FASTER ALGORITHMS FOR MULTIMARGINAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study constructive discrete multimarginal optimal transport (MOT) among m distributions on n points, where the cost tensor has $N = n^m$ entries. Let $\kappa = \max\{1, (\max C - \min C)/\varepsilon\}$. Classically, we give two algorithms that return exactly feasible additive-ε couplings. A deterministic box–simplex method with rounding runs in $O(m^2 N\kappa \log N) = \widetilde{O}(m^3 N\kappa)$ time, while an exact reduction to positive packing followed by rank-one completion runs in randomized time $\widetilde{O}(m^2 N\kappa)$. At fixed κ, both match the $\Omega(N)$ cost of writing a dense coupling, up to factors in m and logarithms. Quantumly, in the general entry-access model, tensor Sinkhorn followed by sparse recovery returns an exactly feasible additive-ε coupling as a classical list of $\widetilde{O}(m^2 n\kappa^2)$ atoms, using $\widetilde{O}(m^4\sqrt{Nn}\,\kappa^3)$ coherent cost queries without materializing the tensor. Finally, for fixed m and constant normalized accuracy, we prove sparse-output lower bounds of $\widetilde{\Omega}(N)$ randomized classical queries, even with unrestricted output size, and $\widetilde{\Omega}(\sqrt{Nn})$ quantum queries for outputs with at most $n\,\mathrm{polylog}(n)$ atoms. Hence, for $\kappa \le \mathrm{polylog}(n)$, explicit sparse MOT construction has tight query complexities $\widetilde{\Theta}(N)$ classically and $\widetilde{\Theta}(\sqrt{Nn})$ quantumly.

1 INTRODUCTION

Optimal transport (OT) compares two probability measures by asking how cheaply mass can be rearranged from one to the other (Villani, 2003). A transport plan is a joint distribution with the prescribed marginals, and OT seeks the plan that minimizes the expected transportation cost. When the cost is a suitable power of the distance between support points, one obtains the Wasserstein distance. Today, OT is a standard tool in computer vision (Rubner et al., 2000; Bonneel et al., 2011; Solomon et al., 2015), statistics (Flamary et al., 2018; Panaretos & Zemel, 2016; Székely et al., 2004; Bigot et al., 2013), and machine learning (Sandler & Lindenbaum, 2011; Mueller & Jaakkola, 2015; Lloyd & Ghahramani, 2015; Jitkrittum et al., 2016; Arjovsky et al., 2017).

Multimarginal optimal transport (MOT) asks the same question for m measures at once. Given discrete marginals μ_k on $n_k \le n$ points and a cost tensor $C \in \mathbb{R}^{n_1 \times \cdots \times n_m}$, it solves

$$\mathsf{MOT}(C, \boldsymbol{\mu}) = \min_{X \in \mathcal{U}(\boldsymbol{\mu})} \langle C, X \rangle, \quad \mathcal{U}(\boldsymbol{\mu}) = \{X \ge 0 : \mathrm{Proj}_k(X) = \mu_k \ \forall k \in [m]\}. \tag{MOT}$$

Here $\mathrm{Proj}_k(X)$ is the kth marginal. Every feasible X has unit mass, so adding a constant to every entry of C changes all feasible costs by the same amount. The relevant scale for additive accuracy ε is therefore the range $\Delta_C = \max C - \min C$. We write $\kappa = \max\{1, \Delta_C/\varepsilon\}$ for the normalized inverse accuracy.

MOT underlies Wasserstein barycenters (Agueh & Carlier, 2011; Cuturi & Doucet, 2014; Benamou et al., 2015; Carlier et al., 2015; Srivastava et al., 2018), density-functional models in statistical physics (Seidl et al., 2007; Buttazzo et al., 2012; Cotar et al., 2013; Mendl & Lin, 2013), matching models in economics (Ekeland, 2005; Carlier & Ekeland, 2010), robust pricing and hedging in quantitative finance (Dolinsky & Soner, 2014; Galichon et al., 2014; Henry-Labordère, 2017), and multi-domain generative modeling (Choi et al., 2018; Cao et al., 2019).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The computational price of adding marginals is severe. If all supports have size n, then a coupling has $N = n^m$ entries. Merely writing a dense answer takes $\Omega(N)$ time. For $m = 2$, discrete OT is a minimum-cost flow problem and admits specialized algorithms (Schrijver, 2003; Orlin, 1997; Lee & Sidford, 2014). That structure disappears for $m \geq 3$ (Lin et al., 2022). As an ancillary observation, we record a rank-sensitive generic-LP baseline that, to our knowledge, has not appeared in prior MOT complexity comparisons. Writing the marginal constraints as $A\,\mathrm{vec}(X) = b$, one has $\mathrm{rank}(A) = mn - m + 1$; see Lemma B.2. Applying the method of Lee & Sidford (2019) to this formulation gives an exactly feasible ε-optimal coupling in $\widetilde{O}(\mathrm{poly}(m)N\sqrt{n})$ arithmetic operations (Remark B.8), with only polylogarithmic dependence on κ and the inverse smallest positive marginal mass. For context, a direct application of current-matrix-multiplication-time LP solvers gives the much larger bound $\widetilde{O}((N + mn)^\omega)$ (Cohen et al., 2021; van den Brand, 2020), where ω denotes the matrix-multiplication exponent. Throughout, $\widetilde{O}(\cdot)$ hides polylogarithmic factors in m, n, κ, and failure probability.

For large-scale OT and MOT, first-order methods are often the most practical alternative to IPMs. Sinkhorn-type scaling is especially prominent: its iterations use only marginalization and elementwise rescaling, and its runtime can be linear in the explicit tensor size N, up to factors in m and logarithms (Cuturi, 2013; Lin et al., 2022); at fixed regularization, it and its greedy variants converge linearly (Carlier, 2022; Kostic et al., 2022). Existing MOT bounds trade dependence on N against dependence on κ. Accelerated alternating minimization (Tupitsa et al., 2020) is linear[1] in κ but costs $\widetilde{O}(m^3 N\sqrt{n}\,\kappa)$, whereas multimarginal Sinkhorn (Lin et al., 2022) is linear in N but quadratic in κ. Accelerated Sinkhorn (Lin et al., 2022) interpolates with runtime $\widetilde{O}(m^3 N n^{1/3}\kappa^{4/3})$. Our central classical question is thus:

Is there a constructive algorithm for MOT with $\widetilde{O}(\mathrm{poly}(m)N\kappa)$ arithmetic cost?

We give two routes linear in both N and κ. The first route starts from the simplex of all nonnegative unit-mass tensors and penalizes the total marginal residual $E(X) = \sum_k \|\,\mathrm{Proj}_k(X) - \mu_k\|_1$ by $2\Delta_C E(X)$. This penalty is the support function of an mn-dimensional box, so the problem is a box–simplex game. We first compute an approximate saddle point, which balances transport cost against marginal violation. We then round its tensor component to exact feasibility. Multimarginal rounding changes the tensor by at most $2E(X)$ in ℓ_1 (Lin et al., 2022); because the shifted cost lies in $[0, \Delta_C]$, its cost increase is at most $2\Delta_C E(X)$, exactly the amount charged by the penalty. Applying Jambulapati & Tian (2023) therefore gives a deterministic, exactly feasible algorithm in $O(m^2 N\kappa \log N) = O(m^3 N\kappa \log n)$ work.

The second route uses positive packing and rank-one completion, giving the sharper randomized bound

$$O(N + mn + mN\kappa \log(mnN\kappa)\log(e\kappa)) = \widetilde{O}(m^2 N\kappa).$$

Thus both routes match the $\Omega(N)$ dense-output cost at fixed κ, while packing saves one explicit factor of m. It also improves on the rank-sensitive IPM bound for $\kappa \lesssim \sqrt{n}$; the IPM is preferable at higher accuracy (Remark B.8).

We obtain this bound by reducing MOT to a positive packing LP, for which nearly-linear first-order solvers are available (Allen-Zhu & Orecchia, 2019). Let $U = \max_{\boldsymbol{i}} C_{\boldsymbol{i}}$ and replace the cost by the nonnegative benefit tensor $B = U\mathbf{1} - C$. Relaxing the marginal constraints from $A\,\mathrm{vec}(X) = b$ to $A\,\mathrm{vec}(X) \leq b$ turns a feasible packing solution into a partial coupling. All of its marginal deficits have the same total mass, so their appropriately normalized tensor product fills every deficit simultaneously and completes the coupling exactly.

The shift from cost to benefit is what protects the objective during this completion. Because $B \geq 0$, adding the nonnegative completion tensor cannot decrease the packing benefit. Once the completed tensor has unit mass, its benefit is exactly U minus its original transport cost. Consequently, the packing optimum equals $U - \mathsf{MOT}(C, \boldsymbol{\mu})$, and a relative packing approximation translates directly into additive MOT accuracy. This extends the bimarginal reductions of Blanchet et al. (2024) and Quanrud (2019) to arbitrary m.

[1] These bounds are commonly stated for nonnegative costs using $\max\{1, \|C\|_\infty/\varepsilon\}$; after the harmless shift $C \mapsto C - (\min C)\mathbf{1}$, this parameter is exactly κ.

Table 1: Classical arithmetic complexity for an ε-optimal coupling.

Method	Reference	Arithmetic runtime
Interior point	Lee & Sidford (2019)	$\widetilde{O}(\mathrm{poly}(m)\, n^{m+1/2})$
Alternating minimization	Tupitsa et al. (2020)	$\widetilde{O}(m^3 n^{m+1/2} \kappa)$
Sinkhorn	Lin et al. (2022)	$\widetilde{O}(m^3 n^m \kappa^2)$
Accelerated Sinkhorn	Lin et al. (2022)	$\widetilde{O}(m^3 n^{m+1/3} \kappa^{4/3})$
Box–simplex + rounding	**Theorem. 2.1**	$\boldsymbol{\widetilde{O}(m^3 n^m \kappa)}$
Packing + completion	**Theorem 2.3**	$\boldsymbol{\widetilde{O}(m^2 n^m \kappa)}$

Quantum algorithms make a different goal possible. Quantum speedups for linear programming (van Apeldoorn & Gilyén, 2019; Bouland et al., 2023; Apers & Gribling, 2023; Gao et al., 2023; Kim et al., 2026) and matrix scaling (van Apeldoorn et al., 2021; Gribling & Nieuwboer, 2022; Nieuwboer, 2024) sometimes run sublinearly in the ambient dimension, making a genuine quantum–classical separation conceivable. But no existing quantum approach settles this question for constructive MOT. Quantum LP solvers retain polynomial overhead in the mn dual variables and do not directly return a feasible coupling. Quantum scaling algorithms are currently limited to matrices and, when applied to OT, compute scaling potentials that implicitly specify an approximately feasible solution to entropy-regularized OT. Standard rounding to an explicit, exactly feasible coupling for unregularized OT materializes the dense plan and loses the sublinear advantage. The quantum question is thus:

Is there a constructive, sublinear-in-N quantum algorithm for MOT?

We answer this question affirmatively in two stages. The first extends quantum Sinkhorn from matrices to order-m tensors. Each update estimates n fiber sums, with n^{m-1} terms per fiber, by a quantum log-sum-exp routine (Gribling & Nieuwboer, 2022; Nieuwboer, 2024). The estimates are noisy and depend on all previous estimates. As a result, the usual exact Sinkhorn identity no longer proves convergence. We replace it with a noisy one-step descent bound and a drift argument showing that most iterates have small marginal error.

The resulting potentials still describe only an implicit, approximately feasible Gibbs tensor. To obtain an actual coupling, we first trim that tensor below every target marginal. We then sample from the trimmed tensor by amplitude amplification, trim the empirical sample once more, and greedily couple the remaining deficits. The final list has exact marginals and can be reused without further cost queries. The whole procedure uses $\widetilde{O}(m^4\sqrt{Nn}\,\kappa^3)$ coherent cost queries and never writes an N-entry tensor.

Two obstacles prevent this result from following directly from prior work. Quantum matrix-scaling algorithms treat two marginals, whose row and column sums have n terms; tensor fibers contain n^{m-1} entries, and their adaptive estimation errors require the noisy randomized descent analysis developed here. Conversely, the multimarginal rounding of Lin et al. (2022) first writes the Gibbs tensor, so it cannot preserve sublinear query complexity. Our trim–sample–complete conversion replaces that dense step by a sparse classical output while restoring the marginals exactly.

These classical and quantum upper bounds do not by themselves settle the intrinsic cost of constructive MOT. The classical method writes a dense tensor, so its linear dependence on N might reflect only serialization. Yet every MOT instance has an optimal coupling with at most $\sum_k n_k - m + 1$ nonzero entries (Lemma B.2). We therefore permit algorithms that return only a sparse list of locations and weights. Under this more permissive output model, we ask the following:

Can our classical and quantum upper bounds be improved in a sparse-output model?

Every MOT instance has an optimal coupling with at most $\sum_k n_k - m + 1$ atoms, so the $\Omega(N)$ cost of writing a dense tensor gives no lower bound for sparse outputs. We prove lower bounds for the sparse task itself. For every fixed m and constant accuracy, there are Boolean-cost instances with uniform marginals on which any randomized classical algorithm needs $\Omega(N/\log^2 n)$ queries, whatever the size of its output, and any quantum algorithm returning at most $n\,\mathrm{polylog}(n)$ atoms

Table 2: Output- and model-specific efficiency estimates; query rows count only cost-entry queries. The last two columns hold for fixed m and $\kappa \le \mathrm{polylog}(n)$.

Output	Model and resource	Upper bound	Lower bound	Tight scale
Dense tensor	Classical arithmetic time	$\widetilde{O}(m^2 N\kappa)$	$\Omega(N)$	$\widetilde{\boldsymbol{\Theta}}(\boldsymbol{N})$
Sparse list	Classical cost-entry queries	N	$\Omega(N/\log^2 n)$	$\widetilde{\boldsymbol{\Theta}}(\boldsymbol{N})$
Sparse list	Quantum cost queries	$\widetilde{O}(m^4\sqrt{Nn}\,\kappa^3)$	$\Omega(\sqrt{Nn}/\log n)$	$\widetilde{\boldsymbol{\Theta}}(\boldsymbol{\sqrt{Nn}})$

needs $\Omega(\sqrt{Nn}/\log n)$ queries. Both bounds persist when exact feasibility is relaxed to marginals with small constant ℓ_1 error. Thus, even for fixed m and $\kappa \le \mathrm{polylog}(n)$, the upper and lower query exponents match up to polylogarithmic factors.

Our results characterize constructive MOT for arbitrary cost tensors in the general entry-access model, with exact marginals and explicit coupling output. This worst-case setting imposes no graphical, low-rank, or other succinct structure on the cost. For coherently evaluable structured costs, the quantum algorithm can be implemented without materializing the tensor; such structure may also enable specialized classical methods (Altschuler & Boix-Adserà, 2023; 2021). We state the separation in cost-query complexity and separately account for the coherent interfaces and compiled gate complexity.

The rest of the paper is organized as follows. Section 2 gives the classical algorithms, Section 3 the quantum algorithm, and Section 4 the lower bounds; Section 5 discusses limitations. The appendices contain complete proofs and spell out the computational assumptions, input-access models, finite-precision implementation, and lower-bound reductions.

2 Near-linear classical algorithms for MOT

We give two routes with work linear in both N and κ. Box–simplex keeps unit mass, penalizes marginal violations, and rounds an approximate saddle point back to exact feasibility. Packing relaxes the marginal equalities to inequalities and completes the resulting partial coupling.

2.1 Box–simplex and rounding

The first route preserves unit mass but penalizes marginal error. Let $A \in \{0,1\}^{mn\times N}$ be the marginal-incidence matrix, let b stack the target marginals, and shift the cost to $C^+ = C - (\min C)\mathbf{1} \in [0, \Delta_C]^{[n]^m}$. Consider

$$\min_{X\in\Delta^N} \max_{z\in[-1,1]^{mn}} \left\{ \langle C^+, X\rangle + 2\Delta_C z^\top (A\,\mathrm{vec}(X) - b) \right\}. \tag{1}$$

For fixed X, maximizing over the box gives $\langle C^+, X\rangle + 2\Delta_C E(X)$, where $E(X) = \sum_k \|\,\mathrm{Proj}_k(X) - \mu_k\|_1$. If $(\bar X, \bar z)$ has saddle gap at most ε, comparison with an optimal feasible coupling gives

$$\langle C^+, \bar X\rangle + 2\Delta_C E(\bar X) \le \mathsf{MOT}(C^+, \boldsymbol{\mu}) + \varepsilon. \tag{2}$$

The multimarginal rounding algorithm of Lin et al. (2022, Theorem 4.4) returns $Y \in \mathcal{U}(\boldsymbol{\mu})$ with $\|Y - \bar X\|_1 \le 2E(\bar X)$. Since $0 \le C^+ \le \Delta_C$, rounding increases cost by at most $2\Delta_C E(\bar X)$, which is already paid for in (2). Thus Y is exactly feasible and ε-optimal.

Theorem 2.1 (Deterministic box–simplex MOT)**.** *For every $\varepsilon > 0$, there is a deterministic algorithm that returns $Y \in \mathcal{U}(\boldsymbol{\mu})$ satisfying $\langle C, Y\rangle \le \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$ in $O(m^2 N\kappa \log N + mN) = \widetilde{O}(m^3 N\kappa)$ arithmetic operations. The saddle-point solver uses $O(m\kappa \log N)$ sequential matrix–vector rounds, each with $O(mN)$ work.*

After negating the objective and exchanging the players, (1) has the form solved by Jambulapati & Tian (2023, Theorem 1). Its game matrix is $-2\Delta_C A$: it has mN nonzeros and induced $1 \to 1$ norm $2m\Delta_C$. The cited theorem therefore gives

$$O\left((mN)(2m\Delta_C)\frac{\log N}{\varepsilon}\right) = O(m^2 N\kappa \log N)$$

work and $O(m\kappa \log N)$ matrix–vector rounds. Products by A and $A^\top$, as well as multimarginal rounding, cost $O(mN)$ work. Since $\log N = m \log n$, no polynomial dependence on n is hidden outside N. Appendix B.5 gives the complete proof.

2.2 Packing and rank-one completion

The deterministic route retains unit mass throughout. Relaxing the marginal equalities instead yields the sharper dependence on m. We extend the bimarginal packing reductions of Blanchet et al. (2024) and Quanrud (2019) to arbitrary m. For two marginals, the completion couples the row and column deficits by a normalized outer product. A partial multimarginal plan has m deficit vectors of equal mass, and their normalized tensor product repairs all m marginals simultaneously. This identity makes the packing relaxation exact, converts relative packing error into additive transport error without an additional rounding loss, and exposes an LP with only mN nonzeros.

Let $L = \min_{\boldsymbol{i}} C_{\boldsymbol{i}}$, $U = \max_{\boldsymbol{i}} C_{\boldsymbol{i}}$, $\Delta_C = U - L$, and $B = U\mathbf{1} - C \geq 0$. Consider the packing problem

$$P^\star = \max_{X \geq 0} \{\langle B, X\rangle : \mathrm{Proj}_k(X) \leq \mu_k \ \forall k \in [m]\}. \tag{3}$$

For every coupling X, the unit-mass identity gives $\langle B, X\rangle = U - \langle C, X\rangle$. Since every coupling is feasible for (3), this immediately yields $P^\star \geq U - \mathsf{MOT}(C, \boldsymbol{\mu})$. To prove equality of the optimal values, it remains to establish the reverse inequality $P^\star \leq U - \mathsf{MOT}(C, \boldsymbol{\mu})$. It suffices to show that every packing-feasible partial plan can be completed to a coupling without decreasing its shifted objective. A rank-one deficit correction is also the final step in the multimarginal Sinkhorn rounding of Lin et al. (2022); their input may have marginal excesses and is first scaled to a subcoupling. Packing feasibility gives these submarginal inequalities immediately, while $B \geq 0$ makes direct completion preserve the packing guarantee and yields exact value correspondence. The packing solver itself is used as a black box (Allen-Zhu & Orecchia, 2019).

The next lemma isolates the two facts needed by the algorithm: the deficits admit one common rank-one completion, and this completion converts relative packing accuracy into additive transport accuracy without a second rounding loss.

Lemma 2.2 (Exact completion and value correspondence)**.** *For every packing-feasible X, let $d_k = \mu_k - \mathrm{Proj}_k(X)$. All d_k have the same mass $\tau = 1 - |X|$, and*

$$D = \begin{cases} \tau^{1-m} d_1 \otimes \cdots \otimes d_m, & \tau > 0, \\ 0, & \tau = 0 \end{cases} \tag{4}$$

satisfies $\mathrm{Proj}_k(D) = d_k$ for every k. Thus $Z = X + D \in \mathcal{U}(\boldsymbol{\mu})$. Moreover, $P^\star = U - \mathsf{MOT}(C, \boldsymbol{\mu})$. If $\langle B, X\rangle \geq (1 - \delta_{\mathrm{P}})P^\star$, its completion obeys $\langle C, Z\rangle - \mathsf{MOT}(C, \boldsymbol{\mu}) \leq \delta_{\mathrm{P}} \Delta_C$.

Every marginal of a nonnegative tensor X has mass $|X|$, so all deficits have the common mass $\tau = 1 - |X|$. Taking the kth marginal of (4) gives $\tau^{1-m} d_{k,i_k} \prod_{\ell \neq k} \|d_\ell\|_1 = d_{k,i_k}$. Thus the same tensor product repairs all m marginals simultaneously.

Feasibility alone would not protect the objective: adding mass can be expensive under C. This is why we work with the shifted benefit B. It is nonnegative, so completion can only improve the packing objective. Together with the unit-mass identity, this gives

$$\begin{aligned} \langle B, X\rangle \leq \langle B, X + D\rangle = U - \langle C, X + D\rangle \leq U - \mathsf{MOT}(C, \boldsymbol{\mu}), \\ \langle C, Z\rangle - \mathsf{MOT}(C, \boldsymbol{\mu}) = P^\star - \langle B, Z\rangle \leq \delta_{\mathrm{P}} P^\star \leq \delta_{\mathrm{P}} \Delta_C. \end{aligned} \tag{5}$$

An optimal MOT coupling attains the first line's upper bound, proving exact value correspondence; the second line holds whenever $\langle B, X\rangle \geq (1 - \delta_{\mathrm{P}})P^\star$ and uses $P^\star \leq U - L = \Delta_C$. Thus the completion incurs neither marginal error nor additional objective loss. Appendix B proves these claims and the stronger unequal-support theorem used below.

Theorem 2.3 (Near-linear classical MOT)**.** *For every $\varepsilon > 0$ and $\rho \in (0, 1/2)$, let $\Delta_C = \max C - \min C$ and $\kappa = \max\{1, \Delta_C/\varepsilon\}$. There is a randomized algorithm that, with probability at least $1 - \rho$, returns $Z \in \mathcal{U}(\boldsymbol{\mu})$ satisfying $\langle C, Z\rangle \leq \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$. Its arithmetic runtime is $\widetilde{O}(m^2 n^m \kappa \log(1/\rho))$.*

Algorithm 1 in Appendix B gives the packing and completion pseudocode. The result is obtained as follows. For $\varepsilon < \Delta_C$, set $\delta_{\mathrm{P}} = \varepsilon/\Delta_C$, run the positive packing solver of Allen-Zhu & Orecchia (2019), and apply (4); if $\varepsilon \geq \Delta_C$, any feasible coupling is already accurate enough. Each tensor entry participates in exactly one constraint per marginal, so the marginal-incidence matrix has exactly mn^m nonzeros. The packing call is nearly linear in this representation and in δ_{P}^{-1}, while computing the deficits and writing the completed tensor takes $O(mn^m)$ work. The result has exact marginals, and accuracy dependence governed by the translation-invariant range Δ_C. For fixed normalized accuracy, its n^m dependence matches the cost of writing the dense output. Theorem B.7 in Appendix B is the stronger unequal-support result with exact logarithmic factors; Theorem 2.3 is its equal-support specialization.

