Unified Jarzynski and Sagawa-Ueda relations for Maxwell’s demon

Hal Tasaki

Department of Physics, Gakushuin University, Mejiro, Toshima-ku, Tokyo 171-8588, Japan

(Dated: May 11, 2014)

By using Newtonian mechanics, we construct a general model of Maxwell’s demon, a system in which the engine and the memory interact only through the exchange of information. We show that the Jarzynski relation and the two Sagawa-Ueda relations hold simultaneously, and argue that they are the unique triplet which has a natural decomposition property. The uniqueness provides a strong support to the assertion that mutual information is the key quantity.

PACS numbers: 05.70.Ln, 05.20.-y, 05.45.-a

Recently there has been a considerable renewed interest in the problem of Maxwell’s demon [1,2]. Based on progress in the twentieth century [2-7] which revealed the essential role of information, and more recent progress in nonequilibrium physics [8,10] in particular the Jarzynski relation and similar results, mathematically refined theories related to demon have been developed [11,14]. In particular Sagawa and Ueda have derived a series of general and exact results [15-21] which shed light on the essence of Maxwell’s demon (or, more generally, systems where measurement and feedback are essential) and suggest a fundamental role played by mutual information.

Imagine a (probably small) thermodynamic system, such as the Szilard engine [2,4], which is subject to measurement and feedback. It is well-known that such an “engine” may produce more work than that is allowed by the second law of thermodynamics. Then the key question is how much extra work is needed to operate the device, which may be called a demon, that realizes the measurement/feedback. It is believed that in principle such a device can be made as efficient as possible so that to waste less and less energy, except for a single component, the “memory”, which stores the information about the engine [2,3,4].

This motivates us to study, in the present paper, a composite system of simultaneously evolving “engine” and “memory” [23,26] that behaves (almost) as a normal physical system as a whole. By constructing such a system within classical mechanics, we can analyze the flow of energy and entropy completely, and realize a situation in which the engine and the memory interact only thorough the exchange of information. This construction provides a definite and most strict criterion of which system should be regarded as a Maxwell’s demon, provided that we restrict ourselves to a classical system and allow an external agent who operates on the system.

We then prove the Jarzynski relation and the two Sagawa-Ueda relations which involve mutual information, recovering the known results in the unified setting. These relations yield the standard and the extended second laws as usual. More importantly we show that the above three relations are the unique triplet of integral fluctuation relations which satisfies a natural decomposition property. This uniqueness provides a strong support to the assertion that mutual information plays a fundamental role in the problem of Maxwell’s demon [15,21].

We believe that our results do not only complete the project of Sagawa and Ueda (for a classical [30] non-autonomous demon), but also can be a crucial guide in further studies of a variety of systems which share certain aspects of Maxwell’s demon [23,28].

Setup and time-evolution.—We consider a system of classical particles which consists of two subsystems, the engine and the memory. The state of the engine is collectively denoted as η = (p_1, . . . , p_N, r_1, . . . , r_N) ∈ E, the state of the memory as θ = (p̃_1, . . . , p̃_N, ˜r_1, . . . , ˜r_N) ∈ M, and the state of the whole system as (η, θ) ∈ E × M. We also write dη = ∏_j=1^N dp_j d^3r_j and dθ = ∏_j=1^N d^3 ˜p_j d^3 ˜r_j.

Physically speaking the “engine” consists of the main body of the engine and a heat bath associated with it, and the “memory” consists of the memory itself and another bath. We have prepared separate heat baths so that to precisely trace the interaction between the engine and the memory. In what follows we shall not explicitly mention about the baths, but we always understand that they are included in the engine or the memory.

Both the engine and the memory are isolated from the external world, and evolve according to the Newtonian mechanics. We assume however that the engine and the memory are operated by an outside agent, and their Hamiltonians are varied in time according to protocols which are fixed in advance. The protocols are designed so that to realize measurement in the first period with t ∈ [0, t_1], and feedback (and memory erasure) in the second period with t ∈ [t_1, t_2]. See Fig. 1 We denote by H and ˜H the Hamiltonians of the engine and the memory, respectively, at the initial time t = 0.

