arXiv:2305.02410v1 [math.OC] 3 May 2023

非平衡最优传输问题的熵正则化

M. BUZE 和 M. H. DUONG

Abstract. 我们构建了非平衡最优传输问题熵正则化的数学理论。我们聚焦于静态表述，并借鉴针对未正则化情形所发展的形式体系，证明非平衡最优传输问题可以通过两种本质上不同的方式进行正则化——要么 *on the original space* ，要么 *on the extended space*。我们推导了这两种正则化问题的若干种等价表述，尤其引入了 *regularised induced marginal perspective cost function* 的概念，从而得以推导出 *extended space formulation of the original space regularisation*。我们还证明，在扩展空间正则化的情形下，所得问题收敛至未正则化问题，并探讨了基于 *higher order liftings* 建立统一框架的相关工作进展，在该框架中可以直接比较这两种正则化方法。我们也简要讨论了这些概念如何转化为相应的动态表述，并说明了为何应优先采用扩展空间正则化。本文是手稿的初步版本，近期将予以更新。

1. Introduction

给定某个完备可分度量空间 X、两个质量有限的正拉东测度 μ_0、 $\mu_1 \in \mathcal{M}(X)$, 以及一个下半连续代价函数 $c\colon X\times X \to \overline{\mathbb{R}}$，标准的平衡最优传输问题可表示为

$$\mathrm{OT}(\mu_0,\mu_1) = \inf_{\gamma\in\Gamma(\mu_0,\mu_1)} (c,\gamma), \quad (c,\gamma) := \int_{X\times X} c(x_0,x_1)d\gamma(x_0,x_1),$$

$$\Gamma(\mu_0,\mu_1) := \Big\{\gamma \in \mathcal{M}(X\times X) \mid \gamma_i = \mu_i\Big\},$$

其中 γ_i 是第 i 个边缘分布，

$$\gamma_i := \pi^i_{\#}\gamma \in \mathcal{M}(X), \quad \pi^i(x_0,x_1) = x_i.$$

为了实现高效算法，同时也出于理论研究的兴趣，平衡最优传输问题通常会进行熵正则化 [Nut21], 即

$$\mathrm{OT}_\varepsilon(\mu_0,\mu_1) = \inf_{\gamma\in\Gamma(\mu_0,\mu_1)} \Big\{(c,\gamma) + \varepsilon\mathcal{F}(\gamma \mid \nu)\Big\},$$

Date：2023年5月5日。2020 *Mathematics Subject Classification.* 49Q22、28A33、46E27、58E30、90C25、49N05。*Key words and phrases.* 熵正则最优传输、非平衡最优传输、正则化。两位作者谨此感谢英国工程与自然科学研究委员会的支持。MHD 获研究基金 EP/V038516/1 资助。本文大部分工作期间，MB 获同一研究基金资助；目前则获研究基金 EP/V00204X/1 资助。

其中 $\nu \in \mathcal{M}(X \times X)$ 是某个已知的参考测度（例如 $\nu = \mu_0 \otimes \mu_1$），熵泛函 $\mathcal{F}$ 定义为

$$\mathcal{F}(\gamma \mid \nu) = \int_{X \times X} F(\varsigma) d\nu + F'_\infty \gamma^\perp(X \times X), \quad \gamma = \varsigma\nu + \gamma^\perp,$$

其中，$F(s) = s \log s - s + 1$ 是 Kullback-Leibler 散度（也称为相对熵），而 F'_∞ 是衰退常数；在此情形下，它等于 $+\infty$，从而确保极小化解满足 $\gamma \ll \nu$。

不难看出，目前介绍的两个问题只有在输入测度的质量平衡时才有意义，也就是说，当 $\mu_0(X) = \mu_1(X)$ 时才有意义；否则，集合 $\Gamma(\mu_0, \mu_1)$ 为空。

部分借助 [LMS18]，中发展的形式体系，我们可以等价地将未正则化问题表述为

$$\mathrm{OT}(\mu_0, \mu_1) = \inf_{\gamma \in \mathcal{M}(X \times X)} \Big\{ \sum_i \overline{\mathcal{F}}(\gamma_i \mid \mu_i) + (c, \gamma) \Big\},$$

其中，若 $\gamma_i = \mu_i$，则 $\overline{\mathcal{F}}(\gamma_i \mid \mu_i) = 0$；否则为 $+\infty$。特别地，这一表达式刻画了测度质量不同时问题不可行的情形，因为

$$\mu_0(X) \neq \mu_1(X) \implies \mathrm{OT}(\mu_0, \mu_1) = +\infty.$$

由三个研究团队 [Chi+16；KMV16；LMS18] 独立提出的非平衡最优传输问题，其核心思想是放宽 $\overline{\mathcal{F}}$ 的严格约束，并以熵泛函取而代之，仅对 γ 与 μ_i 的边缘分布之间的偏差进行惩罚，

$$\mathcal{F}(\gamma_i \mid \mu_i) = \int_X F(\sigma_i) d\mu_i, \quad \gamma_i = \sigma_i \mu_i + \gamma_i^\perp,$$

其中，熵函数 F 不一定是上文介绍的 KL散度。因此，非平衡最优传输问题可表示为

$$\mathrm{UOT}(\mu_0, \mu_1) = \inf_{\gamma \in \mathcal{M}(X \times X)} \Big\{ \sum_i \mathcal{F}(\gamma_i \mid \mu_i) + (c, \gamma) \Big\}. \tag{1}$$

值得注意的是，可以证明 [LMS18, 第 5] 节表明，该问题存在一种扩展空间形式

$$\mathrm{UOT}(\mu_0, \mu_1) = \inf_{\alpha \in S(\mu_0, \mu_1)} (H, \alpha), \quad (H, \alpha) := \int_{Y \times Y} H(y_0, y_1) d\alpha(y_0, y_1), \tag{2}$$

其中 $Y := X \times \mathbb{R}_+$，记号为 $y_i = (x_i, s_i)$，且 $H : Y \times Y \to \overline{\mathbb{R}}$ 是相应的边缘视角代价函数。下确界取遍满足齐次边缘约束的测度集合

$$S(\mu_0, \mu_1) = \Big\{ \alpha \in M(Y \times Y) \mid \pi^{x_i}_\# (s_i \alpha) = \mu_i. \Big\}$$

迄今为止引入的概念将在第 2 节中作全面介绍。

本文旨在研究非平衡最优传输问题的熵正则化。由上文讨论可知，这可以通过两种方式实现——从（1）出发，我们可以在原空间上进行正则化，并考虑

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0, \mu_1) = \inf_{\gamma \in \mathcal{M}(X \times X)} \Big\{ \sum_i \mathcal{F}(\gamma_i \mid \mu_i) + (c, \gamma) + \varepsilon \mathcal{F}(\gamma \mid \nu_X) \Big\},$$

对于某个参考测度 $\nu_X \in \mathcal{M}(X \times X)$。

另一方面，从 (2) 出发，我们可以在扩展空间上进行正则化，并考虑

$$\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1)=\inf_{\alpha\in S(\mu_0,\mu_1)}\Big\{(H,\alpha)+\varepsilon\mathcal{F}(\alpha\mid\nu_Y)\Big\},$$

对于某个参考测度 $\nu_Y\in\mathcal{M}(Y\times Y)$。

论文的提纲及其贡献。本文首先回顾非平衡最优传输静态表述的理论，该理论改编自[LMS18] ，并在第 2 节中介绍。特别地，我们强调，为了证明齐次形式和扩展形式的重新表述，需要同时利用对偶表述和反向表述。我们还通过示例说明，即使在平衡情形下，扩展描述也未必是平凡的。

本文的贡献在第 3–5 节中阐述。在第 3.1 节中，我们发展基于原空间的正则化理论。我们密切遵循第 2 节的结构，并推导出反向、对偶和齐次表述及相关结果。有趣的是，虽然在典型的未正则化情形下（以 KL散度作为熵函数），诱导的边缘视角代价函数为

$$H(x_0,s_0,x_1,s_1)=s_0+s_1-2\sqrt{s_0s_1}\exp\left(\frac{-c(x_0,x_1)}{2}\right),$$

我们证明，相应的正则化诱导边缘透视代价函数为

$$H_\varepsilon(x_0,s_0,x_1,s_1,S)=s_0+s_1+\varepsilon S-(2+\varepsilon)\left((s_0s_1)^{\frac{1}{2+\varepsilon}}S^{\frac{\varepsilon}{2+\varepsilon}}\exp\left(\frac{-c(x_0,x_1)}{2+\varepsilon}\right)\right),$$

其关于（s_0， s_1， S）是 1-齐次的。基于此，我们推导出扩展表述，将其从 $\mathcal{M}(X\times X)$ 提升到空间 $\mathcal{M}(X^2\times\mathbb{R}^3_+)$。

在第 3.2 节中，我们部分地发展了基于扩展空间的正则化理论。特别是，我们证明，当正则化参数 $\varepsilon\to 0$ 时，该正则化收敛于原非平衡最优传输问题。为实现这一点，我们利用了将非平衡最优传输问题重新表述为扩展空间上最优传输问题的下确界，以及平衡熵正则最优传输的已知理论。

在第 4 节中，我们讨论统一扩展框架的关键要素。一方面，该框架将确保能够直接比较两类正则化，从而为原空间正则化的收敛性证明铺平道路。另一方面，该框架确立了高阶提升的原则，可用于表述高阶非平衡最优传输问题及其熵正则化。

本文最后在第 5 节概述了静态表述中原始空间正则化与扩展空间正则化之间的关键区别如何体现在动态表述中，并提供证据表明，扩展空间正则化似乎能使收敛渐近地更快，如 $\varepsilon\to 0$。

关于当前版本稿件的说明。稿件目前仍存在若干尚待补足之处，我们希望在不久的将来弥补这些不足。文中已用蓝色批注清楚标出这些内容。这些不足分为两类：（i）我们确信结论成立，但其中仍有一些技术细节尚待完善；（ii）需要投入大量工作才能实现的、更深层的新颖构想，甚至可能留待未来的论文再行探讨，但我们认为应先将其公之于众。这也是目前仍存在第一类不足的原因——我们希望能及时以某种形式发表

这篇稿件。具体而言，我们知道目前的扩展空间表述尚未充分处理径向分量为零的集合所涉及的问题。欢迎对此项工作感兴趣的人士与我们联系，共同探讨可能存在的错误及后续工作。

2. Unbalanced Optimal Transport

我们沿用 [LMS18]中的表述。设（X，d）为一个完备可分度量空间，并令 $\mathcal{M}$（X）表示 X上的有限非负拉东测度空间。令 μ_0，$\mu_1 \in \mathcal{M}$（X）。 [在当前版本的手稿中，我们隐含地假设 $X = \Omega \subset \mathbb{R}^n$是 $\mathbb{R}^n$的一个紧子集，但尽可能保持讨论的一般性。]

2.1. 静态原始形式。我们所考虑的非平衡最优传输问题类由下式给出

$$\mathrm{UOT}(\mu_0, \mu_1) := \inf_{\gamma \in \mathcal{M}(X \times X)} \mathcal{E}(\gamma \mid \mu_0, \mu_1), \tag{3}$$

其中，原始非平衡最优传输泛函 $\mathcal{E}$ 由下式给出

$$\mathcal{E}(\gamma \mid \mu_0, \mu_1) := \sum_{i=0,1} \mathcal{F}(\gamma_i \mid \mu_i) + (c, \gamma). \tag{4}$$

这里，$\gamma_i = \pi^i_{\#}\gamma \in \mathcal{M}(X)$表示 $\gamma \in \mathcal{M}(X \times X)$的第 i个边缘分布，而相对熵泛函 $\mathcal{F}$ 由下式给出

$$\mathcal{F}(\gamma_i \mid \mu_i) := \int_X F(\sigma_i) d\mu_i + F'_\infty \gamma_i^\perp(X), \tag{5}$$

其中 $F\colon [0, +\infty) \to [0, +\infty]$ 是一个合适的凸熵函数，

$$F'_\infty = \lim_{s \to \infty} \frac{F(s)}{s}$$

称为衰退常数，而 σ_i 是通过勒贝格分解 [LMS18,Lemma 2.3]得到的密度

$$\gamma_i = \sigma_i \mu_i + \gamma_i^\perp, \quad \mu_i = \rho_i \gamma_i + \mu_i^\perp. \tag{6}$$

最后，$c\colon X \times X \to \overline{\mathbb{R}}$ 是代价函数，假设其为下半连续。耦合

$$(c, \gamma) := \int_{X \times X} c(x_0, x_1) d\gamma(x_0, x_1) \tag{7}$$

表示线性成本泛函。

定理 2.1（[LMS18, Theorem 3.3]）。在熵泛函 $\mathcal{F}$ 和代价函数 c满足自然假设的条件下，至少存在一个使（3）式右侧达到最小值的 $\gamma \in \mathcal{M}(X \times X)$。

我们尤其感兴趣的是熵函数 F为 Kullback-Leibler（KL）散度的情形

$$F(s) = s \log s - s + 1, \quad \left(\implies F'_\infty = +\infty \right) \tag{8}$$

且代价函数 c 定义为

$$c(x_0, x_1) := \begin{cases} -\log(\cos^2(\mathsf{d}(x_0, x_1))) & \text{if } \mathsf{d}(x_0, x_1) < \pi/2 \\ +\infty & \text{otherwise.} \end{cases} \tag{9}$$

在此情形下，非平衡最优传输问题 (3) 与 $\mathcal{M}(X)$ [LMS18, Part II] 上的 Hellinger-Kantorovich 距离一致。

2.2. X 上的等价表述。式 (3) 中的非平衡最优传输问题在空间 X 上有三种等价表述，下面将简要介绍。第 3.1 节将会说明，经过必要调整后，类似结果也适用于熵正则化的非平衡最优传输问题。

2.2.1. 反向表述。给定式（5）中的某个熵函数 F，相应的反向熵函数 $R\colon [0,\infty) \to [0,\infty]$ 定义如下:

$$R(s) := \begin{cases} sF(1/s) & \text{if } s > 0, \\ F'_\infty & \text{if } s = 0. \end{cases} \tag{10}$$

注意，若 F 是 KL散度（8），则

$$R(s) = s - \log s - 1. \tag{11}$$

定理2.2（[LMS18，定理 3.11]）。非平衡最优传输问题（3）可等价地表述为

$$\mathrm{UOT}(\mu_0, \mu_1) = \inf_{\gamma \in \mathcal{M}(X\times X)} \mathcal{R}(\mu_0, \mu_1 \mid \gamma),$$