3 Quantum Sinkhorn and Sparse Coupling Recovery

Writing all $N = n^m$ entries already costs N time, so a sublinear quantum algorithm must return a smaller object. We use a classical list $\{(\boldsymbol{i}_r, w_r) : r \in [s']\}$ representing $X = \sum_r w_r e_{\boldsymbol{i}_r}$. Here "sparse" means list-form rather than dense-tensor output; the list is sublinear in N only in the regime stated after Theorem 3.3. The list provides an ε-optimal coupling: $w_r \geq 0$, $\mathrm{Proj}_k(X) = \mu_k$ for every k, and $\langle C, X\rangle \leq \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$. Exact feasibility holds in exact arithmetic on the input masses, because every output weight lies on a common finite rational grid (Appendix D).

A value or dual certificate reports only the optimum, and scaling potentials describe a Gibbs law implicitly. A list is different: after $O(s')$ preprocessing it supports repeated sampling, statistics $\sum_r w_r f(\boldsymbol{i}_r)$, and arbitrary pushforwards without further cost queries. For example, pushing its atoms through a barycentric map gives a discrete Wasserstein barycenter (Agueh & Carlier, 2011; Carlier et al., 2015); in dependence models it records a joint law on which new payoffs can be evaluated (Dolinsky & Soner, 2014; Galichon et al., 2014). This reusability makes exact sparse recovery a distinct algorithmic problem, not an output-formatting step, as also emphasized for structured costs by Altschuler & Boix-Adserà (2023).

Access and output model

Costs, marginals, and accuracy are given as finite binary numbers, and exact feasibility refers to these input marginals. We assume coherent cost access

$$\mathcal{O}_C : |\boldsymbol{i}\rangle\,|w\rangle \longmapsto |\boldsymbol{i}\rangle\,|w \oplus C_{\boldsymbol{i}}\rangle\,. \tag{6}$$

We also assume coherent lookup of the $O(mn)$ entries of the classical working vectors (potentials and trimming factors) and marginals, and preparation of $\sum_a \sqrt{\mu_{k,a}}\,|a\rangle$. The stated bounds count calls to these interfaces, and we separately account for their compiled gate costs. A coherently evaluable structured cost can implement (6) coherently without storing all n^m entries. Appendix C.1 gives the exact encoding and approximation conventions, and Remark D.5 gives compiled gate costs, including ordinary-table access.

3.1 Stage I: quantum multimarginal Sinkhorn

Sinkhorn is useful here because entropy regularization compresses an N-entry plan into mn scaling potentials. For $\eta > 0$ and $H(X) \triangleq -\sum_{\boldsymbol{i}} X_{\boldsymbol{i}} \log X_{\boldsymbol{i}}$, with $0 \log 0 = 0$, the regularized problem is

$$\mathsf{MOT}_\eta(C, \boldsymbol{\mu}) \triangleq \min_{X \in \mathcal{U}(\boldsymbol{\mu})} \big\{\langle C, X\rangle - \eta H(X)\big\}. \tag{7}$$

Entropic regularization enables a compact Gibbs representation of the N-entry transportation plan. Set $K_{\boldsymbol{i}} = \exp(-C_{\boldsymbol{i}}/\eta)$. For mn potentials $\beta = (\beta_1, \ldots, \beta_m)$, define

$$Z(\beta) = \sum_{\boldsymbol{i}} K_{\boldsymbol{i}} \exp\left(\sum_{k=1}^m \beta_{k,i_k}\right), \quad Q_\beta(\boldsymbol{i}) = \frac{K_{\boldsymbol{i}} \exp(\sum_{k=1}^m \beta_{k,i_k})}{Z(\beta)}, \quad q_k(\beta) = \mathrm{Proj}_k(Q_\beta). \tag{8}$$

The potentials are the compact description: Q_β may have N entries, but it is determined by only mn numbers together with the cost oracle. The partition function $Z(\beta)$ merely normalizes this tensor to unit mass, and $q_k(\beta)$ records the marginal that the current potentials induce.

After restricting to positive marginal support as in Lemma B.1, the scaled dual of (7) is

$$\Phi_\eta(\beta) = \log Z(\beta) - \sum_{k=1}^{m} \mu_k^\top \beta_k, \qquad \nabla_{\beta_k} \Phi_\eta(\beta) = q_k(\beta) - \mu_k. \tag{9}$$

Thus a dual minimizer $\beta^\star$ makes $Q_{\beta^\star}$ the unique optimizer of (7). Define the total marginal residual $E(\beta) \triangleq \sum_k \|q_k(\beta) - \mu_k\|_1$. Exact target feasibility is $E(\beta) = 0$; Stage I instead computes $E(\beta) \le \delta$ and passes the resulting compact Gibbs description to sparse recovery.

The catch is evaluating the residual. One coordinate of $q_k(\beta)$ is already a sum over an n^{m-1}-entry fiber, so computing one full marginal by enumeration costs order N time.

Sinkhorn performs block-coordinate minimization of (9). It chooses a mode k and updates

$$\beta_{k,a}^{+} = \beta_{k,a} + \log \mu_{k,a} - \log q_{k,a}(\beta), \qquad a \in [n]. \tag{10}$$

With exact log-marginal values, this update makes the kth marginal equal to μ_k in one shot. Quantumly, we estimate each $\log q_{k,a}$ as a fiber log-sum-exp minus the global log-normalizer. A fiber contains n^{m-1} terms, so the quantum primitive pays its square root rather than enumerating it. Estimating all n coordinates of one marginal to log-accuracy α uses $\widetilde{O}(n^{(m+1)/2}\alpha^{-1})$ coherent cost queries and $\widetilde{O}(mn^{(m+1)/2}\alpha^{-1})$ potential lookups and unit-cost arithmetic operations.

This square-root fiber cost is the source of the quantum saving, but the estimates are necessarily noisy and depend on all preceding estimates. The exact Sinkhorn identity is therefore not enough to prove convergence. We replace it with a one-step inequality that tolerates estimation error: an α-accurate log-marginal update still decreases the entropic dual by at least $\mathsf{KL}(\mu_k \| q_k(\beta)) - \alpha^2/2$. We choose the mode k uniformly at random. Averaging over k and applying Pinsker's inequality gives expected dual decrease at least $E(\beta)^2/(2m^2) - \alpha^2/2$, which is positive whenever $E(\beta) > m\alpha$. Rather than certifying every iterate, we run a fixed number of updates and output the potentials at a uniformly random iteration; a drift argument shows that most iterates have small residual. Appendix C handles small marginal weights, potential ranges, and adaptive failures.

Theorem 3.1 (Quantum multimarginal Sinkhorn; informal version of Theorem C.11)**.** *Under the access model above, quantum tensor Sinkhorn computes potentials whose Gibbs tensor can be rounded to an exactly feasible additive-ε coupling. With constant success probability, it uses $\widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)$ coherent cost queries and $\widetilde{O}(m^5 n^{(m+1)/2}\kappa^3)$ working-vector lookups. The formal theorem gives the finite-precision and gate bounds.*

Since $H(X) \le m \log n$, regularization changes the cost by at most $\eta m \log n$. The exact rounding interface from Lin et al. (2022, Theorem 4.4) says that, for a coupling X with marginals $\boldsymbol{\mu}$ and target marginals $\boldsymbol{\nu}$, some coupling Y with marginals $\boldsymbol{\nu}$ obeys $\|Y - X\|_1 \le 2\sum_k \|\nu_k - \mu_k\|_1$. Apply this to an optimal target coupling and use entropic optimality of Q_β for its own marginals, $H(Y) \ge 0$, and $H(Q_\beta) \le m \log n$ to obtain

$$\langle C, Q_\beta \rangle \le \mathsf{MOT}(C, \boldsymbol{\mu}) + 2\Delta_C E(\beta) + \eta m \log n. \tag{11}$$

We therefore take $\eta = \Theta(\varepsilon/(m \log n))$ and drive $E(\beta)$ below $\delta = \Theta(\varepsilon/\Delta_C)$.

But the potentials still do not give us the requested output. The tensor Q_β is implicit, and standard rounding first materializes all N entries. That would turn a sublinear quantum computation into an $\Omega(N)$ classical one. The remaining question is therefore concrete: can we recover a small, exactly feasible coupling from the potentials without ever writing down Q_β?

3.2 Stage II: sparse exact recovery

We convert the Gibbs potentials to a list describing an ε-optimal coupling in three steps.

Trim. Starting from $E(\beta) \le \delta$, we estimate each marginal of Q_β from above and scale its entries downward, one correction factor per mode. The resulting tensor $Q^{\downarrow}$ lies below every target marginal coordinatewise, while the residual bound ensures that trimming removes only $O(\delta)$ mass.

This one-sided invariant is what makes completion possible later: after trimming, every missing marginal mass is a nonnegative deficit. The total mass removed remains small because Stage I already made the Gibbs tensor nearly feasible.

Sample. We cannot enumerate $Q^{\downarrow}$, but we can sample from it. Use the proposal $P_{\boldsymbol{i}} = \mu_{1,i_1}/n^{m-1}$: draw the first coordinate from μ_1 and the others uniformly. The domination $Q^{\downarrow}_{\boldsymbol{i}} \le \mathrm{Proj}_1(Q^{\downarrow})_{i_1} \le \mu_{1,i_1}$ makes rejection sampling valid. Its acceptance probability is of order $n^{-(m-1)}$: the proposal spreads its mass uniformly over the remaining $m-1$ coordinates, while $Q^{\downarrow}$ retains a constant total mass. Amplitude amplification (Brassard et al., 2000) therefore produces one accepted sample in $\widetilde{O}(n^{(m-1)/2})$ weight queries. This cost does not depend on the smallest marginal weight. If N_{samp} samples are drawn, their expected total marginal error is $O(m\sqrt{n/N_{\mathrm{samp}}})$; choosing $N_{\mathrm{samp}} = O(m^2 n\theta^{-2}\log(m/\zeta))$ makes this error at most θ with failure probability ζ. The same samples control the transport cost.

Complete. Sampling introduces small marginal errors, so we trim the empirical tensor coordinate by coordinate until all of its marginals again lie below their targets. The remaining m deficit vectors are nonnegative and have equal total mass. A greedy procedure couples them using at most $mn-m+1$ additional atoms, restoring every marginal exactly. At no point do we materialize an N-entry tensor.

Lemma 3.2 (Trim–sample–complete conversion; informal version of Lemma D.2)**.** *An implicitly represented Gibbs tensor with total marginal residual at most δ can be converted, without dense materialization, into an exactly feasible classical list. With sampling tolerance θ, the list has $O(m^2 n\theta^{-2}\log(m/\zeta) + mn)$ atoms, its cost under any fixed nonnegative cost tensor C exceeds that of the Gibbs tensor by at most $O((\delta+\theta)\|C\|_\infty)$, and its sampling stage uses $\widetilde{O}(m^2 n^{(m+1)/2}\theta^{-2}\log\zeta^{-1})$ coherent weight queries. For the final choices $\delta, \theta = \Theta(1/\kappa)$, recovery is lower order than Stage I.*

Theorem 3.3 (Quantum sparse coupling recovery; informal version of Theorem D.3)**.** *Under the access model above, a quantum algorithm returns, with probability at least $2/3$, a classical list representing an exactly feasible additive-ε coupling. The list has $O(m^2 n\kappa^2\log(mn\kappa)+mn)$ atoms, and the algorithm uses $\widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)$ coherent cost queries and $\widetilde{O}(m^5 n^{(m+1)/2}\kappa^3)$ working-vector lookups and unit-cost arithmetic operations.*

Because the list is exactly feasible and its cost can be evaluated with one cost query per atom, independent repetitions boost the success probability, and solving the MOT program restricted to its support leaves at most $mn-m+1$ atoms without increasing the cost (Proposition D.4). Algorithm 3 in Appendix D gives the complete pipeline.

The parameter choices in Theorem D.3 make the entropic bias, pre-recovery marginal error, and recovery error at most $\varepsilon/4$, $\varepsilon/8$, and $17\varepsilon/64$, respectively.

For fixed $m \ge 3$ and constant κ, the input has $N = n^m$ entries, whereas the algorithm uses $\widetilde{O}(n^{(m+1)/2})$ queries and returns $\widetilde{O}(n)$ atoms: it needs $\widetilde{O}(n)$ samples, each costing $\widetilde{O}(n^{(m-1)/2})$ queries. More generally, the list is subdense precisely when $m^2 n\kappa^2\log(mn\kappa) + mn = o(n^m)$; "sparse" below refers to list representation and does not assert this inequality for arbitrary precision.

4 Lower Bounds

The small output in Theorem 3.3 makes a serialization lower bound irrelevant. Instead, we force any successful algorithm to reveal many low-cost entries. Draw independent marks with probability $p = \log^2(n)/n^{m-1}$ and set $C_{\boldsymbol{i}} = 1$ on unmarked entries and $C_{\boldsymbol{i}} = 0$ on marked entries. For fixed m, the marked m-partite hypergraph has degree $(1+o(1))\log^2 n$ and codegree $o(\log^2 n)$ with high probability. Pippenger's theorem (Ehard et al., 2020, Theorem 1.1) then supplies a marked matching of size at least $(1-\varepsilon_0)n$, so the MOT value is at most ε_0.

Every atom in a coupling with uniform marginals has mass at most $1/n$; if the marginals are only approximately uniform, the atoms sharing a first coordinate carry at most $1/n$ plus that coordinate's marginal error. A low-cost output must therefore list $\Theta(n)$ distinct marked atoms. Classically, no verification is needed: unqueried entries are marked only with probability p, and q inspected locations contain only pq marks in expectation, which gives $q = \Omega(n/p) = \Omega(n^m/\log^2 n)$. Quantumly, querying the reported support verifies those marks and reduces recovery to multi-solution Bernoulli search, whose lower bound (Cojocaru et al., 2023, Theorem 3.1) gives $q = \Omega(n/\sqrt{p}) = \Omega(n^{(m+1)/2}/\log n)$.

Theorem 4.1 (Sparse coupling-recovery lower bound). *Fix a constant $m \geq 2$, a constant $c \geq 0$, and a constant normalized accuracy $\varepsilon_0 \in (0, 1/10)$. Consider algorithms that, for every Boolean cost tensor $C \in \{0,1\}^{[n]^m}$ with uniform marginals $\boldsymbol{\mu}$, return with probability at least $2/3$ a list of at most $n(\log n)^c$ support atoms and masses describing a tensor $X \geq 0$ with $\langle C, X\rangle \leq \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon_0$ and $\sum_k \|\operatorname{Proj}_k(X) - \mu_k\|_1 \leq \varepsilon_0$; exactly feasible outputs are the special case of zero marginal error. For all sufficiently large n, every such randomized classical algorithm requires $\Omega(n^m/\log^2 n)$ cost-entry queries, even without the bound on its output size, and every such quantum algorithm requires $\Omega(n^{(m+1)/2}/\log n)$ cost-entry queries. The constants and the threshold for n may depend on m, c, and ε_0.*

Appendix D proves Theorem 4.1, including the output-size quantifiers and finite-domain embedding needed by the quantum search bound. On its nonconstant Boolean hard instances, $\Delta_C = 1$, so additive and normalized errors coincide.

This lower bound matches both constructive upper exponents for every fixed m. The proof does not extend uniformly to growing m: the constants in the almost-perfect matching theorem depend on the hypergraph uniformity. Establishing the same lower bounds for $m \leq \operatorname{polylog}(n)$ would require either quantitative growing-uniformity matching estimates or a different hard distribution.

Corollary 4.2 (Sparse-output query complexity; informal version of Corollary D.6). *For fixed m and normalized inverse accuracy $\kappa \leq \operatorname{polylog}(n)$, constructing an exactly feasible additive-ε coupling with $n\operatorname{polylog}(n)$ atoms has query complexity $\widetilde{\Theta}(n^m)$ classically and $\widetilde{\Theta}(n^{(m+1)/2})$ quantumly, under the access models above.*

The formal statement tracks the accuracy range and output-size polylogarithms explicitly.

The matched recovery exponents have a direct interpretation. On the hard instances, exact uniform marginals cap every atom at mass $1/n$, so a low-cost output must expose $\Theta(n)$ marked tensor entries. Up to polylogarithmic factors, marks have density $n^{-(m-1)}$; producing the required list therefore costs $n \cdot n^{m-1} = n^m$ queries classically and $n \cdot \sqrt{n^{m-1}} = n^{(m+1)/2}$ quantumly. This constructive burden is absent from value estimation, which can be hard even when the instance hides only one low-cost block.

Theorem 4.3 (Value-query lower bounds). *For every $\varepsilon \in (0, 1/2)$, additive-ε estimation of the MOT value requires $\Omega(n^m/(1+\varepsilon n))$ randomized classical cost queries and $\Omega(\sqrt{n^m/(1+\varepsilon n)})$ bounded-error quantum cost queries.*

The proof splits n^{m-1} perfect matchings into blocks of $\lfloor 2\varepsilon n \rfloor + 1$ entries and reduces to unstructured search. At constant accuracy the bounds become $\Omega(n^{m-1})$ and $\Omega(n^{(m-1)/2})$, an n-factor gap below exact sparse recovery in the respective exponents.

5 Discussion and Limitations

Writing a dense coupling costs $\Omega(N)$ time, and both classical routes match this bound for fixed κ, up to factors in m and logarithms: packing has sharper m-dependence, while box–simplex is deterministic and uses $O(m\kappa \log N)$ matrix–vector rounds. For explicit $n\operatorname{polylog}(n)$-atom outputs, the query complexities are $\widetilde{\Theta}(N)$ classically and $\widetilde{\Theta}(\sqrt{Nn})$ quantumly for fixed m and $\kappa \leq \operatorname{polylog}(n)$; the lower bounds survive constant marginal error, and the classical bound permits arbitrary output size. Value estimation is a factor n cheaper. The quantum gain is multi-solution Grover amplification of fiber sums and rejection loops (Remark D.7), yielding the optimal exponents in the general entry-access model. Compilation depends on the coherent interfaces: without qRAM, table lookup adds $\widetilde{O}(n)$ gates and preserves sublinearity for $m \geq 4$ (Remark D.5), while the query bound carries $m^4\kappa^3$ rather than the packing route's $m^2\kappa$.

Additional structure changes the baseline: graphical, set-optimization, and low-rank costs can admit classical $\operatorname{poly}(n, m)$ algorithms with exact sparse outputs (Altschuler & Boix-Adserà, 2023). Our Boolean hard instances represent the general worst case and require $\Theta(N)$ oracle bits; on subclasses with a classical trimmed-Gibbs sampler, that sampler is the relevant baseline. The lower bounds assume fixed m and constant accuracy, leaving growing m, cubic κ-dependence, and the complexity of value estimation or approximate tensor scaling open.

Reproducibility Statement

This work is theoretical and uses no datasets or empirical results. The main text states the computational and output models. Appendices A–E contain complete proofs, pseudocode for the constructive algorithms, the trim–sample–complete recovery argument, finite-precision assumptions, oracle implementations, parameter choices, and explicit logarithmic-factor bounds needed to verify every claim.

AI Use Statement

Generative AI tools were used as research assistants in developing and checking proof arguments for the convergence of noisy quantum multimarginal Sinkhorn, the trim–sample–complete recovery procedure, and the sparse-output lower bound. They were also used to formulate and refine mathematical claims, explore proof strategies, edit portions of the manuscript, and assist with algebraic, finite-precision, failure-probability, and consistency checks. The authors reconstructed and independently verified every AI-assisted argument, checked the cited sources, reviewed all generated text, and take full responsibility for the paper's claims and contents. No datasets, experiments, or empirical results were generated or interpreted using AI.

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. doi: 10.1137/100805741. URL https://doi.org/10.1137/100805741.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Nearly linear-time packing and covering LP solvers: Achieving width-independence and-convergence. *Mathematical Programming*, 175(1):307–353, 2019.

Jason M. Altschuler and Enric Boix-Adserà. Hardness results for Multimarginal Optimal Transport problems. *Discrete Optimization*, 42:100669, 2021. ISSN 1572-5286. doi: https://doi.org/10.1016/j.disopt.2021.100669. URL https://www.sciencedirect.com/science/article/pii/S1572528621000487.

Jason M. Altschuler and Enric Boix-Adserà. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *Mathematical Programming*, 199(1):1107–1178, 2023.

Simon Apers and Sander Gribling. Quantum speedups for linear programming via interior point methods. *SIAM Journal on Computing*, 55(1):93–134, 2023.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Ryan Babbush, Craig Gidney, Dominic W Berry, Nathan Wiebe, Jarrod McClean, Alexandru Paler, Austin Fowler, and Hartmut Neven. Encoding electronic spectra in quantum circuits with linear T complexity. *Physical Review X*, 8(4):041015, 2018.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015. doi: 10.1137/141000439. URL https://doi.org/10.1137/141000439.

Charles H. Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5):1510–1523, 1997. doi: 10.1137/S0097539796300933. URL https://doi.org/10.1137/S0097539796300933.

Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo Lopez. Geodesic PCA in the Wasserstein space. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53, 07 2013. doi: 10.1214/15-AIHP706.

Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *Operations Research Letters*, 52:107054, 2024. ISSN 0167-6377. doi: https://doi.org/10.1016/j.orl.2023.11.007. URL https://www.sciencedirect.com/science/article/pii/S0167637723001852.

Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pp. 1–12, 2011.

Adam Bouland, Yosheb Getachew, Yujia Jin, Aaron Sidford, and Kevin Tian. Quantum speedups for zero-sum games via improved dynamic Gibbs sampling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Michel Boyer, Gilles Brassard, Peter Høyer, and Alain Tapp. Tight bounds on quantum searching. *Fortschritte der Physik: Progress of Physics*, 46(4-5):493–505, 1998.

Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. *arXiv preprint quant-ph/0005055*, 2000.

Richard P. Brent and Paul Zimmermann. *Modern Computer Arithmetic*, volume 18. Cambridge University Press Cambridge, 2011.

Giuseppe Buttazzo, Luigi De Pascale, and Paola Gori-Giorgi. Optimal-transport formulation of electronic density-functional theory. *Physical Review A*, 85(6), 2012.

Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multi-marginal wasserstein gan. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Guillaume Carlier. On the linear convergence of the multimarginal Sinkhorn algorithm. *SIAM Journal on Optimization*, 32(2):786–794, 2022.

Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic Theory*, 42:397–418, 2010.

Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6): 1621–1642, 2015.

Y. Choi, M. Choi, M. Kim, J-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pp. 8789–8797, 2018.

Michael B. Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

Alexandru Cojocaru, Juan Garay, Aggelos Kiayias, Fang Song, and Petros Wallden. Quantum multi-solution bernoulli search with applications to bitcoin's post-quantum security. *Quantum*, 7:944, 2023.