In the period [0, t_1] of measurement, the engine evolves according to a fixed protocol. We denote by T^{ms} : E → E the corresponding time-evolution map (which brings the state at t = 0 to that of t = t_1). The memory also evolves according to a protocol, but the choice of the protocol is affected by the state of (the “main body” of) the engine in [0, t_1]. Mathematically we can assume that the protocol
is specified by the state of the engine at \( t = 0 \), which we write \( \Gamma \). The corresponding time-evolution map is \( \mathcal{T}^{ms}_{t_1} : \mathcal{M} \to \mathcal{M} \). We assume that the Hamiltonian \( H' \) of the memory at \( t = t_1 \) is independent of \( \Gamma \). The idea is that the state, not the Hamiltonian, of the memory at \( t = t_1 \) records information about \( \Gamma \).

In the period \([t_1, t_2]\) of feedback (and erasure), the engine and the memory switch their roles. The engine now evolves according to a protocol which depends on the state of the memory at \( t = t_1 \), which we write \( \Upsilon' \). This dependence represents the feedback \([37]\). The time-evolution map is denoted as \( T^{fb}_{t_1} : \mathcal{E} \to \mathcal{E} \). The memory evolves according to a fixed protocol. We suppose that the time-evolution \( T^{fb} : \mathcal{M} \to \mathcal{M} \) finally erases the information stored in the memory \([38]\). We assume that the whole process is cyclic in the sense that the Hamiltonians of the system and the memory at \( t = t_2 \) return to \( H \) and \( \dot{H} \), respectively \([39]\).

Finally we denote by \( T^{ms}_{t_1} = T^{fb}_{t_1} \circ T^{ms} \) and \( \mathcal{T} = \mathcal{T}^{ms} \circ \mathcal{T}^{fb} \) the time-evolution maps of the engine and the memory, respectively, for the whole time interval.

**Basic properties of the system.**—Recall that the Liouville theorem is valid when the Hamiltonian changes according to a fixed protocol. Thus each of the maps \( \mathcal{T}^{ms}, \mathcal{T}^{ms}_{t_1} \) (with any fixed \( \Gamma \)), \( \mathcal{T}^{fb}_{t_1} \) (with any fixed \( \Upsilon' \)), and \( \mathcal{T}^{fb} \) preserves the phase space volume. We further assume that each of them is a one-to-one map \([40]\).

Let \((\Gamma, \Upsilon)\) be the state at \( t = 0 \), and denote the corresponding states at \( t = t_1 \) as \((\Gamma', \Upsilon')\), and at \( t = t_2 \) as \((\Gamma'', \Upsilon'')\), i.e.,

\[
\begin{align*}
\Gamma' &= T^{ms}(\Gamma), \\
\Upsilon' &= T^{ms}(\Upsilon), \\
\Gamma'' &= T^{fb}(\Gamma'), \\
\Upsilon'' &= T^{fb}(\Upsilon').
\end{align*}
\]

We remark that the map from \((\Gamma, \Upsilon)\) to \((\Gamma', \Upsilon')\) is one-to-one. To see this, take an arbitrary \((\Gamma', \Upsilon')\), and note that \( \Gamma = (T^{ms})^{-1}(\Gamma') \) uniquely determines \( \Gamma \), and then \( \Upsilon = (T^{ms})^{-1}(\Upsilon') \) uniquely determines \( \Upsilon \). The map from \((\Gamma, \Upsilon)\) to \((\Gamma', \Upsilon')\) also preserves the phase space volume since both \( T^{ms} \) and \( T^{ms}_{t_1} \) (for a fixed \( \Gamma \)) do. Since the same observation is valid for the map from \((\Gamma', \Upsilon')\) to \((\Gamma'', \Upsilon'')\), we find that the map from the initial state \((\Gamma, \Upsilon)\) to the final state \((\Gamma'', \Upsilon'')\) is also one-to-one and preserves the phase space volume.

