NON-CONSERVATIVE SINKHORN: OPTIMAL TRANSPORT WITH DISSIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal transport (OT) assumes mass is conserved, yet in many settings, including portfolio rebalancing with transaction costs, Mixture-of-Experts routing, and logistics with spoilage, mass *dissipates* in transit. We develop NC-SINKHORN, the first scalable Sinkhorn-type solver for discrete *non-conservative* OT (NC-OT), where an edge-dependent preservation map $m_{ij} \leq 1$ encodes edge-dependent mass loss while hard marginal constraints are maintained. Edge-dependent mass loss breaks the rank-one factorization that makes classical Sinkhorn efficient; we identify a *dual consistency* condition unique to NC-OT and enforce it via a coupled Newton step that preserves $O(NM)$ per-iteration cost. We prove convergence to the unique optimum with an $O(\delta^{-2})$ iteration complexity bound for feasibility tolerance δ (empirically linear), and show that structured preservation maps admit progressively cheaper updates, collapsing to a closed-form $O(N+M)$ solution in the separable case. On benchmarks up to N=5,000, NC-SINKHORN achieves over 600× speedup versus LP solvers. We demonstrate two applications: portfolio rebalancing with transaction costs and taxes, where the separable bid–ask structure yields real-time solves for large portfolios; and Mixture-of-Experts routing, where NC-OT provides *exact* quality-weighted load balance without auxiliary losses or token dropping while matching or exceeding standard baselines.

1 INTRODUCTION

Optimal transport (OT) is a principled framework for comparing and transforming probability distributions at minimal cost that dates back centuries (Monge, 1781; Kantorovich, 1942). Though one can model and solve OT using standard methods for linear programming (LP), the previous decade saw incredible progress, both in theory and in practice, using methods based on entropic regularization and matrix scaling (Cuturi, 2013; Altschuler et al., 2017; Dvurechensky et al., 2018). Through these developments, OT has become a foundational tool across machine learning, with applications ranging from domain adaptation (Courty et al., 2017) and generative modeling (Arjovsky et al., 2017; Genevay et al., 2018) to computational biology (Schiebinger et al., 2019) and NLP (Kusner et al., 2015).

A defining feature of standard OT is *mass conservation*: all mass leaving the source(s) arrives at the destination(s) (Villani, 2009; Peyré & Cuturi, 2019). This assumption is violated in many real-world settings where one would like to apply the OT framework. In logistics, goods spoil during transit; for communication networks, channel noise and packet drops dissipate a source–destination-dependent fraction of transmitted information; in portfolio rebalancing, transaction costs dissipate value; and in MoE routing, mismatched expert–token pairs degrade representation quality. In each case, network edges (i, j) have an associated *preservation factor* $m_{ij} \leq 1$ specifying the fraction of mass (or quality) that survives transport (Figure 1). The total mass received is strictly less than the total mass sent, yet the target distribution must be matched exactly in *shape*.

Existing frameworks for mass-varying transport, including unbalanced OT (Liero et al., 2018; Chizat et al., 2018b; Séjourné et al., 2023) and partial transport (Caffarelli & McCann, 2010; Chapel et al., 2020; Bonneel & Coeurjolly, 2019), handle mass variation through soft marginal penalties or exogenous budgets but do not model *which edges* cause loss (Section 1.1). We study *non-conservative OT* (NC-OT) (Kováčová et al., 2025), where the marginal constraints remain *hard* but each edge (i, j) dissipates a specified fraction $1 - m_{ij}$ of the mass it carries; standard OT is recovered when $m_{ij} = 1$ for all (i, j). Kováčová et al. (2025) established existence, optimality, and strong duality for

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.

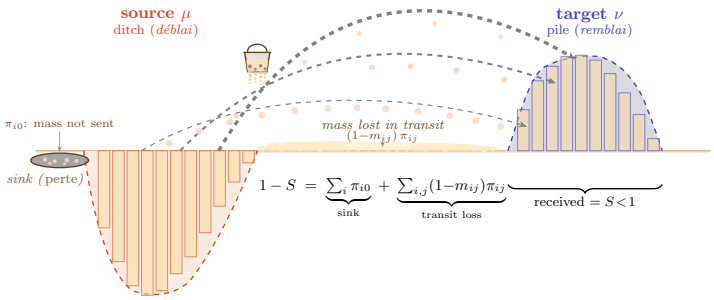


Figure 1: Non-conservative OT (NC-OT), illustrated via Leaky Monge's earth-moving metaphor. Mass is carried from the *source* (ditch) to the *target* (pile) through leaky buckets: each edge (i, j) preserves only a fraction m_{ij} of its load. Two dissipation mechanisms reduce the total mass received ($S < 1$): the *sink* (π_{i0}, mass the optimizer chose not to send) and *transit loss* ($(1-m_{ij})\pi_{ij}$, mass lost en route). Thicker arcs carry more mass; the trail of fallen grains grows with the loss factor $1-m_{ij}$.

NC-OT in the continuous setting, but did not develop computational algorithms. We lift the entropic regularization paradigm (Cuturi, 2013) to NC-OT, developing NC-SINKHORN (Algorithm 1), the first scalable solver with $O(NM)$ per-iteration cost.

Challenges. LP solvers apply directly to NC-OT, but scale empirically as $\Theta(N^{2.6})$ to $\Theta(N^{3.2})$ (Table 7), becoming prohibitive beyond a few thousand variables. Extending Sinkhorn-type methods is also non-trivial: the edge-dependent exponent m_{ij} breaks the rank-one plan factorization, and a *dual consistency constraint* $\sum_j \nu_j \log v_j = 0$, a binding requirement with no analogue in balanced OT, must be actively enforced; ignoring it causes convergence to a *wrong* fixed point (Figure 3). Alternative frameworks such as generalized network flows also do not apply, since the global coupling through S breaks flow-conservation locality (Section 1.1).

We address these challenges by developing NC-SINKHORN, the first scalable solver for entropically regularized NC-OT, and prove its convergence. Our contributions are:

- A *dual consistency constraint* $\sum_j \nu_j \log v_j = 0$ with no analogue in balanced OT (Lemma 3). We show that without it the Sinkhorn iteration converges to a suboptimal fixed point (Figure 3), and enforce it via a coupled column Newton step using the Sherman–Morrison formula at $O(NM)$ per iteration (Section 3). The independent column iteration, a natural first attempt, serves as a baseline throughout.
- *Convergence guarantees*: NC-SINKHORN converges to the unique optimum, which we prove by recognizing the iteration as alternating I-projections under KL divergence (Theorem 1). We establish an $O(\delta^{-2})$ iteration complexity bound (Theorem 2); empirically, convergence is linear.
- A *structural hierarchy* of three regimes based on the factored form of m: when $m \equiv 1$ (balanced OT), the classical Sinkhorn update is recovered; when $m_{ij} = m_j$ (destination-only losses), the column update admits a closed-form power law; and for separable or general m a Newton step is required (Table 1).
- Two applications demonstrating breadth: *portfolio rebalancing* with transaction costs, where $m_{ij} = (P_i^b/q_i)(q_j/P_j^a)$ captures bid-ask spread losses and NC-SINKHORN finishes the measured N=5,000 instance in seconds while both LP solvers exceed a 3-hour cap (Figure 2a, Section 4.1); and *Mixture-of-Experts* (MoE) token routing, where NC-OT provides *exact* quality-weighted load balance by construction, without auxiliary losses, capacity caps, or token dropping, outperforming Switch Transformer capacity constraints (Fedus et al., 2022), expert-choice routing (Zhou et al., 2022), and balanced Sinkhorn baselines (Lewis et al., 2021) (Section 4.2).

