QUALITATIVE STABILITY FOR A FAMILY OF TRACE SOBOLEV INEQUALITIES

ROBIN NEUMAYER

Abstract. The goal of this short note is to prove qualitative stability for a family of trace Sobolev inequalities first proven by Carlen & Loss for $p = 2$ and by Maggi and the author for $p \in (1, n)$. This answers an open problem raised in a recent paper of Fan, Li & Zhang, and in conjunction with their local analysis, yields sharp quantitative stability for this family of inequalities when $p = 2$.

1. Introduction

Fix $n \geq 2$, $p \in (1, n)$, and a half space $H = \{x \in \mathbb{R}^n : x \cdot e_1 > 0\}$. In [MN17], Maggi & Neumayer used a mass transportation argument to establish a one-parameter family of trace Sobolev inequalities on H, which encode the classical Sobolev and Escobar inequalities as special cases. More specifically, consider the variational problem

$$\Phi_H(T) = \inf\{\|\nabla u\|_{L^p(H)} : u \in \mathcal{A}_T\} \qquad T \geq 0, \tag{1.1}$$

where the competitor class $\mathcal{A}_T$ is given by

$$\mathcal{A}_T = \{u \in \dot{W}^{1,p}(H) : \|u\|_{L^{p^*}(H)} = 1, \|u\|_{L^{p^\sharp}(\partial H)} = T\}\,. \tag{1.2}$$

Here the critical Sobolev exponents $p^* = \frac{np}{n-p}$ and $p^\sharp = \frac{(n-1)p}{n-p}$ are determined by scaling. Minimizers of $\Phi_H(T)$ were characterized in [MN17] for each $T > 0$: the family of minimizers $\mathcal{M}_T$ comprises dilations and horizontal translations, and multiples by ± 1 of an explicit profile U_T (recalled in section 2.1). Thus, in the resulting sharp Sobolev trace inequality

$$\|\nabla u\|_{L^p(H)} \geq \Phi_H(T) \qquad \text{for all } u \in \mathcal{A}_T, \tag{1.3}$$

equality holds if and only if $u \in \mathcal{M}_T$. The case $T = 0$ of (1.3) encodes the classical Sobolev inequality $\|\nabla u\|_{L^p(\mathbb{R}^n)} \geq S_{n,p}\|u\|_{L^{p^*}(\mathbb{R}^n)}$ for $u \in \dot{W}^{1,p}(\mathbb{R}^n)$, whose optimal constant $\Phi_H(0) = S_{n,p}$ and extremals on $\mathbb{R}^n$ were characterized in [Aub76b, Tal76]. The Escobar inequality $\|\nabla u\|_{L^p(H)} \geq E_{n,p}\|u\|_{L^{p^\sharp}(\partial H)}$ for $u \in \dot{W}^{1,p}(H)$, whose sharp constant $E_{n,p}$ and extremals were given in [Esc88, Naz06] (see also [Bec93]), implies the linear lower bound $\Phi_H(T) \geq E_{n,p}T$ for all $T \geq 0$. This lower bound is saturated for exactly one value $T_E > 0$ depending on n and p.

When $p = 2$, Carlen & Loss [CL94] first characterized minimizers of (1.1) using their method of competing symmetries [CL90]. In this case, the Sobolev and Escobar inequalities are closely linked to the Yamabe problem [Yam60, Tru68, Aub76a, Sch84, Esc92a, Esc92b].

With extremals of (1.3) characterized, stability is the natural next question: if $u \in \mathcal{A}_T$ almost achieves equality in (1.3), then is u close, in a suitable sense, to some $v \in \mathcal{M}_T$? Closeness to equality is quantified by the deficit

$$\delta_T(u) = \|\nabla u\|^p_{L^p(H)} - \Phi_H(T)^p\,,$$

while the strongest distance of a function $u \in \mathcal{A}_T$ to the nearest extremal that one expects to control is

$$d_T(u) := \inf_{v \in \mathcal{M}_T} \|\nabla(u - v)\|_{L^p(H)}.$$

In the recent paper [FLZ26], Fan, Li & Zhang gave a local quantitative analysis of the problem in the case $p = 2$, showing that there exists $\alpha_T > 0$ such that $\delta_T(u) \geq \alpha_T d_T(u)^2 + o(d_T(u)^2)$. The obstruction to turning this estimate into a global quantitative stability theorem was the absence of a *qualitative* stability result for (1.3). Here we fill this gap by establishing such a qualitative stability result. As a byproduct, we resolve the open problem of global quantitative stability for this problem raised in [FLZ26, Remark 1.1].

Theorem 1.1. *Fix $p \in (1, n)$ and $T > 0$. Given $\{u_k\} \subset \mathcal{A}_T$, if $\delta_T(u_k) \to 0$, then $d_T(u_k) \to 0$.*

Combining Theorem 1.1 with [FLZ26, Theorem 1.1] yields global quantitative stability when $p = 2$.

Corollary 1.2. *Fix $p = 2$ and $T > 0$. There exists $\alpha'_T > 0$ such that $\delta_T(u) \geq \alpha'_T d_T(u)^2$ for all $u \in \mathcal{A}_T$.*

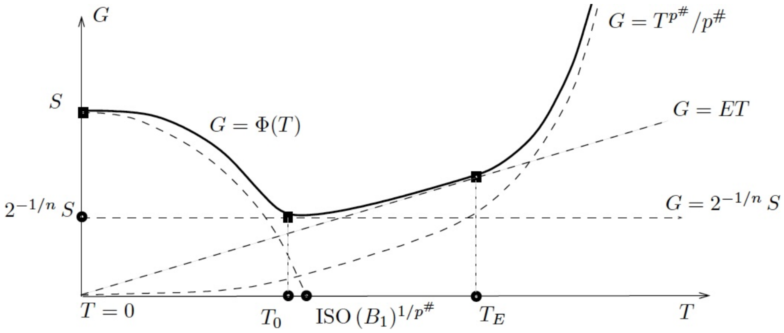

Figure 1. A plot illustrating known properties of $T \mapsto \Phi_H(T)$; see section 2.1 for discussion.

Quantitative stability for the Escobar inequality with $p = 2$ was shown in [Ho22]. For the Sobolev inequality on $\mathbb{R}^n$, sharp quantitative stability was shown in [BE91] for $p = 2$, see also [DEF$^+$25], and for $p \in (1, n)$ in [FZ22], after [CFMP09, FN19, Neu20].

At first glance, one might expect Theorem 1.1 to follow directly from standard concentration compactness and scaling methods [Lio85], as is the case for many functional inequalities, including the classical Sobolev inequality on $\mathbb{R}^n$ and the Escobar inequality on H. The fundamental difference here is that the sharp constant $\Phi_H(T)$ depends nontrivially on T. This means that splitting of mass cannot be ruled out through homogeneity and concave scaling.

More explicitly, concentration compactness arguments essentially reduce the proof of Theorem 1.1 to ruling out the possibility that a sequence $\{u_k\} \subset \mathcal{A}_T$ with $\delta_T(u_k) \to 0$ fails to have $d_T(u_k) \to 0$ because it splits into two asymptotically non-interacting profiles u_k^1 and u_k^2. Suppose this happens, and u_k^1 and u_k^2 have $L^{p^*}(H)$ norms $m_1, m_2 > 0$ and $L^{p^\sharp}(\partial H)$ norms $t_1, t_2 \geq 0$ respectively, with $m_1^{p^*} + m_2^{p^*} = 1$ and $t_1^{p^\sharp} + t_2^{p^\sharp} = T^{p^\sharp}$.

For the classical Sobolev inequality, for instance, scaling easily shows such splitting is energetically too expensive: applying the Sobolev inequality to u_k^1 and u_k^2 separately shows that the total energy is at least $S_{n,p}^p(m_1^p + m_2^p)$, which by strict concavity of $s \mapsto s^{p/p^*}$ is *strictly* larger than the infimal energy $S_{n,p}^p = S_{n,p}^p(m_1^{p^*} + m_2^{p^*})^{p/p^*}$.

Instead, ruling out splitting in the present setting requires comparing the energy lower bound obtained by applying (1.3) to each profile separately, namely $m_1^p \Phi_H({}^{t_1}/_{m_1})^p + m_2^p \Phi_H({}^{t_2}/_{m_2})^p$, with the infimal energy $\Phi_H(T)^p$. These quantities cannot be related by scaling and have no clear strict ordering a priori. Even if T is restricted to a particular interval, the ratios ${}^{t_i}/_{m_i}$ may take any value in $[0, \infty)$.