其中，回顾式（6）中的勒贝格分解，反向非平衡最优传输泛函 $\mathcal{R}$ 定义为

$$\mathcal{R}(\mu_0, \mu_1 \mid \gamma) := \int_{X\times X} \left(\sum_{i=0,1} R_i(\rho_i(x_i)) + c(x_0, x_1) \right) d\gamma(x_0, x_1) + R'_\infty \sum_{i=0,1} (\mu_i - \rho_i \gamma_i)(X), \tag{12}$$

注意 $R'_\infty = F(0)$。

2.2.2. 对偶形式。我们先简要回顾凸分析，内容改编自 [Chi+18 的附录 A]。设 E 和 E^* 为拓扑配对向量空间，它们之间的对偶配对记为 $\langle \cdot, \cdot \rangle\colon E \times E^* \to \mathbb{R}$。函数 $f\colon E \to \overline{\mathbb{R}}$ 的勒让德对偶定义为：对每个 $x^* \in E^*$，

$$f^*(x^*) := \sup_{x\in E} \langle x^*, x \rangle - f(x). \tag{13}$$

定理 2.3（Fenchel-Rockafeller[Roc67]）。设 (E, E^*) 和 (F, F^*) 是两对拓扑配对空间，且 $A\colon E \to F$ 是一个连续线性算子，$A^*\colon F^* \to E^*$ 是其伴随算子。进一步假设 f 和 g 是分别定义在 E 和 F 上的下半连续真凸函数。若存在 $x \in E$ 使得 $f(x) < \infty$，且 g 在 Ax 处连续，则

$$\sup_{x\in E} -f(-x) - g(Ax) = \min_{y^* \in F^*} f^*(A^* y^*) + g^*(y^*)$$

且该最小值能够达到。

为推导对偶形式，我们通过设置$y^* \equiv \gamma$，　$A^* y^* \equiv (\gamma_0, \gamma_1)$　　$f^*(A^* y^*) \equiv \mathcal{F}(\gamma_0 \mid \mu_0) + \mathcal{F}(\gamma_1 \mid \mu_1)$　　$g^*(y^*) \equiv (c, \gamma)$将该定理应用于我们的设置中，这进而意味着

$$x \equiv (\phi_0, \phi_1), \quad Ax \equiv \phi_0 \oplus \phi_1, -f(-x) = \sum_i -\mathcal{F}^*(-\phi_i \mid \mu_i).$$

特别地，由 [Chi+18，命题 A.3] 可知，熵泛函 $\mathcal{F}(\cdot \mid \mu_i)$ 的勒让德对偶为

$$\mathcal{F}^*(\phi_i \mid \mu_i) = \int_X F^*(\phi_i(x)) d\mu_i(x) + \int_X \mathbb{I}^{\leq F'_\infty}(\phi_i(x)) d\mu_i(x).$$

其中，对于凸标量函数 $f\colon\ [0,+\infty) \to [0,+\infty]$，其勒让德对偶 $f^*\colon\ \mathbb{R} \to (\infty,+\infty]$可由式（13）看出为

$$f^*(\phi) = \sup_{s>0} \left(s\phi - f(s)\right). \tag{14}$$

同样，由标量情形可得，$c \in \mathbb{R}$,

$$\eta:\ \mathbb{R} \to \mathbb{R}, \quad \eta(s) := cs, \quad \eta^*(\phi) := \sup_{s>0} \left(s\phi - \eta(s)\right) \begin{cases} +\infty, & \text{if } \phi > c \\ 0, & \text{if } \phi \le c \end{cases}$$

that

$$-g(Ax) = 0 \iff \phi_0 \oplus \phi_1 \le c \quad \text{and} \quad -g(Ax) = -\infty \ \text{otherwise.}$$

因此，为了形式化对偶形式，我们定义

$$\Phi := \{\phi = (\phi_0, \phi_1) \in C_b(X) \times C_b(X) \ \mid\ -F^*(-\phi_i) \in C_b(X),\ \phi_0 \oplus \phi_1 \le c\}. \tag{15}$$

并且，由于变量替换

$$\phi_i := R^*(\psi_i) \iff \psi_i = -F^*(-\phi_i), \tag{16}$$

我们进一步定义

$$\Psi := \{\psi = (\psi_0, \psi_1) \in C_b(X) \times C_b(X) \ \mid\ R^*(\psi_i) \in C_b(X),\ R^*(\psi_0) \oplus R^*(\psi_1) \le c\}. \tag{17}$$

定理 2.4（[LMS18，命题 4.3，定理 4.11]）。非平衡最优传输问题 (3) 可等价地表述为

$$\mathrm{UOT}(\mu_0, \mu_1) = \sup_{\psi \in \Psi} \mathcal{D}_\psi(\psi \mid \mu_0, \mu_1) = \sup_{\phi \in \Phi} \mathcal{D}_\phi(\phi \mid \mu_0, \mu_1), \tag{18}$$

其中，两个对偶 UOT 泛函 $\mathcal{D}_\psi$ 和 $\mathcal{D}_\phi$ 定义为

$$\mathcal{D}_\psi(\psi \mid \mu_0, \mu_1) = \sum_{i=0,1} \mu_i\big(\psi_i\big), \quad \mathcal{D}_\phi(\phi \mid \mu_0, \mu_1) = \sum_{i=0,1} \mu_i\big(-F^*(-\phi_i)\big). \tag{19}$$

这里可以考虑的一个具体示例是 KL 情形，其中 F 由式 (8) 给出，而 R 由式 (11) 给出，其勒让德对偶函数为

$$F^*(\phi) = \exp(\phi) - 1, \quad R^*(\psi) = -\log(1-\psi), \tag{20}$$

对于该情形，可以直接验证式 (16) 中的变量替换公式成立。

2.2.3. 齐次形式。最后一个、也可以说最引人入胜的等价形式，是引入相应的边缘视角代价函数

$$H:\ (X \times \mathbb{R}_+) \times (X \times \mathbb{R}_+) \to [0,+\infty],$$

定义为

$$H(x_0, s_0, x_1, s_1) := \inf_{t>0} \left\{ t\left(R\left(\frac{s_0}{t}\right) + R\left(\frac{s_1}{t}\right) + c(x_0, x_1)\right)\right\}. \tag{21}$$

可以证明 [LMS18，引理 5。3] H 具有类似于第 2.2.2 节所述的对偶表示，即对于固定的 $c(x_0, x_1)$,

$$\begin{aligned} H(x_0, s_0, x_1, s_1) &= \sup\left\{s_0\psi_0 + s_1\psi_1 \ \mid\ R^*(\psi_0) + R^*(\psi_1) \le c(x_0, x_1)\right\} \\ &= \sup\left\{ -s_0F^*(-\phi_0) - s_1F^*(-\phi_1) \ \mid\ \phi_0 + \phi_1 \le c(x_0, x_1)\right\}. \end{aligned}$$

通过这一对偶刻画，可以相当容易地证明以下结果。

定理 2.5（[LMS18，定理 5.5]）。非平衡最优传输问题 (3) 可等价地表述为

$$\mathrm{UOT}(\mu_0,\mu_1)=\inf_{\gamma\in\mathcal{M}(X\times X)}\mathcal{H}(\mu_0,\mu_1\mid\gamma) \tag{22}$$

其中，回顾式 (6) 中的勒贝格分解，齐次非平衡最优传输泛函$\mathcal{H}$ 定义为

$$\mathcal{H}(\mu_0,\mu_1\mid\gamma):=\int_{X\times X}H(x_0,\rho_0(x_0),x_1,\rho_1(x_1)d\gamma(x_0,x_1)+F(0)\sum_{i=0,1}(\mu_i-\rho_i\gamma_i)(X). \tag{23}$$

注意，在 KL 情形 (8)、(11) 中，诱导的边缘视角代价函数定义为

$$H(x_0,s_0,x_1,s_1)=s_0+s_1-2\sqrt{s_0s_1}\exp\left(\frac{-c(x_0,x_1)}{2}\right). \tag{24}$$

2.3. 扩展空间中的齐次形式。勒贝格分解 (6) 中的一对 (γ_i,ρ_i) 可导出定义在扩展空间上的测度 $\beta\in\mathcal{M}(Y)$

$$Y:=X\times\mathbb{R}_+,\quad y=(x,s)\in Y \tag{25}$$

via

$$\beta:=(x,\rho_i(x))_{\#}\gamma_i.$$

这只是 Y 上一族非负测度中的一个示例，它们属于空间

$$\mathcal{M}_p(Y):=\left\{\beta\in\mathcal{M}(Y)\;\middle|\;\int_Y s^p d\beta<\infty\right\} \tag{26}$$

满足第 p个齐次边缘约束（其中 $p=1$ 如上所述），该约束由 $\mathsf{h}^p\beta=\rho_i\gamma_i$ 给出，其中

$$\mathsf{h}^p:\mathcal{M}_p(Y)\to\mathcal{M}(X),\quad \mathsf{h}^p\beta:=\pi^x_{\#}(s^p\beta)\in\mathcal{M}(X).$$

类似地，任意测度 $\gamma\in\mathcal{M}(X\times X)$ 以及勒贝格分解 (6) 中由此得到的各对 (γ_0,ρ_0) 和 (γ_1,ρ_1)，都会产生一个测度

$$\alpha\in\mathcal{M}(Y\times Y),\quad \alpha:=(x_0,\rho_0(x_0),x_1,\rho_1(x_1)_{\#}\gamma$$

并且，使用记号 $(y_0,y_1)\in Y\times Y$, $y_i=(x_i,s_i)$，我们可以定义

$$\mathsf{h}^p_i:\mathcal{M}_p(Y\times Y)\to\mathcal{M}(X),\quad \mathsf{h}^p_i\alpha:=\pi^{x_i}_{\#}(s_i^p\alpha)\in\mathcal{M}(X),$$

其中

$$\mathcal{M}_p(Y\times Y):=\left\{\alpha\in\mathcal{M}(Y\times Y)\;\middle|\;\int_{Y\times Y}(s_0^p+s_1^p)\,d\alpha<\infty\right\}. \tag{27}$$

由于根据勒贝格分解，我们有 $\mu_i=\rho_i\gamma_i+\mu_i^\perp$，且奇异部分 $\mu_i^\perp$ 非负，因此，考虑满足 $\mathsf{h}^p_i\alpha=\rho_i\gamma_i$ 的测度族，等价于要求 $\mathsf{h}^p_i\alpha_i\le\mu_i$，因此我们关注的集合是

$$S^p_{\le}(\mu_0,\mu_1):=\{\alpha\in\mathcal{M}(Y\times Y)\mid \mathsf{h}^p_i\alpha\le\mu_i\},$$
$$S^p_{=}(\mu_0,\mu_1):=\{\alpha\in\mathcal{M}(Y\times Y)\mid \mathsf{h}^p_i\alpha=\mu_i\},$$

其中，显然，

$$S^p_{=}(\mu_0,\mu_1)\subset S^p_{\le}(\mu_0,\mu_1)\subset\mathcal{M}_p(Y\times Y).$$

定理 2.6（[LMS18，定理 5.8]）。非平衡最优传输问题 (3) 可以等价地表述为

$$\begin{aligned}\mathrm{UOT}(\mu_0,\mu_1) &= \inf_{\alpha\in S^p_{\leq}(\mu_0,\mu_1)}\Big\{(H_p,\alpha)+F(0)\sum_{i=0,1}\left(\mu_i-\mathsf{h}^p_i\alpha\right)(X)\Big\},\\ &= \inf_{\alpha\in S^p_{=}(\mu_0,\mu_1)}(H_p,\alpha)\end{aligned}$$

其中

$$(H_p,\alpha):=\int_{Y\times Y}H_p(y_0,y_1)d\alpha(y_0,y_1),\quad H_p(y_0,y_1)=H(x_0,s_0^p,x_1,s_1^p).$$

2.3.1. 缩放不变性。以下关键的缩放不变性结果成立 .

命题 2.7（[LMS18，第 5.2] 节）。固定 $p>0$, $\mu_0,\mu_1\in\mathcal{M}(X)$ 并令

$$s_*:=(\mu_0(X)+\mu_1(X))^{\frac{1}{p}}.$$

考虑由以下公式定义的函数 $\theta\colon\ X\times Y\to[0,\infty)$ 和 $\mathbf{prd}_\theta\colon\ X\times Y\to X\times Y$。

$$\theta(y_0,y_1)=\frac{1}{s_*}(s_0^p+s_1^p)^{\frac{1}{p}},\quad \mathrm{prd}_\theta(x_0,s_0,x_1,s_1):=\left(x_0,\frac{s_0}{\theta},x_1,\frac{s_1}{\theta}\right).\tag{28}$$

进一步假设 $\alpha\in S^p_{\leq}(\mu_0,\mu_1)$ 并定义

$$\tilde\alpha:=\alpha\llcorner(Y\times Y)\setminus\{y\in Y\times Y\mid s_0=s_1=0\},\quad \hat\alpha:=(\mathrm{prd}_\theta)_\#(\theta^p\tilde\alpha).$$

Then

- $(H_p,\alpha)=(H_p,\tilde\alpha)=(H_p,\hat\alpha)$,
- $\hat\alpha\in S^p_{\leq}(\mu_0,\mu_1)$ *(and if* $\alpha\in S^p_{=}(\mu_0,\mu_1)$ *then* $\hat\alpha\in S^p_{=}(\mu_0,\mu_1)$*)*,
- $\hat\alpha\in\mathcal{P}(Y\times Y)$,
- $\hat\alpha((Y\times Y)[s_*]')=0$, *where*

$$Y[s_*]:=\{(x,s)\in Y\mid s\leq s_*\},\quad Y[s_*]'=Y\setminus Y[s_*],\tag{29}$$

$$(Y\times Y)[s_*]:=Y[s_*]\times Y[s_*],\qquad (Y\times Y)[s_*]'=Y\times Y\setminus(Y\times Y)[s_*].\tag{30}$$

命题 2.7 保证了以下结果。

定理 2.8。非平衡最优传输问题 (3) 可等价地表述为

$$\mathrm{UOT}(\mu_0,\mu_1)=\inf_{\alpha\in\overline{S}^p_{=}(\mu_0,\mu_1)}(H_p,\alpha),\tag{31}$$

其中

$$\overline{S}^p_{=}(\mu_0,\mu_1):=\{\alpha\in\mathcal{P}(Y\times Y)\mid \mathsf{h}^p_i\alpha=\mu_i,\ \alpha((Y\times Y)[s_*]')=0\}.$$