We believe that we have defined an ideal class of mechanical systems which captures the essence of Maxwell’s demon (or, more precisely, Szilard’s interpretation of the demon) in the following two senses.

First the engine and the memory are carefully designed so that to interact with each other only through the “exchange of information”. Since the engine and the memory evolve separately as isolated systems, they exchange energy only with the external agent, and not with each other. Moreover the fact that the time-evolutions of the engine and the memory separately preserve their phase space volumes implies that there are no mechanical exchange of entropy between them. The only interaction between the engine and the memory arises from the choice of the protocol by the external agent.

Secondly the time-evolution of the whole system (but not that of the engine or the memory) is one-to-one and preserves the phase space volume. This means that our system, as a whole, behaves (almost) as a normal Newtonian mechanical system.

**Main results.**—We assume that at \( t = 0 \) the state \((\Gamma, \Upsilon)\) is drawn from the probability distribution \( \rho_{0}(\Gamma, \Upsilon) = \rho_{0}(\Gamma) \rho_{0}(\Upsilon) \), where

\[
\rho_{0}(\Gamma) := \frac{e^{-\beta H(\Gamma)}}{Z}, \quad \rho_{0}(\Upsilon) := \frac{e^{-\beta H(\Upsilon)}}{Z}
\]

are the canonical distributions.

For any function \( F(\Gamma, \Upsilon) \) of the initial state \((\Gamma, \Upsilon)\), we define its average as

\[
\langle F(\Gamma, \Upsilon) \rangle := \int d\Upsilon d\Gamma F(\Gamma, \Upsilon) \rho_{0}(\Gamma, \Upsilon).
\]

Let us define (with \( \Upsilon' \) being a free variable)

\[
\tilde{\rho}(\Upsilon'|\Gamma) := \int d\Upsilon \delta[\Upsilon' - \{ T^{ms}(\Upsilon) \}] \rho_{0}(\Upsilon),
\]

which is the probability density to get \( \Upsilon' \) in the memory at \( t = t_1 \) given the condition that the engine was in \( \Gamma \) at \( t = 0 \). We also write the unconditioned probability density as

\[
\tilde{\rho}(\Upsilon') := \int d\Gamma \tilde{\rho}(\Upsilon'|\Gamma) \rho_{0}(\Gamma).
\]

We then define the mutual information function as

\[
I(\Gamma, \Upsilon') := \log \frac{\tilde{\rho}(\Upsilon'|\Gamma)}{\rho(\Upsilon'}),
\]

whose average

\[
\bar{I} := \langle I(\Gamma, \tilde{T}^{ms}(\Upsilon)) \rangle
\]

\[
= \int d\Upsilon d\Upsilon' \tilde{\rho}(\Upsilon'|\Gamma) \rho_{0}(\Gamma) \log \frac{\tilde{\rho}(\Upsilon'|\Gamma)}{\rho(\Upsilon')}, \quad \geq 0
\]

is the mutual information between the state of the engine at \( t = 0 \) and that of the memory at \( t = t_1 \) \([41]\).

We also define

\[
W(\Gamma, \Upsilon) := H(\Gamma) - H(T^{ms}(\Upsilon)), \quad W(\Gamma, \Upsilon) := H(\Upsilon) - H(T^{fb}(\Upsilon)).
\]
which are the works done by the engine and by the memory, respectively, to the agent during the whole process.

Our main results are the three equalities

\[ \langle e^{\beta(W(\Gamma, \Upsilon) + \tilde{W}(\Gamma, \Upsilon))} \rangle = 1, \tag{11} \]
\[ \langle e^{\beta W(\Gamma, \Upsilon) - I(\Gamma, \Upsilon')} \rangle = 1, \tag{12} \]
\[ \langle e^{\beta \tilde{W}(\Gamma, \Upsilon) + I(\Gamma, \Upsilon')} \rangle = 1, \tag{13} \]

where \( \Upsilon' \) in the expectations should be replaced by \( \tilde{T}^{\text{fin}}(\Upsilon) \).

