Classical Simulation of Non-Classical Systems: A Large Deviation Analysis*

Adam Brandenburger † Pierfrancesco La Mura ‡

Preliminary Version

January 30, 2025

Abstract

Any quasi-probability representation of a no-signaling – in particular, quantum – system can be simulated via a classical scheme that involves signed events and cancellation. This poses the question: What properties of the non-classical system does such a classical simulation fail to reproduce? We employ large deviation theory to establish that the probability of a large fluctuation under the classical simulation may be greater than under the system being simulated. The key to our finding is the observation that the data processing inequality of information theory can be reversed in the presence of underlying negative probability.

1 Introduction

The use of signed (or quasi-)probabilities has a long pedigree in quantum mechanics (QM), going back to Wigner (1932), Dirac (1942), Feynman (1987), and others. Their use today continues as a calculational tool – for example, in the field of quantum optics (Kenfack and Życzkowski, 2004). The set of all no-signaling measurement-outcome models (Popescu and Rohrlich, 1994) is characterized by the set of all signed-probability distributions on phase space (Abramsky and Brandenburger, 2011, Theorem 5.9). This makes signed probabilities a useful starting point for foundational investigations of QM (e.g., Ferrie, 2011; Brandenburger, La Mura, and Zoble, 2022; Onggadinata, Kurzynski, and Kaszlikowski, 2023; Gherardini and De Chiara, 2024).

The fact that signed probabilities can describe the no-signaling – and, therefore, quantum – set of systems leads to a puzzle. It is easy to **simulate signed probabilities** via a classical setup (Abramsky and Brandenburger, 2014). Essentially, we just "push the minus sign in" from probabilities to events. We then sample such signed events while, in calculating empirical frequencies, taking care to cancel plus and minus occurrences of the same event. (Details are provided in Section 2.) Since the quantum world is not classical, there must be at least one aspect of this classical simulation that differs from the underlying non-classical system. The puzzle is to identify such an aspect.

A clue to a resolution can be found in Feynman's famous 1982 paper, where he put forward the idea of a computer that would use quantum effects to simulate physical systems obeying the laws of quantum mechanics (Feynman, 1982, 1985). In exploring this theme, Feynman asked whether a computer operating on classical principles could simulate quantum systems. He wrote:

> [W]e could imagine and be perfectly happy, I think, with a probabilistic simulator of a probabilistic nature, in which the machine doesn't exactly do what nature does, but if you repeated a particular type of experiment a sufficient number of times to determine nature's probability, then you did the corresponding experiment on the computer, you'd get the corresponding probability with the corresponding accuracy (with the same kind of accuracy of statistics) . . . .
>
> The only difference between a probabilistic classical world and the equations of the quantum world is that somehow or other it appears as if the probabilities would have to go negative, and that we do not know, as far as I know, how to simulate. (Feynman, 1982, p.473 and p.480)

*This paper owes a big debt to Samson Abramsky for prior joint work, to Pierre Tarres for a very careful read and important improvements, and to Vered Kurtz-David and Stuart Zoble for valuable input. Financial support from NYU Stern School of Business, NYU Shanghai, J.P. Valles, and the HHL - Leipzig Graduate School of Management is gratefully acknowledged.

†Stern School of Business, Tandon School of Engineering, NYU Shanghai, New York University, New York, NY 10012, U.S.A., adam.brandenburger@stern.nyu.edu

‡HHL - Leipzig Graduate School of Management, 04109 Leipzig, Germany, plamura@hhl.de

Interestingly, Feynman pinpoints the difficulty of classical simulation of quantum systems as coming from **negativity of probability**. Yet this is easily simulated, as we have just sketched. The clue is where Feynman says that the simulation would obtain the same kind of **accuracy**. Here, he means accuracy of order $1/\sqrt{n}$ (op.cit, pp.472-473), so this is the behavior of "small" deviations. What is left open is accuracy in the sense of large deviation theory. Specifically, we can ask: What is the **large deviation behavior** of a classical simulation of a no-signaling system relative to the large deviation behavior of the underlying system?

To establish a baseline, consider a simple scenario – say, the throw of a die. Let $\mu = (1/2, 1/2)$ give the probability of obtaining an odd or even number on a throw, and let $f = (f_1, f_2)$ be the empirical distribution of odd vs. even obtained after a large number N of throws. By Sanov's theorem (Sanov, 1957), the probability of obtaining f is approximately $\exp(-N \times D(f||\mu))$, where $D(f||\mu)$ is the **relative entropy** or KL divergence of f relative to μ.

