Comments on and Comments on Comments on Verlinde's paper "On the Origin of Gravity and the Laws of Newton"

Sabine Hossenfelder [*]

NORDITA, Roslagstullsbacken 23, 106 91 Stockholm, Sweden

Abstract

We offer some, hopefully clarifying, comments on Verlinde's recent claim that gravity is an entropic force. A suitable identification of quantities shows that both formulations of Newtonian gravity, the classical and the thermodynamical one, are actually equivalent. It turns out that some additional assumptions made by Verlinde are unnecessary. However, when it comes to General Relativity there remain some gaps in the argument. We comment on whether this identification can be done also for electrostatics. Finally, some thoughts on the use of this reinterpretation are offered.

1 Introduction

The purpose of this brief note is to offer some clarifying words on the logic of Verlinde's recent paper [1]. In his paper Verlinde showed that (time-independent) Newtonian gravity is an entropic force and claimed it to follow from a thermodynamical description that has certain holographic properties. Below, we will make this statement more precise. We will further show that the reverse is also true, i.e. it follows from Newtonian gravity that it can be described as an entropic force with holographic properties. We are thus lead to conclude both descriptions are equivalent and, after a suitable identification of quantities, the thermodynamical character of gravity is a reinterpretation.

Let us first state precisely what we mean. Let there be given a charge distribution ρ which is a density. Then, static Newtonian gravity can be characterized as follows:

> **A:** There is a scalar field ϕ which obeys the Poisson equation $\nabla^2\phi = 4\pi G\rho$. A test-mass m in the background field of a mass M with field ϕ_M experiences a force $\vec{F} = m\vec{\nabla}\phi_M$.

---

[*]hossi@nordita.org

Here, G is some coupling constant. Verlinde's proposal instead can be cast as follows

> **B:** There are two scalar quantities S and T and a continuous set of non-intersecting surfaces $\mathcal{S}$, the 'holographic screens,' whose union covers all of space $\mathbb{R}^3 = \cup \mathcal{S}$. The theory is defined by $2G \int_{(\mathcal{S})} \rho dV = \int_{\mathcal{S}} T dA \ \forall \mathcal{S}$, and the force acting on a particle with test-mass m is given by $F \delta x = \int_{\mathcal{S}} T \delta dS$, where the integral is taken over a screen that does not include the test-mass.

Here, $(\mathcal{S})$ denotes the volume with surface $\mathcal{S}$. The volume integral $\int_{(\mathcal{S})} dV \rho$ is of course just the total mass M, inside that volume, and the quantities S and T are interpreted as the entropy respectively temperature on the holographic screens. The δx is a virtual variation on the location on the particle which induces a change in the entropy on the screen, the details will become clear later.

The statement **B** is not actually exactly what Verlinde claimed in [1], in particular he did not use the function S itself but merely its differential dS. Verlinde further defined another quantity s, a surface entropy density, with help of which $\delta dS = \delta s$, and another quantity $N = GA$, the "number of bits on the screen," with help of which $2M = \int_{\mathcal{S}} T dN$, the "equipartition theorem." We will not use the quantities because we don't actually need them, and our argument becomes leaner without them. The other difference is that Verlinde defines the screens as being equipotential surfaces for ϕ. However, in the thermodynamical approach ϕ should not enter the formulation of the theory since it's what one wants to get rid of. Thus the above formulation **B** avoids referring to the screens as equipotential surfaces. We'll come back to this later.

Before we get to the derivation, let's have a brief look at electrostatics, since it is apparent that **A** could equally have been about electric charges. To avoid having to constantly distinguish two cases, we convert the charge to mass dimension one, and the potential to mass dimension zero by multiplying with powers of the Planck mass m_{Pl} and subsume the remainder in the coupling constant. For the case of gravity there's nothing to do. For electrostatics we have $Q = m_{\text{Pl}} \tilde{Q}$, $\phi = \tilde{\phi}/m_{\text{Pl}}$, where the quantities with tildes are the usual ones. This is just a rescaling, also used in [2], that will make the following apply for both cases. We will thus generally refer to a charge and label it q, but this charge could be a mass. We will on some places comment on the differences between electrostatics and gravity. Note that with the sign convention of **A**, which we use to be in agreement with [1], a graviational potential ϕ is actually negative.