Codina Cotar, Gero Friesecke, and Claudia Klüppelberg. Density functional theory and optimal transportation with Coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4): 548–599, 2013.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NeurIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Marco Cuturi and Arnaud Doucet. Fast Computation of Wasserstein Barycenters. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 685–693, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/cuturi14.html.

Y. Dolinsky and M. H. Soner. Robust hedging and martingale optimal transport in continuous time. *Probability Theory and Related Fields*, 160:391–427, 2014.

Christoph Dürr and Peter Høyer. A quantum algorithm for finding the minimum. *arXiv preprint quant-ph/9607014*, 1996.

Stefan Ehard, Stefan Glock, and Felix Joos. Pseudorandom hypergraph matchings. *Combinatorics, Probability and Computing*, 29(6):868–885, 2020.

Ivar Ekeland. An optimal matching problem. *ESAIM: Control, Optimisation and Calculus of Variations*, 11(1):57–71, 2005.

Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy. Wasserstein discriminant analysis. *Machine Learning*, 107:1923–1945, 2018.

A. Galichon, P. Henry-Labordere, and N. Touz. A stochastic control approach to non-arbitrage bounds given marginals, with an application to lookback options. *The Annals of Applied Probability*, 24:312–336, 2014.

Minbo Gao, Zhengfeng Ji, Tongyang Li, and Qisheng Wang. Logarithmic-regret quantum learning algorithms for zero-sum games. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Sander Gribling and Harold Nieuwboer. Improved Quantum Lower and Upper Bounds for Matrix Scaling. In Petra Berenbrink and Benjamin Monmege (eds.), *39th International Symposium on Theoretical Aspects of Computer Science (STACS 2022)*, volume 219 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 35:1–35:23, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-222-8. doi: 10.4230/LIPIcs.STACS.2022.35. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.STACS.2022.35.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 1988.

Pierre Henry-Labordère. *Model-free hedging: A martingale optimal transport viewpoint*. CRC Press, 2017.

Arun Jambulapati and Kevin Tian. Revisiting area convexity: Faster box-simplex games and spectrahedral generalizations. *Advances in Neural Information Processing Systems*, 36:57583–57596, 2023.

Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NeurIPS'16, pp. 181–189, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Junhyung Lyle Kim, Brandon Augustino, Dylan Herman, Enrico Fontana, Jacob Watkins, Marco Pistoia, and Shouvanik Chakrabarti. Fast zeroth-order convex optimization with quantum gradient methods. *Advances in Neural Information Processing Systems*, 38:84556–84604, 2026.

Vladimir R. Kostic, Saverio Salzo, and Massimiliano Pontil. Batch greenkhorn algorithm for entropic-regularized multimarginal optimal transport: Linear rate of convergence and iteration complexity. In *International Conference on Machine Learning*, pp. 11529–11558. PMLR, 2022.

Yin Tat Lee and Aaron Sidford. Path finding methods for linear programming: Solving linear programs in $\mathcal{O}(\sqrt{\text{rank}})$ iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 424–433. IEEE, 2014.

Yin Tat Lee and Aaron Sidford. Solving linear programs with $\mathcal{O}(\sqrt{\text{rank}})$ linear system solves. *arXiv preprint arXiv:1910.08033*, 2019.

Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I. Jordan. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23(1), January 2022. ISSN 1532-4435.

James Robert Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NeurIPS'15, pp. 829–837, Cambridge, MA, USA, 2015. MIT Press.

C. B. Mendl and L. Lin. Kantorovich dual solution for strictly correlated electrons in atoms and molecules. *Physical Review B*, 87:125106, 2013.

Jonas Mueller and Tommi Jaakkola. Principal differences analysis: interpretable characterization of differences between distributions. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NeurIPS'15, pp. 1702–1710, Cambridge, MA, USA, 2015. MIT Press.

Harold Nieuwboer. *Classical and Quantum Algorithms for Scaling Problems*. PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, January 2024. URL https://hdl.handle.net/11245.1/0c64f620-5420-4978-845a-f95987c18384.

James B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78:109–129, 1997.

Victor M. Panaretos and Yoav Zemel. Amplitude and phase variation of point processes. *The Annals of Statistics*, 44(2):771–812, 2016. ISSN 00905364. URL http://www.jstor.org/stable/43818628.

Kent Quanrud. Approximating optimal transport with linear programs. In Jeremy T. Fineman and Michael Mitzenmacher (eds.), *2nd Symposium on Simplicity in Algorithms, SOSA 2019, San Diego, CA, USA, January 8-9, 2019*, volume 69 of *OASIcs*, pp. 6:1–6:9. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/OASICS.SOSA.2019.6. URL https://doi.org/10.4230/OASIcs.SOSA.2019.6.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

Roman Sandler and Michael Lindenbaum. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1590–1602, 2011.

Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer, 2003.

M. Seidl, P. Gori-Giorgi, and A. Savi. Strictly correlated electrons in density-functional theory: A general formulation with applications to spherical densities. *Physical Review A*, 75:75:042511, 2007.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Sanvesh Srivastava, Cheng Li, and David B. Dunson. Scalable Bayes via Barycenter in Wasserstein Space. *Journal of Machine Learning Research*, 19(8):1–35, 2018. URL http://jmlr.org/papers/v19/17-084.html.

Gábor J. Székely, Maria L. Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.

Nazarii Tupitsa, Pavel Dvurechensky, Alexander Gasnikov, and César A. Uribe. Multimarginal optimal transport by accelerated alternating minimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 6132–6137. IEEE, 2020.

Joran van Apeldoorn and András Gilyén. Improvements in Quantum SDP-Solving with Applications. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi (eds.), *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, volume 132 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 99:1–99:15, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-109-2. doi: 10.4230/LIPIcs.ICALP.2019.99. URL http://drops.dagstuhl.de/opus/volltexte/2019/10675.

Joran van Apeldoorn, András Gilyén, Sander Gribling, and Ronald de Wolf. Quantum SDP-solvers: Better upper and lower bounds. *Quantum*, 4:230, 2020.

Joran van Apeldoorn, Sander Gribling, Yinan Li, Harold Nieuwboer, Michael Walter, and Ronald de Wolf. Quantum Algorithms for Matrix Scaling and Matrix Balancing. In Nikhil Bansal, Emanuela Merelli, and James Worrell (eds.), *48th International Colloquium on Automata, Languages, and Programming (ICALP 2021)*, volume 198 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 110:1–110:17, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-195-5. doi: 10.4230/LIPIcs.ICALP.2021.110. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2021.110.

Jan van den Brand. A deterministic linear program solver in current matrix multiplication time. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 259–278. SIAM, 2020.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI, 2003.

Christof Zalka. Grover's quantum searching algorithm is optimal. *Physical Review A*, 60(4):2746, 1999.

APPENDIX CONTENTS

A Preliminaries **15**

B Solving Multimarginal Optimal Transport with LP Packing **15**

B.1 Finite-dimensional linear programming formulation 16

B.2 The packing-LP black box 17

B.3 From a packing solution to a multimarginal coupling 18

B.4 Exact reduction of MOT to a packing LP 18

B.5 A deterministic box–simplex route 20

C Quantum Sinkhorn for multimarginal OT **21**

C.1 Finite-input conventions and log-domain Gibbs weights 21

C.2 Entropic MOT and its associated dual 22

C.3 Quantum Multimarginal Sinkhorn 23

C.4 Comparison with existing approaches 33

D Sparse coupling recovery from Gibbs potentials **34**

E Value-estimation lower bounds **42**

A PRELIMINARIES

We write $[n] \triangleq \{1, \dots, n\}$. We represent the $n \times 1$ all-ones vector by $\mathbf{1}_n \in \mathbb{R}^n$. We let $\mathbb{R}^n_{\geq 0}$ (resp. $\mathbb{R}^n_{>0}$) denote the nonnegative (resp. positive) orthant, and the probability simplex is $\Delta^n \triangleq \{x \in \mathbb{R}^n_{\geq 0} : \sum_{i\in[n]} x_i = 1\}$. We define $\Delta^n_{>0} \triangleq \mathbb{R}^n_{>0} \cap \Delta^n$.

For a vector $x \in \mathbb{R}^n$ and $p \in [1, \infty]$, we denote by $\|x\|_p$ the ℓ_p-norm of x. We define the n-dimensional ℓ_p-ball of radius R centered at the origin as $\mathcal{B}^n_p(R) \triangleq \{x \in \mathbb{R}^n : \|x\|_p \leq R\}$.

For vectors $x, y \in \mathbb{R}^n$, $\langle x, y\rangle$ denotes their inner product while $x \odot y$ denotes their Hadamard (or, element-wise) product. For matrices A and B of arbitrary sizes, $A \otimes B$ denotes their Kronecker product. For a tensor $A = (A_{i_1,\dots,i_m}) \in \mathbb{R}^{n_1\times\cdots\times n_m}$, $\mathrm{vec}(A)$ is its $\prod_{k\in[m]} n_k \times 1$ vectorization. When all tensor modes have size n, we write a tensor coordinate as a multi-index $\boldsymbol{i} = (i_1, \dots, i_m) \in [n]^m$, and use notation such as $C_{\boldsymbol{i}}$, $x_{\boldsymbol{i}}$, and $A_{j,\boldsymbol{i}}$. For two tensors A and B of equal dimension, their inner product is $\langle A, B\rangle = \mathrm{vec}(A)^\top \mathrm{vec}(B)$. We also adopt the convention that the ℓ_p-norm of a tensor A is the ℓ_p-norm of its vectorization, i.e., $\|A\|_p \triangleq \|\mathrm{vec}(A)\|_p$.

B SOLVING MULTIMARGINAL OPTIMAL TRANSPORT WITH LP PACKING

In this section, we solve discrete MOT by reducing it to a positive packing LP, then completing the resulting partial coupling. The packing solver below is not applied to the standard-form equality LP; rather, it uses the same marginal-incidence data with inequalities and a shifted nonnegative objective. This reduction generalizes the approach used in state-of-the-art algorithms for bimarginal OT (Blanchet et al., 2024; Quanrud, 2019).

In order to extend to the multimarginal setting, we rely on the observation that if a partial coupling X has deficit vectors $d_k \triangleq \mu_k - \mathrm{Proj}_k(X) \in \mathbb{R}^{n_k}_{\geq 0}$ for $k \in [m]$, then all deficits have the same total mass τ, and

$$D = \tau^{1-m} d_1 \otimes \cdots \otimes d_m$$

has exactly those deficits as its marginals. Therefore $X + D$ is an exact coupling. This reduces dense MOT to a packing LP whose constraint matrix has exactly mN nonzero entries.

B.1 Finite-dimensional linear programming formulation

We begin by recalling the discrete MOT problem and its dual:

Problem.

$$\mathsf{MOT}(C, \boldsymbol{\mu}) \triangleq \min_{X \in \mathcal{U}(\boldsymbol{\mu})} \langle C, X \rangle \tag{MOT-P}$$

where, with $I_k \triangleq [n_k]$ and $I \triangleq I_1 \times \cdots \times I_m$, $C = [C_{\boldsymbol{i}}]_{\boldsymbol{i} \in I}$ denotes a real-valued cost tensor with m modes, $\mu_k = [\mu_{k,i}]_{i \in [n_k]}$ is the probability vector representing the k-th marginal and

$$\mathcal{U}(\boldsymbol{\mu}) \triangleq \{X \geq 0 : \mathrm{Proj}_k(X) = \mu_k \ \forall k \in [m]\}.$$

The dual problem is

$$\begin{aligned} \mathsf{MOT}(C, \boldsymbol{\mu}) = \max_{y_1 \in \mathbb{R}^{n_1}, \ldots, y_m \in \mathbb{R}^{n_m}} \quad & \sum_{k=1}^{m} \mu_k^\top y_k \\ \text{subject to} \quad & \sum_{k=1}^{m} y_{k,i_k} \leq C_{i_1, \ldots, i_m}, \qquad i_k \in [n_k] \ \forall k \in [m]. \end{aligned} \tag{MOT-D}$$

Lemma B.1 (Pruning zero marginal coordinates)**.** *For each mode, let $I_k^+ \triangleq \{a \in I_k : \mu_{k,a} > 0\}$. Restrict C to $I_1^+ \times \cdots \times I_m^+$ and restrict each marginal to I_k^+. Extension by zero gives an objective-preserving bijection between couplings of the restricted and original instances. The same statement holds for the packing relaxation obtained by replacing every marginal equality by an inequality. The support sets can be found in $O(M)$ time, and restriction or zero extension never increases M, N, or the number of incidence nonzeros.*

Proof. If $\mu_{k,a} = 0$ and X is a nonnegative feasible coupling, then

$$0 = \mathrm{Proj}_k(X)_a = \sum_{\boldsymbol{i} : i_k = a} X_{\boldsymbol{i}},$$

so every entry in that fiber is zero. The same conclusion follows from $\mathrm{Proj}_k(X)_a \leq 0$ for a packing-feasible tensor. Deleting all such forced zero fibers therefore preserves feasibility and objective values in both directions; extension by zero is the inverse map. □

One can cast (MOT-P) and its dual in standard form using a vectorized representation. Write $N = \prod_{k=1}^{m} n_k$, $M = \sum_{k=1}^{m} n_k$, and define $x = \mathrm{vec}(X)$. Let $A \in \{0,1\}^{M \times N}$ be the marginal-incidence matrix with rows indexed by (k, r) and columns indexed by $\boldsymbol{i} = (i_1, \ldots, i_m)$, so that $A_{(k,r),\boldsymbol{i}} = \mathbb{1}[i_k = r]$. Upon setting

$$c = \mathrm{vec}(C) \in \mathbb{R}^N, \quad b = \mu \triangleq \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix} \in \mathbb{R}^M, \quad A = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} \in \mathbb{R}^{M \times N}, \tag{12}$$

the vectorized primal is the standard-form LP

$$\min_{x \geq 0} \{c^\top x : Ax = b\}, \tag{P}$$

and its dual is

$$\max_{y \in \mathbb{R}^M} \{b^\top y : A^\top y \leq c\}. \tag{D}$$

The $M \times N$ constraint matrix A contains $\mathsf{nnz}(A) = mN$ nonzero entries: for each column $\boldsymbol{i}$, exactly one row in each marginal block is active.

For the packing reduction, the important object is the same incidence matrix A, but with marginal inequalities $Ax \leq b$ rather than the equality constraint in the standard-form LP.

Lemma B.2 (Sparse optimal coupling)**.** *The MOT linear program has an optimal coupling supported on at most* $\operatorname{rank}(A) = \sum_{k=1}^{m} n_k - m + 1$ *tensor entries. In particular, when every mode has size* n*, the support has size at most* $mn - m + 1$*.*

Proof. The transportation polytope is nonempty and compact, so a linear objective attains its minimum at a basic feasible solution. The positive columns of a basic feasible solution are linearly independent and hence number at most $\operatorname{rank}(A)$.

It remains to compute this rank. The sum of the rows in each marginal block is the all-ones row, which gives $m - 1$ independent dependencies and therefore

$$\operatorname{rank}(A) \le \sum_{k=1}^{m} n_k - m + 1.$$

For the reverse inequality, retain every row in the first block and omit one row from each later block. Suppose a linear combination of the retained rows vanishes. Evaluating it on a tensor index whose later coordinates are all the omitted coordinates first forces every coefficient in the first block to vanish. Varying one later coordinate at a time then forces every remaining coefficient to vanish. The retained rows are linearly independent, proving equality and the support bound. □

B.2 The packing-LP black box

Allen-Zhu & Orecchia (2019, Corollary 3.5) state their solver for the standard form $\max\{\mathbf{1}^\top x' : A'x' \le \mathbf{1},\ x' \ge 0\}$ with $A' \ge 0$ normalized so that $\min_j \|A'_{:j}\|_\infty = 1$. For accuracy $\varepsilon' \in (0, 1/30]$, it returns, with probability at least $2/3$, a feasible $(1 - O(\varepsilon'))$-approximate solution in expected time $O(\mathsf{nnz}(A')\varepsilon'^{-1}\log(MN/\varepsilon')\log(1/\varepsilon'))$. We use it through the following weighted, worst-case form; cf. (Blanchet et al., 2024, Theorem 3).

Theorem B.3 (Packing LP solver)**.** *Consider*

$$P^\star \triangleq \max\{w^\top x : Ax \le b, x \ge 0\}, \tag{13}$$

where $A \in \mathbb{R}^{M\times N}_{\ge 0}$ *has no zero column,* $b \in \mathbb{R}^{M}_{\ge 0}$*, and* $w \in \mathbb{R}^{N}_{\ge 0}$*. For every* $\delta_{\mathrm{P}} \in (0, 1)$ *and* $\rho \in (0, 1/2)$*, there is a randomized algorithm which, with probability at least* $1 - \rho$*, returns* $x \ge 0$ *satisfying* $Ax \le b$ *and* $w^\top x \ge (1 - \delta_{\mathrm{P}})P^\star$ *in worst-case arithmetic time*

$$O\left(\left[M + N + \frac{\mathsf{nnz}(A)}{\delta_{\mathrm{P}}}\log\left(\frac{MN}{\delta_{\mathrm{P}}}\right)\log\left(\frac{e}{\delta_{\mathrm{P}}}\right)\right]\log\frac{1}{\rho}\right).$$

Proof. Delete every row r with $b_r = 0$ together with every column meeting it, and then every remaining column j with $w_j = 0$. Columns of the first kind vanish in every feasible solution. Setting columns of the second kind to zero preserves feasibility, because $A \ge 0$, and preserves the objective. The reduced LP therefore has optimum $P^\star$, and zero extension maps its feasible solutions to feasible solutions of (13) with the same objective. For (packing-MOT), the first deletion is exactly Lemma B.1. If no column remains, $x = 0$ is optimal. Otherwise substitute

$$x'_j \triangleq w_j x_j, \qquad A'_{r,j} \triangleq \frac{A_{r,j}}{b_r w_j}.$$

This bijection of nonnegative vectors maps $\{Ax \le b\}$ onto $\{A'x' \le \mathbf{1}\}$ and $w^\top x$ onto $\mathbf{1}^\top x'$, so relative guarantees transfer exactly. Every retained column of A' is nonzero, and multiplying A' by the positive scalar $1/\min_j \|A'_{:j}\|_\infty$ imposes the column normalization while rescaling all solutions and the optimum by the same factor. The reduction does not increase the dimensions or the number of nonzeros, and it and its inverse take $O(M + N + \mathsf{nnz}(A))$ time.

Run the solver with $\varepsilon' = \min\{c\delta_{\mathrm{P}}, 1/30\}$ for a sufficiently small absolute constant $c > 0$, and stop it after three times its expected-time bound. By Markov's inequality and a union bound, each such run returns a feasible $(1 - \delta_{\mathrm{P}})$-approximate solution with probability at least $1/3$. Checking feasibility and the objective value of an output costs $O(M + N + \mathsf{nnz}(A))$. Therefore $\lceil \log(1/\rho)/\log(3/2)\rceil$ independent runs, keeping the best feasible output, fail with probability at most ρ and obey the stated worst-case bound. □

B.3 From a packing solution to a multimarginal coupling

Let $L \triangleq \min_{\boldsymbol{i}\in I} C_{\boldsymbol{i}}$ and $U \triangleq \max_{\boldsymbol{i}\in I} C_{\boldsymbol{i}}$, both computable in $O(N)$ time, so that $\Delta_C = U - L$. Define the nonnegative tensor

$$B \triangleq U\mathbf{1} - C, \qquad 0 \le B_{\boldsymbol{i}} \le \Delta_C \quad (\boldsymbol{i} \in I). \tag{14}$$

Consider the packing LP

$$P^\star \triangleq \max_{X\in\mathbb{R}^I_{\ge 0}} \{\langle B, X\rangle : \mathrm{Proj}_k(X) \le \mu_k \ \forall k \in [m]\}. \tag{packing-MOT}$$

We begin by showing how feasible solutions of (packing-MOT) can be rounded to feasible solutions for (MOT-P) using $O(mN)$ operations.

Lemma B.4. *Let $X \in \mathbb{R}^I_{\ge 0}$ be a feasible solution to* (packing-MOT)*, and define the deficit vectors*

$$d_k \triangleq \mu_k - \mathrm{Proj}_k(X) \in \mathbb{R}^{n_k}_{\ge 0}.$$

Then, all deficits have the same total mass

$$\|d_k\|_1 = 1 - |X| \triangleq \tau.$$

Moreover, there is a nonnegative rank-one tensor D, satisfying $\mathrm{Proj}_k(D) = d_k$ for every k, and hence, $Z = X + D$ belongs to $\mathcal{U}(\boldsymbol{\mu})$. In particular,

$$D = \begin{cases} \tau^{1-m} d_1 \otimes \cdots \otimes d_m & \tau > 0, \\ 0, & \tau = 0 \end{cases} \tag{15}$$

can be computed using $O(mN)$ operations.

Proof. For every k, note that

$$\|\mathrm{Proj}_k(X)\|_1 = \sum_{a\in I_k} \sum_{\boldsymbol{i}: i_k = a} X_{\boldsymbol{i}} = \sum_{\boldsymbol{i}\in I} X_{\boldsymbol{i}} = |X|.$$

Since $\|\mu_k\|_1 = 1$ and $d_k \ge 0$,

$$\|d_k\|_1 = 1 - |X| = \tau.$$

If $\tau = 0$, then every nonnegative d_k has zero mass, so $d_k = 0$ and $D = 0$ suffices. Now assume $\tau > 0$. For $a \in I_k$, the ath entry of the kth marginal of the tensor in (15) is

$$(\mathrm{Proj}_k(D))_a = \tau^{1-m} d_k(a) \prod_{\ell\ne k} \left(\sum_{a'\in I_\ell} d_\ell(a') \right) = \tau^{1-m} d_k(a) \tau^{m-1} = d_k(a).$$

Therefore, $\mathrm{Proj}_k(X + D) = \mathrm{Proj}_k(X) + d_k = \mu_k$ for every k. The constraint matrix has $O(mN)$ nonzeros, so computing $\mu - Ax$ and forming the dense tensor product both take $O(mN)$ operations. □

B.4 Exact reduction of MOT to a packing LP

We have the following.

Theorem B.5. *For the MOT value $\mathsf{MOT}(C, \boldsymbol{\mu})$ and the optimal packing value $P^\star$,*

$$P^\star = U - \mathsf{MOT}(C, \boldsymbol{\mu}).$$

Proof. Suppose that $X^\star$ is an optimal MOT coupling. Then it is feasible for (packing-MOT), has mass one, and

$$\langle B, X^\star\rangle = U - \langle C, X^\star\rangle = U - \mathsf{MOT}(C, \boldsymbol{\mu}).$$

Hence, $P^\star \ge U - \mathsf{MOT}(C, \boldsymbol{\mu})$.

Conversely, suppose that X is a feasible solution to (packing-MOT). By Lemma B.4, there is $D \geq 0$ such that $Z = X + D \in \mathcal{U}(\boldsymbol{\mu})$. Since $B \geq 0$,

$$\langle B, X \rangle \leq \langle B, Z \rangle .$$

Since Z has mass one and is feasible for (MOT-P),

$$\langle B, Z \rangle = U - \langle C, Z \rangle \leq U - \mathsf{MOT}(C, \boldsymbol{\mu}).$$

Thus, every feasible solution to (packing-MOT) has packing value at most $U - \mathsf{MOT}(C, \boldsymbol{\mu})$, so $P^\star \leq U - \mathsf{MOT}(C, \boldsymbol{\mu})$. This completes the proof. □

Finally, we convert a multiplicative guarantee on the packing value to an additive error guarantee on the MOT value.