Eq. (11) is nothing but the original Jarzynski relation applied to the whole system. The relations (12) and (13) are the Sagawa-Ueda relations for feedback and measurement, respectively. See also (22, 30).

We recall that, combined with the Jensen inequality for several different \( X \) or \( Y \) including \( Y = 0 \). But if we further demand that \( X = Y \) so that (18) corresponds to a decomposition of the total work, our choice is essentially unique (in a certain weak sense to be read off from the derivation) and we have \( X = Y = I(\Gamma, \Upsilon') \).

\[ \langle e^{\beta(W(\Gamma, \Upsilon) + \tilde{W}(\Gamma, \Upsilon))} \rangle = 1, \]
\[ \langle e^{\beta W(\Gamma, \Upsilon) - I(\Gamma, \Upsilon')} \rangle = 1, \]
\[ \langle e^{\beta \tilde{W}(\Gamma, \Upsilon) + I(\Gamma, \Upsilon')} \rangle = 1, \]

where \( \Upsilon' \) in the expectations should be replaced by \( \tilde{T}^{\text{fin}}(\Upsilon) \).

Eq. (11) is nothing but the original Jarzynski relation applied to the whole system. The relations (12) and (13) are the Sagawa-Ueda relations for feedback and measurement, respectively. See also (22, 30).

We recall that, combined with the Jensen inequality for several different \( X \) or \( Y \) including \( Y = 0 \). But if we further demand that \( X = Y \) so that (18) corresponds to a decomposition of the total work, our choice is essentially unique (in a certain weak sense to be read off from the derivation) and we have \( X = Y = I(\Gamma, \Upsilon') \).

\[ \langle e^{\beta W(\Gamma, \Upsilon) - I(\Gamma, \Upsilon')} \rangle = 1, \]
\[ \langle e^{\beta \tilde{W}(\Gamma, \Upsilon) + I(\Gamma, \Upsilon')} \rangle = 1, \]

where \( \Upsilon' \) in the expectations should be replaced by \( \tilde{T}^{\text{fin}}(\Upsilon) \).

Eq. (11) is nothing but the original Jarzynski relation applied to the whole system. The relations (12) and (13) are the Sagawa-Ueda relations for feedback and measurement, respectively. See also (22, 30).

We recall that, combined with the Jensen inequality for several different \( X \) or \( Y \) including \( Y = 0 \). But if we further demand that \( X = Y \) so that (18) corresponds to a decomposition of the total work, our choice is essentially unique (in a certain weak sense to be read off from the derivation) and we have \( X = Y = I(\Gamma, \Upsilon') \).

\[ \langle e^{\beta W(\Gamma, \Upsilon) - I(\Gamma, \Upsilon')} \rangle = 1, \]
\[ \langle e^{\beta \tilde{W}(\Gamma, \Upsilon) + I(\Gamma, \Upsilon')} \rangle = 1, \]

where \( \Upsilon' \) in the expectations should be replaced by \( \tilde{T}^{\text{fin}}(\Upsilon) \).

Eq. (11) is nothing but the original Jarzynski relation applied to the whole system. The relations (12) and (13) are the Sagawa-Ueda relations for feedback and measurement, respectively. See also (22, 30).

We recall that, combined with the Jensen inequality for several different \( X \) or \( Y \) including \( Y = 0 \). But if we further demand that \( X = Y \) so that (18) corresponds to a decomposition of the total work, our choice is essentially unique (in a certain weak sense to be read off from the derivation) and we have \( X = Y = I(\Gamma, \Upsilon') \).

\[ \langle e^{\beta W(\Gamma, \Upsilon) - I(\Gamma, \Upsilon')} \rangle = 1, \]
\[ \langle e^{\beta \tilde{W}(\Gamma, \Upsilon) + I(\Gamma, \Upsilon')} \rangle = 1, \]
which shows that the entropy of the engine may decrease but not more than by $\overline{I}$. We can say that the mutual information $I$ (generated during the measurement process) may be used as a resource to reduce the entropy of the engine (in the feedback process). From [19] (which indeed is rigorous) and the nonnegativity of relative entropy one can generalize the second law [15] [17]. This is reasonable if we realize that the decrease in entropy by $I$ is equivalent to the increase in the free energy by $\overline{I}/\beta$, which may be converted into work.