1.1 RELATED WORK

Unbalanced OT (U-OT) (Liero et al., 2018; Chizat et al., 2018b; Séjourné et al., 2023) handles mass variation by *softening* marginal constraints via divergence penalties. Mass can appear or vanish

anywhere, controlled by global penalty weights, but U-OT does not model *which edges* cause the loss: two plans with identical marginal violations can produce very different preserved masses S. U-OT thus has soft marginals and global, edge-agnostic mass variation, whereas NC-OT has hard marginals and edge-specific mass dissipation. One cannot recover NC-OT solutions by tuning U-OT penalty weights (Appendix E.4). Partial transport (Caffarelli & McCann, 2010; Chapel et al., 2020; Bonneel & Coeurjolly, 2019) also differs on three axes: the total transported mass is an exogenous parameter (not endogenous as in NC-OT), mass is moved losslessly or not at all, and pointwise bounds replace NC-OT's proportional shape constraint $\sum_i m_{ij}\pi_{ij} = \nu_j S$.

The classical Sinkhorn algorithm (Sinkhorn, 1964; Cuturi, 2013) exploits the rank-one plan factorization $\pi_{ij} \propto u_i K_{ij} v_j$, reducing optimal transport to alternating diagonal rescaling. Extensions to martingale OT (De March & Henry-Labordère, 2019; Eckstein & Kupper, 2021) break this factorization but retain *row-local* constraints: each source point's update depends only on its own conditional. NC-OT introduces a *global* coupling because every column involves the total preserved mass S, which depends on the entire plan. This coupling requires the Newton step we develop. NC-OT can be cast as a min-cost flow with non-unit gain factors (Ahuja et al., 1993), but standard algorithms (push-relabel, network simplex) assume unit gains and, more fundamentally, exploit *flow-conservation locality*: each node's balance depends only on incident edges. In NC-OT the target constraints are globally coupled: every column involves the total preserved mass S, which depends on the *entire* plan. This breaks flow-conservation locality.

2 Background

Optimal transport. Given discrete distributions $\mu \in \mathbb{R}_+^N$ and $\nu \in \mathbb{R}_+^M$ with $\sum_i \mu_i = \sum_j \nu_j = 1$, and a cost matrix $C \in \mathbb{R}^{N\times M}$, the discrete OT problem (Monge, 1781; Kantorovich, 1942; Peyré & Cuturi, 2019) is

$$\min_{\pi\geq 0}\langle C, \pi\rangle \quad \text{subject to} \quad \pi\mathbf{1} = \mu, \quad \pi^\top\mathbf{1} = \nu. \tag{1}$$

Problem (1) is the discrete Kantorovich relaxation (Kantorovich, 1942); the infinite-dimensional formulation optimizes over couplings of probability measures, but all applications in this paper involve discrete distributions supported on finitely many points, so (1) and its non-conservative extension (2) are the natural objects. The constraints $\pi\mathbf{1} = \mu$ and $\pi^\top\mathbf{1} = \nu$ conserve mass: every unit leaving a source arrives at a destination. General purpose LP solvers (simplex, interior point) apply directly but scale super-quadratically in N. For problems going beyond a few hundred variables, the dominant practical approach is rooted in entropic regularization (discussed below), which transforms the LP into a smooth, strictly convex program that is amenable to fast matrix-scaling iterations (Cuturi, 2013; Altschuler et al., 2017).

Non-conservative optimal transport. In many applications mass is lost during transport: goods spoil, signals attenuate, or transaction costs erode value. NC-OT (Kováčová et al., 2025) introduces a *preservation map* $m_{ij} \in [0, 1]$ specifying the fraction of mass surviving the transfer from source i to destination j. The source marginal is matched exactly, while the target is matched *in shape*: the arriving mass, after losses, is distributed as ν. For N source nodes and M destination nodes, augmented with a sink node 0, the NC-OT problem minimizes transport cost subject to source and target balance:

$$\begin{aligned} &\min_{\pi\geq 0} \ \textstyle\sum_{i=1}^{N}\sum_{j=0}^{M} c_{ij}\,\pi_{ij} \\ &\text{subject to} \quad \textstyle\sum_{j=0}^{M}\pi_{ij} = \mu_i, && i = 1,\dots,N, \\ &\qquad\qquad \textstyle\sum_{i=1}^{N} m_{ij}\,\pi_{ij} = \nu_j\,S(\pi), && j = 1,\dots,M, \end{aligned} \tag{2}$$

where $S(\pi) := \sum_{i=1}^{N}\sum_{j=1}^{M} m_{ij}\,\pi_{ij}$ is the total preserved mass, and sink edges carry cost $c_{i0} \geq 0$. This is not a free parameter: summing target balance over j and using $\sum_{j=1}^{M}\nu_j = 1$ recovers the definition of $S(\pi)$, so no separate preserved-mass constraint is needed. The sink column absorbs unallocated source mass as a residual: $\pi_{i0} = \mu_i - \sum_{j=1}^{M}\pi_{ij}$. The total non-arriving mass $1 - S$ decomposes into mass sent to the sink (never dispatched) plus mass dissipated in transit: $\sum_{i=1}^{N}\pi_{i0} +$

$\sum_{i=1}^{N}\sum_{j=1}^{M}(1-m_{ij})\,\pi_{ij} = 1-S$. In the sequel, bare summations $\sum_i$ run over sources $\{1,\dots,N\}$, $\sum_j$ over destinations $\{1,\dots,M\}$, and $\sum_{ij}$ over $\{1,\dots,N\}\times\{1,\dots,M\}$.

Balanced OT (1) is recovered when $m \equiv 1$, in which case $S = 1$ and the sink carries no mass. The total preserved mass $S(\pi)$ is *endogenous*, meaning it is determined by the plan itself rather than fixed before optimization. This distinguishes NC-OT from both standard OT and partial transport (Caffarelli & McCann, 2010; Chapel et al., 2020).

The sink node escapes the target constraint entirely, making the total disposed mass a free decision variable; when all mass must be routed (e.g., MoE token assignment), the sink can be dropped.

In the special case $c_{ij} = 1 - m_{ij}$, minimizing cost is equivalent to maximizing $S(\pi)$ (Appendix B.4); in general, cost and preservation are distinct and the cost-optimal plan sacrifices preservation to avoid expensive edges (Appendix E.5).

Table 1: Structural hierarchy of the preservation map and its effect on the column update. The plan form $\pi^\star_{ij} \propto u_i K_{ij} v_j^{m_{ij}}$ is always known analytically; what varies is whether v_j admits a closed-form update or requires Newton iteration. *Balanced* ($m \equiv 1$) and *destination-only* ($m_{ij} = m_j$) admit closed-form v_j updates; *separable* ($m_{ij} = m'_i m_j$) decouples into M independent Newton problems; *general* m_{ij} requires a coupled Newton step. When $c = 1 - m$, the LP collapses to an $O(N{+}M)$ threshold problem (Appendix B.4).

Structure of m	**Plan form**	**v_j update**
$m \equiv 1$ (balanced)	$u_i K_{ij} v_j$	closed form
$m_{ij} = m_j$ (dest.-only)	$u_i K_{ij} v_j^{m_j}$	closed form
$m_{ij} = m'_i m_j$ (sep.)	$u_i K_{ij} v_j^{m'_i m_j}$	Newton
general m_{ij}	$u_i K_{ij} e^{m_{ij}(\cdots)/\varepsilon}$	Newton

Entropic regularization. Adding an entropic penalty $\varepsilon\sum_{ij}\pi_{ij}(\log\pi_{ij} - 1)$ to the objective in (1) makes the problem strictly convex and yields a factorized optimum $\pi^\star_{ij} = u_i K_{ij} v_j$ with $K_{ij} = e^{-C_{ij}/\varepsilon}$ (Cuturi, 2013). The factorization is what makes Sinkhorn fast: row and column constraints decouple into independent normalizations $u_i \leftarrow \mu_i / \sum_j K_{ij} v_j$ and $v_j \leftarrow \nu_j / \sum_i u_i K_{ij}$, each costing $O(NM)$ (Sinkhorn, 1964; Cuturi, 2013). Convergence is linear in $1/\varepsilon$ (Altschuler et al., 2017), and the unregularized LP (1) is recovered as $\varepsilon \to 0$.

