D. M. Zuckerman, J. Chem. Phys. **120**, 10876 (2004).
[11] C. Jarzynski, Phys. Rev. E **73**, 046105 (2006).
[12] H. Ahlers and A. Engel, Eur. Phys. J. B **62**, 357 (2008).
[13] H. Then and A. Engel, Phys. Rev. E **77**, 041105 (2008).
[14] P. Geiger and C. Dellago, Phys Rev. E **81**, 021127 (2010).
[15] A. Engel, Phys. Rev. E **80**, 021120 (2009).
[16] X.-L. Meng and S. Schilling, J. Comput. Graph. Statist. **11**, 552 (2002).
[17] C. Jarzynski, Phys. Rev. E **65**, 046122 (2002).
[18] H. Oberhofer, C. Dellago, and S. Boresch, Phys. Rev. E **75**, 061106 (2007).
[19] S. Vaikuntanathan and C. Jarzynski, Phys. Rev. Lett. **100**, 190601 (2008).
[20] A. M. Hahn and H. Then, Phys. Rev. E **79**, 011113 (2009).
[21] N. Lu and T. B. Woolf, in *Free Energy Calculations*, Springer Series in Chemical Physics Vol. 86, edited by Ch. Chipot and A. Pohorille (Springer, Berlin, 2007), pp. 199–247.
[22] P. Maragakis, F. Ritort, C. Bustamante, M. Karplus, and G. E. Crooks, J. Chem. Phys. **129**, 024102 (2008).
[23] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[24] X.-L. Meng and W. H. Wong, Stat. Sinica **6**, 831 (1996).
[25] A. Kong, P. McCullagh, X.-L. Meng, D. Nicolae, and Z. Tan, J. R. Stat. Soc. Ser. B (Methodol.) **65**, 585 (2003).
[26] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
[27] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[28] M. Campisi, P. Talkner, and P. Hänggi, Phys. Rev. Lett. **102**, 210401 (2009).
[29] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[30] G. E. Crooks, Phys. Rev. E **61**, 2361 (2000).
[31] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[32] A. M. Hahn and H. Then, Phys. Rev. E **80**, 031111 (2009).
[33] D. Wu and D. A. Kofke, J. Chem. Phys. **121**, 8742 (2004).
[34] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[35] B. Widom, J. Chem. Phys. **39**, 2808 (1963).

Curriculum Vitae und Publikationsliste

Persönliche Daten

Aljoscha Maria Hahn
Seeburger Chaussee 2
14476 Groß Glienicke
Geb. am 20. Februar 1977 in Hagios Antonios, Kreta,
Staatsangehörigkeit deutsch,
verheiratet, vier Kinder

Lebensabschnitte

1984 – 1988: Grundschule Ilshofen (Baden-Württemberg)

1988 –1997: Albert-Schweitzer-Gymnasium Crailsheim

1997 – 1998: Zivildienst in der Landwirtschaft der Sozialtherapeutischen Gemeinschaften Weckelweiler

1998 – 2002: Studium der Physik an der Humboldt Universität zu Berlin

2002 – 2006: Studium der Physik an der Technischen Universität Berlin

2006: Beginn einer Promotion bei Prof. Dr. Andreas Engel in Oldenburg

seit 2009 Wissenschaftlicher Koordinator des Graduiertenkollegs GRK 1558 an der TU Berlin

seit 2009 Wissenschaftlicher Mitarbeiter von Prof.Dr. Holger Stark, TU Berlin

Publikationsliste

A.M. Hahn and H. Then, Using bijective maps to improve free-energy estimates, Phys. Rev. E 79, 011113 (2009)

A.M. Hahn and H. Then, Characteristic of Bennett's acceptance ratio method, Phys. Rev. E **80**, 031111 (2009).

A.M. Hahn and H. Then, Measuring the convergence of Monte-Carlo free-energy calculations, Phys. Rev. E **81**, 041117 (2010).

Erklärungen

Eigenständigkeitserklärung

Ich erkläre hiermit, die vorliegende Arbeit selbständig verfasst und keine anderen, als die angegebenen Hilfsmittel verwendet zu haben. Ferner erkläre ich, die Arbeit weder in Gänze noch in Teilen einer anderen Universität zur Promotion vorgelegt zu haben.

Berlin, den 13. September 2010 (gez. Aljoscha Hahn)

Erklärung über die Anteile der Autoren

Die vorliegende Arbeit wurde bereits in Gänze in Form von drei Zeitschriftenartikeln veröffentlicht, die in Zusammenarbeit mit Holger Then verfasst wurden. Sein Anteil an dem Entstehen dieser Arbeit bestand in seiner ständigen Bereitschaft über die Fragen, mit denen ich befasst war, zu diskutieren, meine Überlegungen und Rechnungen kritisch zu prüfen, alternative Vorschläge zu unterbreiten, sowie sich in die Details der damit zusammenhängenden Probleme hineinzudenken. Seinen Ermutigungen und seiner Begeisterungsfähigkeit verdanke ich viel. Ferner hat er mich unermüdlich bei allen Belangen die Publikation der Arbeit betreffend unterstützt. Diese Umstände machen es nicht leicht, die Anteile der Autoren eindeutig zu trennen. Dennoch glaube ich, mir die wissenschaftlich neuen Errungenschaften dieser Arbeiten zuschreiben zu dürfen. Ferner habe ich alle konkreten Berechnungen und Simulationen unabhängig selbst durchgeführt.

Berlin, den 13. September 2010 (gez. Aljoscha Hahn)

Bestätigung durch den Erstgutachter

Ich bestätige die Richtigkeit der oben gemachten Angaben über die Anteile der Autoren.

Oldenburg, den .... (gez. Andreas Engel)