**Derivation.**—Jarzynski relation [11] for the whole system is derived as in the original [8] by noting that the time-evolution is one-to-one and measure-preserving. Let $f(\Gamma, \Upsilon')$ be an arbitrary function of $\Gamma$ and $\Upsilon'$. We find

$$\langle e^{\beta W(\Gamma, \Upsilon)} f(\Gamma, \Upsilon) \rangle = \int d\Gamma d\Upsilon e^{\beta W(\Gamma, \Upsilon)} f(\Gamma, \Upsilon) \rho_0(\Gamma, \Upsilon)$$

where we used (5). This uniquely determines $\rho(\Gamma, \Upsilon)$.

This is still a very complicated integral where the integrand depends nontrivially both on $\Gamma$ and $\Upsilon'$. However, the integral becomes tractable if the integrand depends only through $\Upsilon''$. Let $g(\Upsilon'')$ be an arbitrary function of $\Upsilon''$. We thus get

$$\langle e^{\beta W(\Gamma, \Upsilon)} f(\Gamma, \Upsilon) \rangle = \int d\Gamma d\Upsilon e^{\beta W(\Gamma, \Upsilon)} g(\Upsilon'') \rho_0(\Gamma, \Upsilon)$$

where we have used the Liouville theorem $d\Gamma = d\Upsilon''$ (for each fixed $\Gamma$). Again this is still a hardly tractable integral, but simplifies in general if $g$ is chosen to satisfy

$$\int d\Upsilon' g(\Upsilon', \Upsilon) \rho_0(\Gamma) = 1,$$

for any $\Upsilon'$. An obvious choice is $g = 1$. For $g$ satisfying [23], the integral in (23) is easily evaluated and one gets

$$\langle e^{\beta W} \rangle = 1.$$

To require $X = Y$ in [18] corresponds to requiring $g = 1/f$. By substituting (21) into (24), we find

$$1 = \int d\Upsilon' \rho(\Upsilon') \rho_0(\Gamma) = \frac{\rho(\Upsilon')}{\nu(\Upsilon')}$$

where we used (6). This uniquely determines $\nu(\Upsilon')$ to be $\rho(\Upsilon')$, and hence that

$$\rho(\Gamma, \Upsilon') = \frac{1}{f(\Gamma, \Upsilon')} = e^{I(\Gamma, \Upsilon')}.\] (26)

**Discussion.**—As for a classical system operated by an outside agent, we have clarified which system should be called a Maxwell’s demon in the most strict sense. For such a system, we have established that the three relations [11], [12], and [17] form a unique triplet corresponding to the Sagawa-Ueda decomposition. We believe that, as far as we concentrate on classical simple “nonautonomous” demons, these observations complete the project of Sagawa and Ueda to understand the essence of Maxwell’s demon.

A remaining quite interesting challenge is to investigate whether similar results are possible for an “autonomous Maxwell’s demon”, a composite system which evolves under a fixed Hamiltonian without external operation [24] [25] [27] [28]. It is likely that our criterion that “the engine and the memory exchange only information” may be realized only in certain limiting sense. Even though such a criterion is expected to be quite useful in the analysis of demon-like engineering in nature (such as biological machines) or in the future technology.

It is a pleasure to thank Takahiro Sagawa, whom I regard almost as a coauthor, for discussions and suggestions which made the present work possible. I also thank
Takayuki Ariga, Sosuke Ito, and Shin-ichi Sasa for useful discussions.