Applying entropic regularization to NC-OT (2) breaks the rank-one factorization and column independence: the chain rule through $S(\pi)$ introduces an edge-dependent exponent on v_j, and a dual consistency constraint $\sum_j \nu_j \log v_j = 0$ emerges as a binding structural requirement rather than a gauge freedom. Section 3 develops NC-SINKHORN, which addresses both difficulties.

3 NC-SINKHORN: SINKHORN FOR NON-CONSERVATIVE OT

We derive the plan structure of entropically regularized NC-OT, identify a dual consistency constraint unique to the non-conservative setting, and show how to enforce it via a coupled Newton step that yields the NC-SINKHORN iteration (Algorithm 1).

3.1 PLAN DERIVATION AND THE COUPLED NEWTON ITERATION

Adding an entropic regularizer $\varepsilon\sum_{ij}\pi_{ij}(\log\pi_{ij} - 1)$ to the NC-OT objective and forming the Lagrangian with multipliers f_i (source), g_j (target), and the chain rule through $S(\pi) = \sum_{ij} m_{ij}\pi_{ij}$ (see Appendix E.1 for the full derivation), the first-order conditions for $j \in \{1,\dots,M\}$ yield

$$\pi_{ij} = u_i\,K_{ij}\,\exp\!\Big(\frac{m_{ij}(g_j - G)}{\varepsilon}\Big), \quad u_i := e^{f_i/\varepsilon}, \quad K_{ij} := e^{-c_{ij}/\varepsilon}, \quad G := \textstyle\sum_l g_l \nu_l. \tag{3}$$

For the sink node, $\pi_{i0} = u_i K_{i0}$ since sink edges do not contribute to S. In balanced OT ($m \equiv 1$), the plan reduces to $u_i K_{ij} v_j$ and row/column updates decouple. In NC-OT, the exponent carries an m_{ij}-dependent power, which breaks the rank-one structure. Defining $v_j := \exp((g_j - G)/\varepsilon)$, the

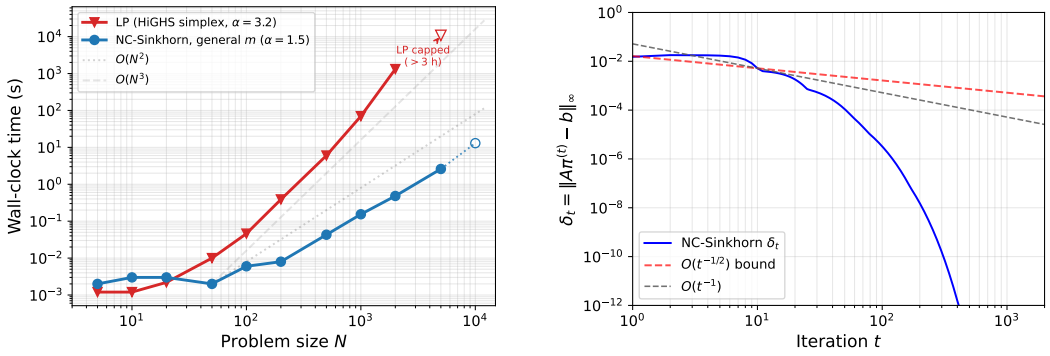


Figure 2: Empirical performance of NC-SINKHORN ($\varepsilon = 0.01$). **(a)** Wall-clock time vs. problem size N for the algorithm-scaling experiment in Table 7: NC-Sinkhorn with general m scales as $N^{1.5}$ vs. $N^{3.2}$ for HiGHS simplex (Huangfu & Hall, 2018) over the common measured range $N \in [100, 1000]$. At N=5,000, LP exceeds the 3-hour cap while NC-Sinkhorn finishes in 2.6 s, a lower-bound speedup above 4,100×. The open NC-Sinkhorn marker and dotted segment denote its extrapolated N=10,000 value. **(b)** Feasibility residual $\delta_t := \|A\pi^{(t)} - b\|_\infty$ vs. outer iteration t, where $A\pi = b$ encodes the row-marginal and column-preservation constraints (Section 3.2). The empirical decay is geometric ($\delta_t \approx C\rho^t$, $\rho \approx 0.95$), far outpacing the $O(t^{-1/2})$ worst-case bound (Theorem 2).

Algorithm 1 NC-SINKHORN

Input: $\mu, \nu, K, m, \varepsilon$, outer tol. τ, inner tol. τ_{in}, inner cap T=5
$u \leftarrow \mathbf{1}_N$, $\ell \leftarrow \mathbf{0}_M$
while $\max_j |\delta_j|/S > \tau$ **do** ▷ outer Sinkhorn loop
  $v_j \leftarrow e^{\ell_j}$ for all j
  $u_i \leftarrow \mu_i / [\sum_j K_{ij} v_j^{m_{ij}} + K_{i0}]$ ▷ row update (exact rescaling)
  for $t = 1$ to T **do** ▷ coupled column Newton (inner loop)
    $v_j \leftarrow e^{\ell_j}$; $C_j \leftarrow \sum_i m_{ij} u_i K_{ij} v_j^{m_{ij}}$; $S \leftarrow \sum_j C_j$
    $\delta_j \leftarrow C_j - \nu_j S$ ▷ col. residuals
    Coupled Newton step (Equation (5)) $\rightarrow \Delta\ell$; $\ell \leftarrow \ell + \Delta\ell$
    if $\max_j |\delta_j|/S < \tau_{\text{in}}$ **then break** ▷ inner stopping rule
  end for
end while
return $\pi_{ij} = u_i K_{ij} e^{m_{ij}\ell_j}$, S

plan becomes $\pi_{ij} = u_i K_{ij} v_j^{m_{ij}}$. The exponent m_{ij} couples v_j with the source index i, so v_j cannot be factored out of the row sum.

This coupling has a second, more subtle consequence. From the definition of $v_j = \exp((g_j - G)/\varepsilon)$, any valid dual must satisfy

$$\sum_j \nu_j \log v_j = \tfrac{1}{\varepsilon} \sum_j \nu_j (g_j - G) = 0. \tag{4}$$

In balanced Sinkhorn, the shift $\log v_j \rightarrow \log v_j + \delta$ is absorbed by $u_i \rightarrow u_i e^{-\delta}$ (a harmless gauge freedom). In NC-OT the same shift produces $\pi'_{ij} = \pi_{ij} \cdot e^{m_{ij}\delta}$: since m_{ij} varies across edges, no scalar correction to u_i can compensate, and constraint (4) is therefore a binding structural requirement rather than a normalization convention (Lemma 3). With these two ingredients, the non-rank-one plan (3) and the dual consistency constraint (4), the iterates of NC-SINKHORN alternate as:

1. **Row update** (closed form): $u_i \leftarrow \mu_i / [\sum_j K_{ij} v_j^{m_{ij}} + K_{i0}]$.
2. **Column update** (coupled Newton): solve for $\ell_j := \log v_j$, $j = 1, \ldots, M$.

The column equations are $F_j(\ell) := C_j - \nu_j S = 0$ for $j = 1, \ldots, M$, where $C_j = \sum_i m_{ij} u_i K_{ij} e^{m_{ij}\ell_j}$ and $S = \sum_j C_j$. Since $\sum_j F_j \equiv 0$, only $M-1$ equations are independent; adding (4) restores M equations in M unknowns.