Theorem B.6. *Let $\delta_{\mathrm{P}} \in (0, 1)$, and suppose that X is feasible for* (packing-MOT) *and satisfies*

$$\langle B, X \rangle \geq (1 - \delta_{\mathrm{P}}) P^\star.$$

Let $Z = X + D$ be its completion from Lemma B.4. Then, $Z \in \mathcal{U}(\boldsymbol{\mu})$ and

$$\langle C, Z \rangle \leq \mathsf{MOT}(C, \boldsymbol{\mu}) + \delta_{\mathrm{P}} \Delta_C.$$

Proof. Since $D \geq 0$ and $B \geq 0$,

$$\langle B, Z \rangle \geq \langle B, X \rangle \geq (1 - \delta_{\mathrm{P}}) P^\star.$$

Applying Theorem B.5 and using the fact that $|Z| = 1$, we have

$$\langle C, Z \rangle - \mathsf{MOT}(C, \boldsymbol{\mu}) = [U - \langle B, Z \rangle] - (U - P^\star) = P^\star - \langle B, Z \rangle \leq \delta_{\mathrm{P}} P^\star.$$

Since $\mathsf{MOT}(C, \boldsymbol{\mu}) \geq L$, the exact equivalence gives $P^\star = U - \mathsf{MOT}(C, \boldsymbol{\mu}) \leq U - L = \Delta_C$. The proof is complete. □

Combining these results, we obtain the following.

Theorem B.7. *Let $m \geq 2$ and $\Delta_C \triangleq \max_{\boldsymbol{i}} C_{\boldsymbol{i}} - \min_{\boldsymbol{i}} C_{\boldsymbol{i}}$. For every $\varepsilon > 0$ and $\rho \in (0, 1/2)$, one can, with probability at least $1 - \rho$, determine a tensor $Z \in \mathcal{U}(\boldsymbol{\mu})$ satisfying $\langle C, Z \rangle \leq \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$. Set $\kappa \triangleq \max\{1, \Delta_C/\varepsilon\}$. This can be done in randomized arithmetic time*

$$O\left([N + M + mN\kappa \log(MN\kappa) \log(e\kappa)] \log \frac{1}{\rho} \right).$$

Proof. Let $L = \min_{\boldsymbol{i}} C_{\boldsymbol{i}}$ and $U = \max_{\boldsymbol{i}} C_{\boldsymbol{i}}$, so that $\Delta_C = U - L$. If $\varepsilon \geq \Delta_C$, return the product coupling $Z^0 = \mu_1 \otimes \cdots \otimes \mu_m$, which can be formed in $O(N)$ time. Since $Z^0 \in \mathcal{U}(\boldsymbol{\mu})$, $\langle C, Z^0 \rangle \leq U$ and $\mathsf{MOT}(C, \boldsymbol{\mu}) \geq L$, so

$$\langle C, Z^0 \rangle - \mathsf{MOT}(C, \boldsymbol{\mu}) \leq U - L = \Delta_C \leq \varepsilon.$$

Assume $\varepsilon \in (0, \Delta_C)$ and set $\delta_{\mathrm{P}} = \varepsilon / \Delta_C \in (0, 1)$. The packing LP (packing-MOT) has M constraints, N variables, and a coefficient matrix with mN nonzero entries and no zero column. The black box packing algorithm of Theorem B.3 returns, with probability at least $1 - \rho$, a feasible solution X to (packing-MOT) in time

$$O\left(\left[N + M + mN \frac{\Delta_C}{\varepsilon} \log\left(MN \frac{\Delta_C}{\varepsilon} \right) \log\left(\frac{e\Delta_C}{\varepsilon} \right) \right] \log \frac{1}{\rho} \right).$$

By Theorem B.6, the completion tensor Z is an exactly feasible solution to (MOT-P), and has cost not exceeding $\mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$. □

Proof of Theorem 2.3. Apply Theorem B.7 with $n_1 = \cdots = n_m = n$, so $N = n^m$ and $M = mn$. Expressing $\log N = m \log n$ explicitly yields

$$O\big(m^2 n^m \kappa \log(emn\kappa) \log(e\kappa) \log(1/\rho) \big),$$

and hence the stated $\widetilde{O}(m^2 n^m \kappa)$ arithmetic bound. □

Algorithm 1 MOT via packing and rank-one completion

Require: Marginals $\boldsymbol{\mu}$, cost access to C, accuracy ε
1: Compute L, U and $B = U\mathbf{1} - C$
2: **if** $\varepsilon \geq U - L$ **then**
3: **return** $\mu_1 \otimes \cdots \otimes \mu_m$
4: **end if**
5: $X \leftarrow \textsc{Packing}(B, \boldsymbol{\mu}, \varepsilon/(U-L))$
6: $d_k \leftarrow \mu_k - \mathrm{Proj}_k(X)$ for every k and $\tau \leftarrow \|d_1\|_1$
7: $D \leftarrow 0$ if $\tau = 0$, else $D \leftarrow \tau^{1-m} \bigotimes_{k=1}^{m} d_k$
8: **return** $X + D$

B.5 A DETERMINISTIC BOX–SIMPLEX ROUTE

We now prove Theorem 2.1. If $\Delta_C \leq \varepsilon$, any feasible coupling suffices. Otherwise shift the cost to $C^+ = C - (\min C)\mathbf{1} \in [0, \Delta_C]^{[n]^m}$, let $A \in \{0,1\}^{mn \times N}$ be the marginal-incidence matrix, and stack the target marginals in $b \in \mathbb{R}^{mn}$. Consider

$$\min_{X \in \Delta^N} \max_{z \in [-1,1]^{mn}} \mathcal{G}(X, z), \qquad \mathcal{G}(X, z) \triangleq \langle C^+, X \rangle + 2\Delta_C z^\top (A \operatorname{vec}(X) - b). \tag{16}$$

After negating the objective and exchanging the players, this is the box–simplex form of Jambulapati & Tian (2023, Theorem 1), with box variable $z \in [-1,1]^{mn}$, simplex variable $\operatorname{vec}(X) \in \Delta^N$, and game matrix $-2\Delta_C A \in \mathbb{R}^{mn \times N}$.

Let $(\bar{X}, \bar{z})$ have saddle gap at most ε, and let $X^\star$ be an optimal MOT coupling. Since $A \operatorname{vec}(X^\star) = b$,

$$\begin{aligned} \max_{z \in [-1,1]^{mn}} \mathcal{G}(\bar{X}, z) &\leq \min_{X \in \Delta^N} \mathcal{G}(X, \bar{z}) + \varepsilon \\ &\leq \mathcal{G}(X^\star, \bar{z}) + \varepsilon = \mathsf{MOT}(C^+, \boldsymbol{\mu}) + \varepsilon. \end{aligned}$$

Maximizing coordinatewise over z shows that

$$\langle C^+, \bar{X} \rangle + 2\Delta_C \sum_{k=1}^{m} \| \mathrm{Proj}_k(\bar{X}) - \mu_k \|_1 \leq \mathsf{MOT}(C^+, \boldsymbol{\mu}) + \varepsilon. \tag{17}$$

The rounding algorithm of Lin et al. (2022, Theorem 4.4) returns $Y \in \mathcal{U}(\boldsymbol{\mu})$ satisfying

$$\|Y - \bar{X}\|_1 \leq 2 \sum_{k=1}^{m} \| \mathrm{Proj}_k(\bar{X}) - \mu_k \|_1.$$

Because $0 \leq C^+ \leq \Delta_C$, this bound and (17) give

$$\langle C^+, Y \rangle \leq \langle C^+, \bar{X} \rangle + \Delta_C \|Y - \bar{X}\|_1 \leq \mathsf{MOT}(C^+, \boldsymbol{\mu}) + \varepsilon.$$

Shifting back proves the objective guarantee in Theorem 2.1.

It remains to expose all dimension dependence. Every column of A has exactly m ones, so

$$\mathsf{nnz}(-2\Delta_C A) = mN, \qquad \| -2\Delta_C A \|_{1 \to 1} = 2m\Delta_C.$$

The cited box–simplex theorem therefore takes

$$O\left(\mathsf{nnz}(A) \, \| -2\Delta_C A \|_{1 \to 1} \frac{\log N}{\varepsilon} \right) = O(m^2 N \kappa \log N)$$

work and $O(m\kappa \log N)$ matrix–vector rounds. Each product by A or $A^\top$ costs $O(mN)$ operations, and multimarginal rounding costs another $O(mN)$. Finally, $\log N = m \log n$, so no polynomial dependence on n is hidden outside N.

Remark B.8 (Interior point baseline)**.** *We record the following consequence of the rank identity in Lemma B.2; to our knowledge, this specialization has not previously appeared in MOT complexity comparisons. For $m \geq 3$ the MOT linear program is not a minimum-cost flow problem (Lin et al., 2022), but general interior point methods do not need that structure. Lee & Sidford (2019, Theorem 1) consider linear programs $\min\{c^\top x : A^\top x = b,\ l \leq x \leq u\}$ in which A has full column*

rank and no zero row. Given a point x_0 with $A^\top x_0 = b$ and $l < x_0 < u$, their algorithm returns, with constant probability, a point x with $A^\top x = b$, $l < x < u$, and $c^\top x \le \mathrm{OPT} + \epsilon$. It uses $\widetilde{O}(\sqrt{\mathrm{rank}(A)}\log(U_{\mathrm{LS}}/\epsilon))$ iterations, where U_{LS} bounds $\|c\|_\infty$, $\|u-l\|_\infty$, and the inverse distance from x_0 to the bounds. Each iteration performs $\widetilde{O}(1)$ solves with $A^\top DA$ for positive diagonal D, plus nearly linear work.

For MOT, take $x = \mathrm{vec}(X)$, keep all marginal equations of the first block and all but one of every later block, so that the retained $mn - m + 1$ equations are independent (Lemma B.2), and use the nonbinding bounds $0 \le X \le 2$. After pruning zero-mass points (Lemma B.1), the product coupling $\mu_1 \otimes \cdots \otimes \mu_m$ is a valid starting point, so $\log U_{\mathrm{LS}} = O(m\log(1/\mu_{\min}) + \log(1 + \|C\|_\infty))$, where $\mu_{\min}$ is the smallest positive marginal mass. The output is an exactly feasible coupling, in the same exact-arithmetic model as Theorem B.7, with cost at most $\mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$. The system matrix has dimension $mn - m + 1$. It is formed in $O(m^2 N)$ operations and factored in $O((mn)^\omega) = O(\mathrm{poly}(m)N)$ operations for $m \ge 3$, and the algorithm's polylogarithmic factors in the number N of variables are polynomial in m. Repeating $O(\log(1/\rho))$ times and keeping the cheapest feasible output gives failure probability ρ in

$$\widetilde{O}\left(\mathrm{poly}(m)\, n^{m+1/2} \log \frac{1}{\mu_{\min}}\right)$$

arithmetic operations. For $m = 2$, where the system matrices are equivalent to bipartite-graph Laplacians, the same approach gives the $\widetilde{O}(n^{2.5})$ interior point bound that Blanchet et al. (2024) list for optimal transport. Compared with Theorem B.7, the packing algorithm is faster when $\kappa \lesssim \sqrt{n}$, up to $\mathrm{poly}(m)$ and logarithmic factors, and the interior point method is faster at higher accuracy. In particular, the $\widetilde{O}(\sqrt{mn}\, n^m)$ bound that Lin et al. (2022) state conditionally on a min-cost-flow formulation holds without that structure, up to $\mathrm{poly}(m)$ factors. These factors are large: the work bound of Lee & Sidford (2019) carries a factor $\log^{13} N = (m \log n)^{13}$. For moderate κ, first-order methods such as AAM (Tupitsa et al., 2020), whose dependence on m is m^3, therefore remain faster in m and are far simpler to implement; the interior point bound is chiefly of theoretical interest.

The packing algorithm constructs a dense tensor and therefore naturally pays its $\Omega(N)$ serialization cost. The next two sections contain the constructive quantum route analyzed in the main text: quantum Sinkhorn potentials are converted into a classical sparse coupling list with exact marginals.

C Quantum Sinkhorn for multimarginal OT

We develop a quantum multimarginal Sinkhorn method for entropy-regularized MOT. Entropic regularization converts the constrained primal problem into a smooth scaling problem with a compact description. The output scaling potentials specify a pre-rounding Gibbs tensor, so a quantum algorithm can avoid writing down all $N = n^m$ entries.

C.1 Finite-input conventions and log-domain Gibbs weights

An (s,t) signed fixed-point number is $z2^{-t}$ with $z \in \mathbb{Z}$ and $|z| < 2^{s+t}$, stored using one sign bit, s integer bits, and t fractional bits. A b-bit number has total length at most b. For the constructive quantum upper bounds, each $C_{\boldsymbol{i}}$ is an exact signed b_C-bit fixed-point number, each $\mu_{k,a}$ is a nonnegative b_μ-bit dyadic rational satisfying $\sum_a \mu_{k,a} = 1$ exactly, and $\varepsilon > 0$ is an exact b_ε-bit dyadic rational. All optimization and exact-feasibility guarantees refer to these represented data. If $\|C - C^\circ\|_\infty \le \xi_C$, the same output is $(\varepsilon + 2\xi_C)$-optimal for C° with the represented marginals; approximating the marginals instead requires a separate stability or rounding argument.

Assume $n \ge 2$, since the one-point case is trivial. For $x > 0$, define $\mathrm{dyad}(x) \triangleq 2^{\lfloor \log_2 x \rfloor}$, so $x/2 < \mathrm{dyad}(x) \le x$. Let $q_n \triangleq \lceil \log_2 n \rceil$ and fix the exactly represented full-support distribution

$$u^{\mathrm{dyad}}_{n,a} \triangleq \begin{cases} 2^{-q_n}, & a < n, \\ 1 - (n-1)2^{-q_n}, & a = n, \end{cases} \qquad \min_a u^{\mathrm{dyad}}_{n,a} \ge \frac{1}{2n}.$$

In (6), $\oplus$ is bitwise XOR with the canonical fixed-width encoding. Elementary functions, controlled rotations, and state preparation are approximated to the tolerances assigned below. If

$C_{\boldsymbol{i}} = \mathbf{c}(z_{1,i_1}, \ldots, z_{m,i_m})$, the oracle can load the represented support points, evaluate $\mathbf{c}$ coherently to $C_{\boldsymbol{i}}$, and uncompute them. Thus the tensor itself need not be stored in qRAM.

The generic scaling theorem in this section accepts coherent entry access to a positive tensor K with known bounds $K_{\min}$ and $K_{\max}$. For MOT we use log-domain access instead: the kernel $K_{\boldsymbol{i}} = \exp(-C_{\boldsymbol{i}}/\eta)$ is never evaluated, because every regularization parameter used below is a power of two and the Gibbs exponent is therefore exact.

Lemma C.1 (Exact log-domain Gibbs exponents)**.** *Let each $C_{\boldsymbol{i}}$ be an exact signed fixed-point number with at most b_C bits, let $\eta = 2^{-e}$ for an integer e, and write $b_\eta \triangleq |e| + 1$. Let β be fixed-point potentials on a common grid whose entries use at most ℓ_β bits. Then $O(1)$ calls to* (6)*, $O(m)$ potential lookups, and $O(m(b_C + b_\eta + \ell_\beta))$ elementary gates compute*

$$w_{\boldsymbol{i}}(\beta) \triangleq \sum_{k=1}^{m} \beta_{k,i_k} - \frac{C_{\boldsymbol{i}}}{\eta}$$

exactly, as a fixed-point number with $O(b_C + b_\eta + \ell_\beta + \log m)$ bits. In particular, the Gibbs weight $K_{\boldsymbol{i}} e^{\sum_k \beta_{k,i_k}} = e^{w_{\boldsymbol{i}}(\beta)}$ of the exact kernel $K_{\boldsymbol{i}} = e^{-C_{\boldsymbol{i}}/\eta}$ has an exactly represented exponent.

Proof. Dividing by $\eta = 2^{-e}$ multiplies $C_{\boldsymbol{i}}$ by 2^e, which moves the binary point by e positions. Hence $C_{\boldsymbol{i}}/\eta$ is exact and fits in a register with $|e|$ additional integer or fractional positions; forming it is a relabeling of wires together with zero padding. Adding the m potentials and $-C_{\boldsymbol{i}}/\eta$ on a common grid, in a register with $\lceil \log_2(m+1) \rceil$ additional integer bits, costs $O(m(b_C + b_\eta + \ell_\beta))$ gates with ripple-carry adders. Compute–copy–uncompute makes the circuit reversible with $O(1)$ cost-oracle calls and $O(m)$ lookups. □

Because the exponent is exact, Lemma C.2 below can be applied with $a \equiv 1$ and the exponents $w_{\boldsymbol{i}}(\beta)$. Its register widths grow only logarithmically with the range of these exponents, which is dominated by $\|C\|_\infty/\eta$ and the potential range, whereas an entrywise fixed-point representation of $K_{\boldsymbol{i}}$ would need $\Theta(\|C\|_\infty/\eta)$ fractional bits. Errors are therefore measured against the exact Gibbs tensor, and all approximation is confined to the log-sum-exp estimates of Lemma C.2. Using $r = \mu_{k,a}$ as in Lemma C.3 incorporates the target log-mass without evaluating it separately, and subtraction and recentering are exact on the common fixed-point grid.

C.2 Entropic MOT and its associated dual

Define the discrete Shannon entropy $H(X) \triangleq -\sum_{\boldsymbol{i}} X_{\boldsymbol{i}} \log X_{\boldsymbol{i}}$, with the convention $0 \log 0 = 0$. This differs from the shifted entropy convention used in Lin et al. (2022) by an additive constant on probability tensors, and therefore leaves the scaling potentials and all marginal equations unchanged. Fix a regularization parameter $\eta > 0$ and consider the *entropic regularized* MOT problem:

$$\min_{X \in \mathcal{U}(\boldsymbol{\mu})} \{\langle C, X \rangle - \eta H(X)\}. \tag{MOT$_\eta$}$$

Let $\lambda = (\lambda_1, \ldots, \lambda_m)$ be the Lagrange multipliers for the marginal constraints. Keeping the redundant mass constraint $\|X\|_1 = 1$ while dualizing the marginal equations, set

$$\beta_k \triangleq \eta^{-1} \lambda_k, \qquad k \in [m].$$

With this change of variables, and after changing sign and dividing the dual objective by η, the entropic dual is the unconstrained smooth convex problem

$$\min_{\beta_1 \in \mathbb{R}^{n_1}, \ldots, \beta_m \in \mathbb{R}^{n_m}} \Phi_\eta(\beta) \triangleq \log\left(\sum_{\boldsymbol{i}} \exp\left(\sum_{k=1}^{m} \beta_{k,i_k} - \frac{C_{\boldsymbol{i}}}{\eta} \right) \right) - \sum_{k=1}^{m} \mu_k^\top \beta_k. \tag{MOT$_\eta$-D}$$

Up to the irrelevant additive constant coming from the entropy convention, this is the dual formulation used in (Lin et al., 2022, Section 2.2).

Assuming the listed support points have positive marginal masses, the entropy term in (MOT$_\eta$) makes the primal objective strictly convex and ensures that the unique optimizer $X^\star$ is strictly positive on the product support. The Karush-Kuhn-Tucker (KKT) optimality conditions imply that this

solution admits the Gibbs representation

$$X^\star_{\boldsymbol{i}} = \widehat{K}(\beta)_{\boldsymbol{i}} \triangleq \frac{K_{\boldsymbol{i}} \exp(\sum_{k=1}^m \beta_{k,i_k})}{\sum_{\boldsymbol{i}'} K_{\boldsymbol{i}'} \exp\Big(\sum_{k=1}^m \beta_{k,i'_k}\Big)}, \qquad K_{\boldsymbol{i}} \triangleq \exp(-C_{\boldsymbol{i}}/\eta), \tag{18}$$

for log-scaling potentials $\beta_k \in \mathbb{R}^{n_k}$ chosen so that $\mathrm{Proj}_k(X) = \mu_k$ for all $k \in [m]$. Equivalently, if

$$K(\beta)_{\boldsymbol{i}} \triangleq K_{\boldsymbol{i}} \exp\left(\sum_{k=1}^m \beta_{k,i_k}\right),$$

then solving (MOT_η) is exactly the problem of scaling the positive Gibbs tensor K so that $\widehat{K}(\beta) = K(\beta)/\|K(\beta)\|_1 \in \mathcal{U}(\boldsymbol{\mu})$. After a harmless gauge shift of the potentials, one may also impose $\|K(\beta)\|_1 = 1$ and view the optimizer as the unnormalized scaled tensor $K(\beta)$. Note that the potentials λ, β are unrelated to the LP dual variables y from the preceding section.

The dual perspective leads to the same observation. Since $\Phi_\eta(\beta) = \log \|K(\beta)\|_1 - \sum_k \mu_k^\top \beta_k$, we have

$$\nabla_{\beta_k} \Phi_\eta(\beta) = \mathrm{Proj}_k(\widehat{K}(\beta)) - \mu_k, \qquad \widehat{K}(\beta) \triangleq \frac{K(\beta)}{\|K(\beta)\|_1}.$$

Thus the marginal errors are exactly the block gradients of the entropic dual. The multimarginal Sinkhorn iterate of (Lin et al., 2022, Algorithm 1) is a block coordinate step on (MOT_η-D): it chooses the next block by a greedy KL-residual rule and updates that block so that the corresponding marginal is corrected. Below we use the same one-block scaling update in normalized, gauge-equivalent notation, but choose the block randomly for the quantum implementation. The quantum task is therefore to implement these block updates without summing explicitly over n^{m-1} entries in every tensor fiber.

C.3 Quantum Multimarginal Sinkhorn

We now spell out the randomized coordinate-descent iteration that we want to implement quantumly. For the complexity statements in this subsection, assume $n_1 = \cdots = n_m = n$. Let $K \in \mathbb{R}_{>0}^{[n]^m}$ be a positive tensor and let $\mu_k \in \Delta_{>0}^n$ be the target marginals. In this subsection we write

$$Z(\beta) \triangleq \sum_{\boldsymbol{i}} K_{\boldsymbol{i}} \exp\left(\sum_{\ell=1}^m \beta_{\ell,i_\ell}\right), \qquad \Phi_K(\beta) \triangleq \log Z(\beta) - \sum_{\ell=1}^m \mu_\ell^\top \beta_\ell.$$

For the Gibbs tensor $K_{\boldsymbol{i}} = \exp(-C_{\boldsymbol{i}}/\eta)$, this is exactly the dual objective Φ_η in (MOT_η-D). The normalized scaled tensor is

$$\widehat{K}(\beta)_{\boldsymbol{i}} \triangleq \frac{K_{\boldsymbol{i}} \exp(\sum_\ell \beta_{\ell,i_\ell})}{Z(\beta)}.$$

The current k-th marginal and total marginal error are

$$q_k(\beta) \triangleq \mathrm{Proj}_k(\widehat{K}(\beta)), \qquad E(\beta) \triangleq \sum_{k=1}^m \|q_k(\beta) - \mu_k\|_1.$$

Starting from $\beta^{(0)} = 0$, iteration t samples a mode

$$k_t \sim \mathrm{Unif}([m]).$$

For a selected mode k, the exact one-block update is

$$\beta^+_{k,a} = \beta_{k,a} + \log \mu_{k,a} - \log q_{k,a}(\beta), \qquad \beta^+_\ell = \beta_\ell \quad (\ell \neq k). \tag{19}$$

At iteration t, we apply (19) with $k = k_t$ and $\beta = \beta^{(t)}$, followed by the usual harmless recentering of the potentials. This update is the same one-block scaling step used in multimarginal Sinkhorn, written for the normalized tensor $\widehat{K}(\beta)$; it corrects the selected marginal to μ_{k_t}, while the other marginals are left to future random updates. The difference from the greedy implementation of (Lin et al., 2022, Algorithm 1) is the randomized choice of the selected mode.