推论 2.9（[LMS18，推论 7.7]）。非平衡最优传输问题 (3) 的另一种改写形式为

$$\mathrm{UOT}(\mu_0,\mu_1)=\inf_{(\beta_0,\beta_1)}\big\{\mathrm{OT}(\beta_0,\beta_1)\mid\beta_i\in\overline{S}^p_{=}(\mu_i)\big\},$$

其中

$$\mathrm{OT}(\beta_0,\beta_1):=\inf_{\alpha\in\Gamma(\beta_0,\beta_1)}(H_p,\alpha),$$

$$\Gamma(\beta_0,\beta_1):=\{\alpha\in\mathcal{P}(Y\times Y),\ \pi^i_\#\alpha=\beta_i\}$$

是耦合的集合，并且

$$\overline{S}^p_{=}(\mu_0,\mu_1):=\{\beta\in\mathcal{P}(Y)\mid\mathsf{h}^p\beta=\mu_i,\ \beta((Y[s_*]')=0\}.$$

命题 2.10（[LMS18，注 5.10，定理 5.8]）。存在 $\overline{\alpha} \in \overline{S}^p_=(\mu_0, \mu_1)$，使得（31）的右侧达到最小值。

此外，若 $\gamma \in \mathcal{M}(X \times X)$ 是式 (3) 中定义的非平衡最优传输问题原始问题形式的一个极小化解（其存在性由定理 2.1 保证），则 $\overline{\alpha} \in \mathcal{M}(Y \times Y)$ 可表示为

$$\overline{\alpha} = (\mathrm{prd}_\theta)_\#(\theta^p \tilde{\alpha}),$$

其中 prd_θ 定义于式 (28)，而 $\alpha \in S^p_=(\mu_0, \mu_1)$ 定义为

$$\alpha = (x_0, \rho_0^{1/p}(x_0), x_1, \rho_1^{1/p}(x_1))_\#\gamma,$$

其中，密度 ρ_i 由勒贝格分解 (6) 得到。

至此，关于未正则化非平衡最优传输问题的所有相关结果已简要介绍完毕。

2.3.2. 示例：平衡最优传输的提升。本节最后讨论平衡最优传输问题这一特殊情形，并强调即使在这种情况下，提升到高维空间的想法也并非多余。具体而言，此示例将为连续提升这一想法铺垫背景，该想法将在第 4 节中探讨。

第 2.1 节介绍的未经正则化的非平衡最优传输问题，在 (5) 中的熵函数 F 被设为时，会退化为平衡最优传输问题

$$\overline{F}(s) := \begin{cases} 0, & \text{when } s = 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

更确切地说，对于 μ_0、$\mu_1 \in \mathcal{M}(X \times X)$，我们考虑

$$\mathrm{OT}(\mu_0, \mu_1) := \inf_{\gamma \in \mathcal{M}(X \times X)} \overline{\mathcal{E}}(\gamma \mid \mu_0, \mu_1),$$

其中

$$\overline{\mathcal{E}}(\gamma \mid \mu_0, \mu_1) := \sum_{i=0,1} \overline{\mathcal{F}}(\gamma_i \mid \mu_i) + (c, \gamma), \quad \overline{\mathcal{F}}(\gamma_i \mid \mu_i) = \int_X \overline{F}(\sigma_i) d\mu_i + \overline{F}'_\infty \gamma_i^\perp(X).$$

特别地，我们注意到，如果 $\mu_0(X) \neq \mu_1(X)$，那么 $\mathrm{OT}(\mu_0, \mu_1) = \infty$。尽管如此，第 2 节目前为止所呈现的所有结果仍然适用，因此我们观察到

$$\mathrm{OT}(\mu_0, \mu_1) = \inf_{\gamma \in \mathcal{M}(X \times X)} \int_{X \times X} H(x_0, \rho_0(x_0), x_1, \rho_1(x_1)) d\gamma \tag{32}$$

其中，由于边缘分布的熵函数 $\overline{F}$，我们有

$$H(x_0, s_0, x_1, s_1) = \begin{cases} ac(x_0, x_1), & \text{when } s_0 = s_1 = a, \\ +\infty, & \text{otherwise.} \end{cases}$$

特别地，我们注意到，对于 γ 的积分 (32) 为有限值，边缘分布的一个必要条件是：对于某个常数 $a > 0$，有 $a\gamma_i = \mu_i$；并且我们可以将下确界求解分为两个步骤，从而观察到

$$\mathrm{OT}(\mu_0, \mu_1) = \inf_{\gamma \in \Gamma(\mu_0, \mu_1)} \left\{ \inf_{a>0} \int_{X \times X} \frac{1}{a} c(x_0, x_1) d(a\gamma) \right\} = \inf_{\gamma \in \Gamma(\mu_0, \mu_1)} \int_{X \times X} c(x_0, x_1) d\gamma,$$

因此只需化简为通常的表述。

一旦我们提升到扩展空间中，这种完全约化便不再成立。我们有

$$\mathrm{OT}(\mu_0, \mu_1) = \inf_{\alpha \in S^p_=(\mu_0, \mu_1)} \int_{Y \times Y} H(x_0, s_0, x_1, s_1) d\alpha. \tag{33}$$

显然，为了使积分有限， α 的支撑集必须满足的一个必要条件是: t

$$\alpha((Y \times Y) \setminus \Omega) = 0, \quad \text{where } \Omega = \{(y_0, y_1) \in Y \times Y \mid s_0 = s_1\}.$$

换言之， α 必须具有以下形式

$$\alpha = (x_0, s, x_1, s)_{\#}\beta, \quad \beta \in \mathcal{M}(X^2 \times \mathbb{R}_+),\ (x_0, x_1, s) \in X^2 \times \mathbb{R}_+.$$

由于 $\alpha \in S^p_{=}(\mu_0, \mu_1)$，相应的 β 必须属于

$$\widetilde{S}^p_{=}(\mu_0, \mu_1) := \left\{ \beta \in \mathcal{M}(X^2 \times \mathbb{R}_+) \,\middle|\, \pi^{x_i}_{\#}(s^p \beta) = \mu_i \right\},$$

其中，对于任意可测函数 $f:\ X \to \mathbb{R}$,

$$\int_X f(x) d\pi^{x_i}_{\#}(s^p\beta) = \int_{X \times X \times \mathbb{R}_+} s^p f(x_i) d\beta.$$

由此可立即得出

$$\mathrm{OT}(\mu_0, \mu_1) = \inf_{\beta \in \widetilde{S}^p_{=}(\mu_0, \mu_1)} \int_{X \times X \times \mathbb{R}_+} s\, c(x_0, x_1) d\beta.$$

因此我们看到，在平衡情形下，提升策略将定义在 $\mathcal{M}(X \times X)$ 上的最小化问题转化为定义在$\mathcal{M}(X \times X \times \mathbb{R}_+)$ 上的最小化问题。

3. Entropic regularisation of unbalanced optimal transportation

在本文的这一主要章节中，我们研究并获得关于非平衡最优传输问题进行熵正则化的不同方式的理论结果。

3.1. 在原空间上的正则化。对非平衡最优传输问题 (3) 引入熵正则化的标准且看似最自然的方式，是考虑某个适当的参考测度 $\nu_X \in \mathcal{M}(X \times X)$（保留下标以便与第 3.2 节中引入的 ν_Y 区分）以及某个较小的 $\varepsilon > 0$,

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0, \mu_1) := \inf_{\gamma \in \mathcal{M}(X \times X)} \mathcal{E}_\varepsilon(\gamma \mid \mu_0, \mu_1, \nu_X), \tag{34}$$

其中，回顾式 (4) 中原始非平衡最优传输泛函 $\mathcal{E}$ 的定义，其正则化对应形式 $\mathcal{E}_\varepsilon$ 定义为

$$\mathcal{E}_\varepsilon(\gamma \mid \mu_0, \mu_1, \nu_X) := \mathcal{E}(\gamma \mid \mu_0, \mu_1) + \varepsilon \mathcal{F}(\gamma \mid \nu_X). \tag{35}$$

这里，与式 (5) 类似，我们有

$$\mathcal{F}(\gamma \mid \nu_X) := \int_{X \times X} F(\varsigma) d\nu_X + F'_\infty \gamma^\perp(X \times X),$$

其中，无论在此处还是在 $\mathcal{F}(\gamma_i \mid \mu_i)$ 的定义中， F 均取为 KL 散度 (8)；此外，如前所述，$\nu_X \in \mathcal{M}(X \times X)$ 是某个先验固定的参考测度，其密度 ς 与式 (6) 类似，通过勒贝格分解得到

$$\gamma = \varsigma \nu_X + \gamma^\perp, \quad \nu_X = \varrho \gamma + \nu_X^\perp. \tag{36}$$

目前已有大量文献致力于研究非平衡最优传输$_{X,\ \varepsilon}$ 及其密切相关的问题 [Chi +18；S´ej+19；SPV22；Lac22], ，下面我们对此作几点说明。

注释 3.1。为供将来参考，我们明确指出，在 [Chi+18] 中，作者要求参考测度 ν_X 为概率测度，并考虑

$$\widetilde{\mathrm{UOT}}_{X,\varepsilon}(\mu_0,\mu_1) := \inf_{\gamma\in\mathcal{M}(X\times X)} \left\{ \sum_i \mathcal{F}(\gamma_i \mid \mu_i) + (c,\gamma) + \varepsilon\tilde{\mathcal{F}}(\gamma \mid \nu_X) \right\}$$

其中

$$\tilde{\mathcal{F}}(\gamma \mid \nu_X) := \int_{X\times X} \tilde{F}(\sigma) d\nu_X + \tilde{F}'_\infty \gamma^\perp(X\times X), \quad \tilde{F}(s) := s\log s - s,$$

并且，由于引入了新的参考测度 $\overline{\nu}_X := \exp\left(-\frac{c}{\varepsilon}\right)\nu_X$，其中考虑的另一个问题为

$$\overline{\mathrm{UOT}}_{X,\varepsilon}(\mu_0,\mu_1) := \inf_{\gamma\in\mathcal{M}(X\times X)} \left\{ \sum_i \mathcal{F}(\gamma_i \mid \mu_i) + \varepsilon\mathcal{F}(\gamma \mid \overline{\nu}_X) \right\}.$$

可以很容易地证明，

$$\begin{aligned}\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1) &= \widetilde{\mathrm{UOT}}_{X,\varepsilon}(\mu_0,\mu_1) + \varepsilon\nu_X(X\times X)\\ &= \overline{\mathrm{UOT}}_{X,\varepsilon}(\mu_0,\mu_1) - \varepsilon\int_{X\times X} \exp\left(\frac{-c}{\varepsilon}\right) d\nu_X + \varepsilon\nu_X(X\times X).\end{aligned}$$

在最近的一项研究中，[S´ej+19] 的作者考虑了 $\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)$，其中参考测度 ν_X 被选为乘积测度 $\nu_X = \mu_0\otimes\mu_1$。这种做法存在一个明显的局限性，即引入了不必要的质量差异，因为 $\nu_X(X\times X) = \mu_0(X)\mu_1(X)$，并且

$$\pi^i_\# \nu_X = \mu_j(X)\mu_i \neq \mu_i,$$

除非 $\mu_j(X) = 1$。

注释 3.2。在近期一项旨在推导保持齐次性的非平衡传输问题正则化的有趣尝试中 [Lac22], 作者考察了

$$\mathrm{UOT}^{(H)}_{X,\varepsilon}(\mu_0,\mu_1) := \inf_{\gamma\in\mathcal{M}(X\times X)} \left\{ \sum_i \mathcal{F}(\gamma_i \mid \mu_i) + (c,\gamma) + \varepsilon\mathcal{G}(\gamma \mid \mu_0,\mu_1) \right\},$$

其中

$$\mathcal{G}(\gamma \mid \mu_0,\mu_1) := \frac{1}{2}\left(\mathcal{F}(\gamma \mid \hat{\mu}_0\otimes\mu_1) + \mathcal{F}(\gamma \mid \mu_0\otimes\hat{\mu}_1)\right), \quad \hat{\mu}_i := \frac{\mu_i}{\mu_i(X)} \in \mathcal{P}(X).$$

经过仔细改写可知，利用式 (36) 中的勒贝格分解以及 $\nu_X = \mu_0\otimes\mu_1$，泛函 $\mathcal{G}$ 可改写为

$$\mathcal{G}(\gamma \mid \mu_0,\mu_1) = \int_{X\times X} G(\sigma) d(\mu_0\otimes\mu_1) + G'_\infty\gamma^\perp(X\times X),$$

其中

$$G(s) := s\log s + (c_0 - 1)s + \frac{1}{c_1}, \quad c_0 = \frac{1}{2}\log(\mu_0(X)\mu_1(X)),\ c_1 = 2\frac{\mu_0(X)\mu_1(X)}{\mu_0(X)+\mu_1(X)}.$$

c_0 和 c_1 的形式看起来有些类似于将在第 3.1.5 节中讨论的缩放不变性。

据我们所知，目前尚无关于问题 (34) 的极小化子存在性的结果；我们现在将参照 [LMS18 定理 3.3、定理 6.2] 中的论证，证明其存在性。

定理 3.3（ X-空间正则化极小化子的存在性）。在熵泛函 $\mathcal{F}$、代价函数 c 和参考测度 ν_X 满足自然假设的条件下，至少存在一个使 (3) 右端达到最小值的 $\gamma_\varepsilon \in \mathcal{M}(X \times X)$。

证明。 [定理的陈述将在下一版中予以精确表述，并补充证明到该手稿中。] □

仿照第 2 节中概述的方法，我们现在将给出非平衡最优传输的等价表述X, ε包括新的齐次形式和扩展空间表述。这些表述利用了（显然的）反向表述与（已知的 [Chi+18]）对偶形式之间的相互作用。据我们所知，这是首次尝试这样做。

3.1.1. 反向表述。回顾式（12）中反向非平衡最优传输泛函的定义$\mathcal{R}$ ，我们将其正则化对应形式引入为

$$\mathcal{R}_\varepsilon(\mu_0, \mu_1, \nu_X \mid \gamma) := \mathcal{R}(\mu_0, \mu_1 \mid \gamma) + \varepsilon \int_{X\times X} R(\varrho(x_0, x_1))d\gamma(x_0, x_1) \qquad (37)$$
$$+ \varepsilon R'_\infty \left(\nu_X - \varrho\gamma\right)(X \times X),$$