Now **fine-grain** the system by writing $\nu = (1/6, \ldots, 1/6)$ for the probabilities of the six possible outcomes of a throw, and let $g = (g_1, \ldots, g_6)$ be the associated empirical distribution after N throws. By another appeal to Sanov's theorem, the probability of obtaining g is approximately $\exp(-N \times D(g||\nu))$. Suppose g transforms to f, that is: $f_1 = g_1 + g_3 + g_5$ and $f_2 = g_2 + g_4 + g_6$. Then, by the **data processing inequality** of information theory (e.g., Goldfeld, 2020, Lecture 7), we know that $D(f||\mu) \leq D(g||\nu)$. By an appeal to the **contraction principle** of large deviation theory (Dembo and Zeitouni, 1998, Theorem 4.2.1), we can conclude that the total probability of all fine-grainings g that map to f is lower than the original – coarse-grained – probability of f. Intuitively, fine-graining increases distinguishability between distributions and therefore a deviation becomes "sharper" and less likely.

Next, start from an underlying process that involves negative probabilities and consider an associated non-negative probability measure on a set of observable values. We set up our classical simulation of this process, which is akin to fine-graining. More precisely, we establish a many-to-one map from simulation probabilities to actual probabilities. Our key result is Theorem 1 in Section 4, which provides an inequality between relative entropy for the simulation and relative entropy for the underlying process. This new inequality allows the classical data processing inequality to be overturned. We call it a **signed data processing inequality**. Given this reversal, another appeal to the contraction principle tells us that the probability of a given (large) deviation can be higher under the simulation than under the original process. In Section 5, we provide a numerical example to verify that this reversal can indeed happen.

This is our answer to the puzzle. Any non-classical no-signaling system can be simulated via a classical scheme. However, the classical simulation may yield higher probabilities of large fluctuations. Equivalently, the non-classical system will yield lower probabilities of large fluctuations. It exhibits superior performance in this sense.

2 Simulation of Signed Probabilities

Fix a finite **phase space** $X = \{x_1, x_2, \ldots, x_m\}$ and a signed probability measure $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ on X. We assume $\lambda_i \neq 0$ for all i. Fix also a set $Y = \{y_1, y_2, \ldots, y_n\}$ of **observable values** and a map $\chi : X \to Y$. We push forward the signed probability measure λ to a non-negative probability measure μ on Y by setting:

$$\mu_i = \sum_{\{j:\, x_j \in \chi^{-1}(y_i)\}} \lambda_j, \tag{1}$$

for $i = 1, 2, \ldots, n$. The requirement that μ is non-negative ensures that all events in Y are observable. Of course, for a given signed probability measure λ, this imposes a restriction on the map χ. We can now draw i.i.d. samples from Y under the process μ. We write $f = (f_1, f_2, \ldots, f_n)$ for the resulting **empirical distribution**.

Next, we define the **classical simulation** of λ, following Abramsky and Brandenburger (2014). Let X^+ and X^- be copies of X and write $Z = X^+ \sqcup X^-$ for the disjoint union of the copies. Set $\Lambda = \sum_{i=1}^{m} |\lambda_i|$ and define a non-negative probability measure $\nu = (\nu_1^+, \nu_2^+, \ldots, \nu_m^+, \nu_1^-, \nu_2^-, \ldots, \nu_m^-)$ on Z by setting:

$$\nu_i^+ = \begin{cases} \frac{\lambda_i}{\Lambda} \text{ if } \lambda_i > 0, \\ 0 \text{ otherwise;} \end{cases} \tag{2}$$

$$\nu_i^- = \begin{cases} \frac{|\lambda_i|}{\Lambda} \text{ if } \lambda_i < 0, \\ 0 \text{ otherwise.} \end{cases} \tag{3}$$

Since ν is a non-negative probability measure, we can draw i.i.d. samples from Z under the process ν. Write $g = (g_1^+, g_2^+, \ldots, g_m^+, g_1^-, g_2^-, \ldots, g_m^-)$ for the resulting empirical distribution.