Thus a quantum implementation has one main computational task per iteration: estimate the vector $\log q_{k_t}(\beta^{(t)})$ without explicitly summing over every fiber entry. For a fixed mode k and coordinate a, define the unnormalized fiber mass

$$S_{k,a}(\beta) \triangleq \sum_{\boldsymbol{i}:i_k=a} K_{\boldsymbol{i}} \exp\left(\sum_{\ell=1}^{m} \beta_{\ell,i_\ell}\right), \qquad q_{k,a}(\beta) = \frac{S_{k,a}(\beta)}{Z(\beta)}.$$

Therefore

$$\log q_{k,a}(\beta) = \log S_{k,a}(\beta) - \log Z(\beta).$$

The first term is a log-sum-exp over a fiber of size n^{m-1}, and the second is a log-sum-exp over the whole tensor. The algorithmic plan is consequently:

(i) use a quantum subroutine to estimate $\log S_{k_t,a}(\beta^{(t)})$ for every a, and $\log Z(\beta^{(t)})$,
(ii) perform the Sinkhorn update with these noisy log-marginal estimates,
(iii) prove that the resulting noisy randomized coordinate descent reaches $E(\beta) \le \delta$.

The next lemmas carry out these three steps. Lemma C.2 states the quantum LogSumExp primitive for estimating quantities of the form

$$\log\left(\frac{1}{r}\sum_{j=1}^{d} a_j \exp(w_j)\right).$$

Its query algorithm and query bound specialize Nieuwboer (2024, Theorem 13.3.2); the normalization by r and the reversible fixed-point gate accounting are made explicit in the statement and proof below. Lemma C.3 converts this subroutine into a one-mode marginal oracle, and Lemma C.4 proves correctness and robustness of one update. Theorem C.9 combines these ingredients into the final complexity bound.

Lemma C.2 (Approximate log-sum-exp; explicit specialization of Nieuwboer (2024))**.** *Let $d \ge 1$. Let $a \in [\gamma, 1]^d \cap \mathbb{Q}^d$ and $r \in (0,1] \cap \mathbb{Q}$ have exact binary encodings of at most ℓ_a and ℓ_r bits, respectively. Let $w \in [-B, B]^d$ be stored exactly in signed fixed-point format with b_I integer bits (including the sign bit) and b_F fractional bits. Assume reversible value-oracle access*

$$\mathcal{O}_a : |j\rangle|z\rangle \longmapsto |j\rangle|z \oplus a_j\rangle, \qquad \mathcal{O}_w : |j\rangle|z\rangle \longmapsto |j\rangle|z \oplus w_j\rangle.$$

Let $\delta, \zeta \in (0,1) \cap \mathbb{Q}$ have exact encodings of lengths ℓ_δ and ℓ_ζ, and define

$$L \triangleq \log\left(\frac{1}{r}\sum_{j=1}^{d} a_j e^{w_j}\right).$$

Suppose the output register has at least

$$b_I \ge 2 + \left\lceil \log_2\left(1 + B + \log\frac{d}{\gamma r}\right)\right\rceil, \qquad b_F \ge \left\lceil \log_2 \frac{8}{\delta}\right\rceil. \tag{20}$$

Then a quantum algorithm outputs a fixed-point number Λ satisfying $|\Lambda - L| \le \delta$ with probability at least $1 - \zeta$, using

$$O\left(\frac{\sqrt{d}}{\delta}\log\frac{1}{\zeta}\right)$$

total calls to $\mathcal{O}_a$ and $\mathcal{O}_w$. If $\mathsf{Mult}(q)$ is the gate cost of multiplying two q-bit integers and

$$b_{\mathrm{LSE}} \triangleq O\left(\ell_a + \ell_r + \ell_\delta + \ell_\zeta + b_I + b_F + \log\frac{d}{\delta\gamma r}\right), \tag{21}$$

then the non-oracle elementary-gate complexity is

$$\widetilde{O}\left(\frac{\sqrt{d}}{\delta}\log\frac{1}{\zeta}\,\mathsf{Mult}(b_{\mathrm{LSE}})\right).$$

In particular, quadratic fixed-point multiplication gives the same bound with $\mathsf{Mult}(b_{\mathrm{LSE}}) = O(b_{\mathrm{LSE}}^2)$.

Proof. We apply Nieuwboer (2024, Theorem 13.3.2) with the following dictionary between its parameters and ours:

$$\gamma_0 \triangleq \min\left\{\gamma, \frac{r}{2}\right\}, \qquad (n, a, r, y, \text{precision}, \text{failure}, \mu) = (d, a, r, w, \delta, \zeta, \gamma_0).$$

Its hypotheses require exact rational encodings of a and r, fixed-point access to w, and a lower bound μ on every nonzero entry of a; these are precisely the assumptions above. The smaller source lower bound costs at most an additional $O(\log(1/(\gamma r)))$ bits, already included in (21). Moreover,

$$\frac{d\gamma e^{-B}}{r} \le e^L \le \frac{de^B}{r}, \qquad |L| \le B + \log\frac{d}{\gamma r}.$$

Thus (20) supplies the integer and fractional output widths required by the cited theorem. Its query bound is exactly $O(\sqrt{d}\,\delta^{-1}\log(1/\zeta))$.

For completeness, we expose the normalization used by that theorem. Its Algorithm 13.3 first rounds each rational a_j upward using

$$b_a = \left\lceil \log_2 \frac{4}{\delta\gamma_0} \right\rceil$$

fractional bits. Lemma 13.3.8 there shows that this changes the target logarithm by at most $\delta/4$. Approximate quantum maximum finding then returns $j^\star$ such that

$$\frac{a_j}{a_{j^\star}} e^{w_j - w_{j^\star}} \le \frac{3}{2} \qquad (j \in [d]).$$

The algorithm coherently approximates and clips one half of these relative weights into $[0, 1]$, applies quantum approximate summation at relative accuracy $\Theta(\delta)$, and reconstructs

$$L = w_{j^\star} + \log a_{j^\star} + \log\left(\sum_{j=1}^{d} \frac{a_j}{a_{j^\star}} e^{w_j - w_{j^\star}}\right) - \log r.$$

The error allocation in the proof of Theorem 13.3.2 assigns the remaining budget to relative-weight approximation, approximate summation, evaluation of the three logarithms, and final fixed-point truncation, yielding total additive error at most δ.

It remains only to make explicit the bit cost hidden by the cited theorem. The rounded weights use $O(\log(1/(\delta\gamma_0)))$ fractional bits, the approximate sum uses $O(\log(d/\delta))$ bits, and r, w, and the output use their declared encodings. Hence every ratio, range-reduced exponential, product, logarithm, and comparison in Algorithm 13.3 fits in $O(b_{\mathrm{LSE}})$ bits. Division, exponentiation, and logarithm evaluation cost $\widetilde{O}(\mathsf{Mult}(b_{\mathrm{LSE}}))$ gates (Brent & Zimmermann, 2011, Chapter 4), and compute–copy–uncompute makes these circuits reversible with the same asymptotic cost. Multiplying by the cited number of coherent evaluations gives the displayed gate bound. □

Lemma C.3 (Quantum marginal estimation)**.** *Suppose $n_1 = \cdots = n_m = n$, every entry of a positive rational tensor K, as well as the known rational bounds $K_{\min}$ and $K_{\max}$, has an exact encoding of at most ℓ_K bits, and $0 < K_{\min} \le K_{\boldsymbol{i}} \le K_{\max}$ hold for every $\boldsymbol{i} \in [n]^m$. Assume exact coherent entry access to K and coherent lookup of fixed-point potentials β whose entries use at most ℓ_β bits. Let $\|\beta\|_\infty \le B$ and fix $k \in [m]$. For represented rational $\alpha, p \in (0, 1)$ with encoding lengths ℓ_α and ℓ_p, there is a quantum algorithm which outputs numbers $\widetilde{\ell}_{k,a}$ satisfying*

$$\left|\widetilde{\ell}_{k,a} - \log q_{k,a}(\beta)\right| \le \alpha, \qquad a \in [n],$$

with probability at least $1 - p$, using $\widetilde{O}(n^{(m+1)/2}\alpha^{-1})$ coherent kernel queries and $\widetilde{O}(mn^{(m+1)/2}\alpha^{-1})$ potential lookups and unit-cost arithmetic operations. If

$$b_{\mathrm{marg}} \triangleq O\left(\ell_K + \ell_\beta + \ell_\alpha + \ell_p + \log\frac{K_{\max}}{K_{\min}} + m\log n + \log\frac{m}{\alpha p} + \log(1 + B)\right),$$

then its non-oracle elementary-gate complexity is $\widetilde{O}(mn^{(m+1)/2}\alpha^{-1}\mathsf{Mult}(b_{\mathrm{marg}}))$. Given in addition an exact rational target mass $\mu_{k,a} > 0$, the same bounds, with its encoding length

and $\log(1/\mu_{k,a})$ *included in* b_{marg}*, produce an* α*-accurate estimate of the complete correction* $\log q_{k,a}(\beta) - \log \mu_{k,a}$*. Finally, suppose instead that* $K_{\boldsymbol{i}} = e^{-C_{\boldsymbol{i}}/\eta}$ *with the costs and* $\eta = 2^{-e}$ *of Lemma C.1, that* $0 \le C_{\boldsymbol{i}} \le \Delta$*, and that the cost oracle* (6) *replaces the entry oracle. Then the same bounds hold with cost queries in place of kernel queries, and* $\ell_K + \log(K_{\max}/K_{\min})$ *may be replaced by* $b_C + b_\eta + \log(1 + \Delta/\eta)$ *in* b_{marg}*.*

Proof. Rescale K by the known upper bound $K_{\max}$. This adds the same constant $-\log K_{\max}$ to all log-sums and does not change any normalized marginal $q_k(\beta)$. Thus the entries used by Lemma C.2 lie in $[\gamma, 1]$ with $\gamma = K_{\min}/K_{\max}$, and the common shift cancels when a fiber log-sum is subtracted from the full tensor log-sum. We invoke Lemma C.2 with $r = 1$ and exponent bound mB, since $|\sum_\ell \beta_{\ell,i_\ell}| \le mB$.

For each $a \in [n]$, the fiber defining $S_{k,a}(\beta)$ has n^{m-1} terms. Applying Lemma C.2 with $r = 1$, accuracy $\alpha/4$, and failure probability $p/(2n)$ gives an $\alpha/4$-accurate estimate of $\log S_{k,a}(\beta)$ using $\widetilde{O}(n^{(m-1)/2}\alpha^{-1})$ coherent evaluations, each of which makes one kernel query and m potential lookups. Applying the same primitive once to the full tensor with $r = 1$ and failure probability $p/2$ gives an $\alpha/4$-accurate estimate of $\log Z(\beta)$ using $\widetilde{O}(n^{m/2}\alpha^{-1})$ coherent evaluations. The n fiber estimates dominate both resource counts. Subtracting the estimate of $\log Z(\beta)$ from the fiber estimates gives every $\log q_{k,a}(\beta)$ to additive error at most $\alpha/2$, and hence at most α. A union bound over the n fiber estimates and the full-tensor estimate gives success probability at least $1 - p$. Substitution in (21), where $\log d \le m \log n$, gives b_{marg} and the stated gate bound.

To compute an update correction directly, invoke the fiber call with $r = \mu_{k,a}$. It returns an estimate of $\log(S_{k,a}(\beta)/\mu_{k,a})$; subtracting the full-tensor estimate gives $\log q_{k,a}(\beta) - \log \mu_{k,a}$. Thus no separately rounded logarithm of $\mu_{k,a}$ is needed. Only b_{marg} changes, through the exact encoding of r and the term $\log(1/r)$.

For the log-domain form, invoke Lemma C.2 with $a \equiv 1$, so that $\gamma = 1$ and no rescaling is needed, and with the exact exponents $w_{\boldsymbol{i}}(\beta)$ of Lemma C.1. They satisfy $|w_{\boldsymbol{i}}(\beta)| \le mB + \Delta/\eta$ and use $O(b_C + b_\eta + \ell_\beta + \log m)$ bits, so (20) and (21) give the stated replacement. Each coherent evaluation then makes $O(1)$ cost queries and m potential lookups. □

For the descent analysis of (19), set

$$\mathsf{KL}_k(\beta) \triangleq \mathsf{KL}(\mu_k \| q_k(\beta)) = \sum_a \mu_{k,a} \log \frac{\mu_{k,a}}{q_{k,a}(\beta)}.$$

Lemma C.4 (Exact and noisy one-mode Sinkhorn descent)**.** *The exact update* (19) *makes the selected marginal equal to the corresponding target marginal and satisfies*

$$\Phi_K(\beta) - \Phi_K(\beta^+) = \mathsf{KL}_k(\beta). \tag{22}$$

Now suppose instead that the update uses estimates

$$\widetilde{\ell}_a = \log q_{k,a}(\beta) + e_a, \qquad |e_a| \le \alpha,$$

and sets $\widetilde{\beta}^+_{k,a} = \beta_{k,a} + \log \mu_{k,a} - \widetilde{\ell}_a$*, leaving the other blocks fixed. Then*

$$\Phi_K(\beta) - \Phi_K(\widetilde{\beta}^+) \ge \mathsf{KL}_k(\beta) - \frac{\alpha^2}{2}.$$

Proof. Let $Z = Z(\beta)$ and $F_a = S_{k,a}(\beta) = Z q_{k,a}(\beta)$. Under the exact update, the new unnormalized fiber mass is

$$F_a^+ = F_a \frac{\mu_{k,a}}{q_{k,a}(\beta)} = Z\mu_{k,a}.$$

Therefore $Z(\beta^+) = \sum_a F_a^+ = Z$ and the normalized k-th marginal is exactly μ_k. Since only block k changes,

$$\Phi_K(\beta) - \Phi_K(\beta^+) = \sum_a \mu_{k,a}(\beta^+_{k,a} - \beta_{k,a}) = \sum_a \mu_{k,a} \log \frac{\mu_{k,a}}{q_{k,a}(\beta)},$$

proving (22).

For the noisy update, the new fiber mass is $\widetilde{F}_a^+ = Z\mu_{k,a}e^{-e_a}$ and hence

$$Z(\widetilde{\beta}^+) = Z\sum_a \mu_{k,a}e^{-e_a}.$$

A direct substitution in Φ_K gives

$$\Phi_K(\beta) - \Phi_K(\widetilde{\beta}^+) = \mathsf{KL}_k(\beta) - \sum_a \mu_{k,a}e_a - \log\left(\sum_a \mu_{k,a}e^{-e_a}\right).$$

Hoeffding's lemma applied to the random variable $-e_a$, with $a \sim \mu_k$ and range length at most 2α, gives

$$\log\left(\sum_a \mu_{k,a}e^{-e_a}\right) \le -\sum_a \mu_{k,a}e_a + \frac{\alpha^2}{2}.$$

The claimed noisy descent bound follows. □

Define

$$K_{\min} \triangleq \min_{\boldsymbol{i}} K_{\boldsymbol{i}}, \qquad K_{\max} \triangleq \max_{\boldsymbol{i}} K_{\boldsymbol{i}}, \qquad \mu_{\min} \triangleq \min_{k,a} \mu_{k,a},$$

and

$$R_K \triangleq \log\frac{K_{\max}}{K_{\min}} + \log\frac{1}{\mu_{\min}}.$$

For finite control flow we use the represented integer

$$\widehat{R}_K \triangleq 1 + \left\lceil \log_2 \frac{K_{\max}}{K_{\min}} \right\rceil + \left\lceil \log_2 \frac{1}{\mu_{\min}} \right\rceil, \tag{23}$$

which is computable from the rational bounds and satisfies $\widehat{R}_K \ge \max\{1, R_K\}$. The next estimates adapt the range, objective-gap, and progress arguments behind multimarginal Sinkhorn to noisy log-marginal estimates and uniform random block selection. For the exact update, averaging (22) over a uniform random mode and applying Pinsker's inequality gives

$$\mathbb{E}_{k\sim\mathrm{Unif}([m])}\left[\Phi_K(\beta) - \Phi_K(\beta^+)\right] = \frac{1}{m}\sum_{k=1}^m \mathsf{KL}_k(\beta) \ge \frac{E(\beta)^2}{2m^2}. \tag{24}$$

The quantum implementation uses noisy log-marginals, so we also need a bound on the size of the potentials seen by the log-sum-exp primitive. The following elementary range bound is the noisy analogue of the range induction in Lin et al. (2022). It is invariant under adding constants to individual blocks; in the finite implementation we recenter a block by subtracting its represented minimum, which preserves the common fixed-point grid.

Lemma C.5 (Noisy Sinkhorn potential range)**.** *Suppose that an update in mode k is performed with errors $|e_a| \le \alpha$ as in Lemma C.4. After recentering the updated block by an additive constant, the updated potentials satisfy*

$$\max_a \beta_{\ell,a} - \min_a \beta_{\ell,a} \le R_K + 2\alpha, \qquad \ell \in [m],$$

provided the same bound held before the update.

Proof. Adding constants to the blocks changes $K(\beta)$ by a global multiplicative factor and hence leaves $\widehat{K}(\beta)$, all marginals, and $\Phi_K(\beta)$ unchanged. It therefore suffices to control block ranges.

Only block k changes. Write

$$S_{k,a}(\beta) = e^{\beta_{k,a}}T_a, \qquad T_a \triangleq \sum_{\boldsymbol{i}:i_k=a} K_{\boldsymbol{i}} \exp\left(\sum_{\ell\ne k}\beta_{\ell,i_\ell}\right).$$

The noisy update gives

$$\widetilde{\beta}^+_{k,a} = \log\mu_{k,a} - \log T_a + \log Z(\beta) - e_a.$$

For any a, a', the two sums T_a and $T_{a'}$ have the same index set in the remaining $m-1$ modes and termwise weights that differ only through K. Hence

$$\frac{K_{\min}}{K_{\max}} \le \frac{T_a}{T_{a'}} \le \frac{K_{\max}}{K_{\min}}.$$

Also $\mu_{k,a}/\mu_{k,a'} \le 1/\mu_{\min}$. Therefore

$$|\widetilde{\beta}^{+}_{k,a} - \widetilde{\beta}^{+}_{k,a'}| \le \log \frac{K_{\max}}{K_{\min}} + \log \frac{1}{\mu_{\min}} + 2\alpha = R_K + 2\alpha.$$

The other blocks are unchanged except for their harmless recentering constants, so the claimed range bound is preserved. □

Lemma C.6 (Optimal entropic potential range)**.** *There is an optimal solution $\beta^\star$ of Φ_K whose blocks satisfy*

$$\max_a \beta^\star_{k,a} - \min_a \beta^\star_{k,a} \le R_K, \qquad k \in [m].$$

Proof. At an optimum, the normalized tensor $\widehat{K}(\beta^\star)$ has marginals μ_k, so for each block k and coordinate a,

$$\mu_{k,a} = \frac{e^{\beta^\star_{k,a}} T_a}{Z(\beta^\star)}, \qquad T_a \triangleq \sum_{\boldsymbol{i}: i_k = a} K_{\boldsymbol{i}} \exp\left(\sum_{\ell \ne k} \beta^\star_{\ell, i_\ell} \right).$$

For any a, a', the sums T_a and $T_{a'}$ have the same index set in the remaining modes and termwise weights that differ only through K, so

$$\frac{K_{\min}}{K_{\max}} \le \frac{T_a}{T_{a'}} \le \frac{K_{\max}}{K_{\min}}.$$

Also $\mu_{k,a}/\mu_{k,a'} \in [\mu_{\min}, \mu_{\min}^{-1}]$. Therefore

$$|\beta^\star_{k,a} - \beta^\star_{k,a'}| \le \log \frac{K_{\max}}{K_{\min}} + \log \frac{1}{\mu_{\min}} = R_K.$$

The objective is invariant under adding constants to the individual blocks, so we may choose any centered representative of this optimal equivalence class. □

Lemma C.7 (Robust Sinkhorn objective gap)**.** *Let β be any potentials whose blocks satisfy*

$$\max_a \beta_{k,a} - \min_a \beta_{k,a} \le R_K + 2\alpha, \qquad k \in [m].$$

Then

$$\Phi_K(\beta) - \Phi^\star_K \le (R_K + \alpha) E(\beta).$$

Proof. By Lemma C.6, there is an optimizer $\beta^\star$ whose block ranges are at most R_K. Convexity gives

$$\Phi_K(\beta) - \Phi^\star_K \le \sum_{k=1}^{m} (\beta_k - \beta^\star_k)^\top (q_k(\beta) - \mu_k).$$

Since both $q_k(\beta)$ and μ_k have mass one, each vector $q_k(\beta) - \mu_k$ has zero sum. Thus we may subtract the midpoint of each block of β and $\beta^\star$ inside the inner products. The centered β_k has ℓ_∞ norm at most $(R_K + 2\alpha)/2$, and the centered $\beta^\star_k$ has ℓ_∞ norm at most $R_K/2$. Therefore

$$(\beta_k - \beta^\star_k)^\top (q_k(\beta) - \mu_k) \le (R_K + \alpha) \|q_k(\beta) - \mu_k\|_1.$$

Summing over k proves the claim. □

Lemma C.8 (Stochastic drift and occupation estimate)**.** *Let $(G_t)_{0 \le t \le T}$ be a bounded nonnegative process adapted to a filtration $(\mathcal{F}_t)_{0 \le t \le T}$, where $T \ge 1$ is an integer, and let $B_t \in \mathcal{F}_t$ be events. Let $R, \delta, c > 0$ and $\chi \ge 0$, and define*

$$h(s) \triangleq c \max\left\{ \frac{\delta^2}{m^2}, \frac{s^2}{m^2 R^2} \right\}.$$

Suppose that, for every $0 \le t < T$,

$$\mathbb{E}[G_t - G_{t+1} \mid \mathcal{F}_t] \ge h(G_t) \quad \text{on } B_t, \qquad \mathbb{E}[G_{t+1} - G_t \mid \mathcal{F}_t] \le \frac{\chi c\delta^2}{m^2} \quad \text{on } B_t^c.$$

Then

$$\mathbb{E}\Big[\sum_{t=0}^{T-1} \mathbb{1}_{B_t}\Big] \le \frac{2m^2 R}{c\delta} + \chi T.$$