以下命题是对第 2.2.1 节中所述讨论的直接扩展

命题 3.4。式 (34) 中定义的正则化非平衡最优传输问题 UOT$_{X,\ \varepsilon}$，可等价地表述为

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0, \mu_1) = \inf_{\gamma\in\mathcal{M}(X\times X)} \mathcal{R}_\varepsilon(\mu_0, \mu_1, \nu_X \mid \gamma).$$

3.1.2. 对偶形式。回顾第 2.2.2 节中的讨论，应用定理 2.3（如 [Chi+18] 中所述）可得以下结果。

命题 3.5。式 (34) 中定义的正则化非平衡最优传输问题 UOT$_{X,\ \varepsilon}$，可等价地表述为

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0, \mu_1) = \sup_{\phi\in\tilde{\Phi}} \mathcal{D}_{\phi,\varepsilon}(\phi \mid \mu_0, \mu_1, \nu_X),$$

其中

$$\mathcal{D}_{\phi,\varepsilon}(\phi \mid \mu_0, \mu_1, \nu_X) := \sum_i \mu_i(-F^*(-\phi_i)) + \varepsilon\nu_X\left(-F^*\left(\frac{\phi_0 \oplus \phi_1 - c}{\varepsilon}\right)\right) \qquad (38)$$

and

$$\tilde{\Phi} := \{\phi = (\phi_0, \phi_1) \in C_b(X) \times C_b(X) \ \mid \ -F^*(-\phi_i) \in C_b(X)\} \qquad (39)$$

3.1.3. 齐次形式。仿照第 2.2.3 节所述的方法，我们现在引入正则化诱导边缘视角代价函数

$$H_\varepsilon : \ (X \times \mathbb{R}_+) \times (X \times \mathbb{R}_+) \times \mathbb{R}_+ \to [0, +\infty],$$

as

$$H_\varepsilon(x_0, s_0, x_1, s_1, S) := \inf_{t>0} \left\{ t\left(R\left(\frac{s_0}{t}\right) + R\left(\frac{s_1}{t}\right) + \varepsilon R\left(\frac{S}{t}\right) + c(x_0, x_1)\right)\right\}. \qquad (40)$$

我们证明如下结论（未正则化情形参见（24））。

引理 3.6。如果 R 由 (11) 给出，那么 (40) 中引入的正则化诱导边缘透视代价函数可表示为

$$H_\varepsilon(x_0,s_0,x_1,s_1,S)=s_0+s_1+\varepsilon S-(2+\varepsilon)\left((s_0s_1)^{\frac{1}{2+\varepsilon}}S^{\frac{\varepsilon}{2+\varepsilon}}\exp\left(\frac{-c(x_0,x_1)}{2+\varepsilon}\right)\right).$$

证明。对 (40) 的右端求导，直接计算即可得出。□

我们还证明了以下技术性结论。

引理 3.7. 在（40）中引入的正则化诱导边缘透视代价函数具有对偶表示；对于固定的 $c(x_0,x_1)$，该表示为

$$H_\varepsilon(x_0,s_0,x_1,s_1,S)=\sup_{(\phi_0,\phi_1)\in\mathbb{R}_+^2}\Big\{-s_0F^*(-\phi_0)-s_1F^*(-\phi_1)\\ -\varepsilon SF^*\Big(\frac{\phi_0+\phi_1-c(x_0,x_1)}{\varepsilon}\Big)\Big\}.$$

证明。为简洁起见，我们省略 H_ε 和 c 对其变量的依赖。根据式（10）中熵函数 F 对反向熵函数 R 的定义，我们有

$$H_\varepsilon=\inf_{t>0}\left\{s_0F\left(\frac{t}{s_0}\right)+s_1F\left(\frac{t}{s_1}\right)+\varepsilon SF\left(\frac{t}{S}\right)+tc\right\}.$$

集合 $\xi(t):=\varepsilon SF\left(\frac{t}{S}\right)+tc$。$\xi$ 的勒让德对偶，如式 (14) 所示，由下式给出。

$$\xi^*(\phi):=\sup_{t>0}\left(t\phi-\varepsilon SF\left(\frac{t}{S}\right)-tc\right)=:\sup_{t>0}\tilde{\xi}(t).$$

回顾 $F(s)=s\log s-s+1$，不难看出

$$\tilde{\xi}'(\overline{t})=0\implies\overline{t}=S\exp\left(\frac{\phi-c}{\varepsilon}\right),$$

将此公式代入 $\overline{t}$ 后，我们便可得出

$$\xi^*(\phi)=-\varepsilon S\left(-\exp\left(\frac{\phi-c}{\varepsilon}\right)+1\right).$$

为应用定理 2.3，我们先确定

$$y^*\equiv t\in\mathbb{R}_+,\ A^*y^*\equiv\left(\frac{t}{s_0},\frac{t}{s_1}\right)\in\mathbb{R}_+^2,\ f^*(A^*y^*)\equiv s_0F\left(\frac{t}{s_0}\right)+s_1F\left(\frac{t}{s_1}\right),\ g^*(y^*)\equiv\xi(t),$$

这进而意味着

$$x\equiv(\phi_0,\phi_1)\in\mathbb{R}_+^2,\quad Ax\equiv\frac{\phi_0}{s_0}+\frac{\phi_1}{s_1}.$$

我们进一步指出，由于 $f^*:\ \mathbb{R}_+^2\to\mathbb{R}$ 由 $f^*(x^*)=s_0F(x_0^*)+s_1F(x_1^*)$ 给出，因此我们有

$$f(x)=\sup_{x^*\in\mathbb{R}_+^2}x\cdot x^*-f^*(x^*)=\left(\sup_{x_0^*>0}x_0^*x_0-s_0F(x_0^*)\right)+\left(\sup_{x_1^*>0}x_1^*x_1-s_1F(x_1^*)\right)\\ =(s_0F)^*(x_0)+(s_1F)^*(x_1)\equiv s_0F^*\left(\frac{\phi_0}{s_0}\right)+s_1F^*\left(\frac{\phi_1}{s_1}\right)$$

其中，最后一个等式是根据一般情况下的事实得出的，

$$(aF)^*(\phi) = aF^*\left(\frac{\phi}{a}\right),$$

这意味着

$$-f(-x) \equiv -s_0F^*\left(-\frac{\phi_0}{s_0}\right) - s_1F^*\left(-\frac{\phi_1}{s_1}\right).$$

因此，直接应用定理 2.3 可得

$$H_\varepsilon(x_0,s_0,x_1,s_1,S) = \sup_{(\phi_0,\phi_1)\in\mathbb{R}^2_+}\left\{ - s_0F^*\left(-\frac{\phi_0}{s_0}\right) - s_1F^*\left(-\frac{\phi_1}{s_1}\right) - \varepsilon SF^*\Big(\frac{\frac{\phi_0}{s_0}+\frac{\phi_1}{s_1}-c(x_0,x_1)}{\varepsilon}\Big)\right\}.$$

因此，只需注意到我们可以放心地代入 $\frac{\phi_i}{s_i}\mapsto\phi_i$，即可得出该引理的结论。□

现在我们可以将正则化的齐次非平衡最优传输泛函$\mathcal{H}_\varepsilon$ 引入如下

$$\mathcal{H}_\varepsilon(\mu_0,\mu_1,\nu_X\mid\gamma) := \int_{X\times X} H_\varepsilon(x_0,\rho_0(x_0),x_1,\rho_1(x_1),\varrho(x_0,x_1))d\gamma(x_0,x_1) \tag{41}$$

$$+ F(0)\left(\sum_{i=0,1}(\mu_i-\rho_i\gamma_i)(X) + (\nu_X-\varrho\gamma)(X\times X)\right) \tag{42}$$

借助引理 3.7，我们现在将证明本节的主要结果。

定理 3.8。正则化非平衡最优传输问题 $\mathrm{UOT}_{X,\ \varepsilon}$（在 (34) 中定义）可等价地表述为

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1) = \inf_{\gamma\in\mathcal{M}(X\times X)}\mathcal{H}_\varepsilon(\mu_0,\mu_1,\nu_X\mid\gamma).$$

证明。采用 LMS18 定理 5 的证明中所使用的方法，我们将证明 [LMS18, Theorem 5. 5], we will˜show that $\mu_i\in\mathcal{M}(X)$, $\nu_X\in\mathcal{M}(X\times X)$ and $\phi\in\Phi$ (定义见（39）)，

$$\mathcal{D}_{\phi,\varepsilon}(\phi\mid\mu_0,\mu_1,\nu_X) \le \mathcal{H}_\varepsilon(\mu_0,\mu_1,\nu_X\mid\gamma) \le \mathcal{R}_\varepsilon(\mu_0,\mu_1,\nu_X\mid\gamma), \tag{43}$$

这确保了定理的结论，因为此时

$$\underbrace{\sup_{\phi\in\tilde{\Phi}}\mathcal{D}_{\phi,\varepsilon}(\phi\mid\mu_0,\mu_1,\nu_X)}_{=\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)} \le \inf_{\gamma\in\mathcal{M}(X\times X)}\mathcal{H}_\varepsilon(\mu_0,\mu_1,\nu_X\mid\gamma) \le \underbrace{\inf_{\gamma\in\mathcal{M}(X\times X)}\mathcal{R}_\varepsilon(\mu_0,\mu_1,\nu_X\mid\gamma)}_{=\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)}.$$

事实上，(43) 中的第二个不等式显然成立，因为只需在右侧 (40) 的下确界中令 $t=1$，即可恢复 $\mathcal{R}_\varepsilon$。

为证明 (43) 中的第一个不等式，我们利用引理 3.7 中得到的 H_ε 的对偶表述，并注意到

$$\begin{aligned}
\mathcal{H}_\varepsilon(\mu_0,\mu_1,\nu_X \mid \gamma) \geq & \int_{X\times X} -\rho_0(x_0)F^*(-\phi_0(x_0))d\gamma(x_0,x_1) \\
& + \int_{X\times X} -\rho_1(x_1)F^*(-\phi_1(x_1))d\gamma(x_0,x_1) \\
& + \int_{X\times X} -\varepsilon\varrho(x_0,x_1)F^*\left(\frac{\phi_0(x_0)+\phi_1(x_1)-c(x_0,x_1)}{\varepsilon}\right)d\gamma(x_0,x_1) \\
& + F(0)\left(\sum_{i=0,1}\left(\mu_i-\rho_i\gamma_i\right)(X)+\left(\nu_X-\varrho\gamma\right)(X\times X)\right) \\
= & \sum_i \mu_i(-F^*(-\phi_i)) + \varepsilon\nu_X\left(-F^*\left(\frac{\phi_0\oplus\phi_1-c}{\varepsilon}\right)\right) \\
& + \sum_i(\rho_i\gamma_i-\mu_i)(-F^*(-\phi_i)) + \varepsilon(\varrho\gamma-\nu_x)\left(-F^*\left(\frac{\phi_0\oplus\phi_1-c}{\varepsilon}\right)\right) \\
& + F(0)\left(\sum_{i=0,1}\left(\mu_i-\rho_i\gamma_i\right)(X)+\left(\nu_X-\varrho\gamma\right)(X\times X)\right) \\
\geq & \sum_i \mu_i(-F^*(-\phi_i)) + \varepsilon\nu_X\left(-F^*\left(\frac{\phi_0\oplus\phi_1-c}{\varepsilon}\right)\right) \\
= & \mathcal{D}_{\phi,\varepsilon}(\phi \mid \mu_0,\mu_1,\nu_X),
\end{aligned}$$

其中，最后一个不等式由 $\sup -F^* = -F(0)$ 得出（这是勒让德对偶的标准性质，例如 [LMS18 第 2 节3] 中有所讨论）。□

3.1.4. 扩展空间中的齐次形式。测度 $\gamma \in \mathcal{M}(X\times X)$、由边缘分布 γ_i 的勒贝格分解得到的密度三元组 (ρ_0,ρ_1,ρ)（见式 (6)），以及 γ 本身（见式 (36)），共同构成测度 $\eta \in \mathcal{M}(Y\times Y\times\mathbb{R}_+)$（回顾式 (25) 中 Y 的定义），其定义如下：

$$\eta = (x_0,\rho_0(x_0),x_1,\rho_1(x_1),\rho(x_0,x_1))_\#\gamma$$

这只是 $Y\times Y\times\mathbb{R}_+$ 上一族非负测度中的一个例子，它们属于空间

$$\mathcal{M}_p(Y\times Y\times\mathbb{R}_+) := \left\{\eta\in\mathcal{M}(Y\times Y\times R_+) \;\middle|\; \int_Y s_0^p + s_1^p + S^p d\eta < \infty\right\} \tag{44}$$

满足由下式给出的第 p 个齐次约束（其中 $p=1$ 如上）：

$$\mathsf{h}_i^p\eta = \rho_i\gamma_i, \quad \mathsf{H}^p\eta = \rho\gamma,$$

其中

$$\mathsf{h}_i^p:\ \mathcal{M}_p(Y\times Y\times\mathbb{R}_+)\to\mathcal{M}(X), \quad \mathsf{h}_i^p\eta := \pi_\#^{x_i}(s_i^p\eta)\in\mathcal{M}(X),$$

$$\mathsf{H}^p:\quad \mathcal{M}_p(Y\times Y\times\mathbb{R}_+)\to\mathcal{M}(X\times X),\ \mathsf{H}^p\eta := \pi^{(x_0,x_1)}{}_\#(S^p\eta)\in\mathcal{M}(X\times X)。$$

与第 2.3 节中的讨论类似，非平衡最优传输的扩展空间表述 $_{X,\varepsilon}$ 将在集合上定义

$$\begin{aligned}
S^p_{\leq}(\mu_0,\mu_1,\nu_X) &:= \{\eta\in\mathcal{M}(Y\times Y\times\mathbb{R}_+) \mid \mathsf{h}_i^p\eta\leq\mu_i, \mathsf{H}^p\eta\leq\nu_X\}, \\
S^p_{=}(\mu_0,\mu_1,\nu_X) &:= \{\eta\in\mathcal{M}(Y\times Y\times\mathbb{R}_+) \mid \mathsf{h}_i^p\eta=\mu_i, \mathsf{H}^p\eta=\nu_X\},
\end{aligned}$$

其中，显然，

$$S^p_{=}(\mu_0,\mu_1,\nu_X)\subset S^p_{\leq}(\mu_0,\mu_1,\nu_X)\subset\mathcal{M}_p(Y\times Y\times\mathbb{R}_+).$$