In view of this discussion, the main tool to prove Theorem 1.1 is the following strict binding inequality for Φ_H.

Theorem 1.3. *Fix $T > 0$ and let $m_1, m_2 > 0$ and $t_1, t_2 \geq 0$ satisfy $m_1^{p^*} + m_2^{p^*} = 1$ and $t_1^{p^\sharp} + t_2^{p^\sharp} = T^{p^\sharp}$. Then*

$$\Phi_H(T)^p < m_1^p \Phi_H\left(\frac{t_1}{m_1}\right)^p + m_2^p \Phi_H\left(\frac{t_2}{m_2}\right)^p . \tag{1.4}$$

Another plausible approach to proving Theorem 1.1 would be to trace through the mass transportation argument of (1.3) from [MN17] (recalled in section 2.3) to extract information about the optimal transport map $\mathcal{T}$ from $u^{p^*}\mathcal{L}^n$ to $U_T^{p^*}\mathcal{L}^n$ and use this to estimate $d_T(u)$. This has been done quantitatively for the isoperimetric inequality [FMP10], the 1-Sobolev inequality [FMP13], and the Sobolev inequality with $p \in (1, n)$ restricted to radially symmetric functions [CFMP09]. This approach faces serious difficulties in the present context, and already for the Sobolev inequality on $\mathbb{R}^n$, in part because the control on $\mathcal{T}$ degenerates in regions where u is small.

To prove Theorem 1.3, we *do* use the mass transportation proof of [MN17]. The key difference is that we only need to control the optimal transport map $\mathcal{T}$ for one explicit test function $w \in \mathcal{A}_T$. We take w to be the sum of cut off and translated copies of $m_1 U_{T_1}$ and $m_2 U_{T_2}$ where $T_i = {}^{t_i}/_{m_i}$. The energy $\int_H |\nabla w|^p$ can be made arbitrarily close to the right-hand side of (1.4). (Non-strict inequality in (1.4) is immediate from testing against w.)

Say $m_1 \leq m_2$. The cyclical monotonicity of the graph of the optimal transport map $\mathcal{T}$ associated with w forces $\mathcal{T}$ to map most of the mass of $m_1 U_{T_1}$ into a half-space $\{y_n > 0\}$. On the other hand, a basic quantitative estimate obtained from the mass transportation proof (see (2.18)) shows that if $\delta_T(w)$ is small, then $\mathcal{T} - s e_1$ is parallel to $-\nabla w$ on a set with a definite amount of mass for a fixed $s = s_T$. These properties are incompatible, forcing a definite lower bound for $\delta_T(w)$ and thus proving Theorem 1.3.

Acknowledgements. The author thanks Dejan Slepčev for a useful discussion and is grateful for the support of NSF CAREER grant DMS-2340195 and NSF RTG grant DMS-2342349.

2. Preliminaries

In this section, we give some background and notation that will be needed in the rest of the paper. Let $n \geq 2$ and $p \in (1, n)$ be fixed throughout. We let $\mathcal{L}^n$ be the Lebesgue measure on $\mathbb{R}^n$ and $\mathcal{H}^{n-1}$ the $(n-1)$-dimensional Hausdorff measure. Let $\dot{W}^{1,p}(H)$ be the space of $L^1_{\text{loc}}(H)$ functions with distributional gradient in $L^p(H; \mathbb{R}^n)$ that vanish at infinity in the sense that $\mathcal{L}^n(\{|u| > t\}) < \infty$ for every $t > 0$. In a slight abuse of notation, for a function $u \in \dot{W}^{1,p}(H)$ we simply write u to refer to the trace $Tu \in L^{p^\sharp}(\partial H)$ of u.

2.1. **Minimizers for $\Phi_H(T)$.**

In [CL94] for $p = 2$ and [MN17] for $p \in (1, n)$, extremal functions of (1.3) (equivalently, minimizers of (1.1)) were characterized as follows. For each $T > 0$, the family $\mathcal{M}_T$ of extremals is given by

$$\mathcal{M}_T = \Big\{ \pm \alpha^{-n/p^*} U_T((\cdot - x_0)/\alpha) : x_0 \in \partial H, \alpha > 0 \Big\}$$

where the function U_T is defined as follows. For a given function $f : \mathbb{R}^n \to \mathbb{R}$ and $s \in \mathbb{R}$, let $\tau_s f(x) = f(x - se_1)$. Recall from the introduction that $T_E > 0$ is the unique T for which $\Phi_H(T_E) = E_{n,p} T_E$, and thus is the unique T for which the linear lower bound $\Phi_H(T) \geq E_{n,p} T$ implied by the Escobar inequality

$$\|\nabla u\|_{L^p(H)} \geq E_{n,p} \|u\|_{L^{p^\sharp}(\partial H)} \qquad \text{for all } u \in \dot{W}^{1,p}(H) \tag{2.1}$$

is saturated; see Figure 1.

- For $T \in (0, T_E)$, there is a unique $s_T \in \mathbb{R}$ such that
$$U_T = \frac{\tau_{s_T} U_S}{\|\tau_{s_T} U_S\|_{L^{p^\star}(H)}} \chi_{\overline{H}} \qquad \text{where} \qquad U_S(x) = \Big(1 + |x|^{p/(p-1)}\Big)^{(p-n)/p}. \tag{2.2}$$
The function U_S is the unique (modulo symmetries) extremal for the Sobolev inequality on $\mathbb{R}^n$. When $p = 2$, up to a constant multiple, the conformal metric $U_T^{4/(n-2)} g_{\text{euc}}$ on H is isometric to a geodesic ball on the round sphere, with radius tending to zero as $T \to T_E$ and to the diameter of the sphere as $T \to 0$.
- For $T = T_E$, i.e. the point corresponding to the Escobar trace inequality,
$$U_T = \frac{\tau_{s_T} U_E}{\|\tau_{s_T} U_E\|_{L^{p^\star}(H)}} \chi_{\overline{H}} \qquad \text{where} \qquad U_E(x) = |x|^{(p-n)/(p-1)} \tag{2.3}$$
with $s_T = -1$. The function U_{T_E} is the unique (modulo symmetries) extremal function for the Escobar trace inequality. Replacing $s_T = -1$ by any other $s < 0$ in (2.3) gives a dilation of U_T and hence another extremal. When $p = 2$, the conformal metric $U_{T_E}^{4/(n-2)} g_{\text{euc}}$ on H is isometric to a ball in $\mathbb{R}^n$.
- For $T > T_E$, there is a unique $s_T < -1$ such that
$$U_T = \frac{\tau_{s_T} U_{BE}}{\|\tau_{s_T} U_{BE}\|_{L^{p^\star}(H)}} \chi_{\overline{H}}, \qquad \text{where} \qquad U_{BE}(x) = \Big(|x|^{p/(p-1)} - 1\Big)^{(p-n)/p}. \tag{2.4}$$
When $p = 2$, up to a constant multiple, the conformal metric $U_T^{4/(n-2)} g_{\text{euc}}$ on H is isometric to a geodesic ball in hyperbolic space, with radius tending to zero as $T \to T_E$ and to infinity as $T \to \infty$.

No minimizers in (1.1) exist for $T = 0$, since the support of the extremals for the Sobolev inequality is all of $\mathbb{R}^n$.

In [MN17] we establish various properties of the function $T \mapsto \Phi_H(T)$. First, let $T_0 \in (0, T_E)$ be the $L^{p^\sharp}(\partial H)$ norm of (2.2) with $s_{T_0} = 0$. Then Φ_H uniquely achieves its global minimum $\Phi_H(T_0) = S_{n,p}/2^{1/n}$ at T_0, saturating the constant lower bound $\Phi_H(T) \geq S_{n,p}/2^{1/n}$ (see Figure 1) resulting from the inequality

$$\|\nabla u\|_{L^p(H)} \geq \frac{S_{n,p}}{2^{1/n}} \|u\|_{L^{p^*}(H)} \qquad \text{for all } u \in \dot{W}^{1,p}(H)\,. \tag{2.5}$$

The inequality (2.5) is a direct consequence of the Sobolev inequality on $\mathbb{R}^n$ and reflection.

We additionally show in [MN17] that $T \mapsto \Phi_H(T)$ is strictly decreasing on $(0, T_E)$, concave on $(0, T_*)$ for some $T_* \in (0, T_0)$, and strictly increasing and convex on (T_0, ∞). A simple divergence theorem computation shows that

$$\Phi_H(T) > \frac{T^{p^\sharp}}{p^\sharp} \qquad \text{for every } T > 0, \tag{2.6}$$

and direct analysis of (2.4) shows this lower bound is asymptotically saturated as $T \to \infty$ (see Figure 1).