We now show how to transform g to obtain μ in the limit of increasing sample size. To do this, define:

$$\gamma_i = \frac{\sum\limits_{\{j:\, x_j \in \chi^{-1}(y_i)\}} (g_j^+ - g_j^-)}{\sum\limits_{k=1}^{m} (g_k^+ - g_k^-)}. \tag{4}$$

If we sample N times, where N is large, we obtain the approximate equality:

$$\gamma_i \approx \frac{\sum\limits_{\{j:\, x_j \in \chi^{-1}(y_i)\}} (\nu_j^+ - \nu_j^-)}{\sum\limits_{k=1}^{m} (\nu_k^+ - \nu_k^-)}. \tag{5}$$

Substituting in from Equations 2 and 3, we get:

$$\gamma_i \approx \frac{\sum\limits_{\{j:\, x_j \in \chi^{-1}(y_i) \wedge \lambda_j > 0\}} \lambda_j/\Lambda - \sum\limits_{\{j:\, x_j \in \chi^{-1}(y_i) \wedge \lambda_j < 0\}} |\lambda_j|/\Lambda}{\sum\limits_{\{k:\, \lambda_k > 0\}} \lambda_k/\Lambda - \sum\limits_{\{k:\, \lambda_k < 0\}} |\lambda_k|/\Lambda}. \tag{6}$$

Multiplying through by Λ, we conclude:

$$\gamma_i \approx \frac{\sum_{\{j:\, x_j \in \chi^{-1}(y_i)\}} \lambda_j}{1} = \mu_i. \tag{7}$$

The rigorous version of this derivation uses the Strong Law of Large Numbers to obtain Equation 5 as a limit on a set of ν-probability 1. See Abramsky and Brandenburger (2014) for details.

We have now shown that, starting with a signed probability measure λ on phase space X, we can directly sample under the image measure μ to obtain an empirical distribution f on the set of observable values Y. Or, we can first simulate λ via a non-negative probability measure ν on a doubled phase space Z. We then sample under ν to obtain an empirical distribution g. The transformation of g via cancellation of plus-signed and minus-signed events yields an empirical distribution γ that tends to μ as the sample size increases. In the following sections, we compare the performance of these two schemes.

3 Large Deviation Analysis

Consider the empirical distribution f obtained via i.i.d. sampling from Y under μ. The **relative entropy** or Kullback-Leibler (KL) divergence of f relative to μ is given by:

$$D(f||\mu) = \sum_{i=1}^{n} f_i \log \frac{f_i}{\mu_i}. \tag{8}$$

By Sanov's theorem (Sanov, 1957), the probability of obtaining f is approximated for a large number of draws N by:

$$\Pr_\mu(f; N) \approx e^{-N \times D(f||\mu)}. \tag{9}$$

Similarly, we consider the empirical distribution g obtained via i.i.d. sampling from Z under ν. The relative entropy is:

$$D(g||\nu) = \sum_{j=1}^{m} g_j \log \frac{g_j}{\nu_j}, \tag{10}$$

and, again by Sanov's theorem, the probability of obtaining g is approximated for a large number of draws N by:

$$\Pr_\nu(g; N) \approx e^{-N \times D(g||\nu)}. \tag{11}$$

Let Γ denote the map that takes g to γ as in Equation 4. We are interested in the total probability:

$$\int_{\{g:\, \Gamma(g) = f\}} \Pr_\nu(g; N)\, d\lambda_g, \tag{12}$$

where λ_g is Lebesgue measure on the $(2m-1)$-simplex. Using the **contraction principle** (Dembo and Zeitouni, 1998, Theorem 4.2.1), we have:

$$\int_{\{g:\, \Gamma(g) = f\}} \Pr_\nu(g; N)\, d\lambda_g \approx e^{-N \times \inf_{\{g:\, \Gamma(g) = f\}} D(g||\nu)}. \tag{13}$$

We want to compare this probability with that in Equation 9. In other words, we want to sign the relationship:

$$D(f||\mu) \lesseqgtr \inf_{\{g:\Gamma(g)=f\}} D(g||\nu). \tag{14}$$

In the next section, we show how to do this in the case that the departure from negativity in the underlying signed probability measure λ is controlled.

4 Signed Data Processing Inequality

We start by restating the problem so that all five probability measures λ, μ, f, ν, and g live on the same set X. To achieve this, redefine the map χ to take the index set $\{1, 2, \ldots, m\}$ into itself. We can then move μ to X by setting:

$$\mu_i = \sum_{\{j:\, j\in\chi^{-1}(i)\}} \lambda_j. \tag{15}$$

With this redefinition, we let f be the empirical distribution on X obtained via i.i.d. draws under μ.