注 3.9。参考测度 $\nu_X\in\mathcal{M}(X\times X)$ 通常取为乘积测度。

$$\nu_X=\nu_X^0\otimes v_X^1$$

其中 $\mu_i\ll\nu_X^i$。例如，如同 [S´ej+19] 中那样，我们可以取 $\nu_X=\mu_0\otimes\mu_1$，那么，对于 $\eta\in S^p_{=}(\mu_0,\mu_1,\nu_X)$，可以很容易地验证，

$$\pi^i_{\#}(\mathsf{H}^p\eta)=\mu_j(X)\mathsf{h}^p_i\eta.$$

更一般地，如果 $\mu_i\ll\nu_X^i$，根据勒贝格分解（6），我们可以写出

$$\mu_i=m_i\nu_X^i,\quad \nu_X^i=n_i\mu_i+\left(\nu_X^i\right)^{\perp},\tag{45}$$

对于某些密度 m_i、n_i，且如果另外满足 $\mu_i\sim\nu_X^i$，那么

$$m_i>0,\ n_i=\frac{1}{m_i},\ \left(\nu_X^i\right)^{\perp}.$$

一个符合 [Chi+18] 中 $\nu_X\in\mathcal{P}(X\times X)$ 要求的示例是，令

$$\nu_X=\frac{1}{\mu_0(X)\mu_1(X)}\mu_0\otimes\mu_1,$$

即密度 $m_i=\mu_i(X)$ 为常数。在一般情形 (45) 下，我们注意到

$$\eta\in S^p_{=}(\mu_0,\mu_1,\nu_X)\implies\pi^i_{\#}(\mathsf{H}^p\eta)=\left(\left(n_j\mathsf{h}^p_j\eta\right)(X)+\left(\nu_X^j\right)^{\perp}(X)\right)\left(n_i\mathsf{h}^p_i\eta+\left(\nu_X^i\right)^{\perp}\right).$$

定理 3.10. 在 (34) 中定义的正则化非平衡最优传输问题 $\mathrm{UOT}_{X,\ \varepsilon}$，可以等价地表述为

$$\begin{aligned}\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)&=\inf_{\eta\in S^p_{\leq}(\mu_0,\mu_1,\nu_X)}\left\{(H^\varepsilon_p,\eta)+F(0)\left(\sum_{i=0,1}\left(\mu_i-\mathsf{h}^p_i\eta\right)(X)+\left(\nu_X-\mathsf{H}^p\eta\right)(X\times X)\right)\right\},\\&=\inf_{\eta\in S^p_{=}(\mu_0,\mu_1,\nu_X)}(H^\varepsilon_p,\eta)\end{aligned}$$

其中

$$(H^\varepsilon_p,\eta):=\int_{Y\times Y\times\mathbb{R}_+}H^\varepsilon_p(x_0,s_0,x_1,s_1,S)d\eta(x_0,s_0,x_1,s_1,S),$$

and

$$H^\varepsilon_p(x_0,s_0,x_1,s_1,S)=H_\varepsilon(x_0,s_0^p,x_1,s_1^p,S^p).$$

证明。该结果本质上是定理 3.8 的直接推论——具体论证与 [LMS18, 定理 5.8] 的证明完全相同。□

3.1.5. 缩放不变性。第 2.3.1 节讨论的非平衡最优传输问题的缩放不变性，源于第 2.2.3 节引入的诱导边缘视角代价函数 H 满足

$$H\left(x_0,\frac{s_0}{\theta},x_1,\frac{s_1}{\theta}\right)=\frac{1}{\theta}H(x_0,s_0,x_1,s_1).$$

其在第 3.1.3 节中引入的熵正则化对应版本也具有类似的性质，即

$$H_\varepsilon\left(x_0,\frac{s_0}{\theta},x_1,\frac{s_1}{\theta},\frac{S}{\theta}\right)=\frac{1}{\theta}H_\varepsilon(x_0,s_0,x_1,s_1,S),$$

这可以从引理 3.6 中确立的 H_ε 的显式形式直接验证。现在我们将证明命题 2.7 的熵正则化对应版本。

命题 3.11。固定 $p>0$、 μ_0、 $\mu_1\in\mathcal{M}$（X）、 $\nu_X\in\mathcal{M}$（$X\times X$），并设定

$$s_*:=(\mu_0(X)+\mu_1(X)+\nu_X(X\times X))^{\frac{1}{p}}.$$

考虑由以下公式定义的函数 θ： $Y\times Y\times\mathbb{R}_+\to[0,\infty)$ 和 prd_θ： $Y\times Y\times\mathbb{R}_+\to Y\times Y\times\mathbb{R}_+$。

$$\theta(x_0,s_0,x_1,s_1,S)=\frac{1}{s_*}(s_0^p+s_1^p+S^p)^{\frac{1}{p}},\quad \mathrm{prd}_\theta(x_0,s_0,x_1,s_1,S):=\left(x_0,\frac{s_0}{\theta},x_1,\frac{s_1}{\theta},\frac{S}{\theta}\right). \tag{46}$$

进一步假设 $\eta\in S^p_{\leq}(\mu_0,\mu_1,\nu_X)$，并定义

$$\tilde\eta:=\eta(Y\times Y\times\mathbb{R}_+)\setminus\{(x_0,s_0,x_1,s_1,S)\in Y\times Y\times\mathbb{R}_+\mid s_0=s_1=S=0,\}$$

和

$$\hat\eta:=(\mathrm{prd}_\theta)_\#(\theta^p\tilde\eta).$$

Then

- $(H_p^\varepsilon,\eta)=(H_p^\varepsilon,\tilde\eta)=(H_p^\varepsilon,\hat\eta)$,
- $\hat\eta\in S^p_{\leq}(\mu_0,\mu_1,\nu_X)$ *(and if* $\eta\in S^p_{=}(\mu_0,\mu_1.\nu_X)$ *then* $\hat\eta\in S^p_{=}(\mu_0,\mu_1,\nu_X)$*)*,
- $\hat\eta\in\mathcal{P}(Y\times Y\times\mathbb{R}_+)$,
- $\hat\eta((Y\times Y\times\mathbb{R}_+)[s_*]')=0$, *where*

$$\begin{aligned}
Y[s_*]&:=\{(x,s)\in Y\mid s\leq s_*\}, & Y[s_*]'&=Y\setminus Y[s_*],\\
\mathbb{R}_+[s_*]&:=\{s\in\mathbb{R}_+\mid s\leq s_*\}, & \mathbb{R}_+[s_*]'&=\mathbb{R}_+\setminus\mathbb{R}_+[s_*],\\
(Y\times Y)\times\mathbb{R}_+[s_*]&:=Y[s_*]\times Y[s_*]\times\mathbb{R}_+[s_*],\\
(Y\times Y\times\mathbb{R}_+)[s_*]'&:=Y\times Y\times\mathbb{R}_+\setminus(Y\times Y\times\mathbb{R}_+)[s_*].
\end{aligned}$$

证明。 证明源于 H_ε 如上文所述的齐次性。 [将在论文的下一版本中补充证明的详细阐述。] □

至此，我们结束了关于如何处理非平衡最优传输的原空间正则化的一般框架介绍。最后，我们讨论平衡最优传输这一特殊情形，并给出与第 2.3.2 节相对应的内容。

3.1.6. 示例：平衡最优传输的熵正则化的基于原空间的提升。对平衡最优传输进行正则化的标准方法在理论上有充分依据（参见例如[Nut21]），即考虑

$$\mathrm{OT}_{X,\varepsilon}(\mu_0,\mu_1):=\inf_{\gamma\in\mathcal{M}(X\times X)}\overline{\mathcal{E}}_\varepsilon(\gamma\mid\mu_0,\mu_1,\nu_X),$$

其中

$$\overline{\mathcal{E}}_\varepsilon(\gamma\mid\mu_0,\mu_1,\nu_X):=\sum_{i=0,1}\overline{\mathcal{F}}(\gamma_i\mid\mu_i)+(c,\gamma)+\varepsilon\mathcal{F}(\gamma\mid\nu_X)$$

其中

$$\begin{aligned}
\overline{\mathcal{F}}(\gamma_i\mid\mu_i)&=\int_X\overline{F}(\sigma_i)d\mu_i+F'_\infty\gamma_i^\perp(X),\\
\mathcal{F}(\gamma\mid\nu_X)&=\int_{X\times X}F(\sigma)d\nu_X+F'_\infty\gamma^\perp(X\times X).
\end{aligned}$$

第 3.1 节的所有结果均适用，因此

$$\mathrm{OT}_{X,\varepsilon}(\mu_0,\mu_1)=\inf_{\gamma\in\mathcal{M}(X\times X)}\int_{X\times X}H_\varepsilon(x_0,\rho_0(x_0),x_1,\rho_1(x_1),\varrho(x_0,x_1))d\gamma \tag{47}$$

其中，由于边缘分布的熵函数 $\overline{F}$，我们有

$$H_\varepsilon(x_0,s_0,x_1,s_1,S)=\begin{cases}ac(x_0,x_1)+\varepsilon aR\left(\frac{S}{a}\right), & \text{when } s_0=s_1=a,\\ +\infty, & \text{otherwise.}\end{cases}$$

特别地，我们注意到，要使（47）中的积分有限，γ 的边缘分布必须满足一个必要条件：对于某个常数 $a>0$，$a\gamma_i=\mu_i$，因此我们可以再次将下确界求解分为两个步骤，并观察到

$$\begin{aligned}\mathrm{OT}_{X,\varepsilon}(\mu_0,\mu_1)&=\inf_{\gamma\in\Gamma(\mu_0,\mu_1)}\left\{\inf_{a>0}\int_{X\times X}\frac{1}{a}\left(c(x_0,x_1)\varepsilon R\left(\frac{(1/a)\varrho(x_0,x_1)}{1/a}\right)\right)d(a\gamma)\right\}\\&=\inf_{\gamma\in\Gamma(\mu_0,\mu_1)}\int_{X\times X}c(x_0,x_1)+\varepsilon R(\varrho(x_0,x_1))d\gamma,\end{aligned}$$

因此，这仅需归结为通常的表述。

一旦我们提升到扩展空间，类似于第 2.3.2 节中所述的未正则化平衡最优传输的提升，这样的完全约化同样不成立。我们有

$$\mathrm{OT}_{X,\varepsilon}(\mu_0,\mu_1)=\inf_{\alpha\in S^p_{=}(\mu_0,\mu_1,\nu_X)}\int_{Y\times Y\times\mathbb{R}_+}H_\varepsilon(x_0,s_0,x_1,s_1,S)d\alpha.$$

显然，α 的支撑集满足积分有限的必要条件是 tha t

$$\alpha((Y\times Y\times\mathbb{R}_+)\setminus\Omega)=0,\quad\text{where now }\Omega=\{(y_0,y_1,S)\in Y\times Y\times\mathbb{R}_+\mid s_0=s_1\}.$$

换言之 , α 必须具有如下形式

$$\alpha=(x_0,s,x_1,s,S)_\#\beta,\quad\beta\in\mathcal{M}(X^2\times\mathbb{R}^2_+),\quad(x_0,x_1,s,S)\in X^2\times\mathbb{R}^2_+.$$

由于 $\alpha\in S^p_{=}(\mu_0,\mu_1,\nu_X)$，相应的 β 必须属于

$$\widetilde{S}^p_{=}(\mu_0,\mu_1,\nu_X):=\left\{\beta\in\mathcal{M}(X^2\times\mathbb{R}^2_+)\;\middle|\;\int_{X^2\times\mathbb{R}^2_+}s^pf(x_i)d\beta=\int_Xf(x)d\mu_i,\ \int_{X^2\times\mathbb{R}^2_+}S^pf(x_0,x_1)d\beta=\int_{X\times X}f(x_0,x_1)d\nu_X\right\}.$$

由此可直接推出。

$$\mathrm{OT}_\varepsilon(\mu_0,\mu_1)=\inf_{\beta\in\widetilde{S}^p_{=}(\mu_0,\mu_1,\nu_X)}\int_{X^2\times\mathbb{R}^2_+}s\,c(x_0,x_1)+\varepsilon sR\left(\frac{S}{s}\right)d\beta.$$

因此，对于平衡情形的熵正则化，提升策略将我们从 $\mathcal{M}(X\times X)$ 带到 $\mathcal{M}(X^2\times\mathbb{R}^2_+)$。

现在我们转而讨论扩展空间上的正则化。之后我们将说明这两种方法之间的关系。

3.2. 扩展空间上的正则化。本节讨论第 2.3 节所述非平衡最优传输（UOT）扩展空间表述的熵正则化。对于某个适当的参考测度 $\nu_Y \in \mathcal{P}(Y \times Y)$（保留下标以区别于第 3.1 节中引入的 ν_Y 和 ν_X），以及较小的 $\varepsilon > 0$，我们引入

$$\mathrm{UOT}_{Y,\varepsilon}(\mu_0, \mu_1) := \inf_{\alpha \in \overline{S}^p_=(\mu_0,\mu_1)} \Big\{ (H_p, \alpha) + \varepsilon \mathcal{F}(\alpha \mid \nu_Y) \Big\}. \tag{48}$$

据我们所知，除了近期 [Stu+23] 中一项颇有意义的研究尝试外，采用这种方法对非平衡最优传输问题进行正则化的研究几乎完全尚未开展。在该研究中，作者研究了非平衡最优传输$_{Y,\ \varepsilon}$的一种离散化简变体，讨论了该问题由此产生的托里克几何，并提出了一种用于求解该问题的概念验证数值算法，其原理类似于 Sinkhorn 算法。

现在我们将着手建立关于非平衡最优传输$_{Y\text{的}\varepsilon}$ 若干理论结果，并首先引用以下有用的标准结果。

命题 3.12（[Nut21，第 1] 节）。设 $\nu_Y \in \mathcal{P}(Y \times Y)$。该映射

$$\mathcal{P}(Y \times Y) \ni \alpha \mapsto \mathcal{F}(\alpha \mid \nu_Y)$$

在其有限定义域上是非负、凸且严格凸的，且仅当 $\alpha = \nu_Y$ 时 $\mathcal{F}(\alpha \mid \nu_Y) = 0$。此外，该映射

$$\mathcal{P}(Y \times Y) \times \mathcal{P}(Y \times Y) \ni (\alpha, \nu_Y) \mapsto \mathcal{F}(\alpha \mid \nu_Y)$$