Proof. Let $\Psi(u) \triangleq \int_0^u ds/h(s)$ for $u \ge 0$. Since h is positive and nondecreasing, Ψ is nonnegative, increasing, and concave, and $\Psi' \le m^2/(c\delta^2)$. Moreover,

$$\sup_{u \ge 0} \Psi(u) \le \int_0^{R\delta} \frac{m^2}{c\delta^2}\, ds + \int_{R\delta}^{\infty} \frac{m^2 R^2}{cs^2}\, ds = \frac{2m^2 R}{c\delta}.$$

On B_t, the drift assumption gives $\mathbb{E}[G_{t+1} \mid \mathcal{F}_t] \le G_t - h(G_t)$, and $G_t - h(G_t) \ge 0$ because $G_{t+1} \ge 0$. Jensen's inequality, monotonicity of Ψ, and $h(s) \le h(G_t)$ for $s \le G_t$ give

$$\mathbb{E}[\Psi(G_{t+1}) \mid \mathcal{F}_t] \le \Psi(G_t - h(G_t)) \le \Psi(G_t) - 1.$$

On B_t^c, the same argument and the derivative bound give

$$\mathbb{E}[\Psi(G_{t+1}) \mid \mathcal{F}_t] \le \Psi\left(G_t + \frac{\chi c\delta^2}{m^2}\right) \le \Psi(G_t) + \chi.$$

Hence $\mathbb{E}[\Psi(G_{t+1}) - \Psi(G_t) \mid \mathcal{F}_t] \le \chi - \mathbb{1}_{B_t}$ for every t. Summing over $t < T$ and taking expectations yields

$$\mathbb{E}\Big[\sum_{t=0}^{T-1} \mathbb{1}_{B_t}\Big] \le \Psi(G_0) - \mathbb{E}[\Psi(G_T)] + \chi T \le \frac{2m^2 R}{c\delta} + \chi T.$$

□

Theorem C.9 (Quantum randomized multimarginal Sinkhorn)**.** *Let* $K \in \mathbb{Q}_{>0}^{[n]^m}$ *and* $\mu_1, \dots, \mu_m \in \Delta_{>0}^n \cap \mathbb{Q}^n$ *have exact encodings of at most* ℓ_K *and* ℓ_μ *bits per entry, and let* $0 < K_{\min} \le K_{\boldsymbol{i}} \le K_{\max}$ *be known rational bounds. Assume exact coherent entry access to* K *and the marginals. For every represented* $\delta \in (0,1)$*, there is a quantum algorithm which, with probability at least* $8/9$*, outputs fixed-point potentials* $\beta = (\beta_1, \dots, \beta_m)$ *such that*

$$\sum_{k=1}^{m} \| \operatorname{Proj}_k(\widehat{K}(\beta)) - \mu_k \|_1 \le \delta.$$

Let ℓ_δ *be the encoding length of* δ*. The algorithm uses*

$$\widetilde{O}\left(m^3 n^{\frac{m+1}{2}} \frac{\widehat{R}_K}{\delta^2}\right)$$

coherent kernel queries and $\widetilde{O}(m^4 n^{(m+1)/2} \widehat{R}_K \delta^{-2})$ *working-vector lookups and unit-cost arithmetic operations. If*

$$b_{\text{work}} \triangleq O\left(\ell_K + \ell_\mu + \ell_\delta + \widehat{R}_K + m \log n + \log \frac{m\widehat{R}_K}{\delta}\right),$$

then its non-oracle elementary-gate complexity, excluding implementation of the kernel and lookup oracles, is $\widetilde{O}(m^4 n^{(m+1)/2} \widehat{R}_K \delta^{-2} \mathsf{Mult}(b_{\text{work}}))$*. Under the log-domain access of Lemma C.3, where* $K_{\boldsymbol{i}} = e^{-C_{\boldsymbol{i}}/\eta}$ *need not be rational and any computable integer at least* $\max\{1, R_K\}$ *may replace* $\widehat{R}_K$*, the same bounds hold with cost queries in place of kernel queries and with* b_{work} *replaced by*

$$b_{\text{work}}^{\log} \triangleq O\left(b_C + b_\eta + \ell_\mu + \ell_\delta + m \log n + \log \frac{m\widehat{R}_K}{\delta}\right).$$

Algorithm 2 Noisy quantum randomized multimarginal Sinkhorn

Require: Coherent access to $K > 0$, known $K_{\min}, K_{\max}$, marginals $\boldsymbol{\mu} > 0$, $\delta \in (0,1)$

$\beta \leftarrow 0$, $\alpha \leftarrow \delta/(32m)$, $T \leftarrow \lceil 1536\, m^2 \widehat{R}_K/\delta \rceil$
Sample the output time $t^\star \sim \mathrm{Unif}(\{0,\dots,T-1\})$
for $t = 0,\dots,t^\star - 1$ **do**
  Sample $k \sim \mathrm{Unif}([m])$
  Estimate $\widetilde{d}_a \approx \log q_{k,a}(\beta) - \log \mu_{k,a}$ to total accuracy α for every $a \in [n]$
  $\beta_{k,a} \leftarrow \beta_{k,a} - \widetilde{d}_a$ for every $a \in [n]$
  Recenter block β_k by subtracting $\min_a \beta_{k,a}$
  if some block range exceeds $\widehat{R}_K + 2\alpha$ **then**
    return FAILURE
  end if
end for
return β

Proof. Set $\alpha \triangleq \delta/(32m)$, $T \triangleq \lceil 1536\, m^2 \widehat{R}_K/\delta \rceil$, and $p \triangleq 1/(48T)$. Before any update, the algorithm samples $t^\star$ uniformly from $\{0,\dots,T-1\}$, independently of everything else. Starting from $\beta^{(0)} = 0$, for $t = 0,\dots,t^\star - 1$ it samples k_t uniformly from $[m]$ and uses the target-mass assertion of Lemma C.3, with $r = \mu_{k_t,a}$, to produce fixed-point corrections

$$\widetilde{d}_a = \log q_{k_t,a}(\beta^{(t)}) - \log \mu_{k_t,a} + e_a, \qquad |e_a| \le \alpha.$$

It performs the noisy update $\beta_{k_t,a} \leftarrow \beta_{k_t,a} - \widetilde{d}_a$ and recenters the updated block by subtracting its represented minimum. The subtraction and recentering are exact operations on the common fixed-point grid, so e_a contains the full approximation error of the implemented update. Recentering multiplies $K(\beta)$ by a global constant and therefore leaves $\widehat{K}(\beta)$, all marginals, and Φ_K unchanged. The output is $\beta^{(t^\star)}$; no residual certificate is computed.

Every initial block has range zero, so the hypothesis of Lemma C.5 holds at $t = 0$ and is preserved by every update whose estimates are α-accurate. After recentering, the algorithm checks in $O(M)$ time that every block range is at most $\widehat{R}_K + 2\alpha$ and aborts otherwise; since $R_K \le \widehat{R}_K$, Lemma C.5 shows that it can abort only after an estimator failure. Every represented block then lies in $[0, \widehat{R}_K + 2\alpha]$, so every quantum call receives potentials with $\|\beta\|_\infty = O(\widehat{R}_K)$, as required for the worst-case cost, regardless of earlier estimator outputs. Choosing the output widths according to (20) prevents overflow, while $b_F = O(\log(m/\delta))$ suffices for the requested update accuracy.

To avoid conditioning on estimator success, we analyze an auxiliary safe process. It runs all T updates with the same random choices and, whenever an estimator call fails, replaces the estimate by the exact correction rounded to the estimator's fixed-point output grid. We choose that grid with spacing at most $\alpha/4$, so the replacement has error at most α and remains a valid represented input to every later estimator call. Every safe update therefore satisfies Lemma C.4, and every safe iterate satisfies the range hypothesis of Lemma C.7. Each of the at most T estimator calls fails with conditional probability at most p given the complete past, so the tower property and a union bound show that all calls succeed except with probability at most $Tp = 1/48$. On that event, the executable and safe trajectories agree through time $t^\star$.

Let $\mathcal{F}_t$ be the history of the safe process through time t, let $G_t \triangleq \Phi_K(\beta^{(t)}) - \Phi_K^\star$, and call t bad if $E(\beta^{(t)}) > \delta/2$. Lemma C.7, $E \le 2m$, and $R_K + \alpha \le 2\widehat{R}_K$ give $0 \le G_t \le 4m\widehat{R}_K$. Since k_t is uniform given $\mathcal{F}_t$, Lemma C.4 and (24) give

$$\mathbb{E}[G_t - G_{t+1} \mid \mathcal{F}_t] \ge \frac{1}{m}\sum_{k=1}^{m} \mathsf{KL}_k(\beta^{(t)}) - \frac{\alpha^2}{2} \ge \frac{E(\beta^{(t)})^2}{2m^2} - \frac{\delta^2}{2048m^2}. \tag{25}$$

If t is bad, then $\delta^2 < 4E(\beta^{(t)})^2$, so the right-hand side is at least $E(\beta^{(t)})^2/(4m^2)$. Lemma C.7 also gives $E(\beta^{(t)}) \ge G_t/(2\widehat{R}_K)$. With $R \triangleq 2\widehat{R}_K$ and $c \triangleq 1/16$, every bad t therefore satisfies

$$\mathbb{E}[G_t - G_{t+1} \mid \mathcal{F}_t] \ge c \max\left\{ \frac{\delta^2}{m^2}, \frac{G_t^2}{m^2 R^2} \right\}.$$

If t is not bad, (25) and $\mathsf{KL}_k \geq 0$ give $\mathbb{E}[G_{t+1} - G_t \mid \mathcal{F}_t] \leq \delta^2/(2048m^2) = \chi c\delta^2/m^2$ with $\chi \triangleq 1/128$. Lemma C.8 therefore bounds the expected number of bad times in $\{0, \ldots, T-1\}$ by

$$\frac{2m^2 R}{c\delta} + \chi T = \frac{64m^2 \widehat{R}_K}{\delta} + \frac{T}{128} \leq \left(\frac{1}{24} + \frac{1}{128}\right) T.$$

Because $t^\star$ is uniform on $\{0, \ldots, T-1\}$ and independent of the safe process, $\Pr[E(\beta^{(t^\star)}) > \delta/2]$ equals this expectation divided by T. A union bound with the estimator-failure event shows that the executable output satisfies $E(\beta) \leq \delta/2$ except with probability at most $1/24 + 1/128 + 1/48 < 1/9$.

Each update calls Lemma C.3 once at accuracy $\alpha = \Theta(\delta/m)$, which costs $\widetilde{O}(mn^{(m+1)/2}/\delta)$ kernel queries and $\widetilde{O}(m^2 n^{(m+1)/2}/\delta)$ potential lookups. At most $T = O(m^2 \widehat{R}_K/\delta)$ updates are performed, which yields $\widetilde{O}(m^3 n^{(m+1)/2} \widehat{R}_K/\delta^2)$ kernel queries and $\widetilde{O}(m^4 n^{(m+1)/2} \widehat{R}_K/\delta^2)$ lookups. Substituting the exact input lengths, $p = 1/(48T)$, the range bound on β, and $\mu_{\min}$ into b_{marg}, or into its log-domain form, gives b_{work} or $b^{\log}_{\mathrm{work}}$ and the stated gate complexity. Certifying every iterate, as a deterministic stopping rule would require, would cost m marginal estimates per iteration; the random output time removes this factor. □

We now specialize to entropy-regularized MOT. For $K_{\boldsymbol{i}} = \exp(-C_{\boldsymbol{i}}/\eta)$, scaling K to the target marginals gives the optimal solution to (MOT_η).

Corollary C.10 (Quantum entropic MOT via multimarginal Sinkhorn)**.** *Suppose $n_1 = \cdots = n_m = n$, the nonnegative cost entries and positive marginals have exact encodings of at most b_C and b_μ bits per entry, and there is coherent cost access through* (6) *and coherent marginal lookup. Let $\eta = 2^{-e_\eta}$ for an integer e_η, write $b_\eta \triangleq |e_\eta| + 1$, and let $\delta \in (0,1)$ have encoding length b_δ. There is a quantum algorithm which outputs fixed-point scaling potentials specifying the exact mathematical Gibbs tensor $X_{\boldsymbol{i}} \propto e^{-C_{\boldsymbol{i}}/\eta + \sum_k \beta_{k,i_k}}$ satisfying*

$$\sum_{k=1}^{m} \| \mathrm{Proj}_k(X) - \mu_k \|_1 \leq \delta$$

with probability at least $8/9$, using

$$\widetilde{O}\Big(m^3 n^{\frac{m+1}{2}} \widehat{R}_{C,\mu,\eta} \delta^{-2}\Big)$$

coherent cost queries and $\widetilde{O}(m^4 n^{(m+1)/2} \widehat{R}_{C,\mu,\eta} \delta^{-2})$ working-vector lookups and unit-cost arithmetic operations, where

$$\tau_\mu \triangleq \log \frac{1}{\min_{k,a} \mu_{k,a}}, \qquad \widehat{R}_{C,\mu,\eta} \triangleq 2 + \left\lceil \frac{\|C\|_\infty}{\eta \log 2} \right\rceil + \left\lceil \log_2 \frac{1}{\min_{k,a} \mu_{k,a}} \right\rceil.$$

Set

$$b_{\mathrm{ent}} \triangleq O\left(b_C + b_\mu + b_\eta + b_\delta + m \log n + \log \frac{m \widehat{R}_{C,\mu,\eta}}{\delta} \right).$$

Its non-oracle elementary-gate complexity is

$$\widetilde{O}\Big(m^4 n^{\frac{m+1}{2}} \widehat{R}_{C,\mu,\eta} \delta^{-2} \mathsf{Mult}(b_{\mathrm{ent}})\Big).$$

Proof. Let $K_{\boldsymbol{i}} = e^{-C_{\boldsymbol{i}}/\eta}$ denote the exact mathematical kernel. Since $C \geq 0$, $K_{\max} \leq 1$, $K_{\min} \geq e^{-\|C\|_\infty/\eta}$, and hence $\widehat{R}_K \leq \widehat{R}_{C,\mu,\eta}$; the algorithm uses the computable integer $\widehat{R}_{C,\mu,\eta}$ in place of $\widehat{R}_K$. The circuit never evaluates the generally transcendental number $K_{\boldsymbol{i}}$. By Lemma C.1, every Gibbs exponent $\sum_k \beta_{k,i_k} - C_{\boldsymbol{i}}/\eta$ is computed exactly from $O(1)$ cost queries and m potential lookups, so the log-domain form of Lemma C.3, with $\Delta = \|C\|_\infty$ and $r = \mu_{k,a}$, returns corrections whose errors are measured with respect to the exact kernel. Therefore the noisy-descent, range, and stochastic-drift proof of Theorem C.9 applies verbatim to the exact Gibbs tensor, giving the stated residual and query bounds. Substituting $\alpha = \Theta(\delta/m)$, $p^{-1} = \widetilde{O}(m^2 \widehat{R}_{C,\mu,\eta}/\delta)$, and $\log(1 + \|C\|_\infty/\eta) = O(\log \widehat{R}_{C,\mu,\eta})$ into $b^{\log}_{\mathrm{work}}$ gives b_{ent} and the stated gate bound. □

Finally, accounting for the entropic bias and the multimarginal rounding argument of Lin et al. (Lin et al., 2022, Theorems 4.4–4.5) gives the following fully explicit unregularized result. The returned potentials describe an approximately marginal-feasible Gibbs tensor; they do not themselves encode the final feasible tensor produced by dense rounding.

Theorem C.11 (Quantum multimarginal Sinkhorn)**.** *Let $(C, \boldsymbol{\mu}, \varepsilon)$ be a represented instance in the sense of the finite-input model above, and define $\kappa = \max\{1, (\max C - \min C)/\varepsilon\}$. Under coherent cost access* (6)*, marginal lookup, and coherent lookup of the length-mn potentials, a quantum algorithm computes, with probability at least $2/3$, fixed-point potentials whose exact mathematical Gibbs tensor can be rounded to an exactly feasible additive-ε coupling. Computing the range and the potentials uses $\widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)$ coherent cost queries and $\widetilde{O}(m^5 n^{(m+1)/2}\kappa^3)$ working-vector lookups and unit-cost arithmetic operations. Set*

$$b_{\mathrm{sink}} \triangleq O(b_C + b_\mu + b_\varepsilon + m \log n)\,.$$

The non-oracle elementary-gate complexity under quadratic reversible fixed-point arithmetic is

$$\widetilde{O}\Big(m^5 n^{\frac{m+1}{2}} \kappa^3 b_{\mathrm{sink}}^2\Big)\,.$$

In particular, this is $\widetilde{O}(m^7 n^{(m+1)/2}\kappa^3)$ when $b_C + b_\mu + b_\varepsilon = O(m \log(n\kappa))$. Let T_C, T_W, and T_U denote the gate costs of one coherent cost query, one working-vector lookup, and one working-vector write; Remark D.5 gives table-based values. A compiled implementation adds $\widetilde{O}(m^4 n^{\frac{m+1}{2}}\kappa^3 T_C + m^5 n^{\frac{m+1}{2}}\kappa^3 T_W + m^3 n \kappa^2 T_U)$ gates for coherent reads and writes.

Proof. Compute $L = \min C$ and $U = \max C$ by quantum minimum and maximum finding (Dürr & Høyer, 1996; van Apeldoorn et al., 2020), jointly boosted to failure probability at most $1/6$; this uses $\widetilde{O}(\sqrt{N})$ cost queries, which is lower order. Then $\Delta_C = U - L$. If $\Delta_C \le \varepsilon$, set $\eta = 1$ and return the fixed-point potentials $\beta = 0$. Their exact Gibbs tensor can be rounded to the target marginals, and every resulting feasible coupling has cost at most $U \le \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon$. Assume henceforth that $\Delta_C > \varepsilon$, so $\kappa = \Delta_C/\varepsilon$.

Replace C by $C^+ \triangleq C - L\mathbf{1} \in [0, \Delta_C]^{[n]^m}$. The shift changes every feasible objective by L and multiplies the Gibbs kernel by a global constant, so it preserves additive optimality gaps, normalized Gibbs tensors, and scaling potentials. Set

$$\eta \triangleq \mathrm{dyad}\left(\frac{\varepsilon}{4m q_n}\right), \qquad \delta \triangleq \mathrm{dyad}\left(\frac{\varepsilon}{4\Delta_C}\right),$$
$$\lambda \triangleq \mathrm{dyad}\left(\frac{\delta}{8m}\right), \qquad \mu'_k \triangleq (1-\lambda)\mu_k + \lambda u_n^{\mathrm{dyad}}.$$

where u_n^{dyad} is the represented distribution defined in the finite-input model. Then

$$\sum_{k=1}^{m} \|\mu'_k - \mu_k\|_1 \le 2m\lambda \le \frac{\delta}{4}, \qquad \min_{k,a} \mu'_{k,a} \ge \frac{\lambda}{2n} > \frac{\delta}{32mn}.$$

Each $\mu'_{k,a}$ is rational with encoding length $O(b_\mu + b_C + b_\varepsilon + \log(mn\kappa))$, so the smoothed instance remains within the finite-input model. Run the algorithm of Corollary C.10 for C^+ and $\boldsymbol{\mu}'$ with residual $\delta/4$. Except with probability at most $1/9$, its potentials define a Gibbs tensor Q with $E_{\boldsymbol{\mu}}(Q) \le \delta/2$, by the triangle inequality. The rounding procedure of Lin et al. (2022, Theorem 4.4), applied to Q written densely, returns $\widehat{X} \in \mathcal{U}(\boldsymbol{\mu})$ with $\|\widehat{X} - Q\|_1 \le 2E_{\boldsymbol{\mu}}(Q)$. Together with (11) for C^+, this gives

$$\begin{aligned}\langle C^+, \widehat{X}\rangle &\le \langle C^+, Q\rangle + 2\Delta_C E_{\boldsymbol{\mu}}(Q) \\ &\le \mathsf{MOT}(C^+, \boldsymbol{\mu}) + 4\Delta_C E_{\boldsymbol{\mu}}(Q) + \eta m \log n \le \mathsf{MOT}(C^+, \boldsymbol{\mu}) + \frac{3\varepsilon}{4}.\end{aligned}$$

Shifting back adds L to both sides. A union bound over range finding and Sinkhorn gives success probability at least $1 - 1/6 - 1/9 > 2/3$.

The dyadic-floor inequalities give $\Delta_C/\eta < 8mq_n\kappa$ and $\delta^{-1} < 8\kappa$. Moreover, $\tau_\mu \le \log(32mn/\delta) = O(\log(mn\kappa))$, so $\widehat{R}_{C,\mu,\eta} = \widetilde{O}(m\kappa)$. The dyadic formulas also give $b_\eta + b_\delta =$

Table 3: Quantum approaches to multimarginal OT discussed in this paper. Bounds are stated under each method's own access model and are therefore not all end-to-end comparable.

Method and scope	Reference	Output	Stated resource bound
Tall-LP IPM (any m)	(Apers & Gribling, 2023)	Generic LP solution	$\widetilde{O}(n^{m/2}(mn)^5)$ row queries
Tensor Sinkhorn (ours; any m)	This work	Potentials	$\boldsymbol{\widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)}$ cost queries
Sparse recovery (ours; any m)	This work	Exact classical list	$\boldsymbol{\widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)}$ cost queries

$O(b_C + b_\varepsilon + \log(mn\kappa))$; substituting in the definition of b_{ent} yields $b_{\mathrm{ent}} = O(b_{\mathrm{sink}})$. Corollary C.10 therefore uses $\widetilde{O}(m^3 \cdot m\kappa \cdot \kappa^2 \cdot n^{(m+1)/2}) = \widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)$ coherent cost queries and $\widetilde{O}(m^5 n^{(m+1)/2}\kappa^3)$ working-vector lookups and unit-cost arithmetic operations. Its gate bound, including the exact exponent computation of Lemma C.1, is $\widetilde{O}(m^5 n^{(m+1)/2}\kappa^3 b_{\mathrm{sink}}^2)$. Under the word-length condition in the theorem this becomes $\widetilde{O}(m^7 n^{(m+1)/2}\kappa^3)$. Each update writes n potential entries, and there are $O(m^2 \widehat{R}_{C,\mu,\eta}/\delta) = \widetilde{O}(m^3\kappa^2)$ updates, which gives the write term in the theorem. These bounds concern the scaling-potential output; materializing and rounding Q costs an additional $O(mn^m)$ classical operations. □

C.4 Comparison with existing approaches

The quantum multimarginal Sinkhorn result is closest in spirit to quantum algorithms for matrix scaling and matrix balancing van Apeldoorn et al. (2021); Gribling & Nieuwboer (2022); Nieuwboer (2024), but it is not an immediate corollary of those results. In the bimarginal case, entropic OT is a matrix-scaling problem: one scales a positive matrix so that its two marginals match prescribed vectors. For MOT with $m \geq 3$, the object is a positive order-m tensor, the constraints are m tensor marginals, and the natural update requires estimating fiber sums over n^{m-1} entries rather than row or column sums over n entries. Our algorithm reuses the same quantum primitive underlying the matrix-scaling work, namely quantum estimation of log-sum-exp quantities, but combines it with the multimarginal Sinkhorn geometry and rounding analysis of Lin et al. Lin et al. (2022).