We also move ν to X. Note that for each $j = 1, 2, \ldots, m$, either $\nu_j^+ > 0$ or $\nu_j^- > 0$, not both. So, we can transfer ν to X directly. We next set $g_j^+ = 0$ if $\nu_j^+ = 0$ and $g_j^- = 0$ if $\nu_j^- = 0$ (these are zero-probability sampling events) and then move g to X the same way we just moved ν to X. To simplify notation, we now drop the $+$ or $-$ superscripts from the g_j.

The next step is to identify the transformation that takes underlying distribution ν to underlying distribution μ, and empirical distribution g to empirical distribution f. To do so, define the $m \times m$ matrix T by:

$$T_{ij} = \begin{cases} \text{sign}(\lambda_j) \text{ if } j \in \chi^{-1}(i), \\ 0 \text{ otherwise.} \end{cases} \tag{16}$$

We then get the relationships:

$$\mu = \Lambda \times T\nu \text{ and } f = F \times Tg, \tag{17}$$

where:

$$F = 1/\sum_{i=1}^{m} (Tg)_i. \tag{18}$$

Note that T is related to the earlier map Γ by $\Gamma = F \times T$.

The key observation is that every column of T contains exactly one nonzero entry, which is either $+1$ or -1. Call T a **signed column-stochastic matrix**. Following the language of information theory, we can also call T a **signed channel**. We write T as the difference of two non-negative matrices: $T = T^+ - T^-$, where T^+ is obtained from T by changing every -1 entry to $+1$, and T^- is obtained from T by changing every $+1$ entry to 0 and every -1 entry to $+2$. Note that T^+ is an ordinary column-stochastic matrix, that is, a classical channel.

Now calculate:

$$\begin{aligned} D(f||\mu) &= \sum_i f_i \log \frac{f_i}{\mu_i} && (19)\\ &= \sum_i F(Tg)_i \log \frac{(FTg)_i}{(\Lambda T\nu)_i} && (20)\\ &= F\sum_i (Tg)_i [\log \frac{(Tg)_i}{(T\nu)_i} + \log \frac{F}{\Lambda}] && (21)\\ &= F\sum_i (Tg)_i \log \frac{(Tg)_i}{(T\nu)_i} + \log \frac{F}{\Lambda} && (22)\\ &= F\sum_i (Tg)_i \log \frac{(T^+g)_i - (T^-g)_i}{(T^+\nu)_i - (T^-\nu)_i} + \log \frac{F}{\Lambda}. && (23) \end{aligned}$$

Next, control the departure from non-negativity. Assume that $\lambda_1 < 0$ and $\lambda_i > 0$ for all $i = 2, \ldots, m$. It follows that $T_{i1} = -1$ for some i (and all other columns contain an entry of $+1$). From this, we can write $(T\nu)_i = -\nu_1 + K_\nu$, for some $K_\nu \geq \nu_1$. We will further limit the departure from non-negativity by assuming $K_\nu \gg \nu_1$. Note that $(T^+\nu)_i = \nu_1 + K_\nu$ and $(T^-\nu)_i = 2\nu_1$. Next write $(Tg)_i = -g_1 + K_g$. We assume that $K_g \gg g_1$ follows from $K_\nu \gg \nu_1$. (This relies on the approximations $g_1 \approx \nu_1$ and $K_g \approx K_\nu$ when the sample size is large.) Note that $(T^+g)_i = g_1 + K_g$ and $(T^-g)_i = 2g_1$.

Substituting into Equation (23), we get:

$$D(f||\mu) = F(K_g - g_1)\log\frac{K_g - g_1}{K_\nu - \nu_1} + F\sum_{j\neq i}(T^+g)_j\frac{(T^+g)_j}{(T^+\nu)_j} + \log\frac{F}{\Lambda} \tag{24}$$

$$= F(K_g - g_1)\log\frac{K_g - g_1}{K_\nu - \nu_1} - F(K_g + g_1)\log\frac{K_g + g_1}{K_\nu + \nu_1} + F\sum_{j=1}^{m}(T^+g)_j\frac{(T^+g)_j}{(T^+\nu)_i} + \log\frac{F}{\Lambda} \tag{25}$$

$$\leq F(K_g - g_1)\log\frac{K_g - g_1}{K_\nu - \nu_1} - F(K_g + g_1)\log\frac{K_g + g_1}{K_\nu + \nu_1} + FD(g||\nu) + \log\frac{F}{\Lambda}, \tag{26}$$

where the inequality comes from the **classical data-processing inequality** (Goldfeld, 2020, Lecture 7) applied to the column-stochastic matrix T^+.