We'll use the unit convention $\hbar = c = 1$ such that $m_{\text{Pl}} = 1/l_{\text{Pl}}$. We will consider the case with 3 spatial dimensions, such that $G = l_{\text{Pl}}^2$, though the number of

dimensions doesn't matter much[1].

2 It follows from Newtonian Gravity that it's an Entropic Force

Now let us come to the derivation. Verlinde in his paper claimed to have shown **B** $\Rightarrow$ **A**. We will thus look at **A** $\Rightarrow$ **B**, and start with the scalar field ϕ obeying the Poisson equation.

The field will have equipotential surfaces of codimension 1 where $\phi(\vec{x}) = $ constant. We identify these surfaces with the holographic screens. That's the first ingredient to **B**. The values of these surface areas should be a smooth function, in particular it won't have gaps. Each of these surface has a normal vector in every point that we'll denote as $\vec{n}$. This normal vector might be ill-defined in some points. For example a Lagrange-point has an indefinite normal vector. (If a particle was placed exactly at this point, it would stay there.) Note that in the general case these surfaces will not be simply connected, but consist of several pieces.

Let us assign the corresponding surface $A(\phi)$ to every value of ϕ on $\mathcal{S}$ and normalize it to a unit area $A_0 = G$. We then define the following scalar function

$$S(\vec{x}) := -\phi(\vec{x})\frac{A(\phi)}{2G} + S_0 \quad , \tag{1}$$

Here, S_0 is some additive constant. Suggestively named S, in the case of gravity the function reduces to the black hole entropy for $r = 2MG$. We further define another scalar quantity

$$T(\vec{x}) := \frac{1}{2\pi}\nabla_n\phi \quad , \tag{2}$$

which is the derivative in direction of the normal vector on the equipotential surface. For the case of gravity again it reduces to the black hole temperature at the horizon. The quantity T can always be constructed, where the normal vector is well-defined. For gravity where like charges attract and there's only positive charges T is positive definite, but for electric fields this will not generally be the case.

It then follows by use of Gauss' law from the Poisson equation that on any equipotential surface

$$\int_{(S)} \rho dV = \frac{1}{4\pi G}\int_{\mathcal{S}} \nabla_n\phi dA = \frac{1}{2G}\int_{\mathcal{S}} T dA \quad , \tag{3}$$

---

[1] The case with only one dimension is pathological. To begin with because a point doesn't have a surface.

3

which is the second ingredient to **B**. Note that this is not true for general surfaces since the normal used to define the temperature will in general not be the normal of the surface. Let us further recall the potential energy of the system is

$$U = -\int \rho \phi dV \quad , \tag{4}$$

which might contain divergent self-energy contributions that have to be suitably omitted (recall that with our sign convention $\phi < 0$). Now consider a test-particle with charge q at location r_0 in the background of a charge $Q \gg q$ contained within a compact volume. The background field has potential ϕ_Q and the test-particle has the potential ϕ_q which by use of the Poisson-equation is just the usual $\phi_q = -Gq/|r - r_0|$. Now we ask what force we have to apply to change the test-particle's location by δx. It is

$$\vec{F}\delta\vec{x} = \delta U = -\int dV \phi_Q \delta\rho = -\frac{1}{4\pi G}\int dV \phi_Q \nabla^2 \delta\phi_q \quad , \tag{5}$$

where the integral is taken over some volume that it includes the test-particle, but not the background charges, i.e. it is outside some surface dividing both. Note that this has nothing to do with thermodynamics or holography whatsoever.

We will now repeat the steps used in [1]. First, we set the volume the integral is taken over to be the volume inside an equipotential surface of ϕ_Q such that the test-particle is outside that surface. We denote that surface S, the volume inside (S) and the volume outside $\mathbb{R}^3 \backslash (S)$. We start with rewriting the volume integral in an integral over all of space minus the integral over the inside, and then transform the volume over the inside by making use of Green's second identity (see Appendix). One obtains

$$-4\pi G\delta U = \int_{\mathbb{R}^3} \phi_Q \nabla^2 \delta\phi_q dV - \int_{(S)} \delta\phi_q \nabla^2 \phi_Q dV + \int_S \left(\delta\phi_q \nabla\phi_Q - \phi_Q \nabla\delta\phi_q\right) dA \ . \tag{6}$$