In [MV05], Maggi & Villani show that for any open, connected Lipschitz domain $\Omega \subset \mathbb{R}^n$, $\Phi_\Omega(T) \geq \Phi_B(T)$ for $T \in (0, \mathrm{ISO}(B_1)^{1/p^\sharp})$, where $\mathrm{ISO}(B_1) = n|B_1|$ and $\Phi_\Omega(T)$ is the analogous minimization problem to (1.1) with Ω in place of H. The lower bound $\Phi_H(T) \geq \Phi_B(T)$ on this interval is not saturated (see Figure 1); the question of whether other domains can have $\Phi_\Omega(T) = \Phi_B(T)$ was investigated in [MNT23]. Note that any open Lipschitz domain Ω enjoys the complementary upper bound $\Phi_\Omega(T) \leq \Phi_H(T)$.

2.2. **Mass transportation background.** We briefly recall some facts from optimal transport theory, and refer the reader to [Vil03, Mag23] for further introduction. For a μ (Borel) probability measure on $\mathbb{R}^n$ and a Borel measurable map $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$, the pushforward of μ through T is the probability measure $\mathcal{T}_\# \mu$ defined by

$$\mathcal{T}_\#\mu(A) = \mu\left(\mathcal{T}^{-1}(A)\right) \qquad \text{for all } A \subset \mathbb{R}^n. \tag{2.7}$$

By approximation, this means that

$$\int_{\mathbb{R}^n} \xi \, d\mathcal{T}_\#\mu = \int_{\mathbb{R}^n} \xi \circ \mathcal{T} \, d\mu \tag{2.8}$$

for every Borel measurable function $\xi : \mathbb{R}^n \to [0, \infty]$.

Now suppose $\mu = F\mathcal{L}^n$ and $\nu = G\mathcal{L}^n$ are absolutely continuous probability measures on $\mathbb{R}^n$. By the Brenier-McCann theorem [Bre91, McC97] (see [Vil03, Cor. 2.30]), there is a convex function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ such that the map $\mathcal{T} = \nabla\varphi$ is defined μ-a.e. and satisfies

$$\mathcal{T}_\#\mu = \nu.$$

The map $\mathcal{T}$ is uniquely determined μ-a.e., and is called the Brenier map from μ to ν. It is the unique optimal transport map for the quadratic cost, though we will not use this. If φ is C^2, then (2.8) and the area formula give

$$F(x) = G(\nabla\varphi(x)) \det \nabla^2\varphi(x) \qquad \mu\text{-a.e.}; \tag{2.9}$$

the same identity holds in general for the Alexandrov Hessian $\nabla^2\varphi$, i.e. for the absolutely continuous part of the distributional Hessian [McC97].

A key property of the Brenier map is the cyclical monotonicity of its graph. A subset $\Gamma \subset \mathbb{R}^n \times \mathbb{R}^n$ is *cyclically monotone* if

$$\sum_{i=1}^m y_i \cdot (x_{i+1} - x_i) \leq 0$$

for every $m \in \mathbb{N}$ and every collection of m points $(x_1, y_1), \dots, (x_m, y_m)$ in Γ, with the convention that $x_{m+1} = x_1$. There is a set A of full μ measure (i.e. $\mu(A) = 1$) such that the graph $\Gamma = \{(x, \mathcal{T}(x)) : x \in A\}$ of the Brenier map is cyclically monotone, see, e.g., [Vil03, Prop. 2.24][1] Applying this fact for $m = 2$ guarantees that

$$(\mathcal{T}(x_1) - \mathcal{T}(x_2)) \cdot (x_1 - x_2) \geq 0 \qquad \text{for } (x_1, x_2) \in A \times A. \tag{2.10}$$

2.3. **Mass transportation argument.** Let us sketch the mass transportation proof of (1.3) given in [MN17]. It is a variant of the mass transportation arguments used to prove the Sobolev inequality [CENV04] and various other functional inequalities, and is especially inspired by Nazaret's proof of the Escobar inequality [Naz06]. Fix $T > 0$ and let $s = s_T \in \mathbb{R}$ be as in section 2.1. Direct computation verifies the identity

$$p^\sharp \|\nabla U_T\|_{L^p(H)} Y_T + sT^{p^\sharp} = n\int_H U_T^{p^\sharp}\, dx \qquad \text{where} \qquad Y_T = \left(\int_H U_T^{p^*}\, |x - se_1|^{p/(p-1)}\, dx\right)^{(p-1)/p}. \tag{2.11}$$

Now, fix $u \in \mathcal{A}_T \cap C^1_c(\overline{H})$ with $u \geq 0$. We aim to show that

$$\|\nabla u\|_{L^p(H)} \geq \|\nabla U_T\|_{L^p(H)}\,. \tag{2.12}$$

Letting $F = u^{p^*}$ and $G = U_T^{p^*}$, consider the measures

$$\mu = u^{p^*}\,\mathcal{L}^n = F\,\mathcal{L}^n, \qquad\qquad \nu = U_T^{p^*}\,\mathcal{L}^n = G\,\mathcal{L}^n\,,$$

and let $\mathcal{T} = \nabla\varphi$ be the Brenier map from μ to ν. Applying the transport condition (2.8), the identity (2.9), and the arithmetic-geometric mean inequality to the nonnegative eigenvalues of the Alexandrov Hessian $D^2\varphi$,[2] we find

$$\int_H U_T^{p^\sharp} = \int_{\mathbb{R}^n} G^{1-1/n} = \int_{\mathbb{R}^n} G(\nabla\varphi)^{-1/n} F = \int_{\mathbb{R}^n} \left(\det \nabla^2\varphi\right)^{1/n} F^{1-1/n} \leq \frac{1}{n}\int_{\mathbb{R}^n} F^{1-1/n} d(\operatorname{div}\mathcal{T})\,. \tag{2.13}$$

---

[1] More generally, the support of the optimal plan between any two probability measures on $\mathbb{R}^n$ is cyclically monotone, and this fact is used in one proof of the Brenier theorem.

[2] As with (2.9), there is some subtlety when φ is not C^2; in (2.13), AM–GM is applied μ-a.e. to the eigenvalues of the Alexandrov Hessian, and by convexity, the Alexandrov Laplacian of φ is bounded above by its distributional Laplacian.

On the right-hand side, we subtract the divergence-free vector field se_1 from $\mathcal{T}$, letting $\mathcal{S} = \mathcal{T} - se_1$. Then, applying the divergence theorem, we obtain

$$\int_{\mathbb{R}^n} F^{1-1/n} d(\operatorname{div} \mathcal{T}) = \int_H u^{p^\sharp} d(\operatorname{div} \mathcal{S}) = -p^\sharp \int_H u^{p^\sharp - 1} \nabla u \cdot \mathcal{S} dx - \int_{\partial H} u^{p^\sharp} \mathcal{S} \cdot e_1 d\mathcal{H}^{n-1}. \tag{2.14}$$

Since $\mathcal{T}$ transports μ to ν and $\operatorname{spt}(\nu) \subset \overline{H}$, we have $\mathcal{S}(x) \cdot (-e_1) \leq s$ for $\mathcal{H}^{n-1}$-a.e. $x \in \operatorname{spt}(\mu) \cap \partial H$. Thus, in summary, (2.13) and (2.14) yield

$$n \int_H U_T^{p^\sharp} \, dx \leq -p^\sharp \int_H u^{p^\sharp - 1} \nabla u \cdot (\mathcal{T} - se_1) \, dx + s \, T^{p^\sharp} \, . \tag{2.15}$$

Now we bound the first term on the right-hand side. Using Cauchy-Schwarz and Hölder's inequalities and the transport condition (2.8), we find

$$\begin{aligned} -p^\sharp \int_H u^{p^\sharp - 1} \nabla u \cdot (\mathcal{T} - se_1) \, dx &\leq p^\sharp \int_H u^{p^\sharp - 1} |\nabla u| \cdot |\mathcal{T} - se_1| \, dx \\ &\leq p^\sharp \|\nabla u\|_{L^p(H)} \left(\int_H u^{p^\star} \, |\mathcal{T}(x) - se_1|^{p/(p-1)} \, dx \right)^{(p-1)/p} \\ &= p^\sharp \|\nabla u\|_{L^p(H)} \left(\int_H U_T^{p^\star} \, |x - se_1|^{p/(p-1)} \, dx \right)^{(p-1)/p} = p^\sharp \|\nabla u\|_{L^p(H)} \, Y_T . \end{aligned} \tag{2.16}$$

Combining this with (2.11) and (2.15) shows that

$$p^\sharp \|\nabla U_T\|_{L^p(H)} Y_T + s T^{p^\sharp} \leq p^\sharp \|\nabla u\|_{L^p(H)} \, Y_T + s T^{p^\sharp} \tag{2.17}$$

which directly implies (2.12).