For multimarginal OT, one could also apply the tall-LP quantum interior-point method of Apers and Gribling Apers & Gribling (2023) to the dual formulation (MOT-D). A row query to this LP is sparse: the constraint indexed by $\boldsymbol{i}$ contains exactly m nonzero coefficients and one query to $C_{\boldsymbol{i}}$. Thus the Apers–Gribling Lewis-weight IPM applies with M variables and $N + 2M$ inequalities, giving row-query complexity

$$\widetilde{O}(\sqrt{N+2M}\, M^5) = \widetilde{O}\big(n^{\frac{m}{2}}(mn)^5\big) = \widetilde{O}\Big(n^{\frac{m+10}{2}} m^5\Big),$$

and time $\widetilde{O}(\sqrt{N+2M}\ \mathrm{poly}(M))$; their volumetric-barrier bound gives $\widetilde{O}((N+2M)^{\frac{3}{4}} M^{\frac{5}{4}})$ row queries. These IPM bounds have only polylogarithmic dependence on $1/\varepsilon$, but the generic method pays a large polynomial overhead in the smaller dimension $M = mn$. Therefore, the off-the-shelf quantum IPM is not competitive in the low-precision regime targeted here; improving this comparison would require an MOT-specific IPM implementation that substantially reduces the $\mathrm{poly}(M)$ overhead.

Classical methods can also avoid dense iterates when the cost has additional structure. For general m, the structured-cost algorithms of Altschuler & Boix-Adserà (2023) exploit succinct dual-feasibility oracles. The multimarginal rounding of Lin et al. (2022), by contrast, writes the dense plan and therefore costs $\Theta(n^m)$. These structured methods are complementary to ours. In the general entry-access model and for fixed m, Theorem 4.1 rules out classical algorithms that output an approximately optimal coupling with $o(N/\log^2 n)$ cost queries, whatever their output size. When a structured subclass admits a classical sampler for the trimmed Gibbs law, the relevant comparison is against that specialized sampler; our lower bounds characterize the general entry-access setting.

D SPARSE COUPLING RECOVERY FROM GIBBS POTENTIALS

The scaling potentials from Theorem C.11 do not themselves list a coupling. We now state and prove the fully explicit trim–sample–complete conversion. The conversion trims the Gibbs tensor to a subcoupling, samples only from this trimmed tensor using a target-marginal proposal, trims the empirical tensor back below the marginals, and completes the deficits sparsely.

Lemma D.1 (Finite-precision amplified rejection sampling)**.** *Let P be a probability distribution on a finite set Ω, let $A_i \in [0,1]$, and define*

$$a \triangleq \sum_{i\in\Omega} P_i A_i, \qquad R_i \triangleq \frac{P_i A_i}{a}.$$

Suppose $a \geq a_0 > 0$, an ideal unitary U_P prepares $\sum_i \sqrt{P_i}\,|i\rangle$, and a reversible fixed-point circuit computes $\widetilde{A}_i \in [0,1]$ with

$$\max_i |\widetilde{A}_i - A_i| \leq \frac{a_0\varphi}{8}.$$

For every $\varphi, \sigma \in (0, 1/4)$, exponential-search amplitude amplification returns, except with probability σ, a sample from a distribution within total variation distance φ of R using

$$O\left(a_0^{-1/2} \log \frac{1}{\sigma}\right)$$

state-preparation and acceptance-circuit invocations. It suffices to use $O(\log(a_0^{-1}\varphi^{-1}\sigma^{-1}))$ additional fractional bits beyond the input word lengths. The same guarantee holds for an implemented preparation unitary $\widetilde{U}_P$ provided

$$\|\widetilde{U}_P - U_P\|_{\mathrm{op}} \leq c\,\varphi\sqrt{a_0}/\log(1/\sigma)$$

for a sufficiently small absolute constant $c > 0$. Under quadratic reversible arithmetic, if one acceptance probability uses $O(m)$ fixed-point operations on b-bit words, its non-oracle gate cost is $\widetilde{O}(mb^2)$ per invocation.

Proof. Let $\widetilde{a} = \sum_i P_i \widetilde{A}_i$ and $\widetilde{R}_i = P_i\widetilde{A}_i/\widetilde{a}$. Then

$$D \triangleq \sum_i P_i |\widetilde{A}_i - A_i| \leq \frac{a_0\varphi}{8}, \qquad |\widetilde{a} - a| \leq D.$$

Consequently,

$$\|R - \widetilde{R}\|_{\mathrm{TV}} \leq \frac{D}{a} \leq \frac{\varphi}{8}.$$

Amplitude amplification preserves the normalized accepting component, so exact implementations of state preparation and the controlled rotation for $\widetilde{A}_i$ sample exactly from $\widetilde{R}$ when they succeed. The exponential-search procedure of Boyer et al. (1998), combined with amplitude amplification (Brassard et al., 2000), uses $O(a_0^{-1/2}\log(1/\sigma))$ invocations. Implement state preparation and the acceptance rotation in each invocation to operator-norm error at most a constant times $\varphi a_0^{1/2}/\log(1/\sigma)$. A hybrid argument then changes the output distribution by at most $O(\varphi)$; adjusting the constants gives the stated total variation bound. Fixed-point rounding to absolute error $O(a_0\varphi)$ and this rotation accuracy require the displayed number of additional bits. Reversible multiplication, division, square root, and rotation synthesis on b-bit words cost $\widetilde{O}(b^2)$ gates under quadratic arithmetic, proving the final claim. □

Lemma D.2 (Trim–sample–complete conversion)**.** *Let Q be a probability tensor on $[n]^m$ with $E_{\boldsymbol{\mu}}(Q) = \sum_k \|\operatorname{Proj}_k(Q) - \mu_k\|_1 \leq \delta \leq 1/8$. Suppose $Q_{\boldsymbol{i}} = W_{\boldsymbol{i}}/Z_Q$, with coherent fixed-point access to $\log W_{\boldsymbol{i}}$ and the log-sum-exp access used above. Fix a nonnegative cost tensor C; it may be the cost that defines W, but it does not depend on the algorithm's randomness. For every $\theta, \zeta \in (0, 1/4)$, a quantum algorithm returns, with probability at least $1-\zeta$, a classical list for an exactly feasible coupling $\widetilde{X} \in \mathcal{U}(\boldsymbol{\mu})$ such that*

$$|\operatorname{supp}(\widetilde{X})| = O\left(m^2 n\theta^{-2}\log\frac{m}{\zeta} + mn\right), \qquad \langle C, \widetilde{X}\rangle \leq \langle C, Q\rangle + \left(\frac{5}{8}\delta + \frac{3}{2}\theta\right)\|C\|_\infty.$$

Its sampling stage uses $\widetilde{O}(m^2 n^{\frac{m+1}{2}} \theta^{-2} \log \zeta^{-1})$ *coherent weight queries and* $O(m^3 n \theta^{-2} \log(m/\zeta) + mn)$ *additional classical arithmetic operations, under the marginal-state access in the access model above.*

Proof. Write $\nu_k = \mathrm{Proj}_k(Q)$ and set $\alpha \triangleq \mathrm{dyad}(\delta/(8m))$. Run the final marginal-estimation calls with log-accuracy $\alpha/2$ and failure probability $\zeta/(8m)$ for each mode. Let $\widetilde{\ell}_{k,a}$ be the returned fixed-point estimate and store the upper log-marginal

$$u_{k,a} \triangleq \widetilde{\ell}_{k,a} + \alpha/2$$

exactly on their common dyadic grid. Thus, except with probability at most $\zeta/8$, the mathematical numbers $\bar{\nu}_{k,a} \triangleq e^{u_{k,a}}$ satisfy

$$\nu_{k,a} \le \bar{\nu}_{k,a} \le e^{\alpha} \nu_{k,a} \qquad (k \in [m],\ a \in [n]). \tag{26}$$

The algorithm retains the represented upper logs rather than exponentiating them. Define the ideal trimming factors equivalently by

$$f_{k,a} \triangleq \begin{cases} 0, & \mu_{k,a} = 0, \\ \exp\big(\min\{0, \log \mu_{k,a} - u_{k,a}\}\big), & \mu_{k,a} > 0, \end{cases} = \min\left\{1, \frac{\mu_{k,a}}{\bar{\nu}_{k,a}}\right\}, \qquad Q^{\downarrow}_{\boldsymbol{i}} \triangleq Q_{\boldsymbol{i}} \prod_{k=1}^{m} f_{k,i_k}.$$

Since $Q^{\downarrow} \le f_{k,i_k} Q$ for each k, (26) implies

$$\mathrm{Proj}_k(Q^{\downarrow})_a \le f_{k,a} \nu_{k,a} \le \mu_{k,a}. \tag{27}$$

Let $s = \|Q^{\downarrow}\|_1$ and $\tau = 1 - s$. Using $1 - \prod_k t_k \le \sum_k (1 - t_k)$ for $t_k \in [0,1]$, we obtain

$$\tau \le \sum_{k=1}^{m} \sum_{a=1}^{n} \nu_{k,a}(1 - f_{k,a}).$$

If $\bar{\nu}_{k,a} > \mu_{k,a}$, then

$$\nu_{k,a}(1 - f_{k,a}) = \nu_{k,a} - \frac{\mu_{k,a}\nu_{k,a}}{\bar{\nu}_{k,a}} \le (\nu_{k,a} - \mu_{k,a})_+ + \mu_{k,a}(1 - e^{-\alpha});$$

if $\bar{\nu}_{k,a} \le \mu_{k,a}$, the left-hand side is zero. Since the positive deviations of two probability vectors have total mass one half of their ℓ_1 distance,

$$\tau \le \frac{1}{2} E_{\boldsymbol{\mu}}(Q) + m(1 - e^{-\alpha}) \le \frac{1}{2}\delta + m\alpha \le \frac{5}{8}\delta. \tag{28}$$

Set $s_0 \triangleq 1 - 5\delta/8$. On the event above, $0 < s_0 \le s$. The value s is used only to describe the conditional sampling distribution; the algorithm uses the explicit scalar s_0 when assigning masses to the sampled atoms.

We next sample from $R \triangleq Q^{\downarrow}/s$. Let

$$W^{\downarrow}_{\boldsymbol{i}} \triangleq W_{\boldsymbol{i}} \prod_{k=1}^{m} f_{k,i_k}, \qquad Q^{\downarrow}_{\boldsymbol{i}} = W^{\downarrow}_{\boldsymbol{i}} / Z_Q.$$

Estimate $\log Z_Q$ to additive accuracy $\alpha_Z \le 1/16$ and failure probability $\zeta/8$, and store its fixed-point upper confidence endpoint u_Z. Let $\bar{Z}_Q \triangleq e^{u_Z}$ mathematically. On the resulting success event,

$$Z_Q \le \bar{Z}_Q \le e^{2\alpha_Z} Z_Q.$$

Prepare the proposal state

$$\sum_{\boldsymbol{i} \in [n]^m} \sqrt{P_{\boldsymbol{i}}}\, |\boldsymbol{i}\rangle, \qquad P_{\boldsymbol{i}} \triangleq \frac{\mu_{1,i_1}}{n^{m-1}},$$

by preparing $\sqrt{\mu_1}$ in the first register and uniform superpositions in the remaining registers. Since $Q^{\downarrow}$ is nonnegative, Equation (27) gives the entrywise chain

$$Q^{\downarrow}_{\boldsymbol{i}} \le \mathrm{Proj}_1(Q^{\downarrow})_{i_1} \le \mu_{1,i_1},$$

and hence

$$0 \le A_{\boldsymbol{i}} \triangleq \frac{W_{\boldsymbol{i}}^{\downarrow}}{\bar{Z}_Q \mu_{1,i_1}} \le 1$$

whenever $\mu_{1,i_1} > 0$. Set $A_{\boldsymbol{i}} = 0$ when $\mu_{1,i_1} = 0$; then $W_{\boldsymbol{i}}^{\downarrow} = 0$ as well. Coherently compute $A_{\boldsymbol{i}}$ and perform a controlled rotation with accepting probability $A_{\boldsymbol{i}}$. The total acceptance probability is

$$\sum_{\boldsymbol{i}} P_{\boldsymbol{i}} A_{\boldsymbol{i}} = \frac{s Z_Q}{n^{m-1} \bar{Z}_Q} \ge e^{-2\alpha_Z} \frac{s}{n^{m-1}}.$$

Since $\delta \le 1/8$, (28) gives $s \ge 1 - 5/64 > 1/2$. Hence the acceptance probability is at least $a_0 \triangleq (3n^{m-1})^{-1}$. Apply Lemma D.1 to each sample with $\varphi = \sigma = \zeta/(16 N_{\mathrm{samp}})$. We now verify its finite-precision premise. Set $q \triangleq a_0 \varphi$. If $\mu_{1,i_1} = 0$ or some $\mu_{k,i_k} = 0$, the evaluator returns acceptance zero exactly. Otherwise the ideal log-acceptance ratio is

$$g_{\boldsymbol{i}} \triangleq \log W_{\boldsymbol{i}} + \sum_{k=1}^{m} \min\{0, \log \mu_{k,i_k} - u_{k,i_k}\} - u_Z - \log \mu_{1,i_1} = \log A_{\boldsymbol{i}} \le 0.$$

Approximate each of the at most $m+1$ logarithms of positive represented marginal masses to error at most $q/(64(m+1))$, and form the remaining fixed-point sum exactly. Since $x \mapsto \min\{0, x\}$ and projection onto $(-\infty, 0]$ are nonexpansive, the resulting exponent $\widehat{g}_{\boldsymbol{i}} \le 0$ satisfies $|\widehat{g}_{\boldsymbol{i}} - g_{\boldsymbol{i}}| \le q/64$.

Compute a downward-rounded threshold $t_q \le \log(q/32)$ with $t_q \ge \log(q/64)$. If $\widehat{g}_{\boldsymbol{i}} < t_q$, return acceptance zero; then the ideal acceptance probability is at most $q/32 + q/64 = 3q/64$. Otherwise evaluate $e^{\widehat{g}_{\boldsymbol{i}}}$ to absolute error at most $q/64$. The exponential is 1-Lipschitz on $(-\infty, 0]$, so in either branch the computed $\widetilde{A}_{\boldsymbol{i}} \in [0, 1]$ satisfies

$$|\widetilde{A}_{\boldsymbol{i}} - A_{\boldsymbol{i}}| \le \frac{3q}{64} < \frac{a_0 \varphi}{8}.$$

Synthesize the controlled rotation to the operator-norm tolerance required by Lemma D.1; its hybrid error is charged separately there. All comparisons with zero use the exact marginal encodings, so no evaluation of $\log 0$ occurs. Each accepted sample then uses $\widetilde{O}(n^{(m-1)/2} \log(N_{\mathrm{samp}}/\zeta))$ weight queries. A union bound over amplification failures and the tensorization bound for total variation charge at most $\zeta/4$ to the full sampling stage. Since $\log(q^{-1}) = O(m \log n + \log(N_{\mathrm{samp}}/\zeta))$, the directed logarithms, threshold, and exponential require

$$O\left(m \log n + \log \frac{N_{\mathrm{samp}}}{\zeta}\right)$$

additional fractional bits beyond the represented inputs.

Draw

$$N_{\mathrm{samp}} = \left\lceil c_{\mathrm{samp}} m^2 n \theta^{-2} \log \frac{m}{\zeta} \right\rceil$$

independent samples $I_1, \dots, I_{N_{\mathrm{samp}}}$ and set $\widehat{R} = N_{\mathrm{samp}}^{-1} \sum_{t=1}^{N_{\mathrm{samp}}} e_{I_t}$. For $p_k \triangleq \mathrm{Proj}_k(R)$, the standard empirical-distribution estimate gives

$$\mathbb{E} \| \mathrm{Proj}_k(\widehat{R}) - p_k \|_1 \le \sqrt{\frac{n}{N_{\mathrm{samp}}}}.$$

Changing one sample changes $\sum_k \| \mathrm{Proj}_k(\widehat{R}) - p_k \|_1$ by at most $2m/N_{\mathrm{samp}}$. McDiarmid's inequality and the choice of N_{samp} therefore imply

$$\sum_{k=1}^{m} \| \mathrm{Proj}_k(\widehat{R}) - \mathrm{Proj}_k(R) \|_1 \le \theta \tag{29}$$

except with probability at most $\zeta/4$, for a sufficiently large absolute constant c_{samp}. For the fixed tensor C, the same sample size and Hoeffding's inequality give

$$\langle C, \widehat{R} \rangle \le \mathbb{E}_{I \sim R}[C_I] + \theta \|C\|_\infty \tag{30}$$

except with probability at most $\zeta/4$.

Set $V^{(0)} \triangleq s_0\widehat{R}$. For $k = 1, \dots, m$ and each a, compute the excess

$$r_{k,a} \triangleq \big(\mathrm{Proj}_k(V^{(k-1)})_a - \mu_{k,a}\big)_+.$$

Traverse the listed atoms with $i_k = a$ in a fixed order and remove mass from them until exactly $r_{k,a}$ has been removed; the last visited atom may be only partially reduced. Denote the resulting tensor by $V^{(k)}$. At step k, the kth marginal becomes the coordinatewise minimum of $\mathrm{Proj}_k(V^{(k-1)})$ and μ_k. Marginals processed earlier remain below their targets because

$$V^{(m)} \le \cdots \le V^{(0)}$$

entrywise. The mass removed at step k is

$$\begin{aligned}
\|V^{(k-1)}\|_1 - \|V^{(k)}\|_1 &= \sum_a \big(\mathrm{Proj}_k(V^{(k-1)})_a - \mu_{k,a}\big)_+ \\
&\le \sum_a \big(s_0\,\mathrm{Proj}_k(\widehat{R})_a - \mu_{k,a}\big)_+ \\
&\le \sum_a \big(s_0\,\mathrm{Proj}_k(\widehat{R})_a - s_0\,\mathrm{Proj}_k(R)_a\big)_+ \\
&= \frac{s_0}{2}\|\,\mathrm{Proj}_k(\widehat{R}) - \mathrm{Proj}_k(R)\|_1,
\end{aligned}$$

where the second inequality uses $s_0\,\mathrm{Proj}_k(R) = (s_0/s)\,\mathrm{Proj}_k(Q^{\downarrow}) \le \mu_k$. Set $X_0 \triangleq V^{(m)}$. Summing over k and applying (29) gives

$$\mathrm{Proj}_k(X_0) \le \mu_k \quad (k \in [m]), \qquad \|X_0\|_1 \ge s_0\left(1 - \frac{\theta}{2}\right). \tag{31}$$

Moreover, $X_0 \le V^{(0)}$, so (30), $s_0 \le s$, and $Q^{\downarrow} \le Q$ imply

$$\langle C, X_0\rangle \le \frac{s_0}{s}\langle C, Q^{\downarrow}\rangle + s_0\theta\|C\|_\infty \le \langle C, Q\rangle + \theta\|C\|_\infty. \tag{32}$$

It remains to complete X_0 without materializing a tensor product. Set $d_k = \mu_k - \mathrm{Proj}_k(X_0)$. These vectors are nonnegative and all have mass $1 - \|X_0\|_1$. Because the remaining masses stay equal after every subtraction, either all deficits are zero or every d_k has a positive coordinate. Repeatedly choose one such coordinate a_k from every d_k, place weight $w = \min_k d_{k,a_k}$ on $(a_1, \dots, a_m)$, and subtract w from the chosen coordinates. Each iteration zeros at least one active coordinate, and the final iteration zeros at least m of them. The resulting completion D consequently has

$$|\,\mathrm{supp}(D)| \le \sum_{k=1}^{m} |\,\mathrm{supp}(d_k)| - m + 1 \le mn - m + 1, \qquad \mathrm{Proj}_k(D) = d_k.$$

Thus $\widetilde{X} = X_0 + D$ is exactly feasible. Its completion cost is at most

$$(1 - \|X_0\|_1)\|C\|_\infty \le \left(\frac{5}{8}\delta + \frac{\theta}{2}\right)\|C\|_\infty.$$

Combining this bound with (32) proves the claimed cost bound. The empirical tensor has at most N_{samp} atoms, proving the support bound. Multiplying the cost of one amplified sample by N_{samp} proves the sampling-query bound; forming the empirical marginals, applying the m trimming passes, and running the greedy completion take the stated classical time.

Finally, suppose every represented marginal mass is a multiple of 2^{-b_μ}. Choose the constant-factor tolerances δ and θ to be dyadic numbers within a factor two of their prescribed values, and let b_s be the fractional bit length of the dyadic number $s_0 = 1 - 5\delta/8$. All entries of $V^{(0)}$ and all target marginal masses then lie on the common grid

$$\Gamma^{-1}\mathbb{Z}, \qquad \Gamma = N_{\mathrm{samp}}2^{\max\{b_\mu, b_s\}}.$$

Algorithm 3 Quantum Sinkhorn with sparse exact recovery

Require: Marginals $\boldsymbol{\mu}$, coherent cost and marginal-state access, accuracy ε

Compute $C_{\min} \leftarrow \min C$ and $C_{\max} \leftarrow \max C$ by quantum min/max finding (Dürr & Høyer, 1996)
Replace $C \leftarrow C - C_{\min}\mathbf{1}$ and set $\Delta_C \leftarrow C_{\max} - C_{\min}$
if $\Delta_C \le \varepsilon$ **then**
    return GREEDYCOUPLE$(\mu_1, \dots, \mu_m)$
end if
Choose dyadic η, δ, θ as in Theorem D.3
Form full-support marginals $\boldsymbol{\mu}'$ with total perturbation at most $\delta/4$
$\beta \leftarrow$ QUANTUMSINKHORN$(K = e^{-C/\eta}, \boldsymbol{\mu}', \delta/4)$
if $\beta =$ FAILURE **then**
    return GREEDYCOUPLE$(\mu_1, \dots, \mu_m)$
end if
Estimate upper marginals $\bar{\nu}_{k,a}$ of Q_β and set $f_{k,a} \leftarrow \min\{1, \mu_{k,a}/\bar{\nu}_{k,a}\}$
Represent $Q^{\downarrow}_{\boldsymbol{i}} \leftarrow Q_\beta(\boldsymbol{i}) \prod_k f_{k,i_k}$ implicitly and set $s_0 \leftarrow 1 - 5\delta/8$
Set $N_{\mathrm{samp}} \leftarrow \widetilde{\Theta}(m^2 n \theta^{-2})$ and $P_{\boldsymbol{i}} \leftarrow \mu_{1,i_1}/n^{m-1}$
Draw $I_1, \dots, I_{N_{\mathrm{samp}}} \sim Q^{\downarrow}/\|Q^{\downarrow}\|_1$ by amplitude amplification from P
$X_0 \leftarrow$ TRIM$\big((s_0/N_{\mathrm{samp}}) \sum_{t=1}^{N_{\mathrm{samp}}} e_{I_t}, \boldsymbol{\mu}\big)$ so $\mathrm{Proj}_k(X_0) \le \mu_k$ for all k
$d_k \leftarrow \mu_k - \mathrm{Proj}_k(X_0)$ and $D \leftarrow$ GREEDYCOUPLE$(d_1, \dots, d_m)$
return the sparse list representing $X_0 + D$

The additive trimming pass removes minima of grid masses, and the greedy completion does the same, so every output weight remains on this grid. Each weight uses

$$O\big(\max\{b_\mu, b_s\} + \log(mn\theta^{-1}\zeta^{-1})\big)$$

bits. For the choices in Theorem D.3, $b_s = O(b_C + b_\varepsilon + \log(m\kappa))$, which gives its stated output-length bound. This also shows that exact feasibility does not rely on a unit-cost operation on unbounded rationals. □

Theorem D.3 (Quantum sparse coupling recovery; formal version)**.** *Let $(C, \boldsymbol{\mu}, \varepsilon)$ be a represented instance as in Section C.1, and define $\kappa = \max\{1, (\max C - \min C)/\varepsilon\}$. Under coherent cost access* (6)*, coherent lookup of the length-mn working vectors and represented marginal masses, and target-marginal state preparation, a quantum algorithm returns, with probability at least $2/3$, a classical list representing $\widetilde{X}$ such that*

$$\mathrm{Proj}_k(\widetilde{X}) = \mu_k \quad (k \in [m]), \qquad \langle C, \widetilde{X}\rangle \le \mathsf{MOT}(C, \boldsymbol{\mu}) + \varepsilon.$$

The list has $O(m^2 n\kappa^2 \log(mn\kappa) + mn)$ atoms, and the algorithm uses $\widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)$ coherent cost queries and $\widetilde{O}(m^5 n^{(m+1)/2}\kappa^3)$ working-vector lookups and unit-cost arithmetic operations. With b_{sink} as in Theorem C.11, its non-oracle gate complexity under quadratic reversible fixed-point arithmetic is

$$\widetilde{O}\Big(m^5 n^{\frac{m+1}{2}} \kappa^3 b_{\mathrm{sink}}^2\Big),$$

of which the recovery stage contributes only $\widetilde{O}(m^3 n^{(m+1)/2}\kappa^2 b_{\mathrm{sink}}^2)$. This is $\widetilde{O}(m^7 n^{(m+1)/2}\kappa^3)$ when $b_C + b_\mu + b_\varepsilon = O(m \log(n\kappa))$. Compilation adds the access costs in Remark D.5. The output has

$$b_{\mathrm{out}} = O\big((m^2 n\kappa^2 \log(mn\kappa) + mn)(m \log n + b_C + b_\mu + b_\varepsilon + \log(mn\kappa))\big)$$

bits and takes $O(b_{\mathrm{out}})$ classical bit operations to write.