Since we are integrating over an equipotential surface for ϕ_Q, in the second term in the surface integral we can pull out ϕ_Q. Then we rewrite that surface-integral again in a volume integral over (S) by making use of Gauss's law. We then see that it vanishes because ϕ_q does not have sources inside the volume. The second term with the volume integral over the inside we rewrite again in a volume integral over $\mathbb{R}^3$ minus an integral over $\mathbb{R}^3 \backslash (S)$. The integral over the outside vanishes because ϕ_Q does not have sources there. We are then left with

$$-4\pi G\delta U = \int_{\mathbb{R}^3} \left(\phi_Q \nabla^2 \delta\phi_q - \delta\phi_q \nabla^2 \phi_Q\right) dV + \int_S \delta\phi_q \nabla\phi_Q dA \ . \tag{7}$$

For any finite volume, the remaining volume integral can be written as a surface integral making use of Green's 2nd identity again. One then sees that the integral

4

vanishes if the surface is moved to infinity as long as the sources are compact. Thus, we finally arrive at

$$\vec{F}\delta\vec{x} = -\frac{1}{4\pi G}\int_{\mathcal{S}}\delta\phi_q\nabla\phi_Q dA \quad , \tag{8}$$

where the surface integral could have been taken over any equipotential surface for the background field ϕ_Q if the test-particle is not inside the volume enclosed by the surface. At such a fixed surface we have $2G\delta S = -A\delta\phi_q$ and the surface element is[2] $\delta dS = -\delta\phi_q dA$ thus

$$\vec{F}\delta\vec{x} = \int_{\mathcal{S}} T\delta dS . \tag{9}$$

This completes the derivation **A** $\Rightarrow$ **B**.

One should note two points about this derivation. One is that we have not made a variation of the surface. We have instead made a variation of the potential on a the fixed surface, induced by wiggling the location of the test-particle. The other point is that all steps from Eq. (5) to Eq. (8) can be reversed. For the final step from Eq.(8) to Eq.(9), one uses the identification of the thermodynamical quantities with the Newtonian potential.

Since $\delta\phi_q$ is not constant on $\mathcal{S}$ it is somewhat cumbersome but one can work out the integral on the right side of Eq.(8) for the case of two point-masses and obtain the usual Newtonian gravitational force or Coulomb law respectively. This should not surprise us, since all we have done is using integral identities for harmonic functions.

It is apparent that this derivation does not make use of the character of the charge. However, let us look at a very simple case to see why the thermodynamic analogy does not make much sense for electrodynamics. Consider two point-charges q_1 and q_2 very far away from each other, and an equipotential surface with value ϕ_0. If the charges are far enough away from each other, to good precision the area of the equipotential surface is

$$A_{q_1} + A_{q_2} = 4\pi\frac{G^2}{\phi_0^2}\left(q_1^2 + q_2^2\right) . \tag{10}$$

Now we put the two charges very close together, and find that the area of the equipotential surface changes to

$$A_{q_1+q_2} = 4\pi\frac{G^2}{\phi_0^2}\left(q_1 + q_2\right)^2 . \tag{11}$$

---

[2]This is the δs in Verlinde's paper.

Now for gravity, two positive charges (masses) attract. In this process when the masses approach each other one thus has an increase in area of the surface, proportional to $q_1 q_2$. Identifying the change of area for a constant value of ϕ with the change in entropy, the entropy appropriately increases. However, for the electromagnetic interaction it's opposite charges that attract, resulting in a decrease of area or 'entropy' since $q_1 q_2$ is now negative. This is in agreement with our previous observation that for electrodynamics the 'temperature' could be negative.

Thus, while the derivation above can also be made for electrostatics, the interpretation of S and T in thermodynamical terms does not seem meaningful. Only if the force is attractive between like charges the association with an entropic force makes sense. If one starts with assuming that gravity is an entropic force then, as pointed out by Smolin in [3], this means gravity is attractive between like charges. It does not exclude however that there are negative gravitational charges, just that if there were opposite charges one would have to do work to bring them closer together since $dA, dS < 0$ in that process.

3   Newtonian Gravity follows from an Entropic Force Law

For completeness, the following is a repetition of Verlinde's argument with some further remarks. As mentioned previously, absent a gravitational potential we should not define the 'holographic screens' as its equipotential surfaces. One could think of instead defining them as surfaces of constant entropy. However, from our previous identification of quantities $2GS = \phi A(\phi)$, we see that a surface of constant ϕ is a surface of constant S, but the reverse might not necessarily be true since a change in ϕ could be balanced by a change in area.