The inequality (2.17) sandwiches each individual inequality in the proof, including (2.16), so

$$\begin{aligned} \left(\|\nabla u\|_{L^p(H)} - \|\nabla U_T\|_{L^p(H)} \right) Y_T &\geq \int_H u^{p^\sharp - 1} \Big(|\nabla u| \cdot |\mathcal{T} - se_1| - (-\nabla u) \cdot (\mathcal{T} - se_1) \Big) \, dx \\ &= \frac{1}{2} \int_H u^{p^\sharp - 1} |\nabla u| \, |\mathcal{T} - se_1| \left| \frac{-\nabla u}{|\nabla u|} - \frac{(\mathcal{T} - se_1)}{|\mathcal{T} - se_1|} \right|^2 dx. \end{aligned}$$

Since $\|\nabla u\|_{L^p(H)} - \|\nabla U_T\|_{L^p(H)} \leq C \delta_T(u)$ with $C = 1/(p \Phi_H(T)^{p-1})$, this means there is a constant $C_{n,p,T} > 0$ such that

$$C_{n,p,T} \delta_T(u) \geq \int_H u^{p^\sharp - 1} |\nabla u| \, |\mathcal{T} - se_1| \left| \frac{-\nabla u}{|\nabla u|} - \frac{(\mathcal{T} - se_1)}{|\mathcal{T} - se_1|} \right|^2 dx. \tag{2.18}$$

In the proof of Theorem 1.3, we will only use (2.18) and the fact that U_T is radially symmetric and decreasing about a point se_1.

3. Proof of Theorem 1.3

In this section we prove Theorem 1.3. The idea is to construct a function w, which is essentially the sum of a copy of $m_1 U_{T_1}$ centered at $2^{R+1} e_n$ and a copy of $m_2 U_{T_2}$ centered at $-2^{R+1} e_n$ for $R \gg 1$, such that $\int |\nabla w|^p \leq m_1^p \Phi_H(T_1)^p + m_2^p \Phi_H(T_2)^p + \varepsilon$. We are able to estimate the corresponding Brenier map $\mathcal{T}$ in a somewhat explicit manner. In particular, the cyclical monotonicity of the graph of $\mathcal{T}$ in the form (2.10) shows that $\mathcal{T}$ maps most points in the support of the translated copy of $m_1 U_{T_1}$ to $\{y \cdot e_n > 0\}$, forcing a definite lower bound for the right-hand side of the estimate (2.18) and thus showing $\delta_T(w) \geq 2\varepsilon$.

Proof of Theorem 1.3. Step 1: We begin by fixing parameters and notation. Without loss of generality assume $m_1 \leq m_2$, and as above, let $T_1 = t_1/m_1$ and $T_2 = t_2/m_2$. From the explicit form of U_{T_1}, it is not difficult to see that there exists $\bar{c} = \bar{c}(n, p, T_1) > 0$ small enough such that the set

$$\mathcal{G} = \Big\{ x \in H \ : \ U_{T_1}(x) \geq \bar{c} \, \|U_{T_1}\|_{L^\infty(H)} \, , \quad |\nabla U_{T_1}(x)| \geq \bar{c}, \quad \nabla U_{T_1}(x) \cdot e_n \geq \bar{c} \Big\} \tag{3.1}$$

is nonempty. Note that $\mathcal{G} \subset B_\rho$ for some $\rho = \rho(n, T_1) > 0$. We define $\bar{a} = \bar{a}(n, T_1, m_1)$ by

$$3\bar{a} := m_1^{p^*} \int_{\mathcal{G}} U_{T_1}^{p^*} > 0. \tag{3.2}$$

Let $R = R(n, T, m_1, m_2, t_1, t_2) \geq 2\rho > 0$ be a large fixed number to be specified later in the proof. Let

$$\tilde{K} = \left\{ z \in \mathbb{R}^n : |z_n| < \frac{R}{2^R} \, |z'| \right\} . \tag{3.3}$$

Here z' denotes the projection of z onto $\mathbb{R}^{n-1}\times\{0\}\subset\mathbb{R}^n$. Since $|\tilde{K}\cap B_r| = o_R(1)$ for any fixed $r>0$, we can take R large enough so that

$$\bar{b} := \int_{\tilde{K}\cap H} U_T^{p^*}\, dx \leq \bar{a}\,.$$

Here and throughout, $o_R(1)$ is a number whose absolute value can be made arbitrarily small by taking R sufficiently large.

Step 2: Next we construct the main function $w = w_R \in \mathcal{A}_T$. Let $\eta : \mathbb{R}^n \to \mathbb{R}$ be a smooth nonnegative cutoff function supported in B_1 with $0 \leq \eta \leq 1$ on $\mathbb{R}^n$ and $\eta = 1$ in $B_{1/2}$. The function $W_1(x) = m_1 U_{T_1}(x)\eta(\frac{x}{R})\chi_{\overline{H}}(x)$ satisfies

$$\int_H W_1^{p^*}\, dx = m_1^{p^*} + o_R(1), \qquad \int_{\partial H} W_1^{p^\sharp}\, d\mathcal{H}^{n-1} = m_1^{p^\sharp}T_1^{p^\sharp} + o_R(1), \qquad \int_H |\nabla W_1|^p\, dx = m_1^p\Phi_H\left(T_1\right)^p + o_R(1)\,.$$

The analogous estimates hold for $W_2(x) = m_2 U_{T_2}(x)\eta(\frac{x}{R})\chi_{\overline{H}}(x)$. By construction, we have $\int_H W_i^{p^*}\, dx \leq m_i^{p^*}$ and $\int_{\partial H} W_i^{p^\sharp}\, d\mathcal{H}^{n-1} \leq (m_iT_i)^{p^\sharp} = t_i^{p^\sharp}$ for $i = 1, 2$. So, we may choose a nonnegative smooth function $\psi_R : H \to \mathbb{R}$ supported in B_R so that

$$w(x) = W_1\left(x - 2^{R+1}e_n\right) + \left(W_2\left(x + 2^{R+1}e_n\right) + \psi_R\left(x + 2^{R+1}e_n\right)\right)$$

lies in $\mathcal{A}_T$ and satisfies

$$\int_H |\nabla w|^p\, dx = m_1^p\Phi_H\left(T_1\right)^p + m_2^p\Phi_H\left(T_2\right)^p + o_R(1)\,. \tag{3.4}$$

The support of w is contained in $B_R^+\cup B_R^-$ where we let

$$B_R^+ = B_R\left(2^{R+1}e_n\right)\cap\overline{H}, \qquad B_R^- = B_R\left(-2^{R+1}e_n\right)\cap\overline{H}.$$

Recalling that $\psi_R \geq 0$ and $m_1 \leq m_2$, we have

$$\int_{B_R^-} w^{p^*}\, dx \geq \frac{1}{2}. \tag{3.5}$$

Step 3: Let $\mu = w^{p^*}\mathcal{L}^n$ and $\nu = U_T^{p^*}\mathcal{L}^n$ and let $\mathcal{T}$ be the Brenier map from μ to ν. Consider the sets

$$E = \left\{x\in B_R^+ : \mathcal{T}(x)\cdot e_n < 0\right\} = \mathcal{T}^{-1}\left(\{y_n<0\}\right)\cap B_R^+, \tag{3.6}$$

$$F = \left\{x\in B_R^- : \mathcal{T}(x)\cdot e_n \geq 0\right\} = \mathcal{T}^{-1}\left(\{y_n\geq 0\}\right)\cap B_R^- \tag{3.7}$$

of points in $B_R^\pm$ that get mapped across the plane $\{y_n = 0\}$ by $\mathcal{T}$. We claim that

$$\mu(E) \leq \bar{b}. \tag{3.8}$$

As above $\bar{b} = \int_{\tilde{K}\cap H} U_T^{p^*}\, dx = \nu(\tilde{K})$. Suppose not, i.e. $\mu(E) > \bar{b}$. From the symmetry of U_T and the transport condition (2.7), we have

$$\begin{aligned}\frac{1}{2} = \nu\left(\{y_n<0\}\right) &= \mu\left(\mathcal{T}^{-1}\left(\{y_n<0\}\right)\right)\\ &= \mu\left(\mathcal{T}^{-1}(\{y_n<0\})\cap B_R^-\right) + \mu(E) > \mu\left(\mathcal{T}^{-1}\left(\{y_n<0\}\right)\cap B_R^-\right) + \bar{b}\,.\end{aligned}$$