[1] J. C. Maxwell, “Theory of Heat”, (Appleton, London, 1871).
[2] “Maxwell’s demon 2: Entropy, Classical and Quantum Information, Computing”, H. S. Leff and A. F. Rex (eds.), (Princeton University Press, New Jersey, 2003).
[3] K. Maruyama, F. Nori, and V. Vedral, Rev. Mod. Phys. 81, 1 (2009).
[4] L. Szilard, Z. Phys. 53, 840 (1929).
[5] L. Brillouin, J. Appl. Phys. 22, 334 (1951).
[6] R. Landauer, IBM J. Res. Dev. 5, 183 (1961).
[7] C. H. Bennett, Int. J. Theor. Phys. 21, 905 (1982).
[8] C. Jarzynski, Phys. Rev. Lett. 78, 2690 (1997), arXiv:cond-mat/9610209
[9] G. E. Crooks, Phys. Rev. E 60, 2721 (1999), arXiv:cond-mat/9901352
[10] See U. Seifert, arXiv:1205.4176 (2012), which is a recent extensive review.
[11] H. Touchette and S. Lloyd, Phys. Rev. Lett. 84, 1156 (2000), arXiv:chao-dyn/9905039
[12] B. Piechocinska, Phys. Rev. A 61, 062314 (2000).
[13] K. H. Kim and H. Qian, Phys. Rev. E 75, 022102 (2007).
[14] K. Funo, Y. Watanabe, M. Ueda, arXiv:1307.2362
[15] T. Sagawa and M. Ueda, Phys. Rev. Lett. 100, 080403 (2008), arXiv:0710.0956
[16] T. Sagawa and M. Ueda, Phys. Rev. Lett. 102, 250602 (2009); 106, 189901(E) (2011), arXiv:0809.4098
[17] T. Sagawa and M. Ueda, Phys. Rev. Lett. 104, 090602 (2010), arXiv:1105.3262
[18] T. Sagawa and M. Ueda, Phys. Rev. E 85, 021104 (2012), arXiv:1206.2479v1
[19] T. Sagawa and M. Ueda, arXiv:1206.2479v1 (2012). This is an early version of [20].
[20] T. Sagawa and M. Ueda, Phys. Rev. Lett. 109, 180602 (2012), arXiv:1206.2479
[21] T. Sagawa and M. Ueda, arXiv:1307.6092 (2013).
[22] K. Funo, Y. Watanabe, M. Ueda, arXiv:1307.2362 (2013).
[23] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, Nature Physics 6, 988 (2010), arXiv:1009.5287
[24] D. Mandal and C. Jarzynski, Proc. Natl. Acad. Sci. U.S.A., 109 11641 (2012), arXiv:1206.5553
[25] P. Strasberg, G. Schaller, T. Brandes, M. Esposito, Phys. Rev. Lett. 110, 040601 (2013), arXiv:1210.5661
[26] J. M. Horowitz, T. Sagawam, J. M. R. Parrondo Phys. Rev. Lett. 111, 010602 (2013), arXiv:1210.6449
[27] D. Mandal, H. T. Quan, C. Jarzynski, Phys. Rev. Lett. 111, 030602 (2013), arXiv:1307.2208
[28] P. Strasberg, G. Schaller, T. Brandes, M. Esposito, arXiv:1305.6589 (2013).
[29] J. M Horowitz and S. Vaikuntanathan, Physical Review E 82, 061120 (2010), arXiv:1011.4273
[30] S. Ito and T. Sagawa, arXiv:1306.2756 (2013).
[31] J. M. Horowitz and J. M. R. Parrondo, Europhys. Lett. 95, 10005 (2011), arXiv:1104.0332
[32] T. Hatano and S.-I. Sasa, Phys. Rev. Lett. 86, 3463 (2001), arXiv:cond-mat/0010405
[33] T. Speck, U. Seifert J. Phys. A: Math. Gen. 38, L581 (2005), arXiv:cond-mat/0507420
[34] M. Esposito and C. Van den Broeck, Phys. Rev. Lett. 104, 090601 (2010), arXiv:0911.2666
[35] To our knowledge such a composite system was first studied by Sagawa and Ueda in [19].
[36] See [22] for a somewhat similar treatment of quantum systems.
[37] We are, in a sense, assuming that the agent itself has no memory. In the period of feedback, it does not remember which protocol was chosen in the period of measurement.
[38] By only using the deterministic Newtonian mechanics it is impossible to completely delete the information about the state at $t = t_1$. We expect however that the recovery can be made practically impossible by designing a proper dynamics since the heat bath contains many particles whose motion can be complicated. Anyway we shall not make use of any assumptions about memory erasure in the derivation of the results.
[39] The assumption of cyclicity is not at all essential. If the process is not cyclic, one should include the differences of the initial and the final free energies in the main equalities [11], [12], and [19].
[40] The one-to-one property does not follow automatically and should be assumed. In the present context, it means that measurement and feedback are associated with some errors. The similar problem with no errors (in which the one-to-one property no longer holds) can also be treated, both in classical and quantum settings.
[41] The final expression in (8) can be easily derived by proceeding as in [20].
[42] To be rigorous we have to add small errors to the system so as to make it satisfy the conditions of the present work.
[43] More abstractly, a Sagawa-Ueda decomposition is a special case of a decomposition $A = B + C$ with the property that $\langle e^A \rangle = \langle e^B \rangle = \langle e^C \rangle = 1$. Although we still do not know what this exactly implies, we remark that it is a highly nontrivial property which can hardly be realized accidentally. The same type of decomposition is found in a driven nonequilibrium system where one decomposes the total entropy production into the sum of the “housekeeping” part and the remainder. See [22] and also [31].
[44] One should note that the entropies include those of the heat baths.
[45] One easily finds that this is exactly equal to $I$ in (8), which was defined as the mutual information between the engine at $t = 0$ and the memory at $t = t_1$.
[46] The argument which led to $\tilde{S}(t_1) \geq \bar{S}(0)$ is no longer valid since the state of the memory at $t = t_1$ is correlated with the previous state of the engine.
[47] Let $I'' := \langle S(t_2) + \bar{S}(t_2) \rangle - \bar{S}(t_2)$ be the mutual information between the engine and the memory in the final states. Almost by definition we have $S(t_2) = S(0) - I + I''$, from which we get (again rigorously) an improved bound $\langle W \rangle \leq (I - I'')/\beta$. We do not know whether there is a corresponding integral fluctuation relation, but see [22],[31].
Appendix: Error-free system.—Let us discuss the error-free version of the same problem of the engine and the memory.