We now use $F = 1/(1-2g_1)$ and also approximate to first order the first two log terms on the right hand side of Inequality 26:

$$F(K_g - g_1)\log\frac{K_g - g_1}{K_\nu - \nu_1} - F(K_g + g_1)\log\frac{K_g + g_1}{K_\nu + \nu_1} \approx$$
$$F(-2g_1\log\frac{K_g}{K_\nu} - 2g_1 + 2\frac{K_g}{K_\nu}\nu_1) = \frac{1}{1-2g_1}(-2g_1\log\frac{K_g}{K_\nu} - 2g_1 + 2\frac{K_g}{K_\nu}\nu_1) \approx$$
$$(1+2g_1)(-2g_1\log\frac{K_g}{K_\nu} - 2g_1 + 2\frac{K_g}{K_\nu}\nu_1) \approx -2g_1\log\frac{K_g}{K_\nu} - 2g_1 + 2\frac{K_g}{K_\nu}\nu_1. \tag{27}$$

From $\Lambda = 1 + 2|\lambda_1|$ we get:

$$\log\frac{F}{\Lambda} = \log\frac{1}{(1-2g_1)(1+2|\lambda_1|)} \approx 2g_1 - 2|\lambda_1|. \tag{28}$$

Substituting the approximations 27 and 28) into Inequality 26, we get the approximate relation:

$$D(f||\mu) \leq -2g_1\log\frac{K_g}{K_\nu} - 2g_1 + 2\frac{K_g}{K_\nu}\nu_1 + (1+2g_1)D(g||\nu) + 2g_1 - 2|\lambda_1|. \tag{29}$$

Finally, use again the large-sample approximations $g_1 \approx \nu_1$ and $K_g \approx K_\nu$, and also $\nu_1 = |\lambda_1|/\Lambda$. Substituting into Inequality 29 yields Theorem 1 below. Before stating the result, we recap the assumptions on controlled negativity we have made: (i) $\lambda_1 < 0$ and $\lambda_i > 0$ for all $i = 2, \ldots, m$; (ii) $K_\nu \gg \nu_1$; and (iii) $K_g \gg g_1$.

Theorem 1. *The relative entropy $D(f||\nu)$ for the empirical distribution under the actual signed process and the relative entropy $D(g||\mu)$ for the empirical distribution under the classical simulation satisfy:*

$$D(f||\mu) \leq (1 + \frac{2|\lambda_1|}{\Lambda})D(g||\nu) - \frac{4|\lambda_1|^2}{\Lambda}. \tag{30}$$

This is our **signed data processing inequality**, which holds when a small degree of negativity is allowed in the underlying process and we consider a large-sample approximation. Of course, when $\lambda_1 = 0$, we recover the classical inequality $D(f||\mu) \leq D(g||\nu)$.

The key difference from the classical case is that this new inequality allows $D(f||\mu) > D(g||\nu)$, from which it would follow that:

$$D(f||\mu) \geq \inf_{\{g:\,\Gamma(g)=f\}} D(g||\nu), \tag{31}$$

answering the question at the end of Section 3. From this we would conclude that the probability of a given large deviation is lower under the actual process μ than under the simulation ν. Equivalently, large fluctuations are more likely under the simulation than under the actual process. (Strictly speaking, we need to rephrase the large deviation theory of Section 3 to take into account that we have moved all probability measures to live on the set X. But exactly the same formulas will carry over.)

In the next section, we give a numerical example to show that the phenomenon of larger fluctuations under the simulation can actually occur. We end this section with an immediate consequence of Theorem 1, that gives a general condition under which there is a reversal of the classical inequality.

Corollary 2. *Suppose the relative entropy $D(f||\mu)$ saturates the upper bound in Theorem 1. Then, if:*

$$D(g||\nu) > 2|\lambda_1|, \tag{32}$$

we obtain $D(f||\mu) > D(g||\nu)$.

Since any relative entropy $D(g||\nu)$ is non-negative and λ_1 is small by assumption, Corollary 2 provides a mild sufficient condition for reversal under the saturation assumption.