That is, $\mu(\mathcal{T}^{-1}(\{y_n<0\})\cap B_R^-) < \frac{1}{2} - \bar{b}$. Since $\mu\left(B_R^-\right) \geq \frac{1}{2}$ by (3.5), this means the set F defined in (3.7) has $\mu(F) > \bar{b}$ as well. Now, with $\tilde{K}$ as in (3.3), let

$$F_* = \{x\in B_R^- : \mathcal{T}(x)\in\{y\cdot e_n > 0\}\setminus\tilde{K}\}\subset F\,,$$

$$E_* = \{x\in B_R^+ : \mathcal{T}(x)\in\{y\cdot e_n < 0\}\setminus\tilde{K}\}\subset E$$

be the sets of points in $B_R^\pm$ mapping even further to the "wrong side" of the plane $\{y_n = 0\}$. Since $F\setminus F_*\subset\mathcal{T}^{-1}(\{y_n\geq 0\}\cap\tilde{K})$, we have

$$\mu(F\setminus F_*) \leq \mu\big(\mathcal{T}^{-1}(\{y_n\geq 0\}\cap\tilde{K})\big) = \nu(\{y_n\geq 0\}\cap\tilde{K}) = \frac{\bar{b}}{2}\,.$$

The final identity comes from the reflection symmetry of $\tilde{K}$ and U_T across $\{x_n = 0\}$. Since $\mu(F) = \mu\left(F_*\right) + \mu\left(F\setminus F_*\right)$, we find $\mu\left(F_*\right) \geq \frac{\bar{b}}{2}$, and analogous reasoning shows $\mu\left(E_*\right) > \frac{\bar{b}}{2}$. In particular, letting A be the full μ-measure set on which (2.10) holds, the sets $E_*\cap A$ and $F_*\cap A$ are nonempty.

Now, take any $x_1 \in E_* \cap A \subset B_R^+$ and $x_2 \in F_* \cap A \subset B_R^-$. Then $x_1 - x_2$ lies in the positive cone

$$K = \left\{ z \in \mathbb{R}^n : z_n > \frac{2^R}{R} |z'| \right\}$$

Observe that $z \cdot y < 0$ for any $z \in K$ and $y \in \{y_n < 0\} \setminus \tilde{K}$. So, by cyclical monotonicity (2.10),

$$\mathcal{T}(x_1) - \mathcal{T}(x_2) \notin \{y_n < 0\} \setminus \tilde{K}\,. \tag{3.9}$$

On the other hand, from the definitions of E_* we have $\mathcal{T}(x_1) \in \{y_n < 0\} \setminus \tilde{K}$. Similarly, from the definition of F_*, we have $\mathcal{T}(x_2) \in \{y_n > 0\} \setminus \tilde{K}$ and so $-\mathcal{T}(x_2) \in \{y_n < 0\} \setminus \tilde{K}$. Since $\{y_n < 0\} \setminus \tilde{K}$ is a convex cone,

$$\mathcal{T}(x_1) - \mathcal{T}(x_2) \in \{y_n < 0\} \setminus \tilde{K}.$$

This contradicts (3.9). Thus (3.8) holds.

Step 4: Since $\mathcal{T}$ is a transport map and U_T is bounded, for any $\varepsilon > 0$,

$$\mu\left(\{x : |\mathcal{T}(x) - se_1| < \varepsilon\}\right) = \nu(B(se_1, \varepsilon)) \leq C\varepsilon^n \tag{3.10}$$

where $C = C(n, p, T)$. Choose $\varepsilon > 0$ small enough so $C\varepsilon^n \leq \bar{a}$ with $\bar{a}$ as in (3.2). (In the case $s < 0$ we can choose ε so that $\nu(B(se_1, \varepsilon)) = 0$.) Note that since $R \geq 2\rho$,

$$\mu(\mathcal{G} + 2^{R+1}e_n) = \int_{\mathcal{G}+2^{R+1}e_n} w^{p^*}\,dx = \int_{\mathcal{G}} (m_1 U_{T_1})^{p^*} = 3\bar{a}.$$

So, thanks to (3.10), (3.8), and $\bar{b} \leq \bar{a}$, we have $\mu(\mathcal{G}_*) \geq \bar{a}$ where

$$\mathcal{G}_* = \left(\mathcal{G} + 2^{R+1}e_n\right) \setminus \left(\{x : |\mathcal{T}(x) - se_1| < \varepsilon\} \cup \{x : \mathcal{T}(x) \cdot e_n < 0\}\right).$$

From the definition of $\mathcal{G}$ in (3.1), we thus have $|\frac{-\nabla w}{|\nabla w|} - \frac{\mathcal{T} - se_1}{|\mathcal{T} - se_1|}|^2 \geq \bar{c}^2$ on $\mathcal{G}_*$. So, recalling (2.18) and noting that $p^\sharp - 1 = p^* - \frac{n}{n-p}$, we have

$$\begin{aligned}
C_{n,p,T}\delta_T(w) &\geq \int_{\mathcal{G}_*} w^{p^\sharp - 1} |\nabla w| \cdot |\mathcal{T} - se_1| \left| \frac{-\nabla w}{|\nabla w|} - \frac{\mathcal{T} - se_1}{|\mathcal{T} - se_1|} \right|^2 dx \\
&= \int_{\mathcal{G}_*} w^{-n/(n-p)} |\nabla w| \cdot |\mathcal{T} - se_1| \left| \frac{-\nabla w}{|\nabla w|} - \frac{\mathcal{T} - se_1}{|\mathcal{T} - se_1|} \right|^2 d\mu \\
&\geq \|U_{T_1}\|_{L^\infty(H)}^{-n/(n-p)} \cdot \bar{c} \cdot \varepsilon \int_{\mathcal{G}_*} \left| \frac{-\nabla w}{|\nabla w|} - \frac{\mathcal{T} - se_1}{|\mathcal{T} - se_1|} \right|^2 d\mu \geq \bar{a}\bar{c}^3 \|U_{T_1}\|_{L^\infty(H)}^{-n/(n-p)} \varepsilon =: 2c_0.
\end{aligned} \tag{3.11}$$

Up to possibly further increasing R depending on $n, p, T_1, \bar{a}, \bar{c}$ and ε, and thus on $n, p, T, m_1, m_2, t_1, t_2$, in (3.4) we may take the error $o_R(1)$ to be at most c_0, so absorbing it yields

$$m_1^p \Phi_H(T_1)^p + m_2^p \Phi_H(T_2)^p - \Phi_H(T)^p \geq c_0.$$

This completes the proof. □

4. Proofs of Theorem 1.1 and Corollary 1.2

The strict binding inequality Theorem 1.3 is the main tool toward proving Theorem 1.1. To complete the proof of Theorem 1.1, let us recall the first and second concentration-compactness lemmas in the present setting. **Just make the points in $\mathbb{R}^n$**

Lemma 4.1 (Concentration-Compactness Lemma I)**.** *Let ν_k be a sequence of probability measures on $\mathbb{R}^n$. There is a subsequence (unrelabeled) such that one of the following three conditions holds:*

1. *(Compactness) There is a sequence $x_k \in \mathbb{R}^n$ such that for any $\varepsilon > 0$, there is a radius $R > 0$ such that $\nu_k(B_R(x_k)) \geq 1 - \varepsilon$.*
2. *(Vanishing) For all $R > 0$, $\lim_{k\to\infty}(\sup\{\nu_k(B_R(x)) : x \in \mathbb{R}^n\}) = 0$.*
3. *(Dichotomy) There is a number $\lambda \in (0, 1)$ such that for all $\varepsilon > 0$, there exist $R > 0$ and $\{x_k\} \subset \mathbb{R}^n$ such that*

   $$\limsup_{k\to\infty} \left(|\lambda - \nu_k^1(\mathbb{R}^n)| + |(1 - \lambda) - \nu_k^2(\mathbb{R}^n)| \right) \leq \varepsilon$$

   for the measures

   $$\nu_k^1 = \nu_k \llcorner B_R(x_k) \qquad \text{and} \qquad \nu_k^2 = \nu_k \llcorner \mathbb{R}^n \setminus B_{8R}(x_k)\,.$$

Compared to the usual statement (see [Lio84, Lemma I.1] or [Str08, Lemma 4.3]) the statement of Lemma 4.1, we spell out the splitting measures ν_k^1, ν_k^2 in the dichotomy alternative more explicitly. The statement given above is already present in the proof of the standard statement.