For a monopole (or far distance field), ϕ drops slower than the area of its equipotential surfaces increases. The entropy thus increases monotonically with distance and thus the surfaces of constant S do identify the surfaces of constant ϕ. However, when we think of higher multipole moments, this might not generally be true[3]. But we have additional information in T. With it, we can define a screen $\mathcal{S}$ as a surface with constant entropy that fulfils the relation $2G \int_{(\mathcal{S})} \rho dV = \int_{\mathcal{S}} T dA$ in every point. On these screens with area $A(\mathcal{S})$, we can then get back the Newtonian potential by defining it as $\phi = 2GS/A(\mathcal{S})$. Consequently, if we make a small change to S on a constant surface, we have $\delta \phi = 2G\delta S/A(\mathcal{S})$.

Having all quantities necessary for **A**, the rest of the argument is then an exact reversal of the steps in the previous section, with one additional subtlety. If one knows that $4\pi G \int_{(\Sigma)} dV \rho = \int_{\Sigma} \nabla_n \phi dA$ holds for every surface Σ with normal vector $\vec{n}$, then one can easily conclude that ρ must fulfill the Poisson equation. However,

---

[3]At least it is not obvious to me. Comments are welcome.

we only have this relation for equipotential surfaces. The missing information is in the shape of the surfaces themselves. This can be seen as follows. The charge distribution ρ and the normal on the (known) surface constitutes a case of a Neumann boundary condition and if the field fulfills the Poisson equation it is uniquely specified up to a constant. If there was another, different, solution not fulfilling the equation, it would have to differ from the solution to the Poisson equation in at least one point. But this would mean the equipotential surfaces had to be shifted and thus cannot be.

Taken together, we have thus shown that $\mathbf{A} \Leftrightarrow \mathbf{B}$.

4 Remarks

The number N of 'bits' on the screen made no appearance here. In particular, making the "identification for the temperature and the information density on the holographic screens" [1] might be "natural" but is unnecessary. In any case if one uses that $A = GN$ then to obtain Eq.(9), we kept A and thus N constant. Since the missing variation with respect to N was criticized in [4, 5], let us point out there is nothing inconsistent about this, one just has to be careful about stating which quantities are held fix (as usual in thermodynamics). Since we did not need N, the equipartition theorem also does not play a role.

Further, for the above it was unnecessary to use that ΔS is proportional to Δx. This is just a shortcut to arrive at Newton's gravitational force law, but not necessary for the general case which deals with the variation of the entropy on a screen induced by varying the position of the test-particle. Since the relation $\Delta S \propto \Delta x$ does not enter the derivation, it is moot to dissect its meaning, discuss the constant of proportionality, or the appropriateness in general cases. One also does not actually need any knowledge about the Unruh-effect, though this knowledge helps to give meaning to the definition of the temperature.

It should be emphasized that the direction $\mathbf{B} \Rightarrow \mathbf{A}$ is the more complicated direction of the equivalence. This is partly due to the identification of the potential. But mostly it is due to $\mathbf{A}$ being a description available in the whole volume and all its surfaces, while $\mathbf{B}$ makes use only of specific surfaces. That both are equivalent nevertheless is a consequence of the specific properties of solutions to the Poisson equation.

In the later sections of [1] the argument is extended from Newtonian gravity to General Relativity (GR) in a spacetime with a time-like killing vector. From the argument presented there, one sees that $\mathbf{A} \Rightarrow \mathbf{B}$ can be extended to GR. However, the assumptions for the reverse claim, that gravity follows from thermodynamics, rely heavily on knowledge about GR already: Instead of defining the gravitational

potential from thermodynamical quantities, the temperature is defined making use of the GR generalization of the gravitational potential. It enters the (previously defined) killing vector. The definition of screens still has the same problem as in the Newtonian case. Later it is moreover used that the total mass inside a volume can be rewritten in terms of the stress-energy-tensor, and in addition that this stress-energy-tensor is covariantly conserved. One also has to use the equivalence principle when going into a locally Minkowski frame. Finally, one might wonder if, absent something like a space-time, it was appropriate to use derivatives to begin with, or if these should not also 'emerge' in some sense.

5 Discussion

In any case, it presently does not seem impossible to fix the gaps pointed out in the previous section, so let us for now consider there was indeed a thermodynamical description equivalent to GR. This then raises two questions: What does this mean? and What is it good for?