关于弱收敛，它是联合凸且联合下半连续的。

注释 3.13。在 (48) 中，我们限制参考测度 ν_Y 属于 $\mathcal{P}(Y \times Y)$（因此特别地具有有限质量）。可以将该框架推广到 σ-有限测度（[L´eo14；GT19]），也就是存在某个可测函数 $W\colon\ Y \times Y \to \mathbb{R}$ 使得的测度。

$$z_W := \int_{Y \times Y} \exp(-W) d\nu_Y < +\infty.$$

然后，我们可以引入概率测度 $\hat{}\nu_Y := z_W^{-1} e^{-W} \nu_Y \in \mathcal{P}(Y \times Y)$，从而定义相对于 ν_Y 的相对熵。

$$\mathcal{F}(Q \mid \nu_Y) := \mathcal{F}(Q \mid \hat{\nu}_Y) - \int_{Y \times Y} W dQ - \log z_W,$$

其定义良好，当

$$Q \in \mathcal{P}_W(Y \times Y) := \mathcal{M}_W(Y \times Y) \cap \mathcal{P}(Y \times Y),$$

其中 $\mathcal{M}_W(Y \times Y)$ 由以下公式给出

$$\mathcal{M}_W(Y \times Y) := \{\alpha \in \mathcal{M}(Y \times Y) \ \mid \ (W, \alpha) < \infty\}.$$

需要牢记的一个具体示例是由 Y 上的黎曼度量导出的体积测度。

注记 3.14. 我们特别希望设定 $W(y_0,\ y_1)\colon = s_0^p + s_1^p$，以利用其与空间 $\mathcal{M}_p($ $Y \times Y)$ 之间的联系，正如平衡情形中例如[Car+17，命题 2.3 所做的那样。]，但有趣的是，这种选择与由黎曼度量在 Y上导出的体积测度不相容，因为这样我们会得到 $z_W = \infty$，除非 X 是紧的。 [这一点将在手稿的下一版本中进一步阐述。]

重要的是，与未正则化问题在定理 2.9 中给出的最终改写类似，不难看出

$$\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1) = \inf_{(\beta_0,\beta_1)} \{\mathrm{OT}_\varepsilon(\beta_0,\beta_1) \mid \beta_i \in \overline{S}^p_=(\mu_i)\}. \tag{49}$$

其中

$$\mathrm{OT}_\varepsilon(\beta_0,\beta_1) := \inf_{\alpha\in\Gamma(\beta_0,\beta_1)} \{(H_p,\alpha) + \varepsilon\mathcal{F}(\alpha \mid \nu_Y)\}. \tag{50}$$

并回顾

$$\Gamma(\beta_0,\beta_1) := \{\alpha \in \mathcal{P}(Y\times Y),\ \pi^i_\#\alpha = \beta_i\}$$

是耦合的集合。

我们先证明以下结论。

命题 3.15. 在代价函数 c 和参考测度 ν_Y 满足自然假设的条件下，对于每个 $\varepsilon>0$，存在 $\overline{\alpha}^\varepsilon \in \overline{S}^p_=(\mu_0,\mu_1)$，使得

$$\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1) = (H_p,\overline{\alpha}^\varepsilon) + \varepsilon\mathcal{F}(\overline{\alpha}^\varepsilon \mid \nu_Y) = \mathrm{OT}_\varepsilon(\overline{\alpha}^\varepsilon_0,\overline{\alpha}^\varepsilon_1),$$

其中 $\overline{\alpha}^\varepsilon_i := \pi^i_\#\overline{\alpha}^\varepsilon \in \overline{S}^p_=(\mu_i)$ 是第 i 个边缘分布。

证明。由变分法的直接方法可得此结论，因为 $(H_p,\cdot)$ 和 $\mathcal{F}(\cdot \mid \nu_Y)$ 均非负且下半连续（参见 [LMS18，第 7.1] 节及命题 3.12），并且我们可以在 $\overline{S}^p_=(\mu_0,\mu_1)$ 的一个窄紧子集中寻找解，该子集由不在 $(X\times\{0\})\times(X\times\{0\})$ 上赋予质量的测度构成，参见 [LMS18，第 7.3] 节。第二个等式由式 (49) 中非平衡最优传输的等价表述得出$_{Y,\ \varepsilon}$。□

下文还将用到以下标准结果。

命题 3.16（[Nut21，推论 5.4]）。对于任意 $\beta_0,\beta_1 \in \mathcal{P}(Y)$，式 (50) 中引入的熵正则化最优传输问题 $\mathrm{OT}_\varepsilon(\alpha_0,\alpha_1)$ 收敛到推论 2.9 中引入的 $\mathrm{OT}(\beta_0,\beta_1)$，即

$$\lim_{\varepsilon\to 0} \mathrm{OT}_\varepsilon(\beta_0,\beta_1) = \mathrm{OT}(\beta_0,\beta_1).$$

现在我们将证明，扩展空间熵正则化问题确实逼近非平衡最优传输问题。

定理 3.17. 对任意 $\mu_0,\mu_1 \in \mathcal{M}(X)$，在 (48) 中引入的 Y 空间正则化非平衡最优传输问题 $\mathrm{UOT}_{Y,\ \varepsilon}(\mu_0,\mu_1)$ 收敛于在 (3) 中引入的 $\mathrm{UOT}(\mu_0,\mu_1)$，即

$$\lim_{\varepsilon\to 0} \mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1) = \mathrm{UOT}(\mu_0,\mu_1).$$

证明. 命题 2.10 保证存在 $\overline{\alpha} \in \overline{S}^p_=(\mu_0,\mu_1)$，使得

$$\mathrm{UOT}(\mu_0,\mu_1) = (H_p,\overline{\alpha}) = \mathrm{OT}(\overline{\alpha}_0,\overline{\alpha}_1),$$

其中 $\overline{\alpha}_i := \pi^i_\#\overline{\alpha}$ 是第 i 个边缘分布。

由命题 3.16 可知，

$$\lim_{\varepsilon\to 0} \mathrm{OT}_\varepsilon(\overline{\alpha}_0,\overline{\alpha}_1) = \mathrm{OT}(\overline{\alpha}_0,\overline{\alpha}_1)$$

这意味着，当 $\varepsilon\to 0$ 时，

$$\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1) \le \mathrm{OT}_\varepsilon(\overline{\alpha}_0,\overline{\alpha}_1) \to \mathrm{OT}(\overline{\alpha}_0,\overline{\alpha}_1) = \mathrm{UOT}(\mu_0,\mu_1)$$

因此

$$\lim_{\varepsilon\to 0} \mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1) \le \mathrm{UOT}(\mu_0,\mu_1).$$

要得出这一结果，只需证明随着 ε 减小，$\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1)$ 的值单调递减，因为这将意味着

$$\lim_{\varepsilon\to 0}\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1)\geq \mathrm{UOT}(\mu_0,\mu_1),$$

正如 [Nut21 第 5.1] 节所述。

为证明这一点，我们依赖命题 3.15，该命题确立了 $\exists\,\overline{\alpha}^{\varepsilon}\in\overline{S}^{p}_{=}(\mu_0,\mu_1)$，使得

$$\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1)=(H_p,\overline{\alpha}^{\varepsilon})+\varepsilon\mathcal{F}(\overline{\alpha}^{\varepsilon}\mid\nu_Y)=\mathrm{OT}_{\varepsilon}(\overline{\alpha}^{\varepsilon}_0,\overline{\alpha}^{\varepsilon}_1).$$

取 $\varepsilon_2>\varepsilon_1>0$ 并记 $\overline{\alpha}^{i}=\overline{\alpha}^{\varepsilon_i}$。于是有

$$\begin{aligned}\mathrm{UOT}_{Y,\varepsilon_1}(\mu_0,\mu_1)&\leq(H_p,\overline{\alpha}^{2})+\varepsilon_1\mathcal{F}(\overline{\alpha}^{2}\mid\nu_Y)=\mathrm{UOT}_{Y,\varepsilon_2}(\mu_0,\mu_1)+(\varepsilon_1-\varepsilon_2)\mathcal{F}(\overline{\alpha}^{2}\mid\nu_Y)\\&\leq\mathrm{UOT}_{Y,\varepsilon_2}(\mu_0,\mu_1),\end{aligned}$$

其中，最后一个等式由 $\varepsilon_2>\varepsilon_1$ 和 $\mathcal{F}(Q\mid\nu_Y)\geq 0$ 推出（参见命题 3.12）。□

注释 3.18。齐次边缘映射 $\mathrm{h}^{p}_{i}\colon\ \mathcal{M}_2(Y\times Y)\to\mathcal{M}(X)$ 并非单射；这意味着，与其先验地固定一个参考测度 $\nu_X\in\mathcal{P}(Y\times Y)$，也可以将熵正则化引入为

$$\widehat{\mathrm{UOT}}_{Y,\varepsilon}(\mu_0,\mu_1)=\inf_{(\beta_0,\beta_1)}\{\widehat{\mathrm{OT}}_{\varepsilon}(\beta_0,\beta_1)\mid\beta_i\in S^{p}_{=}(\mu_i)\},\tag{51}$$

其中

$$\widehat{\mathrm{OT}}_{\varepsilon}(\beta_0,\beta_1):=\inf_{\alpha\in\Gamma(\beta_0,\beta_1)}\left\{(H_p,\alpha)+\varepsilon\mathcal{F}(\alpha\mid\alpha_0\otimes\alpha_1)\right\},\tag{52}$$

其等价表述为

$$\widehat{\mathrm{UOT}}_{Y,\varepsilon}(\mu_0,\mu_1)=\inf_{\alpha\in\overline{S}^{p}_{=}(\mu_0,\mu_1)}\big\{(H_p,\alpha)+\varepsilon\mathcal{F}(\alpha\mid\alpha_0\otimes\alpha_1)\big\}.$$

$(\alpha,\ \nu_Y)\mapsto\mathcal{F}(\alpha,\ \nu_Y)$ 的联合下半连续性（参见命题 3.12）确保在这种情况下，类似于定理 3.17 的收敛结果仍然成立。 [这一点将在手稿的下一个版本中进一步展开。]

[关于扩展空间正则化的理论将在论文的下一版本中依据第 4 节提出的思路进行大幅扩展。]

3.2.1. 示例：平衡最优传输的熵正则化的基于扩展空间的提升。从（33）出发，我们可以考虑 [本节将在手稿的下一版本中添加，并依赖于第 4 节中提出的思想。]

4. Discussion on deriving a unified extended framework for comparing X space regularisation and Y space regularisation

本节报告一些初步观察，探讨如何推导出一个共同框架，以便直接比较第 3.1 节讨论的非平衡最优传输$_{X,\ \varepsilon}$ 问题与第 3.2 节讨论的非平衡最优传输$_{Y,\ \varepsilon}$ 问题。这一点尤为重要，因为我们认为，建立这种直接对应关系，首先是证明数值收敛的关键。

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)\to\mathrm{UOT}(\mu_0,\mu_1)\ \text{ as }\ \varepsilon\to 0,$$

其次，这也是为了发展一套更为完善的非平衡最优传输问题熵正则化理论。最终，我们还希望将静态理论与熵正则化非平衡最优传输问题的各种可能动态表述联系起来（对此的进一步讨论见第 5 节）。

构建统一扩展框架的基本思路源于扩展描述之间的维数差异：在非平衡最优传输问题$_{X,\ \varepsilon}$ 中，如第 3.1.4 节所述，扩展表述是在一部分测度中寻找极小化子

$$\mathcal{M}(Y \times Y \times \mathbb{R}_+) \equiv \mathcal{M}(X^2 \times \mathbb{R}_+^3),$$

而在非平衡最优传输$_{Y,\ \varepsilon}$ 问题中，如第 3.2 节所述，极小化子是在以下测度中寻求的

$$\mathcal{M}(Y \times Y) \equiv \mathcal{M}(X^2 \times R_+^2).$$

我们认为，建立统一框架的一种有前景的方法，是采用连续提升这一思路，从而将这两个问题都改写为在共同的测度空间 $\mathcal{M}(X^2 \times \mathbb{R}_+^4)$,. 上的两个极小化问题。这一尝试出乎意料地棘手，因此我们首先说明如何将非平衡最优传输问题提升为 $\mathcal{M}(X^2 \times \mathbb{R}_+^3)$, 上的极小化问题，然后概述原空间正则化与扩展空间正则化的策略。[本节内容将在稿件的下一版本中大幅扩充，或推迟至后续论文中。]

4.1. 非平衡最优传输问题的进一步提升。从定理 2.6 出发，我们观察到

$$\mathrm{UOT}(\mu_0, \mu_1) = \inf_{\alpha \in S^p_{\leq}(\mu_0, \mu_1)} (H_p, \alpha) = \inf_{\alpha \in M(Y \times Y)} \sum_i \overline{\mathcal{F}}(h_i^p \alpha \mid \mu_i) + (H_p, \alpha), \tag{53}$$

这里我们回顾一下，$\overline{\mathcal{F}}$ 的定义已在第 2.3.2 节中引入。此表述可视为我们所称的二阶非平衡最优传输问题的一种平衡变体的原始问题形式。按照第 2.2 节和第 3.1 节中详述的步骤，我们可以从 (53) 出发，推导出反向表述、对偶形式、齐次形式和扩展空间表述。

可以很容易地证明，反向表述为

$$\mathrm{UOT}(\mu_0, \mu_1) = \inf_{\alpha \in M(Y \times Y)} \int_{Y \times Y} \left(\sum_i s_i^p \overline{R}(\rho_i(x_i)) + H_p \right) d\alpha,$$

其中，我们采用 $\mathsf{h}_i^p \alpha$ 与 μ_i 之间的勒贝格分解，即

$$\mathsf{h}_i^p \alpha = \sigma_i \mu_i + (\mathsf{h}_i^p \alpha)^\perp, \quad \mu_i = \rho_i \mathsf{h}_i^p \alpha + \mu_i^\perp,$$

在当前情形下，奇异部分显然实际上为零。此外，在当前情形下，显然 $\overline{R}(s) = s\overline{F}\left(\frac{1}{s}\right) = \overline{F}(s)$。

同样，依据定理 2.3，并注意到

$$(\overline{F})^*(\phi) = (\overline{R})^*(\phi) = \phi,$$

不难证明，对偶形式为

$$\mathrm{UOT}(\mu_0, \mu_1) = \sup_{\phi \in \Phi_Y} \sum_i \mu_i(\phi_i),$$

其中，关键的是，

$$\Phi_Y := \{\phi = (\phi_0, \phi_1) \in C_b(X) \times C_b(X) \mid (s_0^p \phi_0) \oplus (s_1^p \phi_1) \leq H\}.$$