Lemma 4.2 (Concentration-Compactness Lemma II)**.** *Let* $n \geq 2$ *and* $p \in (1,n)$*. If* $\{u_k\}_k$ *is a sequence in* $L^1_{loc}(H)$, $\{\nabla u_k\}_k$ *is bounded in* $L^p(H;\mathbb{R}^n)$ *and* $u_k \rightharpoonup u$ *as distributions in* H*, then the Radon measures on* $\overline{H}$ *defined by*

$$\mu_k = |\nabla u_k|^p \, \mathcal{L}^n \llcorner H, \quad \nu_k = |u_k|^{p^\star} \mathcal{L}^n \llcorner H, \quad \tau_k = |u_k|^{p^\sharp} \mathcal{H}^{n-1} \llcorner \partial H$$

have subsequential weak-star limits μ, ν *and* τ *which satisfy*

$$\begin{aligned}
\nu &= |u|^{p^\star} \mathcal{L}^n \llcorner H + \sum_{i\in I} m_i^{p^\star} \delta_{x_i}, \\
\tau &= |u|^{p^\sharp} \mathcal{H}^{n-1} \llcorner \partial H + \sum_{i\in I} t_i^{p^\sharp} \delta_{x_i}, \\
\mu &\geq |\nabla u|^p \mathcal{L}^n \llcorner H + \sum_{i\in I} \mathrm{g}_i^p \delta_{x_i},
\end{aligned}$$

where $\{x_i\}_{i\in I} \subset \bar{H}$ *is an at most countable set,* $m_i > 0$ *and* $t_i \geq 0$ *for every* $i \in I$*, with* $t_i > 0$ *only if* $x_i \in \partial H$*, and*

$$\mathrm{g}_i \geq m_i \Phi_H\left(\frac{t_i}{m_i}\right), \quad \forall i \in I\,.$$

In particular, $\mathrm{g}_i \geq S_{n,p} m_i$ *whenever* $x_i \in H$*.*

This form of the second concentration-compactness lemma, accounting for the trace term, was shown in [MNT23, Lemma 2.1]. Its proof is a basic adaptation of the classical version on $\mathbb{R}^n$, see [Lio85, Lemma I.1] or [Str08, Lemma 4.8]. In [MNT23], the lemma is stated on an open bounded domain Ω with C^1 boundary rather than on H, but the boundedness of the domain is not used in the proof; the only modification is to replace inequalities (A.2) and (A.3) there by the inequalities (2.5) and (2.1) respectively.

Finally, in the proof of Theorem 1.1, we will also use the following simple lemma.

Lemma 4.3. *Fix* $n \geq 2$ *and* $p \in (1,n)$*. Let* $\varepsilon > 0$ *and* $R > 0$*, and suppose* $u \in \dot{W}^{1,p}(H)$ *has* $\|u\|_{L^{p^*}(H)} = 1$ *and* $\int_{H\cap B_{8R}\setminus B_R} |u|^{p^*} \leq \varepsilon$*. Then*

$$\int_{\partial H \cap (B_{7R}\setminus B_{2R})} |u|^{p^\sharp} \, d\mathcal{H}^{n-1} \leq C(\varepsilon^{1/p^*} + \|\nabla u\|_{L^p(H)})\varepsilon^{(p^\sharp - 1)/p^*}$$

for $C > 0$ *independent of* R*. In particular, if* $\|\nabla u\|_{L^p(H)} \leq \tilde{C}$*, the right-hand side tends to zero as* $\varepsilon \to 0$*.*

Proof. Choose a cutoff function $\psi : \mathbb{R}^n \to \mathbb{R}$ with $0 \leq \psi \leq 1$, $\psi = 1$ on $B_7 \setminus B_2$, $\psi = 0$ on $B_1 \cup (\mathbb{R}^n \setminus B_8)$, and let $\psi_R(x) = \psi(x/R)$. Then the function $v = \psi_R u$ has $\int_H |v|^{p^*} \, dx \leq \varepsilon$ and

$$\|\nabla v\|_{L^p(H)} \leq \|\nabla u\|_{L^p(H)} + \|u \nabla \psi_R\|_{L^p(H)} \leq \|\nabla u\|_{L^p(H)} + \|u\|_{L^{p^*}(B_{8R}\setminus B_R)} \|\nabla \psi_R\|_{L^n(H)}\,.$$

By scaling, $\|\nabla \psi_R\|_{L^n(H)} = \|\nabla \psi\|_{L^n(H)}$, so $\|\nabla v\|_{L^p(H)} \leq \|\nabla u\|_{L^p(H)} + C\varepsilon^{1/p^*}$ for a constant C independent of R. Applying (2.6) shows that

$$\|\nabla u\|_{L^p(H)} + C\varepsilon^{1/p^*} \geq \|v\|_{L^{p^*}(H)} \Phi_H\left(\frac{\|v\|_{L^{p^\sharp}(\partial H)}}{\|v\|_{L^{p^*}(H)}}\right) \geq \frac{\|v\|^{p^\sharp}_{L^{p^\sharp}(\partial H)}}{p^\sharp \|v\|^{p^\sharp - 1}_{L^{p^*}(H)}} \geq \frac{\|v\|^{p^\sharp}_{L^{p^\sharp}(\partial H)}}{p^\sharp \varepsilon^{(p^\sharp - 1)/p^*}}\,,$$

which, after rearranging, completes the proof. ☐

With these lemmas in hand, we can now prove Theorem 1.1.

Proof of Theorem 1.1. Let $\{\hat{u}_k\} \subset \mathcal{A}_T$ be a sequence with $\delta_T(\hat{u}_k) \to 0$. For each $k \in \mathbb{N}$, choose $y_k \in \mathbb{R}^n$ and $R_k > 0$ so

$$\sup_{x\in\mathbb{R}^n} \int_{H\cap B_{R_k}(x)} |\hat{u}_k|^{p^*} \, dx = \int_{H\cap B_{R_k}(y_k)} |\hat{u}_k|^{p^*} \, dx = \frac{1}{2}\,. \tag{4.1}$$

It is apparent from set containment that we may choose $y_k \in \overline{H}$. Writing $y_k = (y_k^1, y_k')$, we set $\tilde{u}_k(\cdot) = \hat{u}_k(\cdot - (0, y_k'))$, so that $\tilde{u}_k \in \mathcal{A}_T$ has $\delta_T(\tilde{u}_k) \to 0$ and satisfies (4.1) with $y_k^1 e_1$ in place of y_k. Next, consider the rescaled function $u_k \in \mathcal{A}_T$ defined by $u_k(x) = R_k^{(p-n)/p}\tilde{u}_k(\frac{x}{R_k})$, which has

$$\sup_{x\in\mathbb{R}^n} \int_{H\cap B_1(x)} |u_k|^{p^*}\,dx = \int_{H\cap B_1(d_k e_1)} |u_k|^{p^*}\,dx = \frac{1}{2}\,, \tag{4.2}$$

where $d_k = y_k^1/R_k \in [0,\infty)$. Once again, $\delta_T(u_k) = \delta_T(\hat{u}_k) \to 0$ by scaling.

Apply Lemma 4.1 to the sequence of probability measures $\nu_k = |u_k|^{p^*}\,\mathcal{L}^n \llcorner H$. We will rule out both the vanishing alternative and the dichotomy alternative using Theorem 1.3 as follows.

Vanishing does not occur. We claim that $A := \limsup_{k\to\infty} d_k < \infty$, which rules out the vanishing alternative, since in this case $\nu_k(B_{A+2}(0)) \geq 1/2$ for all k large. Suppose by way of contradiction that, up to an un-relabeled subsequence, $d_k \to \infty$. We will show this forces splitting of the sequence.

Let $J_k = \lfloor \log_2(d_k - 1)\rfloor - 1 \in \mathbb{N}$, so that $2^{J_k+1} \leq d_k - 1 \leq 2^{J_k+2}$. For at least one $1 \leq j \leq J_k$, we have

$$\int_{B_{2^{j+1}}(d_k e_1)\setminus B_{2^j}(d_k e_1)} |u_k|^{p^*}\,dx \leq \frac{2}{\log_2(d_k-1)}, \tag{4.3}$$

as otherwise $\int_H |u_k|^{p^*}\,dx \geq 2J_k/\log_2(d_k - 1) > 1$ for k large enough. Let j_k^* be the first j for which (4.3) holds and let $R_k^* = 2^{j_k^*}$.