5 Numerical Example

We consider the case where λ is "almost uniform," specifically:

$$\lambda = (-\epsilon, \frac{1}{m} + \epsilon, \frac{1}{m}, \ldots, \frac{1}{m}), \tag{33}$$

for small $\epsilon > 0$, so that:

$$\nu = (\frac{\epsilon}{1+2\epsilon}, \frac{1/m+\epsilon}{1+2\epsilon}, \frac{1/m}{1+2\epsilon}, \ldots, \frac{1/m}{1+2\epsilon}). \tag{34}$$

We set:

$$\chi(i) = \begin{cases} 2 \text{ if } i = 1, 2, \\ i \text{ otherwise,} \end{cases} \tag{35}$$

from which:

$$\mu = (0, \frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}). \tag{36}$$

Now calculate (setting $f_1 = 0$ since $\mu_1 = 0$):

$$D(g||\nu) = \sum_{i=1}^{m} g_i \log \frac{g_i}{\nu_i} \tag{37}$$

$$= g_1 \log \frac{g_1}{\epsilon/(1+2\epsilon)} + g_2 \log \frac{g_2}{(1/m+\epsilon)/(1+2\epsilon)} + \sum_{i=3}^{m} g_i \log \frac{g_i}{(1/m)/(1+2\epsilon)} \tag{38}$$

$$= \log(1+2\epsilon) + g_1 \log \frac{g_1}{\epsilon} + g_2 \log \frac{g_2}{1/m+\epsilon} + \sum_{i=3}^{m} g_i \log \frac{g_i}{1/m} \tag{39}$$

$$= \log(1+2\epsilon) + g_1 \log \frac{g_1}{\epsilon} + g_2 \log \frac{g_2}{1/m+\epsilon} + \sum_{i=3}^{m} \frac{f_i}{F} \log \frac{f_i/F}{1/m} \tag{40}$$

$$= \log(1+2\epsilon) + g_1 \log \frac{g_1}{\epsilon} + g_2 \log \frac{g_2}{1/m+\epsilon} + \frac{1}{F} D(f||\mu) - \frac{f_2}{F} \log \frac{f_2}{1/m} - \frac{1-f_2}{F} \log F. \tag{41}$$

Note that $F = 1/(1-2g_1)$ and $f_2 = F(g_2 - g_1)$, so that:

$$D(g||\nu) - D(f||\mu) =$$

$$\log(1+2\epsilon) + g_1 \log \frac{g_1}{\epsilon} + g_2 \log \frac{g_2}{1/m+\epsilon} - (1 - \frac{1}{F}) D(f||\mu) - \frac{f_2}{F} \log(f_2 m) - \frac{1-f_2}{F} \log F \tag{42}$$

$$= \log(1+2\epsilon) + g_1 \log \frac{g_1}{\epsilon} + [g_1 + (1-2g_1)f_2] \log \frac{g_1 + (1-2g_1)f_2}{1/m+\epsilon}$$
$$- 2g_1 D(f||\mu) - (1-2g_1) f_2 \log(f_2 m) + (1-f_2)(1-2g_1) \log(1-2g_1). \tag{43}$$

To control the behavior of $g_1 \log(g_1/\epsilon)$ near $\epsilon = 0$, we set $g_1 = c\epsilon$ where $c \neq 0$, to obtain:

$$D(g||\nu) - D(f||\mu) =$$
$$\log(1+2\epsilon) + c\epsilon \log c + [c\epsilon + (1-2c\epsilon)f_2] \log \frac{c\epsilon + (1-2c\epsilon)f_2}{1/m+\epsilon} - 2c\epsilon D(f||\mu)$$
$$- (1-2c\epsilon) f_2 \log(f_2 m) + (1-f_2)(1-2c\epsilon) \log(1-2c\epsilon). \tag{44}$$

It can be checked that:

$$[D(g||\nu) - D(f||\mu)]\Big|_{\epsilon=0} = 0. \tag{45}$$

We next calculate:

$$\frac{\partial[D(g||\nu) - D(f||\mu)]}{\partial \epsilon} = \frac{2}{1+2\epsilon} + c \log c + (c - 2cf_2) \log \frac{c\epsilon + (1-2c\epsilon)f_2}{1/m+\epsilon}$$
$$+ [c\epsilon + (1-2c\epsilon)f_2][\frac{c - 2cf_2}{c\epsilon + (1-2c\epsilon)f_2} - \frac{1}{1/m+\epsilon}] - 2cD(f||\mu)$$
$$+ 2cf_2 \log(f_2 m) + (1-f_2)[-2c \log(1-2c\epsilon) + (1-2c\epsilon)\frac{-2c}{1-2c\epsilon}]. \tag{46}$$