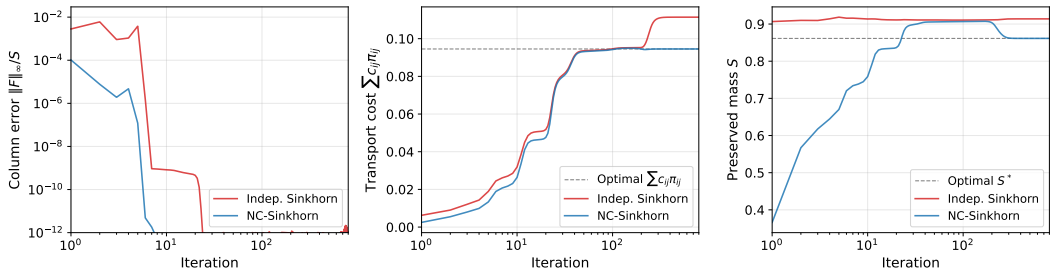


Figure 3: NC-SINKHORN vs. independent column Newton (N=15, ε=0.01). **(a)** Both solvers drive column error to zero, but the independent solver (red) converges to a *wrong* fixed point. **(b)** Transport cost: the independent solver forces mass through expensive routes, producing an 18% gap versus the CVXPY optimum (blue). **(c)** Preserved mass S: the independent solver converges to S=0.914 vs. $S^\star$=0.861 (6% error). Without dual consistency, the iteration drifts toward high-m routes that preserve more mass but at higher cost, an artifact with no analogue in balanced Sinkhorn.

Table 2: Distance to the interior-point optimum ($N = 10$, tol. 10^{-9}). Entries report NC-SINKHORN / Independent Sinkhorn; lower is better.

ε	$\lvert S - S^\star\rvert$	$\lVert\pi - \pi^\star\rVert_\infty$	$\lvert\text{obj} - \text{obj}^\star\rvert$
0.05	$\mathbf{2.3{\times}10^{-6}}$ / $1.7{\times}10^{-2}$	$\mathbf{3.7{\times}10^{-6}}$ / $3.6{\times}10^{-3}$	$\mathbf{2.8{\times}10^{-10}}$ / $7.2{\times}10^{-5}$
0.01	$\mathbf{5.2{\times}10^{-7}}$ / $3.0{\times}10^{-2}$	$\mathbf{2.2{\times}10^{-6}}$ / $1.9{\times}10^{-2}$	$\mathbf{1.7{\times}10^{-10}}$ / $1.5{\times}10^{-4}$
0.005	$\mathbf{1.5{\times}10^{-6}}$ / $2.3{\times}10^{-2}$	$\mathbf{2.2{\times}10^{-6}}$ / $2.0{\times}10^{-2}$	$\mathbf{7.2{\times}10^{-10}}$ / $1.8{\times}10^{-4}$

The Jacobian is $J = D - \boldsymbol{\phi}'\boldsymbol{\nu}^\top$ with the M-th row replaced by $\boldsymbol{\nu}^\top$, where $D = \text{diag}(\phi'_j)$ and $\phi'_j = \sum_i m_{ij}^2 u_i K_{ij} e^{m_{ij}\ell_j}$. This is a rank-one update of a diagonal. The same m_{ij}-dependent coupling that breaks the classical Sinkhorn factorization in (3) makes the Sherman–Morrison formula applicable (derived in Appendix E.2):

$$\Lambda = \frac{\sum_j \nu_j F_j/\phi'_j - \sum_j \nu_j \ell_j}{\sum_j \nu_j^2/\phi'_j}, \quad \Delta\ell_j = \frac{-F_j + \nu_j \Lambda}{\phi'_j}. \tag{5}$$

The per-iteration cost is $O(NM)$, dominated by evaluating the M column residuals δ_j, each a sum over N sources. The coupled column solve is the explicit inner loop of Algorithm 1: in practice 3–5 inner Newton steps suffice, and once warm-started a *single* step already reaches machine precision (Table 6).

A natural simplification is to drop the dual consistency constraint (4) and solve each column independently, removing the rank-one coupling and letting the M Newton equations decouple. However, this *independent* iteration converges to a wrong fixed point. Figure 3 illustrates the failure on an N=15 instance (ε=0.01): both solvers drive column residuals to zero (Figure 3a), yet the independent solver converges to $S = 0.914$ instead of the optimal $S^\star = 0.861$ (a 6% error) and incurs 18% higher transport cost (Figure 3b,c). The coupled Newton resolves this by enforcing $\sum_j \nu_j \ell_j = 0$ at each step via the Sherman–Morrison update (5), matching the CVXPY optimum to six digits. Table 2 quantifies this across $\varepsilon \in \{0.05, 0.01, 0.005\}$ (N=10): NC-SINKHORN matches the interior-point solution to $\|\pi - \pi^\star\|_\infty < 4 \times 10^{-6}$, while Independent Sinkhorn converges to the wrong fixed point.

3.2 CONVERGENCE

The NC-SINKHORN iteration alternates I-projections under the generalized KL divergence onto the row and column constraint sets

$$\mathcal{R} := \{\pi \geq 0 : \textstyle\sum_{j=0}^{M} \pi_{ij} = \mu_i,\ \forall i\}, \quad \mathcal{C} := \{\pi \geq 0 : \textstyle\sum_i m_{ij}\pi_{ij} = \nu_j S(\pi),\ \forall j\}, \tag{6}$$

where $S(\pi) = \sum_{ij} m_{ij}\pi_{ij}$. Both are linear in π and $\mathcal{C}$ incorporates the dual-consistency constraint (4). Writing the combined constraints in matrix form as $A\pi = b$ with $A \in \mathbb{R}^{(N+M)\times N(M+1)}$ and $b = (\mu_1, \ldots, \mu_N, 0, \ldots, 0)^\top$, the entropic NC-OT problem becomes $\min_{\pi\geq 0} \text{KL}(\pi\|\xi)$ s.t. $A\pi = b$, with $\xi_{ij} := e^{-c_{ij}/\varepsilon}$.

Theorem 1 (Convergence of NC-SINKHORN). *Suppose the coupled Newton column update (5) is run to convergence at each outer step. Then* $\pi^{(t)} \to \pi^\star := \arg\min_{\pi \in \mathcal{R} \cap \mathcal{C}} \mathrm{KL}(\pi \| \xi)$*, the unique entropic NC-OT optimum.*

This follows from the Csiszár–Bregman alternating-projection theorem (Bregman, 1967; Csiszár, 1975), since $\mathcal{R}$ and $\mathcal{C}$ are convex sets with non-empty intersection (proof in Appendix E.6). The convergence rate is as follows.

Theorem 2 (Iteration complexity). *The sequential I-projection algorithm initialized at* $\tilde{\pi} := \xi / \|\xi\|_1$*, projecting onto the most-violated constraint at each step, returns* $\pi^{(K)}$ *with* $\|A\pi^{(K)} - b\|_\infty \le \delta$ *in at most* $K \le \frac{2\,\|A\|_\infty^2 \,\mathrm{KL}(\pi^\star \| \tilde{\pi})}{\delta^2}$ *iterations (proof in Appendix E.6.3).*

Since each NC-SINKHORN outer step simultaneously satisfies $N+M$ constraints via Newton, this $O(\delta^{-2})$ bound is conservative; empirically, convergence is geometric with rate $\rho(\varepsilon) < 1$ that degrades as $\varepsilon \to 0$ (Figure 2 and Appendix A.2).

Scope. Theorem 1 covers the deployed coupled-Newton iteration when each inner column solve reaches its root. By contrast, Theorem 2 is a worst-case rate for the sequential most-violated-constraint backbone, not a proved rate for the deployed method. Numerically, warm-started Newton solves the coupled column system to machine precision within the 3–5 inner steps used in all experiments (Table 6).