Fix a smooth nonnegative cutoff function $\psi : \mathbb{R}^n \to [0,1]$ with $\psi = 1$ in B_1, $\psi = 0$ in $\mathbb{R}^n\setminus B_2$. Let $\psi_k(\cdot) = \psi(\cdot/R_k^*)$. Let $(m_k^*)^{p^*} := \int_{B_{R_k^*}(d_k e_1)} |u_k|^{p^*}\,dx$

Dichotomy does not occur. Suppose by way of contradiction that the dichotomy alternative holds with splitting proportion $\lambda \in (0,1)$. Take a sequence $\varepsilon_k \to 0$. By Lemma 4.1 and a diagonal argument, up to a subsequence, we may find $R_k \to \infty$ and $\{x_k\}$ such that the measures $\nu_k^1 = \nu_k \llcorner B_{R_k}(x_k)$ and $\nu_k^2 = \nu_k \llcorner (\overline{H} \setminus B_{8R_k}(x_k))$ satisfy

$$\limsup_{k\to\infty} \left\{ |\nu_k^1(\overline{H}) - \lambda| + |\nu_k^2(\overline{H}) - (1-\lambda)| \right\} = 0\,. \tag{4.4}$$

Now, take a smooth cutoff function φ with $\varphi \in [0,1]$, $\varphi = 1$ on B_2, and $\varphi = 0$ on $\mathbb{R}^n \setminus B_3$. Similarly let η be such that $\eta = 0$ in B_6 and $\eta = 1$ on $\mathbb{R}^n \setminus B_7$. Let $\varphi_k(x) = \varphi(\frac{x-x_k}{R_k})$ and $\eta_k(x) = \eta(\frac{x-x_k}{R_k})$. Then setting $m_1^{p^*} = \lambda$ and $m_2^{p^*} = 1 - \lambda$, from (4.4) we have

$$\int_H |u_k\varphi_k|^{p^*}\,dx = m_1^{p^*} + o_k(1), \qquad \int_H |u_k\eta_k|^{p^*}\,dx = m_2^{p^*} + o_k(1)$$

By (4.4) and Lemma 4.3, we also have $\int_{\partial H\cap B_{7R_k}(x_k)\setminus B_{2R_k}(x_k)} |u_k|^{p^\sharp}\,d\mathcal{H}^{n-1} \to 0$, so in particular,

$$T^{p^\sharp} = \int_{\partial H} |u_k\varphi_k|^{p^\sharp}\,d\mathcal{H}^{n-1} + \int_{\partial H} |u_k\eta_k|^{p^\sharp}\,d\mathcal{H}^{n-1} + o_k(1)\,.$$

Up to passing to a further subsequence, there exist $t_1, t_2 \geq 0$ with $t_1^{p^\sharp} + t_2^{p^\sharp} = T^{p^\sharp}$ such that

$$\int_{\partial H} |u_k\varphi_k|^{p^\sharp}\,d\mathcal{H}^{n-1} = t_1^{p^\sharp} + o_k(1), \qquad \int_{\partial H} |u_k\eta_k|^{p^\sharp}\,d\mathcal{H}^{n-1} = t_2^{p^\sharp} + o_k(1)\,.$$

Similarly, using $1 \geq \varphi^p + \eta^p$ and $\nu_k(B_{8R_k}(x_k) \setminus B_{R_k}(x_k)) \to 0$, and applying Hölder's inequality and scaling as in the proof of Lemma 4.3, we obtain

$$\int_H |\nabla u_k|^p\,dx \geq \int_H |\nabla(u_k\varphi_k)|^p\,dx + \int_H |\nabla(u_k\eta_k)|^p\,dx + o_k(1)\,.$$

Therefore, applying (1.3) to $u_k\varphi_k$ and $u_k\eta_k$ separately and using the continuity of the mapping $T \mapsto \Phi_H(T)$, we have

$$\int_H |\nabla(u_k\varphi_k)|^p\,dx + \int_H |\nabla(u_k\eta_k)|^p\,dx \geq m_1^p\Phi_H\Big(\frac{t_1}{m_1}\Big)^p + m_2^p\Phi_H\Big(\frac{t_2}{m_2}\Big)^p + o_k(1).$$

Since on the other hand, the left-hand side is bounded above by $\int_H |\nabla u_k|^p\,dx = \Phi_H(T)^p + o_k(1)$, we reach a contradiction to Theorem 1.3 for sufficiently large k. Thus the dichotomy alternative cannot occur.

So, the concentration alternative occurs in Lemma 4.1. In particular, we can find $\{x_k\} \subset \overline{H}$ and $R_0 > 0$ such that $\int_{B_{R_0}(x_k)} |u_k|^{p^*} > \frac{1}{2}$. Thanks to (4.2), this means $B_1(0) \cap B_{R_0}(x_k)$ is nonempty, i.e. $|x_k| < R_0 + 1$. So, the concentration alternative guarantees that for any $\varepsilon > 0$, there exists $R > 0$ such that

$$\int_{B_R(0)} d\nu_k \geq 1 - \varepsilon.$$

So, up to a subsequence, $\nu_k \overset{*}{\rightharpoonup} \nu$ for a measure ν on $\overline{H}$ with $\nu(\overline{H}) = 1$. Applying the argument of Lemma 4.3 but now taking $\psi : \mathbb{R}^n \to \mathbb{R}$ to be a cutoff with $\psi = 0$ in B_1 and $\psi = 1$ on $\mathbb{R}^n \setminus B_2$, we also see that the measures $\tau_k = |u_k|^{p^\sharp} \mathcal{H}^{n-1} \llcorner \partial H$ have $\tau_k \overset{*}{\rightharpoonup} \tau$ for a measure τ on ∂H with $\tau(\partial H) = T^{p^\sharp}$.

Now, let $\mu_k = |\nabla u_k|^p \mathcal{L}^n \llcorner H$. Up to a further subsequence, $\mu_k \overset{*}{\rightharpoonup} \mu$ for measure μ on $\overline{H}$. We apply Lemma 4.2. Up to a subsequence, $u_k \rightharpoonup u$ in $\dot{W}^{1,p}(H)$, $L^{p^*}(H)$, and $L^{p^\sharp}(\partial H)$. Then, since $\delta_T(u_k) \to 0$,

$$\begin{aligned} \Phi_H(T)^p = \lim_{k\to\infty} \mu_k(\overline{H}) \geq \mu(\overline{H}) &\geq \|u\|^p_{L^{p^*}(H)} \Phi_H\Big(\frac{\|u\|_{L^{p^\sharp}(\partial H)}}{\|u\|_{L^{p^*}(H)}}\Big) + \sum_{i\in I} m_i^p \Phi_H\Big(\frac{t_i}{m_i}\Big)^p \qquad (4.5)\\ &\geq \|u\|^p_{L^{p^*}(H)} \Phi_H\Big(\frac{\|u\|_{L^{p^\sharp}(\partial H)}}{\|u\|_{L^{p^*}(H)}}\Big) + \Big(\sum_{i\in I} m_i^{p^*}\Big)^{p/p^*} \Phi_H\Big(\frac{(\sum_{i\in I} t_i^{p^\sharp})^{1/p^\sharp}}{(\sum_{i\in I} m_i^{p^*})^{1/p^*}}\Big)^p. \end{aligned}$$

In the last inequality, we use that Theorem 1.3 passes to countable sums at the expense of losing the strict inequality; the proof is a basic analysis exercise. Finally, we apply Theorem 1.3 to bound the right-hand side below by $\Phi_H(T)^p$. The inequality must be strict (a contradiction) unless either $\|u\|_{L^{p^*}(H)}$ or $\sum_{i\in I} m_i^{p^*}$ is zero. If $\|u\|_{L^{p^*}(H)} = 0$, then $\sum_{i\in I} m_i^{p^*} = 1$. Since the normalization (4.2) guarantees that $m_i^{p^*} \leq 1/2$ for each $i \in I$, then instead peeling off the first term in the sum on the right-hand side of (4.5) and again applying Theorem 1.3 yields a contradiction. So, $m_i = 0$ for each $i \in I$ and $\nu = |u|^{p^*} \mathcal{L}^n$. From Lemma 4.2, this means the index set I is empty and thus $\tau = |u|^{p^\sharp} \mathcal{H}^{n-1} \llcorner \partial H$.