Evaluating the derivative at $\epsilon = 0$ yields:

$$\begin{aligned}\frac{\partial[D(g||\nu) - D(f||\mu)]}{\partial\epsilon}\Big|_{\epsilon=0} &= 2 + c\log c + c(1-2f_2)\log(f_2 m) + c(1-2f_2) - f_2 m - 2cD(f||\mu) \\ &\quad + 2cf_2\log(f_2 m) - (1-f_2)2c \\ &= 2 + c\log c + c\log(f_2 m) - f_2 m - 2cD(f||\mu) - c. \end{aligned} \tag{47}$$

For a large sample size N, we write $f_2 \approx \mu_2 = 1/m$ and $g_1 \approx \nu_1$, so that $c \approx 1$. We find:

$$\frac{\partial[D(g||\nu) - D(f||\mu)]}{\partial\epsilon}\Big|_{\epsilon=0} \approx -2D(f||\mu) \leq 0, \tag{48}$$

where the inequality follows from non-negativity of relative entropy. Recall also that any relative entropy $D(f||\mu)$ satisfies: $D(f||\mu) = 0$ if and only if $f = \mu$.

We conclude that if the negative component ϵ in the almost-uniform probability measure λ is small but nonzero, the sample size N is sufficiently large, and $f \neq \mu$, then:

$$D(g||\nu) < D(f||\mu), \tag{49}$$

which is the reversed inequality we were seeking.

References

Abramsky, S., and A. Brandenburger, "The Sheaf-Theoretic Structure of Non-Locality and Contextuality," *New Journal of Physics*, 13, 2011, 113036.

Abramsky, S., and A. Brandenburger, "An Operational Interpretation of Negative Probabilities and No-Signalling Models," in van Breugel, F., E. Kashefi, C. Palamidessi, and J. Rutten (eds.), *Horizons of the Mind: A Tribute to Prakash Panagaden*, Lecture Notes in Computer Science 8464, Springer, 2014, 59-75.

Brandenburger, A, P. La Mura, and S. Zoble, "Rényi Entropy, Signed Probabilities, and the Qubit," *Entropy*, 24, 2022, 1412.

Dembo, A., and O. Zeitouni, *Large Deviations: Techniques and Applications*, Springer, 2nd edition, 1998.

Dirac, P., "The Physical Interpretation of Quantum Mechanics," *Proceedings of the Royal Society of London (Series A: Mathematical and Physical Sciences)*, 180, 1942, 1-40.

Ferrie, C., "Quasi-Probability Representations of Quantum Theory with Applications to Quantum Information Science," *Reports on Progress in Physics*, 74, 2011, 116001.

Feynman, R., "Simulating Physics with Computers," *International Physics with Computers*, 21, 1982, 467-488.

Feynman, R., "Quantum Mechanical Computers," *Optics News*, 11, 1985, 11-20.

Feynman, R., "Negative Probability," in Hiley, B., and F. Peat (eds.), *Quantum Implications: Essays in Honour of David Bohm*, Routledge & Kegan Paul, 1987, pp.235-248.

Gherardini, S., and G. De Chiara, "Quasiprobabilities in Quantum Thermodynamics and Many-Body Systems," *PRX Quantum*, 5, 2024, 030201.

Goldfeld, Z., "Information Theory for Data Transmission, Security, and Machine Learning," ECE 5630, Cornell University, 2020, Lecture 7, at http://people.ece.cornell.edu/zivg/ECE_5630_Lectures7.pdf.

Kenfack, A., and K. Życzkowski, "Negativity of the Wigner Function as an Indicator of Non-Classicality," *Journal of Optics B: Quantum and Semiclassical Optics*, 6, 2004, 396-404.

Onggadinata, K., P. Kurzynski, and D. Kaszlikowski,"Qubits from the Classical Collision Entropy," *Physical Review A*, 107, 2023, 032214.

Popescu, S., and D. Rohrlich, "Quantum Nonlocality as an Axiom," *Foundations of Physics*, 24, 1994, 379-385.

Sanov, I., "On the Probability of Large Deviations of Random Variables," *Mat. Sbornik*, 42, 1957, 11-44.

Wigner, E., "On the Quantum Correction for Thermodynamic Equilibrium," *Physical Review*, 40, 1932, 749-759.