4 EXPERIMENTS

We evaluate NC-SINKHORN on two application domains: portfolio rebalancing with transaction costs and Mixture-of-Experts token routing. The experiments separate two contributions: NC-OT supplies the hard arrived-mass and quality-weighted constraints, while NC-SINKHORN makes those constrained problems scalable. Comparisons with unregularized LP test formulation fidelity; comparisons with alternative routers test the value of the NC-OT constraints.

4.1 PORTFOLIO REBALANCING WITH FRICTIONS

Selling a unit of asset i at bid price P_i^b and buying asset j at ask price P_j^a preserves a fraction $m_{ij} = (P_i^b/q_i)(q_j/P_j^a)$ of value, where q_i is a reference price, capturing bid–ask spread loss. This preservation map is naturally separable ($m_{ij} = m_i' m_j$), so it exercises the fastest regime in our computational hierarchy. The transport cost c_{ij} additionally includes capital-gains tax that depends on the source asset's embedded gains but not on the destination, creating a genuine $c \neq 1 - m$ separation (see Appendix E.5). All marginals are Dirichlet-distributed (details including data generation are in Appendix B.1).

For scaling, we benchmark NC-SINKHORN vs. HiGHS LP (dual simplex) (Huangfu & Hall, 2018) on measured instances from $N = 5$ to 5,000 ($\varepsilon = 0.01$, tol. 10^{-9}) and report an NC-Sinkhorn power-law extrapolation at $N = 10{,}000$. Figure 2a (full tables and exponent-fit protocol in Appendix B.2) shows general-m Sinkhorn scales as $N^{1.5}$ vs. $N^{3.2}$ for LP simplex over the common measured range $N \in [100, 1000]$. The headline $607\times$ speedup is measured within the separable-structure ablation (1968 s LP vs. 3.2 s at N=5,000, Table 8); under the algorithm-scaling setup both LP solvers exceed a 3-hour wall-clock cap at N=5,000 without finishing while NC-SINKHORN ends in seconds (Table 7). We quote speedups only between numbers measured in the same setup.

Closeness to the exact (unregularized) LP. Beyond speed, we measure how close the entropic NC-SINKHORN plan is to the exact NC-OT optimum on this portfolio problem (separable bid–ask m, tax cost $c \neq 1 - m$). Let $\pi^\star$ be the exact *unregularized* LP optimum and π_ε the NC-SINKHORN plan. The arrived-allocation decision, which is the quantity a portfolio manager actually acts on, is the tracking error $\mathrm{TE} = \max_j |\hat{\nu}_j - \nu_j|$, where $\hat{\nu}_j = \frac{1}{S}\sum_i m_{ij}(\pi_\varepsilon)_{ij}$ is the realized arrived shape. Because the target-shape constraint is *hard*, TE is at machine precision ($\le 10^{-14}$) for every N and every ε we tested (Table 9): the decision is exact and ε-independent. The objective gap $\Delta_{\mathrm{obj}} = \langle c, \pi_\varepsilon \rangle - \langle c, \pi^\star \rangle$ shrinks by roughly an order of magnitude per ε halving, reaching $\sim 10^{-4}$ (relative $<1\%$) at ε=0.002. At the scales where the exact LP is affordable the two agree closely;

Table 3: WikiText-103 perplexity (PPL) at default, untuned settings (3 seeds, mean ± std). NC-OT uses $\alpha = 1 - 2/E$; Balanced OT uses $m \equiv 1$ and $\varepsilon = 0.1$. *Alive*: experts with $> 0.1/E$ token share.

	E=8		E=32		E=128	
Method	**PPL↓**	**Alive**	**PPL↓**	**Alive**	**PPL↓**	**Alive**
NC-OT (α=1−2/E)	164.5±0.2	8/8	153.4±0.7	32/32	151.5±0.2	128/128
Balanced OT (ε=0.1)	163.9±1.1	8/8	152.8±0.6	32/32	152.0±0.5	128/128
Top-2	168.0±1.6	8/8	164.2±1.0	19/32	161.2±0.6	43/128
Switch	182.9±2.0	8/8	185.8±1.3	32/32	188.1±0.9	128/128

Table 4: **Matched exploration.** Fixed $\alpha = 0.8$ across E ($\varepsilon = 0.1$, 3 seeds; Δ = Bal. OT − NC-OT, so $\Delta > 0$ favors NC-OT).

E	NC-OT	Bal. OT	Δ
8	164.53±0.82	163.15±0.55	−1.4
32	153.31±0.51	152.49±0.23	−0.8
64	151.18±0.35	150.68±0.29	−0.5
128	151.52±0.13	152.04±0.52	+0.5
256	153.63±0.11	156.24±0.63	**+2.6**

Table 5: **Budget-matched tuning** at $E = 128$ (3 seeds). Each OT method tunes one knob.

Method	PPL↓	balance
GShard ($\lambda = 0.01$)	154.2±0.3	soft
Top-2	161.2±0.6	none
Balanced OT (default)	152.0±0.4	hard
Balanced OT (tuned)	150.9±0.7	hard
NC-OT (default)	151.5±0.2	hard, exact
NC-OT (tuned)	**148.5**±0.2	hard, exact

the regime where NC-SINKHORN is *necessary* is precisely large N, where the LP does not finish (Table 7) yet the arrived allocation must still be matched exactly.

We also study multi-stage rebalancing where transaction costs alternate between liquid and illiquid regimes (Figure 9). Jointly optimizing intermediate portfolios re-routes trades through cheap stages, reducing cost by up to 81% over myopic planning; under stochastic rates, model predictive control (MPC) with look-ahead H=5 captures 79% of the full-information oracle savings (Appendix D).

4.2 SMARTER TOKEN ROUTING IN MIXTURE-OF-EXPERTS

Sparse MoE models route each token to a few experts (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022). NC-OT treats routing as lossy: $m_{ij} \leq 1$ measures representation quality retained by sending token i to expert j, while $\sum_i m_{ij}\pi_{ij} = \nu_j S$ imposes hard quality-weighted balance. NC-OT can improve MoE perplexity (or, *PPL*) by balancing useful expert workload instead of merely sending each token to its highest-scoring Top-k experts or balancing raw token counts via Balanced OT. Top-k extraction preserves sparse computation. Balanced Sinkhorn (Lewis et al., 2021) balances counts but ignores quality; unbalanced top-k may collapse experts, and expert-choice routing (Zhou et al., 2022) may orphan tokens.

We use NC-OT as a gradient-detached assignment step. For gate logits $z_{ij} = \mathbf{x}_i^\top \mathbf{w}_j$, each MoE layer sets $m_{ij} = \alpha + (1-\alpha)\,\mathrm{softmax}_j(z_{ij})$, $\nu_j = 1/E$, and $c_{ij} = -z_{ij}$ shifted to have row minimum zero. The solve chooses expert indices; gradients flow through their softmax dispatch weights, not through the OT solver.

Setup. We train a 4-layer Transformer-MoE on WikiText-103 (Merity et al., 2017) (hidden dimension 256, 4 heads, FFN dimension 128), with top-2 MoE layers 2 and 4. The tied input/output embedding gives 15.2M parameters at $E = 8$ and 31.1M at $E = 128$. We vary only $E \in \{8, 32, 128\}$ and the router: NC-OT, Balanced OT ($m \equiv 1$), raw Top-2, GShard top-2, or Switch. Both OT methods use $\varepsilon = 0.1$, 10 outer iterations, and 5 inner Newton steps; Switch uses auxiliary-loss weight 0.01 and capacity factor 1.25. We report a parameter-free NC-OT default $\alpha = 1 - 2/E$ and a tuned schedule; Balanced OT receives the same one-knob tuning budget over ε. All models train for 3,000 AdamW steps (lr 3×10^{-4}, cosine schedule, 1,000-step warmup, batch size 32), averaged over 3 seeds; we report validation perplexity (PPL; exponentiated mean next-token cross-entropy), with full search grids in Appendix C.3.