Proof of Theorem D.3. Let $\Delta_C = \max C - \min C$ and $\kappa = \max\{1, \Delta_C/\varepsilon\}$. We combine Theorem C.11 with Lemma D.2. Compute $L = \min C$ and $U = \max C$ by quantum minimum and maximum finding (Dürr & Høyer, 1996; van Apeldoorn et al., 2020), jointly boosted to failure probability at most $1/9$, so that $\Delta_C = U - L$, and write $C^+ = C - L\mathbf{1}$. This preprocessing uses $\widetilde{O}(\sqrt{N})$ cost queries. If $\Delta_C \le \varepsilon$, greedily couple the marginals as in the last step of Lemma D.2. This gives

an exactly feasible coupling with at most $mn - m + 1$ atoms whose cost exceeds the optimum by at most $U - L \le \varepsilon$. Otherwise set

$$\eta \triangleq \operatorname{dyad}\left(\frac{\varepsilon}{4mq_n}\right), \qquad \delta = \theta \triangleq \operatorname{dyad}\left(\frac{\varepsilon}{8\Delta_C}\right).$$

Set $\lambda \triangleq \operatorname{dyad}(\delta/(8m))$ and $\mu'_k = (1-\lambda)\mu_k + \lambda u_n^{\mathrm{dyad}}$. Then

$$\sum_{k=1}^{m} \|\mu'_k - \mu_k\|_1 \le 2m\lambda \le \delta/4, \qquad \min_{k,a} \mu'_{k,a} \ge \frac{\lambda}{2n} > \frac{\delta}{32mn}.$$

Run the algorithm of Corollary C.10 for C^+ with residual $\delta/4$ against $\boldsymbol{\mu}'$; it fails with probability at most $1/9$. Otherwise it returns potentials defining a normalized Gibbs tensor Q such that, by the triangle inequality,

$$E_{\boldsymbol{\mu}}(Q) \le \sum_{k=1}^{m} \|\operatorname{Proj}_k(Q) - \mu'_k\|_1 + \sum_{k=1}^{m} \|\mu'_k - \mu_k\|_1 \le \delta/2.$$

A Gibbs tensor with potentials β is the exact optimizer of the entropic MOT problem for the shifted cost C^+ with its own marginals $\nu_k = \operatorname{Proj}_k(Q)$. Let $X^\star$ be an optimal coupling for the target marginals; shifting by L does not change the optimizer. Applying the rounding theorem of Lin et al. (2022, Theorem 4.4) to $X^\star$ with target marginals ν_k gives a tensor Y with those marginals and

$$\|Y - X^\star\|_1 \le 2E_{\boldsymbol{\mu}}(Q).$$

Entropic optimality of Q, nonnegativity of entropy, and $H(Q) \le m \log n$ therefore imply

$$\begin{aligned}\langle C^+, Q\rangle &\le \langle C^+, Y\rangle + \eta m \log n \\ &\le \mathsf{MOT}(C^+, \boldsymbol{\mu}) + 2\Delta_C E_{\boldsymbol{\mu}}(Q) + \eta m \log n.\end{aligned}$$

Apply Lemma D.2 to C^+ with $\zeta = 1/9$ and the choices above. A union bound over range finding, Sinkhorn, and recovery gives overall success probability at least $2/3$. Since $\delta, \theta \le \varepsilon/(8\Delta_C)$, the entropic, pre-recovery marginal, and recovery contributions are at most $\varepsilon/4$, $\varepsilon/8$, and $17\varepsilon/64$, so the total is at most $41\varepsilon/64 < \varepsilon$. Since every feasible coupling has mass one, shifting back adds L to both costs and preserves this guarantee for C.

The potential computation has the query and gate costs of Theorem C.11. The final one-sided marginal and log-normalizer estimates use the same primitives and are lower order. Lemma D.2 adds $\widetilde{O}(m^2 n^{(m+1)/2}\kappa^2)$ weight queries and outputs $O(m^2 n\kappa^2 \log(mn\kappa) + mn)$ atoms. For constant recovery failure probability, Lemma D.1 evaluates each acceptance probability through its logarithm: $b_C + b_\varepsilon$ covers the exact shifted cost and dyadic parameters, b_μ covers marginal lookup and state preparation, the log-acceptance exponent has magnitude $\widetilde{O}(m^2\kappa \log n)$ and therefore needs only logarithmically many integer bits, and $O(m \log n + \log(N_{\mathrm{samp}}/\zeta))$ fractional bits cover exponentiation and sampling error. These word lengths are $O(b_{\mathrm{sink}})$ up to logarithmic factors. The reversible arithmetic of recovery therefore adds $\widetilde{O}(m^3 n^{(m+1)/2}\kappa^2 b_{\mathrm{sink}}^2)$ gates. The $\widetilde{O}(\sqrt{N})$ range computation, $\widetilde{O}(n)$ marginal-state-preparation preprocessing, and recovery query costs are dominated by the displayed bounds. Exact feasibility depends only on common-grid additive trimming and greedy deficit completion; finite-precision state preparation affects the sampling distribution and failure budget, not the marginal identities of the returned list. The final paragraph of Lemma D.2 gives the stated weight and output bit lengths. □

Proposition D.4 (Boosting and support purification)**.** *Let s denote the list size in Theorem D.3.*

(i) For every $\rho \in (0, 1/3)$, running the algorithm of Theorem D.3 $O(\log(1/\rho))$ times and returning the cheapest output gives an exactly feasible additive-ε coupling with probability at least $1 - \rho$, using $O(s \log(1/\rho))$ additional cost queries.

(ii) Let $\widetilde{X}$ be an exactly feasible list with support S. Given the costs $C_{\boldsymbol{i}}$ for $\boldsymbol{i} \in S$, a basic optimal solution X' of the MOT program restricted to S satisfies $\operatorname{Proj}_k(X') = \mu_k$ for every k, $\operatorname{supp}(X') \subseteq S$, $|\operatorname{supp}(X')| \le mn - m + 1$, and $\langle C, X'\rangle \le \langle C, \widetilde{X}\rangle$. It can be computed exactly with no further cost queries in time polynomial in $|S|$, mn, and the encoding lengths of the listed costs and masses.

Proof. For (i), every run returns an exactly feasible list: trimming and deficit completion enforce the marginal identities on every execution path, and a run that aborts can output the greedy coupling of the marginals instead. Querying the cost of every listed atom evaluates each output cost exactly. The cheapest output is additive-ε optimal whenever at least one run succeeds, which happens except with probability $(1/3)^r \le \rho$ after $r = \lceil \log_3(1/\rho) \rceil$ runs.

For (ii), the restricted program

$$\min\Big\{ \langle C, X\rangle :\ X \ge 0,\ \mathrm{supp}(X) \subseteq S,\ \mathrm{Proj}_k(X) = \mu_k\ (k \in [m]) \Big\}$$

is a linear program in $|S|$ variables. Its feasible set is bounded and contains $\widetilde{X}$, so it has an optimal vertex X' with $\langle C, X'\rangle \le \langle C, \widetilde{X}\rangle$. The positive entries of a vertex have linearly independent constraint columns, and Lemma B.2 bounds the rank of these columns by $mn - m + 1$. Only the listed costs enter the program, and an exact polynomial-time linear-programming algorithm that returns a basic optimal solution (Grötschel et al., 1988) computes X' in the stated time. □

Remark D.5 (Compilation without qRAM)**.** *Let T_C, T_W, T_μ, T_U denote the gate costs of one cost query, working-vector lookup, marginal-state preparation or mass lookup, and working-vector write, respectively. If the recovery algorithm makes q_C cost queries and q_W working-vector lookups, compiling these interfaces and maintaining its working data adds*

$$\widetilde{O}\big(q_C T_C + q_W(T_W + T_\mu) + m^3 n \kappa^2 T_U + n b_\mu\big)$$

gates, while writing the output list takes $O(b_{\mathrm{out}})$ classical bit operations. The query bounds do not depend on how classical data are stored, and quantum random-access memory is not needed for sublinear gate counts. Suppose $C_{\boldsymbol{i}} = \mathbf{c}(z_{1,i_1}, \dots, z_{m,i_m})$ and the supports, potentials, and represented marginal masses are stored as classical tables. A multiplexed table-lookup circuit reads one word of an n-word table with $\widetilde{O}(n)$ gates (Babbush et al., 2018), and a potential update only changes the classical description of that circuit. Hence $T_W = \widetilde{O}(n)$ and $T_C = \widetilde{O}(mn) + T_{\mathbf{c}}$, where $T_{\mathbf{c}}$ is the gate cost of one reversible evaluation of $\mathbf{c}$; writes cost $O(n)$ classical work per update instead of T_U; and a multiplexed rotation tree prepares $\sum_a \sqrt{\mu_{1,a}}\,|a\rangle$ and looks up masses with $T_\mu = \widetilde{O}(n)$. With $q_C = \widetilde{O}(m^4 n^{(m+1)/2}\kappa^3)$ and $q_W = \widetilde{O}(m^5 n^{(m+1)/2}\kappa^3)$ from Theorem D.3, compilation therefore adds $\widetilde{O}(m^5 n^{(m+3)/2}\kappa^3 + m^4 n^{(m+1)/2}\kappa^3 T_{\mathbf{c}})$ gates. For fixed m, $\kappa \le \mathrm{polylog}(n)$, $b_\mu \le \mathrm{polylog}(n)$, and $T_{\mathbf{c}} = \mathrm{polylog}(n)$, the total gate count is $\widetilde{O}(n^{(m+3)/2})$, which is sublinear in $N = n^m$ for every $m \ge 4$. For $m \le 3$ this naive compilation is not sublinear, and the sublinear gate bounds require lookups with polylogarithmic cost.

Proof of Theorem 4.1. Draw independent bits $f_{\boldsymbol{i}} \sim \mathrm{Bernoulli}(p)$ for

$$p \triangleq \frac{\log^2 n}{n^{m-1}}, \qquad C_{\boldsymbol{i}} \triangleq 1 - f_{\boldsymbol{i}}.$$

Regard the marked entries as the edges of an m-partite, m-uniform hypergraph $\mathcal{H}$ with n vertices in every part. A vertex degree is binomial with mean $d = \log^2 n$. Chernoff's inequality and a union bound over the mn vertices give

$$\min \deg(\mathcal{H}) = (1 - o(1))d, \qquad \max \deg(\mathcal{H}) = (1 + o(1))d$$

with probability $1 - o(1)$. Two vertices in the same part have codegree zero. Two vertices in different parts have binomial codegree with mean d/n, and the standard binomial tail bound gives

$$\Pr[\mathrm{codeg}_{\mathcal{H}}(u, v) \ge \log n] \le \left(\frac{e \log n}{n}\right)^{\log n}.$$

A union bound over all vertex pairs therefore gives $\max \mathrm{codeg}(\mathcal{H}) \le \log n = o(d)$ with probability $1 - o(1)$.

Apply Pippenger's almost-perfect matching theorem (Ehard et al., 2020, Theorem 1.1) with uncovered fraction ε_0. For fixed m and ε_0, its regularity and codegree constants are fixed, while the preceding estimates give $\min \deg(\mathcal{H}) / \max \deg(\mathcal{H}) = 1 - o(1)$ and $\max \mathrm{codeg}(\mathcal{H}) / \max \deg(\mathcal{H}) = o(1)$. Thus its hypotheses hold for all sufficiently large n, and with probability $1 - o(1)$ the hypergraph $\mathcal{H}$

has a matching of at least $(1-\varepsilon_0)n$ marked edges. Place mass $1/n$ on each edge in this matching and couple the unmatched vertices arbitrarily. This is feasible for the uniform marginals and has cost at most ε_0, so

$$\mathsf{MOT}(C, u_n^{\otimes m}) \le \varepsilon_0 \tag{33}$$

on the same event.

Run a putative algorithm on $C = 1 - f$, and suppose that (33) and the algorithm's guarantee both hold. Let M be the set of marked entries and $\theta \triangleq \sum_k \|\operatorname{Proj}_k(X) - \mu_k\|_1 \le \varepsilon_0$. Then $1-\theta \le \|X\|_1 = \|\operatorname{Proj}_1(X)\|_1 \le 1+\theta$, and $\langle C, X\rangle$ is the mass of X on unmarked entries, so

$$X(M) \ge \|X\|_1 - 2\varepsilon_0 \ge 1 - 3\varepsilon_0.$$

Any set S of entries has at most $|S|$ distinct first coordinates, so

$$X(S) \le \frac{|S|}{n} + \theta. \tag{34}$$

Applying (34) to $S = M \cap \operatorname{supp}(X)$ shows that the output lists at least $(1-4\varepsilon_0)n$ distinct marked entries. Exact feasibility is the special case $\theta = 0$.

For the quantum claim, suppose the algorithm uses q_{rec} queries and returns at most $s \triangleq n(\log n)^c$ atoms. Query f on each listed atom and return any $k \triangleq \lfloor (1-5\varepsilon_0)n \rfloor$ marked ones. This verification uses at most s additional queries. A quantum algorithm may make these final queries in the computational basis, so the composed procedure is a valid quantum query algorithm using $q_{\mathrm{rec}} + s$ queries. Since the guarantee holds on every input, it solves k-solution Bernoulli search with success probability at least $2/3 - o(1)$ over the random choice of f.

The quantum multi-solution Bernoulli-search bound of Cojocaru et al. (2023, Theorem 3.1) states that an algorithm making q queries has success probability at most $(1/(\pi k))(e^2(q/k)^2 p)^k$. Although that result indexes its finite domain by bit strings, it applies here by embedding $[n]^m$ into a power-of-two domain and restricting every query and reported solution to the embedded subset; the Bernoulli law on that subset is unchanged. Constant success thus requires $q = \Omega(k/\sqrt{p})$. Applying this with $q = q_{\mathrm{rec}} + s$ gives $q_{\mathrm{rec}} + s = \Omega(n^{(m+1)/2}/\log n)$. For fixed $m \ge 2$ and fixed c, the allowed support size s is lower order, proving the quantum claim.

For the classical claim, neither an output bound nor verification is needed. Let the algorithm make q_{rec} adaptive queries, let $\mathcal{Q}$ be the set of queried locations, and let $\widehat{X} \triangleq X\,\mathbb{1}[\|X\|_1 \le 2]$; both are functions of the algorithm's randomness and the answers it has seen. Conditional on this view, the unqueried bits remain independent $\mathrm{Bernoulli}(p)$ variables, so

$$\mathbb{E}\big[\widehat{X}(M \setminus \mathcal{Q})\big] = p\,\mathbb{E}\big[\widehat{X}([n]^m \setminus \mathcal{Q})\big] \le 2p = o(1).$$

By Markov's inequality, $\widehat{X}(M \setminus \mathcal{Q}) < \varepsilon_0$ except with probability $o(1)$. On the event above, $\widehat{X} = X$ because $\|X\|_1 \le 1+\theta$, so $X(M \cap \mathcal{Q}) > 1 - 4\varepsilon_0$, and (34) shows that the algorithm has queried at least k marked locations. Every newly queried location is marked with conditional probability p, so the expected number of marked locations among the queried ones is at most pq_{rec}. By Markov's inequality, finding k of them with probability $2/3 - o(1)$ requires $q_{\mathrm{rec}} = \Omega(k/p) = \Omega(n^m/\log^2 n)$, proving the classical claim. □

Corollary D.6 (Tight sparse-output query complexity)**.** *Fix $d > 0$, $c > 0$, a constant $m \ge 2$, and $\varepsilon_0 \in (0, 1/10)$. If $\Delta_C \le 1$ and $(\log n)^{-d} \le \varepsilon \le \varepsilon_0$, then constructing an exactly feasible additive-ε coupling with $O(n(\log n)^c)$ atoms has query complexity $\widetilde{\Theta}(n^m)$ classically and $\widetilde{\Theta}(n^{(m+1)/2})$ quantumly, under the access models above.*

Proof. The classical upper bound queries all n^m costs and extracts a basic optimum. The quantum upper bound is Theorem D.3 followed by the purification of Proposition D.4, which uses $\widetilde{O}(n)$ additional cost queries and leaves at most $mn - m + 1 \le n(\log n)^c$ atoms for all sufficiently large n. The lower bounds follow from Theorem 4.1, since additive-ε accuracy implies additive-ε_0 accuracy. Thus both bounds count queries to the same cost-entry oracle; exact Gibbs-exponent computation affects the gate count but not the query exponent. □

Theorem D.3 applies to arbitrary marginal weights, while Theorem 4.1 proves hardness already for uniform marginals. They match in their dependence on n, up to logarithms, for fixed m and constant normalized accuracy: after the purification of Proposition D.4, the upper bound outputs at most $mn - m + 1$ atoms and therefore lies in the lower-bound output class for every constant $c > 0$. Both results count cost-entry queries; exact Gibbs-exponent computation appears only in the upper bound's gate complexity. Because the matching theorem's constants depend on hypergraph uniformity, this proof is not uniform in growing m. The quantum exponent is the standard multi-solution Grover exponent. The contribution is a constructive algorithm that attains it, not a speedup beyond Grover scaling.

Remark D.7 (Classical analogue of the pipeline)**.** *None of the recovery steps is inherently quantum. Classically, each exact Sinkhorn update costs $\Theta(n^m)$ kernel evaluations, and ordinary rejection sampling from the proposal P of Lemma D.2 accepts with probability $\Theta(n^{-(m-1)})$, so each accepted sample costs $\Theta(n^{m-1})$ weight evaluations in expectation. The classical version of Theorem D.3 therefore returns the same kind of sparse, exactly feasible coupling in $\widetilde{O}(\mathrm{poly}(m)\, n^m \kappa^2)$ time. This is no better than reading the tensor, as it must be by the classical part of Theorem 4.1. The quantum algorithm replaces every fiber sum and every rejection loop by its amplified counterpart, turning each n^{m-1} search factor into $n^{(m-1)/2}$ while leaving the factor n from marginal coordinates or output atoms unchanged. For fixed m, Theorem 4.1 shows that neither exponent can be improved under unstructured access, so the gain is exactly the quadratic search speedup on these steps.*

E VALUE-ESTIMATION LOWER BOUNDS

We now provide quantum and classical lower bounds for approximating (MOT-P) in the unstructured cost-query model. The reduction below interpolates between constant and fine accuracy. At constant additive accuracy it gives a search lower bound over n^{m-1} hidden blocks; below the single-entry mass scale $1/n$, it recovers search over all n^m tensor entries.

Identify $[n]$ with the cyclic group $\mathbb{Z}_n$. For each shift vector $s = (s_2, \dots, s_m) \in \mathbb{Z}_n^{m-1}$, define the perfect m-partite matching

$$P_s \triangleq \{(i, i + s_2, \dots, i + s_m) : i \in \mathbb{Z}_n\} \subseteq [n]^m,$$

with addition modulo n. The sets $\{P_s : s \in \mathbb{Z}_n^{m-1}\}$ partition $[n]^m$.

Proof of Theorem 4.3. Fix uniform marginals $\mu_1 = \cdots = \mu_m = u_n$; hardness on this subclass proves the theorem. Set $r \triangleq \lfloor 2\varepsilon n \rfloor + 1$, so $1 \le r \le n$ and $r/n > 2\varepsilon$. Split each matching P_s into $N_{\mathrm{seg}} \triangleq \lfloor n/r \rfloor$ disjoint consecutive blocks of size r:

$$P_{s,t} \triangleq \{(i, i + s_2, \dots, i + s_m) : tr \le i < tr + r\}, \qquad t = 0, \dots, N_{\mathrm{seg}} - 1,$$

ignoring any leftover entries. The total number of blocks is $N_{\mathrm{blk}} = n^{m-1} N_{\mathrm{seg}} = \Omega(n^m/(1 + \varepsilon n))$.

We reduce from promised unstructured search on N_{blk} bits indexed by the pairs (s, t), with promise $|z| \in \{0, 1\}$. Given such a search input, define a Boolean cost tensor by setting $C^z_{\boldsymbol{i}} = 0$ if $\boldsymbol{i} \in P_{s,t}$ for the unique indexed block with $z_{s,t} = 1$, and $C^z_{\boldsymbol{i}} = 1$ otherwise. Equivalently, a query to C^z determines the unique matching P_s containing $\boldsymbol{i}$, checks whether $\boldsymbol{i}$ lies in one of the retained length-r blocks, and if so queries the corresponding bit $z_{s,t}$. Thus one cost query is simulated using at most one query to z.

If $|z| = 0$, then $C^z \equiv 1$, so every feasible coupling has cost 1. If $|z| = 1$, let $P_{s,t}$ be the marked block. The feasible coupling that places mass $1/n$ on every entry of the full matching P_s has cost $1 - r/n$, because exactly the r entries in $P_{s,t}$ have zero cost. Therefore the MOT value is at most $1 - r/n$ in the marked case. Since $r/n > 2\varepsilon$, an additive ε value estimate is at least $1 - \varepsilon$ in the unmarked case and at most $1 - r/n + \varepsilon < 1 - \varepsilon$ in the marked case. Thus any such algorithm distinguishes the two search promises. The stated lower bounds follow from the randomized and quantum query lower bounds for promised unstructured search on N_{blk} items Bennett et al. (1997); Zalka (1999). □