This means $\|u_k\|_{L^{p^*}(H)} \to \|u\|_{L^{p^*}(H)}$ and $\|u_k\|_{L^{p^\sharp}(\partial H)} \to \|u\|_{L^{p^\sharp}(\partial H)}$, and so the weak convergence upgrades to strong convergence: $u_k \to u$ in $L^{p^*}(H)$ and $L^{p^\sharp}(\partial H)$. In particular, $u \in \mathcal{A}_T$ and $\int_H |\nabla u|^p \geq \Phi_H(T)^p$. Combining this with lower semicontinuity and the fact that $\int_H |\nabla u_k|^p \to \Phi_H(T)^p$, we have $\|\nabla u_k\|_{L^p(H)} \to \|\nabla u\|_{L^p(H)} = \Phi_H(T)$. This means (a) $u \in \mathcal{M}_T$, and (b) again the weak convergence upgrades to $\nabla u_k \to \nabla u$ in $L^p(H)$. Scaling back, this means that for the original sequence, $d_T(\hat{u}_k) \to 0$. We have passed to various subsequences, but since each subsequence has a further subsequence to which the argument applies, we obtain the conclusion for the entire sequence. □

Finally we show how Theorem 1.1 combined with the local stability result of [FLZ26] implies Corollary 1.2.

Proof of Corollary 1.2. Observe that for any $u \in \mathcal{A}_T$,

$$d_T(u)^2 \leq 2\|\nabla u\|^2_{L^2(H)} + 2\|\nabla U_T\|^2_{L^2(H)} \leq 4\|\nabla u\|^2_{L^2(H)}.$$

So, for any fixed number $\delta_0 \in (0,1)$, the desired inequality holds with constant $\alpha'_T = \frac{\delta_0}{4}$ among those functions $u \in \mathcal{A}_T$ with $\delta_T(u) \geq \delta_0 \|\nabla u\|^2_{L^2(H)}$. It thus suffices to prove the inequality when $\delta_T(u) \leq \delta_0 \|\nabla u\|^2_{L^2(H)}$, or equivalently after rearranging terms, when $\delta_T(u) \leq \frac{\delta_0}{1-\delta_0} \Phi_H(T)^2$ for a small enough fixed δ_0 of our choosing.

Let $\varepsilon > 0$ be chosen so that, in the main result of [FLZ26], we have $\delta_T(u) \geq \frac{\alpha_T}{2} d_T(u)^2$ provided $d_T(u)^2 \leq \varepsilon$. By Theorem 1.1, there exists $\delta_0 > 0$ such that if $\delta_T(u) \leq \frac{\delta_0}{1-\delta_0} \Phi_H(T)^2$, then $d_T(u)^2 \leq \varepsilon$. Thus the desired stability inequality holds with $\alpha'_T = \frac{\alpha_T}{2}$ for $u \in \mathcal{A}_T$ with $\delta_T(u) \leq \delta_0 \|\nabla u\|^2_{L^2(H)}$, and therefore for all $u \in \mathcal{A}_T$ with $\alpha'_T = \min\{\frac{\alpha_T}{2}, \frac{\delta_0}{4}\}$. □

REFERENCES

[Aub76a] T. Aubin. Équations différentielles non linéaires et problème de Yamabe concernant la courbure scalaire. *J. Math. Pures Appl. (9)*, 55(3):269–296, 1976.

[Aub76b] T. Aubin. Problèmes isopérimétriques et espaces de Sobolev. *J. Differential Geom.*, 11(4):573–598, 1976.

[BE91] G. Bianchi and H. Egnell. A note on the Sobolev inequality. *J. Funct. Anal.*, 100(1):18–24, 1991.

[Bec93] W. Beckner. Sharp Sobolev inequalities on the sphere and the Moser-Trudinger inequality. *Ann. of Math. (2)*, 138(1):213–242, 1993.

[Bre91] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[CENV04] D. Cordero-Erausquin, B. Nazaret, and C. Villani. A mass-transportation approach to sharp Sobolev and Gagliardo-Nirenberg inequalities. *Adv. Math.*, 182(2):307–332, 2004.

[CFMP09] A. Cianchi, N. Fusco, F. Maggi, and A. Pratelli. The sharp Sobolev inequality in quantitative form. *J. Eur. Math. Soc.*, 11(5):1105–1139, 2009.

[CL90] Eric A. Carlen and Michael Loss. Extremals of functionals with competing symmetries. *J. Funct. Anal.*, 88(2):437–456, 1990.

[CL94] E. A. Carlen and M. Loss. On the minimization of symmetric functionals. *Rev. Math. Phys.*, 6(5A):1011–1032, 1994. Special issue dedicated to Elliott H. Lieb.

[DEF$^+$25] J. Dolbeault, M. J. Esteban, A. Figalli, R. L. Frank, and M. Loss. Sharp stability for Sobolev and log-Sobolev inequalities, with optimal dimensional dependence. *Camb. J. Math.*, 13(2):359–430, 2025.

[Esc88] J. F. Escobar. Sharp constant in a Sobolev trace inequality. *Indiana Univ. Math. J.*, 37(3):687–698, 1988.

[Esc92a] J. F. Escobar. Conformal deformation of a Riemannian metric to a scalar flat metric with constant mean curvature on the boundary. *Ann. of Math. (2)*, 136(1):1–50, 1992.

[Esc92b] J. F. Escobar. The Yamabe problem on manifolds with boundary. *J. Differential Geom.*, 35(1):21–84, 1992.

[FLZ26] S. Fan, G.-D. Li, and J.J. Zhang. A note on the Sobolev–Escobar bridge inequality. *arXiv:2604.12677*, 2026.

[FMP10] A. Figalli, F. Maggi, and A. Pratelli. A mass transportation approach to quantitative isoperimetric inequalities. *Invent. Math.*, 182(1):167–211, 2010.

[FMP13] A. Figalli, F. Maggi, and A. Pratelli. Sharp stability theorems for the anisotropic Sobolev and log-Sobolev inequalities on functions of bounded variation. *Adv. Math.*, 242:80–101, 2013.

[FN19] A. Figalli and R. Neumayer. Gradient stability for the Sobolev inequality: the case $p \geq 2$. *J. Eur. Math. Soc. (JEMS)*, 21(2):319–354, 2019.

[FZ22] A. Figalli and Y. R.-Y. Zhang. Sharp gradient stability for the Sobolev inequality. *Duke Math. J.*, 171(12):2407–2459, 2022.

[Ho22] P. T. Ho. A note on the Sobolev trace inequality. *Proc. Amer. Math. Soc.*, 150(3):1257–1267, 2022.

[Lio84] P.-L. Lions. The concentration-compactness principle in the calculus of variations. The locally compact case. I. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 1(2):109–145, 1984.

[Lio85] P.-L. Lions. The concentration-compactness principle in the calculus of variations. The limit case. I. *Rev. Mat. Iberoamericana*, 1(1):145–201, 1985.

[Mag23] F. Maggi. *Optimal mass transport on Euclidean spaces*, volume 207 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2023.

[McC97] R.J. McCann. A convexity principle for interacting gases. *Adv. Math.*, 128(1):153–179, 1997.

[MN17] F. Maggi and R. Neumayer. A bridge between Sobolev and Escobar inequalities and beyond. *J. Funct. Anal.*, 273(6):2070–2106, 2017.

[MNT23] F. Maggi, R. Neumayer, and I. Tomasetti. Rigidity theorems for best Sobolev inequalities. *Adv. Math.*, 434:Paper No. 109330, 43, 2023.

[MV05] F. Maggi and C. Villani. Balls have the worst best Sobolev inequalities. *J. Geom. Anal.*, 15(1):83–121, 2005.

[Naz06] B. Nazaret. Best constant in Sobolev trace inequalities on the half-space. *Nonlinear Anal.*, 65(10):1977–1985, 2006.

[Neu20] R. Neumayer. A note on strong-form stability for the Sobolev inequality. *Calc. Var. Partial Differential Equations*, 59(1):Paper No. 25, 8, 2020.

[Sch84] R. Schoen. Conformal deformation of a Riemannian metric to constant scalar curvature. *J. Differential Geom.*, 20(2):479–495, 1984.

[Str08] M. Struwe. *Variational methods*, volume 34 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, fourth edition, 2008. Applications to nonlinear partial differential equations and Hamiltonian systems.

[Tal76] G. Talenti. Best constant in Sobolev inequality. *Ann. Mat. Pura Appl. (4)*, 110:353–372, 1976.

[Tru68] N. S. Trudinger. Remarks concerning the conformal deformation of Riemannian structures on compact manifolds. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)*, 22:265–274, 1968.

[Vil03] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[Yam60] H. Yamabe. On a deformation of Riemannian structures on compact manifolds. *Osaka Math. J.*, 12:21–37, 1960.

Department of Mathematical Scienes, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States of America

Email address: `neumayer@cmu.edu`