Results. At the default, untuned settings (Table 3) NC-OT and Balanced OT are *statistically indistinguishable* across $E \in \{8, 32, 128\}$ (within 0.5 PPL, inside seed error): the parameter-free

α=1−2/E heuristic sits on a plateau, so we do not claim a perplexity win here. Both OT methods dramatically outperform Top-2 and Switch. Top-2 suffers from expert collapse, retaining only 19/32 experts at E=32 and 43/128 at E=128 and wasting model capacity. Switch keeps all experts alive via its auxiliary loss but at a steep quality cost (183–188 PPL, 18–37 points worse than the OT methods, with the gap growing with E). Its PPL *degrades* as E grows, showing that forcing uniform load *without* accounting for token–expert affinity hurts the model.

Two controlled comparisons separate NC-OT from Balanced OT where the default table cannot. First, **matched exploration** (Table 4): holding α *fixed and equal* across expert counts removes the confound between E and the α=1−2/E schedule, and under this apples-to-apples condition NC-OT's advantage grows monotonically with E, from a slight deficit at $E\leq$64 to +0.5 at E=128 and +**2.6** PPL at E=256 (with NC-OT's seed variance simultaneously shrinking 0.82→0.11). Since α is constant, this is a property of the NC-OT mechanism, not the schedule. Second, **budget-matched tuning** (Table 5): giving Balanced OT the same one-knob search over ε improves it to a best of 150.9; tuned NC-OT (148.5) still leads by 2.4 PPL, while the GShard top-2 + auxiliary-loss baseline (the standard strong balanced router) trails both at 154.2. Perplexity is a *token-averaged* metric, so where both methods route the bulk of tokens identically and differ only on the tail, averaging dilutes the effect; the quantities NC-OT is designed to control are *extremal*, and there the two are not close. On a synthetic routing study with known block-diagonal affinity (Appendix C.2), NC-OT delivers *exact* quality-weighted load balance (0.000 imbalance) with higher worst-expert accuracy (0.983 vs. Balanced OT's 0.896) and full effective throughput at a fixed per-expert budget (100% vs. 99.6%, vs. 81.4% for Top-1); over a multi-step simulation the gap compounds (worst-expert 0.997 vs. 0.71, while Top-1 collapses). We treat the exact balance as a *guarantee* of the constrained solve, not an empirical result, so we do not headline it.

Batched vs. single-token routing. The quality-weighted balance constraint $\sum_i m_{ij}\pi_{ij} = \nu_j S$ is a *batch-level* property, well-defined over any token batch; we apply the identical coupled batched solve at training and evaluation, so no separate inference path is needed for batched serving. For true single-token serving (one input, no batch to balance across) the constraint is undefined, and one falls back to the learned top-k gate (or a cached routing map), exactly as Switch/GShard do at inference. This is cheap, since the gate and preservation map are already learned.

Structural certificate. To verify the balance gap is structural, we solve the NC-OT LP exactly (CVXPY/Clarabel, 5 seeds; Figure 5): the LP cost is a hard floor that Top-k methods fall below, certifying that their lower cost comes from violating QW-balance. Full seven-method synthetic breakdowns (128 token types, 16 experts, 4 specialization blocks) and the multi-step dynamics are in Appendices C.2 and C.4.

Additional results. Appendix A.2 verifies linear convergence of NC-SINKHORN across $\varepsilon \in \{0.001, \ldots, 0.1\}$ and fits the spectral-gap scaling $(1-\rho) \propto \varepsilon^{1.84}$; Appendix B.2 gives full scaling tables. Detailed synthetic MoE breakdowns, routing heatmaps, and expert learning dynamics appear in Appendices C.2 and C.4; multi-stage experiment tables are in Appendix D.

5 CONCLUSION

We introduced NC-SINKHORN, a Sinkhorn-type solver for edge-dependent mass dissipation. Its coupled column Newton step enforces the NC-OT dual consistency constraint via Sherman–Morrison in $O(NM)$ per iteration; without this coupling, the iteration converges to a wrong fixed point. We proved convergence to the unique entropic optimum through alternating KL I-projections (Theorem 1) and identified structural regimes culminating in an $O(N+M)$ closed form when m is separable and $c = 1 - m$. Experiments confirm $O(N^2)$ per-iteration scaling and over 600× speedup over LP at N=5,000; in MoE routing, NC-OT provides exact quality-weighted load balance and improves over capacity-capped, expert-choice, and balanced-Sinkhorn baselines.

Limitations and future work. The framework assumes known m; learning it under uncertainty remains open. Entropic plans are dense, although quadratic regularization yields sparse plans with the same column coupling (Appendix E.7). Our $O(\delta^{-2})$ bound applies to the sequential backbone, whereas the deployed coupled-Newton method is empirically geometric; closing this rate gap is open.

Reproducibility Statement

The algorithm, stopping criteria, and hyperparameters are specified in Algorithm 1 and Section 4. The appendix provides the data-generation procedures, complete scaling measurements, model architecture and training protocol, tuning grids, convergence diagnostics, and proofs and derivations (Appendices A to C and E). Random seeds, hardware, solver tolerances, and the distinction between measured, capped, and extrapolated timings are reported with the corresponding experiments.

Ethics Statement

The experiments use a public language-modeling dataset and synthetic portfolio and routing instances; they involve no human subjects or private data. The portfolio study is a methodological simulation and should not be interpreted as financial advice or as a validated trading system. Mixture-of-Experts routing can inherit biases present in training data and learned representations. Deployment in consequential settings would require application-specific evaluation of fairness, privacy, reliability, and financial or operational risk.

AI Use Statement

Generative AI tools were used to assist with code implementation and debugging and manuscript editing and organization. The authors formulated and verified the reported numerical and mathematical claims, reviewed all AI-assisted work, and take responsibility for the final content of the paper.

References

Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. *Network flows: theory, algorithms, and applications*. Prentice Hall, 1993.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 880–889, 2018.

Nicolas Bonneel and David Coeurjolly. SPOT: Sliced partial optimal transport. *ACM Trans. Graph.*, 38(4), July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323021. URL https://doi.org/10.1145/3306346.3323021.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Luis Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge–Ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010. ISSN 0003-486X. doi: 10.4007/annals.2010.171.673. URL https://doi.org/10.4007/annals.2010.171.673.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2903–2913, 2020.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018a.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018b.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 685–693, 2014.

Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4): 941–965, 2018.

Hadrien De March and Pierre Henry-Labordère. Building arbitrage-free implied volatility: Sinkhorn's algorithm and variants. *arXiv preprint arXiv:1902.04456*, 2019.

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning*, pp. 1367–1376, 2018.

Stephan Eckstein and Michael Kupper. Computation of optimal transport and related hedging problems via penalization and neural networks. *Applied Mathematics & Optimization*, 83(2): 639–667, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

L. V. Kantorovich. On the translocation of masses (in Russian). *Dokl. Akad. Nauk SSSR*, 37:227–229, 1942.

Gabriela Kováčová, Georg Menz, and Niket Patel. Non-conservative optimal transport. *arXiv preprint arXiv:2510.03332*, 2025.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pp. 957–966, 2015.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pp. 6265–6274, 2021.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211 (3):969–1117, 2018.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, pp. 666–704, 1781.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berber, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. In *Numerical Control: Part B*, volume 24 of *Handbook of Numerical Analysis*, pp. 407–471. Elsevier, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964.