因此，我们可以引入二阶边缘透视函数

$$\tilde{H} : X^2 \times \mathbb{R}_+^3 \to \mathbb{R},$$

~ ~其定义如下，省略对 $H \equiv H(x_0, x_1, s_0, s_1, w_0, w_1)$ 的依赖，

$$\begin{aligned}\tilde{H} &:= \inf_{t>0}\left\{t\left(s_0^p\overline{R}\left(\frac{w_0}{t}\right)+s_1^p\overline{R}\left(\frac{w_1}{t}\right)+H(x_0,s_0,x_1,s_1)\right)\right\}\\ &= \inf_{t>0}\left\{w_0 s_0^p\overline{F}\left(\frac{t}{w_0}\right)+w_1 s_1^p\overline{F}\left(\frac{t}{w_1}\right)+tH(x_0,s_0,x_1,s_1)\right\}\\ &= \sup_{(\phi_0,\phi_1)\in\mathbb{R}_+^2}\left\{-s_0^p w_0(\overline{F})^*(-\phi_0)-s_1^p w_1(\overline{F})^*(-\phi_1)\,\middle|\, s_0^p\phi_0+s_1^p\phi_1\le H(x_0,s_0,x_1,s_1)\right\},\end{aligned}$$

后两个等式是通过对本文第 2.2 节和第 3.1 节中所述论证作显而易见的调整得到的。

因此，可以证明齐次形式为

$$\mathrm{UOT}(\mu_0,\mu_1)=\inf_{\alpha\in M(Y\times Y)}\int_{Y\times Y}\tilde{H}(x_0,x_1,s_0,s_1,\rho_0(x_0),\rho_1(x_1))d\alpha.$$

与第 2.3.2 节一样，熵泛函 $\overline{F}$ 确保

$$\tilde{H}(x_0,x_1,s_0,s_1,w_0,w_1)=\begin{cases}aH(x_0,s_0,x_1,s_1), & \text{when } w_0=w_1=a,\\ +\infty, & \text{otherwise.}\end{cases}$$

特别地，要使积分有限，α 的齐次边缘分布必须满足一个必要条件：存在某个常数 $a>0$，使得 $a\mathrm{h}_i^p\alpha=\mu_i$。

二阶扩展形式为

$$\mathrm{UOT}(\mu_0,\mu_1)=\inf_{\eta\in\mathcal{S}_=^p(\mu_0,\mu_1)}\int_{X^2\times\mathbb{R}_+^4}\tilde{H}(x_0,x_1,s_0,s_1,w_0,w_1)d\eta,$$

其中

$$\mathcal{S}_=^p(\mu_0,\mu_1):=\left\{\eta\in M(X^2\times\mathbb{R}_+^4)\mid \pi_\#^{x_i}(s_i^p w_i\eta)=\mu_i\right\}.$$

特别地，对于任意满足 $\mu_i=\rho_i\mathrm{h}_i^p\alpha$ 的 $\alpha\in M(Y\times Y)$（如上所述，这里的设定实际上只允许密度为常数），我们有

$$(x_0,s_0,\rho_0(x_0),x_0,s_0,\rho_1(x_1))_\#\alpha\in\mathcal{S}_=^p(\mu_0,\mu_1).$$

最后，与第 2.3.2 节中所述平衡最优传输问题的提升示例相同，若要使积分有限，η 的支撑集必须满足一个必要条件：

$$\eta((X^2\times\mathbb{R}_+^4)\setminus\Omega)=0,\quad \text{where } \Omega=\{(x_0,x_1,s_0,s_1,w_0,w_1)\in X^2\times\mathbb{R}_+^4\mid w_0=w_1\}.$$

换言之，η 必须具有如下形式：

$$\eta=(x_0,x_1,s_0,s_1,w,w)_\#\xi,\quad \xi\in\mathcal{M}(X^2\times\mathbb{R}_+^3),\ (x_0,x_1,s_0,s_1,w)\in X^2\times\mathbb{R}_+^3.$$

由于 $\eta\in\mathcal{S}_=^p(\mu_0,\mu_1)$，相应的 ξ 必须属于

$$\widetilde{\mathcal{S}}_=^p(\mu_0,\mu_1):=\left\{\xi\in\mathcal{M}(X^2\times\mathbb{R}_+^3)\,\middle|\,\pi_\#^{x_i}(s^p w_i\xi)=\mu_i\right\},$$

其中，对于任意可测函数 $f:\ X\to\mathbb{R}$，

$$\int_X f(x)d\pi_\#^{x_i}(s^p w_i\xi)=\int_{X^2\times\mathbb{R}_+^3}s^p w_i f(x_i)d\xi.$$

由此很容易得出

$$\mathrm{UOT}(\mu_0,\mu_1)=\inf_{\xi\in\widetilde{\mathcal{S}}_=^p(\mu_0,\mu_1)}\int_{X^2\times\mathbb{R}_+^3}w\,H(x_0,s_0,x_1,s_1)d\xi,$$

这是一个在空间 $\mathcal{M}(X^2 \times \mathbb{R}^3_+)$ 的子集上进行的最小化问题，这正是我们要阐述的内容。

注 4.1. 一个完整的二阶非平衡最优传输问题有待未来研究，其关注点是放宽齐次边缘分布约束，即将（53）中的 $\overline{\mathcal{F}}$ 替换为一般的熵泛函 $\mathcal{F}$，例如使用 KL散度作为熵函数的熵泛函。

4.2. 将非平衡最优传输$_{Y,\varepsilon}$ 问题提升至 $\mathcal{M}(X^2 \times \mathbb{R}^4_+)$. [本节将在手稿的下一个版本中补充。基本观察是，经过显而易见的调整，上述构造同样适用于非平衡最优传输$_{Y,\varepsilon}$ 。]

4.3. 将非平衡最优传输$_{X,\varepsilon}$ 问题提升至 $\mathcal{M}(X^2 \times \mathbb{R}^4_+)$。为了在非平衡最优传输$_{X,\varepsilon}$ 问题中采用类似的提升方法，我们需要分两个步骤进行。首先，按照第 3.1.4 节中所述进行一阶提升，从而使我们能够将该问题表述为在空间 $\mathcal{M}(X^2 \times \mathbb{R}^3_+)$ 的一个子集上进行的最小化问题，即

$$\begin{aligned}\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1) &= \inf_{\eta \in S^p_=(\mu_0,\mu_1,\nu_X)} (H^\varepsilon_p, \alpha) \\ &= \inf_{\eta \in M(X^2 \times \mathbb{R}^3_+)} \sum_i \overline{\mathcal{F}}(\mathsf{h}^p_i \eta \mid \mu_i) + \overline{\mathcal{F}}(\mathsf{H}^p \eta \mid \nu_X) + (H^\varepsilon_p, \alpha), \end{aligned} \tag{54}$$

按照第 4.1 节中详述的步骤，我们可以很容易地推导出一个扩展表述，其中问题被表述为在 $\mathcal{M}(X^2 \times \mathbb{R}^6_+)$ 中某个测度子集上的最小化；不过，熵泛函 $\overline{\mathcal{F}}$ 的尖锐性确保我们可以将问题简化为在 $\mathcal{M}(X^2 \times \mathbb{R}^4_+)$ 的某个子集上进行最小化。[本节将在手稿的下一版本中大幅扩展。]

5. Discussion on the dynamic formulation

到目前为止，我们一直专注于非平衡最优传输问题的静态表述，以及对其进行熵正则化的两种根本不同的方法。在接下来的启发式讨论中，我们将尝试阐明，如何从动态表述 [BB00], [LMS18，第 8] 节出发，作出类似的区分；并说明在这种情况下，鉴于非平衡最优传输问题的整体设置，将扩展空间正则化视为更优方案似乎是自然的。

我们先简要回顾非平衡最优传输问题 $\mathrm{UOT}(\mu_0,\mu_1)$ 的动态表述，然后讨论可能的熵正则化方式。

5.1. 非平衡最优传输的动态表述。第 2.1 节中引入的非平衡最优传输问题 $\mathrm{UOT}(\mu_0,\mu_1)$ 还可以用另一种方式重新表述，通常称为动态表述。

命题 5.1（[LMS18，定理 8.18，定理 8.19]）。对于代价函数和边缘分布熵函数的特定选择，式 (3) 中的非平衡最优传输问题可表述为动态形式，即

$$\mathrm{UOT}(\mu_0,\mu_1) := \inf_{(\rho,v,r)\in \mathrm{Adm}(\mu_0,\mu_1)} \left\{ \int \int_0^1 (|v|^2 + |r|^2)\rho \, dx dt \right\} \tag{55}$$

其中，(ρ, v, r) 属于适当定义的函数空间（特别地，v 是一个向量场），

$$\mathrm{Adm}(\mu_0,\mu_1) := \{\rho, v, r \mid \partial_t \rho + \mathrm{div}(\rho v) = \rho r, \quad \rho|_{t=i} = \mu_i\}.$$

特别地，由此可知，使（55）达到最小值的三元组 (ρ, v, r) 中，向量场v 是标量场 r 的梯度，即 $v = \nabla r$。

5.2. 熵正则化非平衡最优传输问题的动态表述。促使我们开展这项工作的关键观察，源自将非平衡最优传输问题重新表述为扩展空间上的一个最优传输问题，见推论 2.9——由于在这一表述中，我们允许沿径向分量进行传输，而这在原问题中对应于质量的产生/湮灭，因此，熵正则化自然也应同时处理传输和质量的产生/湮灭（反应）。在动力学表述中，这意味着熵正则化连续性方程除了包含扩散项，还应包含熵反应项。

然而，这与已知的熵正则化非平衡最优传输的动态表述形成了对比，参见例如 [BL21], 其为

$$\mathrm{UOT}^{\mathrm{D}}_{X,\varepsilon}(\mu_0,\mu_1) := \inf_{(\rho,v,r)\in \mathrm{Adm}_{X,\varepsilon}(\mu_0,\mu_1)} \left\{ \int \int_0^1 (|v|^2+|r|^2)\rho\, dxdt \right\} \tag{56}$$

其中，对于属于适当函数空间的 (ρ, v, r)，

$$\mathrm{Adm}_{X,\varepsilon}(\mu_0,\mu_1) := \left\{ \rho, v, r \mid \partial_t\rho + \mathrm{div}(\rho v) = \rho r + \frac{\varepsilon}{2}\Delta\rho, \quad \rho|_{t=i} = \mu_i \right\}.$$

在这一表述中，所有的 ε-项都与扩散而非反应相关。该表述使人联想到原空间上的静态熵正则化，即在第 3.1 节中讨论的非平衡最优传输 $_{X,\ \varepsilon}(\mu_0, \mu_1)$，因此采用下标 X。

我们将采用与 [LLW20] 中讨论的类似论证，为 Wasserstein 梯度流构造高阶（Fisher信息）正则化，并阐述这种正则化方式如何被视为导致一阶误差。从与 $\mathrm{Adm}_{X,\varepsilon}(\mu_0, \mu_1)$相关的连续性方程出发，

$$\partial_t\rho + \mathrm{div}(\rho v) = \rho r + \frac{\varepsilon}{2}\Delta\rho$$

并利用 $\Delta\rho = \mathrm{div}(\rho\nabla \log\ \rho)$，我们定义

$$\tilde{v} := v - \frac{\varepsilon}{2}\nabla\log\rho, \quad \tilde{r} = r - \varepsilon s, \tag{57}$$

其中函数 s有待确定。此时连续性方程变为

$$\partial_t\rho + \mathrm{div}(\rho\tilde{v}) = \rho(\tilde{r} + \varepsilon s). \tag{58}$$

由此可知，对于任意三元组 $(\rho, v, r) \in \mathrm{Adm}_{X,\ \varepsilon}(\mu_0, \mu_1)$，并代入式（57）后，式（56）中的积分为

$$\begin{aligned}\int_0^1 \int_X \left(|v|^2+|r|^2\right) \mathrm{d}\rho\mathrm{d}t = \int_0^1 \int_X \left(|\tilde{v}|^2 + \frac{\varepsilon^2}{4}|\nabla\log\rho|^2 + \varepsilon\tilde{v}\cdot\nabla\log\rho\right) \mathrm{d}\rho\mathrm{d}t \\ + \int_0^1 \int_X \left(|\tilde{r}|^2 + \varepsilon^2|s|^2 + 2\varepsilon\tilde{r}s\right) \mathrm{d}\rho\mathrm{d}t.\end{aligned} \tag{59}$$

与 [LLW20], 中处理项 $\int_0^1 \int_X \varepsilon\tilde{v}\cdot\nabla \log\ \rho\mathrm{d}\rho\mathrm{d}t$ 的方法相同，我们定义 $H[\rho] := \int_X \rho\ \log\ \rho dx$，并注意到，由于 ρ 满足 (58),

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}H[\rho] &= \int (\log\rho+1)\partial_t\rho dx = \int (\log\rho+1)(-\mathrm{div}(\rho\tilde{v}) + \rho(\tilde{r}+\varepsilon s))dx \\ &= \int (\nabla\log\rho\cdot\tilde{v})\rho dx + \int (\log\rho+1)(\tilde{r}+\varepsilon s)\rho dx,\end{aligned}$$

其中，第二个等号用到了 (58)，第三个等号则用到了分部积分。因此可得

$$\varepsilon\int_0^1\int_X \tilde{v}\cdot\nabla\log\rho\mathrm{d}\rho\mathrm{d}t = \varepsilon\int_0^1\frac{\mathrm{d}}{\mathrm{d}t}H[\rho]\mathrm{d}t \\ -\varepsilon\int_0^1\int_X(\log\rho+1)\tilde{r}\rho-\varepsilon^2\int_0^1\int_X(\log\rho+1)s\rho.$$

回想一下，函数 s 尚待确定，因此自然可以设定为

$$s=\frac{1}{2}(\log\rho+1),$$

因为这会使 ε 中的一些线性项相互抵消，即（59）变为

$$\int_0^1\int_X(|v|^2+|r|^2)\mathrm{d}\rho\mathrm{d}t=\int_0^1\int_X\left(|\tilde{v}|^2+|\tilde{r}|^2+\frac{\varepsilon^2}{4}\left(|\nabla\log\rho|^2+|\log\rho+1|^2\right)\right)\mathrm{d}\rho\mathrm{d}t \\ +\varepsilon\left(H[\mu_1]-H[\mu_0]\right)$$