In the previous sections we avoided an interpretation of the reformulation suggested by Verlinde. It remains to be clarified whether there is more justification for referring to the quantities in thermodynamical terms. Of course one wonders what it's supposed to mean that any point in spacetime is now associated with a temperature. For the case of a black hole horizon there is an actual particle emission associated to the temperature. Should we now expect arbitrary screens to also emit particles?

Well, first a screen that is not a black hole horizon does not have a problem emitting particles. It is clear where they come from: they come from the next screen closer to a source and can be traced back to the source which is a sink of field lines. The problem with a horizon is just that all paths crossing it are classically forbidden. However, the temperature of any such screen is ridiculously small and entirely negligible for all practical purposes. Consider a proton with a mass of $m_p \approx$ GeV. Assuming it has a has a typical size of $\approx 1/m_p$, it has a surface temperature of $T \approx m_p(m_p/m_{Pl})^2 \approx 10^{-25}$eV. Asking how long it would take for the proton to noticeably lose a quantum of energy, we can as usual integrate $dM/dt \approx R^2 T^4$. One finds that it would take about 10^{100} times the current age of the universe for the proton to decay in that fashion. We can safely forget about that[4].

In any case, if observable consequences of the reinterpretation are such remote,

---

[4]That this number is so large compared to the black hole temperature is because T scales with the third power of the mass rather than with the inverse of the mass as is the case for a black hole horizon.

what is it good for? Clearly, the equivalence of two formulations for the gravitational interaction is not the interesting aspect. As long as both descriptions are the same claiming "the redshift must be seen as a consequence of the entropy gradient and not the other way around" [1] is merely words. The interesting aspect of the reformulation would be to make use of it in regimes where the equivalence might no longer hold. The thermodynamical description might provide a bridge to a statistical mechanics description of a possibly underlying theory.

We know that gravity is not easy to quantize, and there are cases we cannot describe by use of GR. The question is whether in these cases classical gravity might cease to be meaningful, but the thermodynamic relations continue to hold, thus making them accessible to us. However, before one starts thinking about quantum gravity one should maybe try to first reproduce at least quantum mechanics in a classical background field. The next step could then be asking for the description of the gravitational force exerted by a superposition state, which is already a case where a classical description of the particle's gravitational field fails.

Another remark, mentioned before in [4], is that we do have a fundamental theory for electrodynamics that is extremely well tested. Since formally the identification with thermodynamics can also be made for electrostatics it might be illuminating to find out whether it continues to hold under quantization. And then it should be clarified what to do about solutions of the Laplace solutions, i.e. how to reproduce gravitational waves.

Finally, let's come back to the claim that according to this reformulation, gravity has a holographic character. Since the thermodynamical interpretation is just an equivalent rewriting there's no more holography in it then there was in Newtonian gravity or electrostatics till last year. Arguably one could say though that harmonic functions do have 'holographic' properties, encoded by Gauss's and Stoke's integral identities that we heavily made use of. Rather than speaking of holography, it might be more appropriate to refer to such a theory that does have bulk degrees of freedom but allows for a holographic description as 'holographic friendly,' meaning it might possibly be extendable to regimes where bulk degrees of freedom are indeed not available[5].

Acknowledgements

---

[5]Lee Smolin, private communication.

Appendix

For convenience, here is Gauss's law

$$\int_{(\Sigma)} \nabla^2 \psi \, dV = \int_{\Sigma} \nabla_n \psi \, dA \ . \tag{12}$$

and Green's 2nd identity

$$\int_{(\Sigma)} \left(\psi \nabla^2 \phi - \phi \nabla^2 \psi \right) dV = \int_{\Sigma} \left(\psi \nabla_n \phi - \phi \nabla_n \psi \right) dA \ . \tag{13}$$

where, as previously, (Σ) denotes the volume inside a surface Σ, and ∇_n is the derivative in direction of the surface's normal vector. ψ and ϕ are both scalar fields.

References

[1] E. P. Verlinde, arXiv:1001.0785 [hep-th].

[2] T. Wang, arXiv:1001.4965 [hep-th].

[3] L. Smolin, arXiv:1001.3668 [gr-qc].

[4] S. Gao, arXiv:1002.2668 [gr-qc].

[5] H. Culetu, arXiv:1002.3876 [hep-th].