Cédric Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

SUPPLEMENTARY MATERIAL

This appendix follows the same two-application arc as the main text and adds one extension. We first look *under the hood of the solver* (Appendix A): why the coupled Newton step is necessary, how fast the outer iteration converges as ε varies, and how numerically exact the inner column solve is. We then develop the two applications end to end. Appendix B treats *portfolio rebalancing*: how the friction-aware instance is generated, how the solver scales against LP, and, finally, the quality question a practitioner ultimately cares about, namely how close its plan lands to the exact unregularized LP. Appendix C treats *Mixture-of-Experts routing*: the language-modeling protocol and tuning budgets, the controlled synthetic study that isolates *why* quality-weighted balance helps, and the learning dynamics it induces. Appendix D then extends NC-OT to the *multi-stage* setting, and Appendix E collects the proofs and derivations. Each application subsection is self-contained: the problem setup, what the preservation map m means in that domain, and how to read each result appear together, so the two studies can be read independently.

A UNDER THE HOOD OF THE SOLVER

Before the applications, we establish three facts about the solver itself: the coupled Newton step is *necessary* (dropping it converges to the wrong answer), the outer iteration converges *fast* (geometrically, across regularization levels), and the inner column solve is *numerically exact* (a single warm-started Newton step reaches machine precision). Together these justify the deployed configuration used throughout the paper.

A.1 WHY THE COUPLED NEWTON STEP IS NECESSARY

See Figure 3 and Table 2 in the main text for the convergence diagnostic and the comparison with the interior-point optimum.

The independent column update solves N independent equations $C_j = \nu_j S$ using S computed *before* the update. This ignores two couplings: (i) each column's update changes S, invalidating the other columns' targets, and (ii) the dual consistency constraint $\sum_j \nu_j \ell_j = 0$ is never enforced. The independent solver converges confidently to a wrong preserved mass ($\Delta S/S^* \approx 6\%$) while the dual consistency violation stabilizes at a large nonzero value (≈ 6). The coupled Newton resolves both issues simultaneously via the Sherman–Morrison step (5) derived in Appendix E.2.

A.2 HOW FAST THE OUTER ITERATION CONVERGES

Having established that the coupling is necessary, we ask how quickly the correct iteration converges, and how that depends on the regularization ε. Figure 4 extends the convergence analysis from the main text to multiple regularization levels. We define the empirical contraction ratio $\rho(\varepsilon)$ as the per-iteration multiplicative decay of the feasibility residual: $\delta_t \approx C\,\rho^t$, estimated via least-squares on $\log \delta_t$ vs. t. Panel (a) shows that the empirical rate is strictly linear (exponential decay of δ_t) for all tested ε, far outpacing both the $O(t^{-1/2})$ worst-case bound of Theorem 2 and even $O(t^{-1})$. Panel (b) reveals that the spectral gap $(1-\rho)$ scales as $(1-\rho) \propto \varepsilon^{1.84}$: smaller ε tightens the contraction but convergence remains linear for all values tested.

A.3 HOW EXACT THE INNER COLUMN SOLVE IS

The outer convergence above assumes each column update is an exact projection onto $\mathcal{C}$. This is what the inner Newton loop delivers. Each column solve is a rank-one–coupled root find for $\ell_j = \log v_j$ and converges quadratically. Table 6 reports the worst-case inner residual $\mathrm{err}_F = \max_j |C_j - \nu_j S|/S$ after each Newton step over 20 seeds (N=200, ε=0.1). From a cold start the residual roughly squares each step, reaching $\sim 10^{-17}$ by step 6. Once warm-started (every outer iteration after the first), a

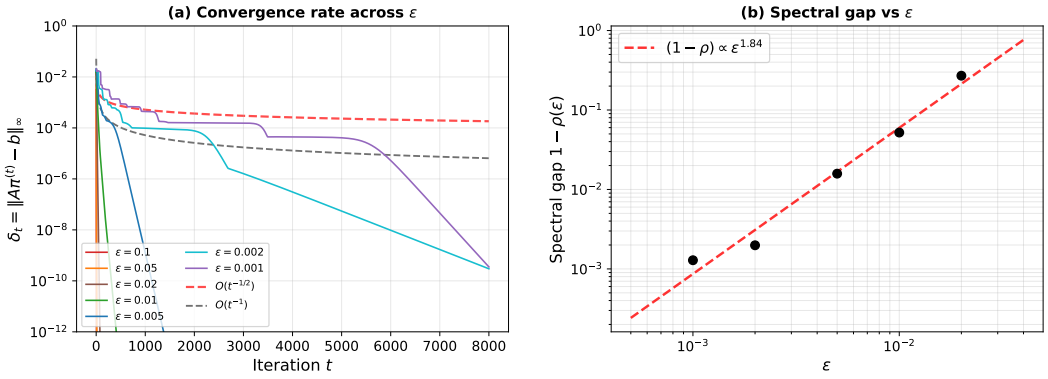


Figure 4: Convergence rate analysis (N=100, portfolio problem). **(a)** δ_t vs. iteration for $\varepsilon \in \{0.001, \ldots, 0.1\}$; all exhibit strict linear convergence beating both $O(t^{-1/2})$ and $O(t^{-1})$ rates. The slower convergence at small ε (e.g. 0.001, 0.002) is expected: the Gibbs kernel $K_{ij} = e^{-c_{ij}/\varepsilon}$ develops entries spanning $\sim e^{-\|c\|_\infty/\varepsilon}$ orders of magnitude, making the fixed-point map increasingly ill-conditioned. **(b)** Spectral gap $(1-\rho)$ vs. ε: the fitted exponent $(1-\rho) \propto \varepsilon^{1.84}$ quantifies how the convergence rate degrades as $\varepsilon \to 0$.

single step already reaches $\sim 10^{-17}$, so the deployed 3–5 inner steps are more than sufficient to make the projection exact to machine precision, matching the exact-projection assumption of Theorem 1.

Table 6: Inner coupled-Newton convergence: worst-case column-balance residual $\mathrm{err}_F = \max_j |C_j - \nu_j S|/S$ after each Newton step over 20 seeds (N=200, ε=0.1). Cold start = first outer iteration; warm start = later outer iterations. Warm-start cells for steps $\geq$3 are blank because a single warm step already reaches machine precision; the cold curve is run to step 6 to display the full quadratic decay.

Newton step	cold start (first outer iter)	warm start (later outer iters)
0	1.5×10^{-2}	3.5×10^{-10}
1	8.6×10^{-3}	2.3×10^{-17}
2	1.2×10^{-3}	1.4×10^{-17}
3	3.4×10^{-5}	n/a
4	1.6×10^{-7}	n/a
5	1.5×10^{-11}	n/a
6	1.8×10^{-17}	n/a

B Portfolio Rebalancing in Depth

This section develops the portfolio application end to end: how the friction-aware instance is generated (Appendix B.1), how NC-Sinkhorn scales against LP solvers and why the separable structure only helps the *algorithm* and not the LP (Appendices B.2 to B.4), and how close the resulting plan lands to the exact unregularized LP on the decision a portfolio manager actually acts on (Appendix B.5).

B.1 How the friction-aware instance is generated

All portfolio experiments (Section 4.1 and Appendices B.2 and D) share the following data-generation process, which models a multi-asset rebalancing problem with bid–ask spreads, capital-gains taxes, and multiple tax lots.

Source and target portfolios are drawn as $\mu, \nu \sim \mathrm{Dirichlet}(5 \cdot \mathbf{1}_N)$, representing moderately concentrated holdings across N assets (a concentration of 5 yields realistic portfolio weights with no extreme sparsity). The preservation map is *separable*, $m_{ij} = m'_i \cdot m_j$ with

$$m'_i \sim \mathrm{Uniform}(0.92, 1.0), \qquad m_j \sim \mathrm{Uniform}(0.92, 1.0),$$

where m'_i and m_j represent half-spreads at the bid (source) and ask (destination). The round-trip loss on any trade $(i \to j)$ is therefore $1 - m_{ij} \in [0.003, 0.154]$ (median $\approx 7\%$), capturing typical bid–ask friction for a diversified portfolio spanning liquid and moderately illiquid names. Holding positions retain full value, $m_{ii} = 1$ (diagonal override).