这非常清楚地表明，通过在（58）中反向执行代换（57），可以得到一个更精确的连续性方程，即

$$\partial_t\rho+\mathrm{div}(\rho v)=\rho r+\frac{\varepsilon}{2}\mathrm{div}(\rho\nabla\log\rho)-\frac{\varepsilon}{2}\rho(\log\rho+1). \tag{60}$$

鉴于命题 5.1 的第二部分证明了在未正则化的公式中，最小化三元组满足$v=\nabla r$，这一点并不令人意外。当我们使用 (60) 进行正则化时，这一性质仍然保持，因为当我们设定

$$\tilde{v}=v-\frac{\varepsilon}{2}\nabla\log\rho,\quad \tilde{r}=r-\frac{\varepsilon}{2}(\log\rho+1),$$

我们得到标准的连续性方程。

$$\partial_t\rho+\mathrm{div}(\rho\tilde{v})=\rho\tilde{r}$$

˜ ˜并且重要的是， $\tilde{v}=\nabla\xi$ 其中 $\xi=\xi-\frac{\varepsilon}{2}$（log $\rho+C$）对于任意常数 C 以及˜同样地， $\tilde{r}=\xi$ 对于 $C=1$。

为使上述讨论形式化，我们引入

$$\mathrm{UOT}^{\mathrm{D}}_{Y,\varepsilon}(\mu_0,\mu_1):=\inf_{(\rho,v,r)\in\mathrm{Adm}_{Y,\varepsilon}(\mu_0,\mu_1)}\left\{\int\int_0^1(|v|^2+|r|^2)\rho\,dxdt\right\} \tag{61}$$

其中，对于属于适当函数空间的（ρ， v， r），

$$\mathrm{Adm}_{Y,\varepsilon}(\mu_0,\mu_1):=\{\rho,v,r\mid\partial_t\rho+\mathrm{div}(\rho v)-\frac{\varepsilon}{2}(\rho\nabla\log\rho)=\rho r-\frac{\varepsilon}{2}(\log\rho+1),\ \rho|_{t=i}=\mu_i\}.$$

我们使用下标 Y，因为现在熵项同时出现在扩散项和反应项中，这让人联想到第 3.2 节中讨论的扩展空间静态正则化。

根据上述讨论，我们推测 $\mathrm{UOT}_{Y,\varepsilon}(\mu_0,\mu_1)$ 是 $\mathrm{UOT}(\mu_0,\mu_1)$ 的高阶近似方案；这里的含义是，当$\varepsilon\to0$时，以下内容的收敛为

$$\mathrm{UOT}^{\mathrm{D}}_{Y,\varepsilon}(\mu_0,\mu_1)\to\mathrm{UOT}(\mu_0,\mu_1)$$

渐近速度快于

$$\mathrm{UOT}^{\mathrm{D}}_{X,\varepsilon}(\mu_0,\mu_1)\to\mathrm{UOT}(\mu_0,\mu_1).$$

我们的长期目标是建立此类结果，并推导熵正则化非平衡最优传输问题对应的静态与动态表述之间的直接转换。对于平衡情形，这已在附录 A 中完成。

6. Outlook

本文所呈现的工作发展了非平衡传输问题熵正则化理论的若干方面，并重要地开辟了许多进一步理论研究的方向，下面我们将简要讨论这些方向。

高阶提升理论。正如备注 4.1 中所暗示的，我们可以合理地设想，提出并研究如下形式的高阶非平衡最优传输问题。

$$\mathrm{HUOT}(\mu_0,\mu_1)=\inf_{\alpha\in M(Y\times Y)}\sum_i \mathcal{F}(h_i^p\alpha\mid\mu_i)+(H_p,\alpha),$$

其中，我们放宽齐次边缘分布约束，允许 $\mathrm{h}_i^p\alpha$ 偏离 μ_i。这也可以应用于其熵正则化。一个可能适用的情形是为 $\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)$ 建立收敛性证明——也许在进行高阶提升后，我们必须放宽约束，否则问题可能定义不良?

更广泛地说，如果这类二阶问题（和/或其熵正则化版本）被证明可行且有趣，我们还可以迭代这一过程，研究n阶问题。

据我们所知，目前没有任何物理依据可以说明为何要提出此类高阶问题，但或许可以将其与动态 Schr¨odinger 问题联系起来：根据近期关于分枝布朗运动 [BL21] 的有趣研究，熵最小化是相对于某种随机过程而言的。

不同正则化之间的衔接。要发展出完整的理论，显而易见的下一步是明确推导原空间与扩展空间正则化的静态与动态表述之间的联系，可以通过公式或渐近结果来实现。特别是，为了实现静态与动态表述之间的衔接，我们希望借助平衡情形下有关此类转换的丰富文献（[GLR17；GT19] ，另见附录 A）。

收敛性及收敛速率的证明。我们旨在建立如下收敛结果

$$\mathrm{UOT}_{X,\varepsilon}(\mu_0,\mu_1)\to\mathrm{UOT}(\mu_0,\mu_1),\ \text{as } \varepsilon\to 0.$$

我们希望高阶提升策略能够使我们利用非平衡最优传输$_{Y,\ \varepsilon}(\mu_0,\mu_1)$ 的相应结果来实现这一目标。

基于第 5 节中的讨论，我们也希望能够建立一些关于两类正则化在 ε 中的收敛速率的结果。

新型数值算法。正如前文所述， [Stu+23] 提出了基于扩展空间正则化这一思路的数值算法，例如运用广义迭代缩放 [DR72]。我们希望，目前致力于为非平衡最优传输的熵正则化理论奠定更严谨的数学基础，能够推动进一步的发展，例如运用高阶提升的思路。我们还想指出，即使在平衡情形中，正如第 2.3.2 节所述，以及第 3.2.1 节将要详细说明的，

可以将其视为在扩展空间上逼近平衡最优传输问题的数值算法。

Appendix A. Passage between static and dynamic formulations of entropy regularised balanced optimal transport

基于原空间的最优传输问题熵正则化（见第 3.1.6 节）在文献中以多种密切相关的形式提出，并采用了不同的记号。现在我们用一套统一的记号来介绍它们。令 $X = \mathbb{R}^d$，并省略 X 下标，使记号稍微简洁一些。我们考虑三种静态表述

$$\mathrm{OT}_{\varepsilon,1}(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \{(c,\gamma) + \varepsilon H(\gamma)\} \tag{62}$$

$$\mathrm{OT}_{\varepsilon,2}(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \{(c,\gamma) + \varepsilon H(\gamma \mid \mu\otimes\nu)\} \tag{63}$$

$$\mathrm{OT}_{\varepsilon,3}(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)} \{\varepsilon H(\gamma \mid K)\} \tag{64}$$

其中 $H(\cdot \mid \cdot)$ 是相对熵的通常定义，而 $H(\cdot) = H(\cdot \mid \mathcal{L})$ 是相对于勒贝格测度的相对熵，配对定义为

$$(c,\gamma) := \int c(x,y)d\gamma(x,y), \quad c(x,y) = |x-y|^2,$$

而测度 K 对应于在时刻 $t = \varepsilon/2$ 取值的热核，并且 is 定义为

$$dK(x,y) = (2\pi\varepsilon)^{-d/2} \exp\left(\frac{-c(x,y)}{2\varepsilon}\right) dxdy.$$

形式 $W_{\varepsilon,\,1}$ 与 [Car+17] 中相同，$W_{\varepsilon,\,2}$ 与 [Nut21] 中相同，也与 [Chi +20] 中相同（但将 ε 替换为 $\lambda = \varepsilon/2$；在协调这些表述时，这个前因子将证明至关重要）。形式 $W_{\varepsilon,\,3}$ 用于 [GT19]，以确立静态与动态表述之间的明确联系。

直接计算可得，

$$W^2_{\varepsilon,2}(\mu,\nu) = W^2_{\varepsilon,1}(\mu,\nu) - \varepsilon\left(H(\mu) + H(\nu)\right)$$
$$W^2_{\varepsilon,3}(\mu,\nu) = \frac{1}{2}W^2_{2\varepsilon,1}(\mu,\nu) + \frac{d\varepsilon}{2}\log(2\pi\varepsilon),$$

因此，我们可以在所有变体之间建立清晰的对应关系（注意第二个等式中下标由 ε 变为 2ε）。

就动态表述而言，通常采用两种变体。 d, 即 乐彩

$$\widetilde{W}^2_{\varepsilon,a}(\mu,\nu) := \inf_{\rho,v}\Big\{\int_0^1\int |v|^2 d\rho(x)dt \mid \partial_t\rho + \mathrm{div}(\rho v) = \frac{\varepsilon}{2}\Delta\rho,\ \rho_0 = \mu,\ \rho_1 = \nu\Big\}$$

$$\widetilde{W}^2_{\varepsilon,b}(\mu,\nu) := \inf_{\rho,v}\Big\{\int_0^1\int \Big(|v|^2 + \frac{\varepsilon^2}{4}|\nabla\log\rho|^2\Big)d\rho(x)dt$$
$$\partial_t\rho + \mathrm{div}(\rho v) = 0,\ \rho_0 = \mu,\ \rho_1 = \nu\Big\}$$

而根据 [LLW20] 中的方法进行直接计算可得

$$\widetilde{W}^2_{\varepsilon,a}(\mu,\nu) = \widetilde{W}^2_{\varepsilon,b}(\mu,\nu) + \varepsilon\left(H(\nu) - H(\mu)\right).$$

由 [GT19，命题 3.6] 可得。

$$\begin{aligned} W_{\varepsilon,3}^2(\mu,\nu) &= \frac{1}{2}\widetilde{W}_{\varepsilon,b}^2(\mu,\nu) + \frac{\varepsilon}{2}\left(H(\mu)+H(\nu)\right) \\ &= \frac{1}{2}\widetilde{W}_{\varepsilon,a}^2(\mu,\nu) + \varepsilon H(\mu) \end{aligned}$$

因此，两种最常用变体之间的公式为

$$W_{\varepsilon,1}^2(\mu,\nu) = \widetilde{W}_{\frac{\varepsilon}{2},a}^2(\mu,\nu) + \varepsilon H(\mu) - \frac{d\varepsilon}{2}\log(\pi\varepsilon)$$

这与 [Chi+20] 一致。

References

[Roc67] R. Rockafellar.《涉及凸函数的极值问题中的对偶性与稳定性》。载于：Pacific J ournal of Mathematics 21.1（1967），第 167–187 页。[DR72] J. N. Darroch 和 D. Ratcliff.《对数线性模型的广义迭代尺度法》。载于：The annals of mathematical statistics（1972），第 1470–1480 页。[BB00] J.-D. Benamou 和 Y. Brenier.《Monge-Kantorovich 质量传输问题的计算流体力学解法》。载于：Numerische Mathematik 84.3（2000 年 1 月），第 375–393 页。 DOI: 10.1007/s002110050002。[L´eo14] C. L´eonard.《路径测度的一些性质》。载于：S´ eminaire de Prob-abilit´es XLVI（2014），第 207–230 页。[Chi+16] L. Chizat 等。《最优传输与 Fisher–Rao 度量之间的一种插值距离》。载于：Foundations of Computational Mathematics 18.1（2016 年 10 月），第 1–44 页。 DOI: 10.1007/s10208-016-9331-y。[KMV 16] S. Kondratyev、L. Monsaingeon 和 D. Vorotnikov.《有限 Radon 测度空间上的一种新最优传输距离》。载于：Advances in Differential Equations 21.11/12（2016），第 1117–1164 页。 DOI: 10.57262/ade/1476369298。[Car+17] G. Carlier 等。《最优传输与梯度流熵方案的收敛》。载于：SIAM Journal on Mathematical Analysis 49.2（2017 年 1 月），第 1385–1418 页。 DOI: 10.1137/15m1050264。[GLR17] I. Gentil、C. L´eonard 和 L. Ripani.《论最优传输与最小熵之间的类比》。载于：Annales de la Facult´e des sciences de Toulouse: Math´ ematiques 26.3（2017），第 569–600 页。[Chi+18] L. Chizat 等。《非平衡最优传输问题的缩放算法》。载于：Mathematics of Computation 87.314（2018），第 2563–2609 页。 DOI: 10.1090/mcom/3303。[LMS18] M. Liero、A. Mielke 和 G. Savar´e.《最优熵传输问题及正测度之间的一种新 Hellinger–Kantorovich 距离》。载于：Inventiones mathematicae 211.3（ 2018），第 969–1117 页。 DOI: 10.1007/s00222-017-0759-8。[GT19] N. Gigli 和 L. Tamanini.《($RCD^*(K,N)$ 空间上熵成本的 Benamou-Brenier 公式与对偶公式》。载于：Probability Theory and Related Fields 176.1-2（2019 年 4 月），第 1–34 页。 DOI: 10.1007/s00440-019-00909-1。[S´ej+19] T. S´ejourn´e 等。《用于非平衡最优传输的 Sinkhorn 散度》。载于：arXiv 预印本 arXiv:1910.12958（2019）。

[Chi+20] L. Chizat 等。《利用 Sinkhorn 散度加速 Wasserstein 距离估计》。载于：神经信息处理系统进展 33（2020），第 2257–2269 页。[LLW20] W. Li、J. Lu 和 L. Wang.《用于 Wasserstein 梯度流的 Fisher 信息正则化方案》。载于：计算物理学杂志 416（2020），第 109449 页。[BL21] A. Baradat 和 H. Lavenant.《将正则化非平衡最优传输视为相对于分枝布朗运动的熵最小化》。载于：arXiv 预印本 arXiv:2111.01666（2021）。[Nut21] M. Nutz.《熵正则最优传输导论》。2021。[Lac22] T. Lacombe.《一种齐次非平衡正则化最优传输模型及其在带边界的最优传输中的应用》。载于：arXiv 预印本 arXiv:2201.02082（2022）。[SPV22] T. S´ejourn´e、G. Peyr´e 和 F.-X. Vialard.《非平衡最优传输：从理论到数值计算》。载于：arXiv 预印本 arXiv:2211.08775（2022）

[Stu+23] B. Sturmfelset al.《熵正则化的托里克几何》。载于：符号计算杂志（2023），第 102221 页。

School of Mathematical and Computer Sciences, Heriot-Watt University, Edin-burgh, Scotland, EH14 4AS, United Kingdom*Email address*: m.buze@hw.ac.uk

School of Mathematics, Watson Building, University of Birmingham, Edgbaston,Birmingham, B15 2TT UK*Email address*: h.duong@bham.ac.uk