We assume here that the state spaces are decomposed into disjoint unions as \( \mathcal{E} = \bigcup_{\mu=1}^{m} \mathcal{E}_\mu \) and \( \mathcal{M} = \bigcup_{\mu=1}^{n} \mathcal{M}_\mu \). The time-evolution rule is basically the same. But \( T_{ms}^{\mu} \) now depends on \( \Gamma \) only through the unique index \( \mu \) such that \( \Gamma \in \mathcal{E}_\mu \), and hence is written as \( T_{ms}^{\mu} \). We assume that \( T_{ms}^{\mu} \) is a one-to-one map from \( \mathcal{M} \) to \( \mathcal{M}_\mu \). Thus the state \( \Upsilon' \) of the memory at \( t = t_1 \) specifies the index \( \mu \) without any errors. Likewise \( T_{fb}^{\mu} \) now depends on \( \Upsilon' \) only through the unique \( \mu' \) such that \( \Upsilon' \in \mathcal{M}_{\mu'} \). But since we already know that \( \Upsilon' \in \mathcal{M}_\mu \), we have \( \mu' = \mu \).

The time-evolution map is then denoted as \( T_{ms}^{\mu} \), which is assumed to be a one-to-one map from \( T_{ms}^{\mu} (\mathcal{E}_\mu) \) to \( \mathcal{E} \). The time-evolution maps for the whole interval is denoted as \( T_{\mu} = T_{ms}^{\mu} \circ T_{ms}^{\mu} \) and \( \tilde{T}_{\mu} = T_{fb}^{\mu} \circ \tilde{T}_{ms}^{\mu} \).

Again the map from \( (\Gamma, \Upsilon) \in \mathcal{E} \times \mathcal{M} \) to \( (\Gamma'', \Upsilon'') := (T_{\mu}(\Gamma), \tilde{T}_{\mu}(\Upsilon)) \in \mathcal{E} \times \mathcal{M} \) is one-to-one and preserves the phase space volume. We defined \( \mu(\Gamma) \) as the unique index such that \( \Gamma \in \mathcal{E}_\mu(\Gamma) \).

Let \( p_{\mu} := \int_{\Gamma \in \mathcal{E}_\mu} \rho_0(\Gamma) \) be the probability that the state of the engine is initially in \( \mathcal{E}_\mu \). Then we can show

\[
\langle e^{\beta W(\Gamma) + \tilde{W}(\Gamma, \Upsilon)} \rangle = 1, \tag{27}
\]

\[
\langle e^{\beta W(\Gamma)} \rangle = 1, \tag{28}
\]

and

\[
\langle e^{\beta W(\Gamma, \Upsilon) - \log p_{\mu}(\Gamma)} \rangle = 1, \tag{29}
\]

which are the Jarzynski relation and the two Sagawa-Ueda relations, respectively. Note that we have the Shannon entropy function \(- \log p_{\mu}(\Gamma)\) instead of the mutual information function \( I(\Gamma, \Upsilon') \).

Let us derive the Sagawa-Ueda relations, and also show the uniqueness of the Sagawa-Ueda decomposition.

First we concentrate on the time-evolution of the engine. Then the only role of the memory is to ensure the correct feedback to the system. For a fixed \( \mu \), we have

\[
\int_{\Gamma \in \mathcal{E}_\mu} d\Gamma e^{\beta W(\Gamma)} \rho_0(\Gamma) = \int_{\Gamma'' \in \mathcal{E}} d\Gamma'' e^{-\beta H(\Gamma'')} \frac{e^{-\beta H(\Gamma'')}}{Z} = 1, \tag{30}
\]

where \( \Gamma'' = T_{\mu}(\Gamma) \) and we noted that \( d\Gamma'' = d\Gamma \). Let \( q_\mu \) be any quantity with \( \sum_{\mu} q_\mu = 1 \). Then by multiplying \( \rho_0(\Gamma) \) by \( q_\mu \) and summing up over \( \mu \), one gets

\[
\int_{\Gamma \in \mathcal{E}_\mu} d\Gamma q_\mu(\Gamma) e^{\beta W(\Gamma)} \rho_0(\Gamma) = 1, \tag{31}
\]

which is nothing but \( \langle e^{\beta W + \log q_\mu} \rangle = 1 \).

Let us fix \( \mu \), and examine the time-evolution of the memory. It is convenient to define \( \tilde{W}_\mu(\Upsilon) = H(\Upsilon) - \tilde{H}(\tilde{T}_{\mu}(\Upsilon)) \), which satisfies \( \tilde{W}(\Gamma, \Upsilon) = \tilde{W}_\mu(\Upsilon) \). Then we get

\[
\int d\Upsilon e^{\beta \tilde{W}_\mu(\Upsilon)} \tilde{\rho}_0(\Upsilon) = \int_{\Upsilon'' \in \tilde{T}_{\mu}(\mathcal{M}_\mu)} d\Upsilon'' e^{-\beta H(\Upsilon'')} \frac{e^{-\beta H(\Upsilon'')}}{Z}. \tag{32}
\]

Summing this over \( \mu \) we get

\[
\sum_{\mu} \int d\Upsilon e^{\beta \tilde{W}_\mu(\Upsilon)} \tilde{\rho}_0(\Upsilon) = 1, \tag{33}
\]

which is rewritten as

\[
\sum_{\mu} p_{\mu} \int d\Upsilon \frac{1}{p_{\mu}} e^{\beta \tilde{W}_\mu(\Upsilon)} \tilde{\rho}_0(\Upsilon) = 1. \tag{34}
\]

This is nothing but the desired Sagawa-Ueda relation \( \langle e^{\beta W - \log p_\mu} \rangle = 1 \). Interestingly the fluctuation relation is essentially unique in this situation. From the requirement corresponding to \( X = Y \), we uniquely determine \( q_\mu \) to be \( p_{\mu} \).