The cost matrix introduces a *capital-gains tax* as an additive source-only surcharge,

$$c_{ij} = \underbrace{(1 - m_{ij})}_{\text{spread loss}} + \underbrace{\tau_i}_{\text{tax cost}}, \qquad \tau_i = 0.20 \cdot \max(g_i, 0),$$

where $g_i \sim \text{Uniform}(-0.10, 0.30)$ is the embedded capital gain fraction for asset i (negative values represent loss lots). The tax rate is 20%, applied only to unrealized gains ($g_i > 0$). This creates the $c \neq 1 - m$ separation central to our paper: two edges with identical spread (the same m) can differ substantially in cost due to tax. Diagonal entries have zero cost, $c_{ii} = 0$, since holding an asset pays no spread or tax. A uniform sink cost $c_{\text{sink}} = 0.05$ represents the option to leave mass unrouted (a cash position); in the portfolio interpretation, routing to the sink corresponds to raising cash rather than investing proceeds.

For the more detailed tax-lot analysis in Appendix E.5, each asset i holds $K_i = 5$ lots with sizes $n_{ik} \sim \text{Uniform}(0, 1)$ (normalized to the current position) and cost bases $B_{ik} = P_i \cdot (1 - g_{ik})$ where $g_{ik} \sim \text{Uniform}(-0.15, 0.35)$. The piecewise-linear tax function $L_i(u)$ selects lots in tax-efficient order (losses first), so the marginal tax rate increases as more of asset i is sold. For the scaling benchmarks (Appendix B.2) we use the linearized τ_i approximation above, equivalent to a single lot per asset. All experiments use a fixed random seed (seed $= 42$) unless otherwise noted.

B.2 How the solver scales against LP

Table 7 gives the full wall-clock data underlying Figure 2a. Two scaling tables appear in this paper and answer *different* questions under *different* setups; we state their roles explicitly to avoid reading a cell of one as a cell of the other. Table 7 (this subsection) is the **algorithm-scaling** comparison, NC-SINKHORN vs. a general-purpose LP solver (HiGHS) under $c = \max(1 - m, 0)$, and is the source of Figure 2a and its fitted exponents. Table 8 (next subsection) is a separate **LP-formulation ablation** (General LP vs. Separable LP vs. NC-SINKHORN) under a different setup (separable m, arbitrary $c \neq 1 - m$); its N=5,000 General-LP cell (1968 s) is *not* the same experiment as the N=5,000 cell of Table 7 and should not be equated with it.

Table 7: **Algorithm scaling.** NC-SINKHORN vs. HiGHS LP wall-clock time under $c = \max(1-m, 0)$ ($\varepsilon = 0.01$, tol. 10^{-9}, median of 3 runs). LP cells at $N \leq 1{,}000$ are measured directly; the N=2,000 and N=5,000 LP cells (previously extrapolated) are now *measured* on the same machine. Both solvers hit our 3-hour wall-clock cap at N=5,000 without finishing, reported as >3 hr. At N=10,000, only the NC-Sinkhorn values are extrapolated (†); LP values are not reported because the N=5,000 observations are right-censored by the cap. Over the common measured range $N \in [100, 1000]$, a single log-log least-squares fit gives NC-SINKHORN $\Theta(N^{2.0})$ for separable m and $\Theta(N^{1.5})$ for general m, versus LP-simplex $\Theta(N^{3.2})$ and LP-IPM $\Theta(N^{2.6})$.

	NC-SINKHORN (sep. m)		NC-SINKHORN (gen. m)		HiGHS LP	
N	Time	Iters	Time	Iters	Simplex	IPM
5	0.3 ms	3	2 ms	26	1.2 ms	1.2 ms
10	0.4 ms	3	3 ms	51	1.2 ms	1.5 ms
20	0.3 ms	3	3 ms	41	2.2 ms	3.3 ms
50	0.6 ms	3	2 ms	18	10 ms	17 ms
100	1.3 ms	3	6 ms	13	46 ms	85 ms
200	5.3 ms	3	8 ms	8	388 ms	407 ms
500	31 ms	3	43 ms	6	6.0 s	4.6 s
1000	122 ms	3	154 ms	5	70 s	35 s
2000	516 ms	3	482 ms	4	22 min	4 min
5000	3.5 s	3	2.6 s	3	>3 hr (capped)	>3 hr (capped)
10000	17.2 s†	3†	13.0 s†	3†	n/r	n/r

The measured-vs-measured speedup within this setup is $454\times$ at N=1,000 (LP-simplex 70 s vs. NC-SINKHORN general 154 ms), and at N=5,000 both LP solvers exceed the 3-hour cap while

NC-SINKHORN finishes in 2.6–3.5 s, a lower bound of >3,000× (>4,100× against the general-m solve quoted in Figure 2a); we quote speedups only between numbers measured in the same setup.

B.3 SEPARABLE LP VS. GENERAL LP

When $m_{ij} = m'_i m_j$, despite the algebraic simplification, the LP still operates over N^2 plan entries because the cost c couples all of them.

Table 8 compares the three formulations on arbitrary c (not $c = 1 - m$) with separable m. The separable LP provides only marginal speedup (1.0–1.6×) over the general LP because both must operate over N^2 plan entries coupled through the cost matrix c. By contrast, NC-SINKHORN exploits the separable structure *algorithmically*, achieving a 607× speedup over the General LP at N=5,000 *within this ablation's setup* (1968 s vs. 3.2 s); this is the source of the headline 607× figure, and it is a different experiment from Table 7.

Table 8: **LP-formulation ablation** (separate experiment from Table 7). General LP vs. Separable LP vs. NC-SINKHORN (arbitrary $c \neq 1 - m$, separable m, HiGHS simplex, median of 3 runs for $N \leq 1{,}000$, single run for $N > 1{,}000$). Although the separable factorization $m_{ij} = m'_i m_j$ simplifies the constraint structure, LP solvers cannot exploit it effectively: the Separable LP is only 1.0–1.6× faster than the General LP because both must operate over N^2 plan entries coupled through the cost matrix c. Small-N NC-Sinkhorn timings are mildly non-monotonic because fixed overhead and timing noise are comparable to the millisecond solve time. By contrast, NC-SINKHORN exploits the separable structure *algorithmically* via the Gibbs kernel, achieving 607× speedup over the General LP at N=5,000; the 1968 s General-LP cell here is not the same experiment as the N=5,000 cell of Table 7.

N	General LP	Separable LP	NC-SINKHORN	Gen/NC-S	Gen/Sep
50	14 ms	9 ms	4 ms	3.4×	1.6×
100	36 ms	35 ms	14 ms	2.6×	1.0×
200	174 ms	152 ms	10 ms	18×	1.1×
500	1.2 s	1.1 s	55 ms	22×	1.1×
1000	6.5 s	6.1 s	176 ms	37×	1.1×
2000	52 s	39 s	0.68 s	77×	1.3×
5000	1968 s	1612 s	3.2 s	607×	1.2×

B.4 CLOSED-FORM LP WHEN $c = 1 - m$

When $m_{ij} = m'_i m_j$ and $c_{ij} = 1 - m_{ij}$, the LP objective simplifies dramatically:

$$\sum_{ij}(1 - m'_i m_j)\,\pi_{ij} = \sum_i \mu_i - S.$$

The cost depends only on the total preserved mass S, not on individual plan entries. The NM variables collapse to an $O(N+M)$ threshold problem: given a target S, each column j independently saturates its capacity $\nu_j S$ using the cheapest source rows (sorted by $c_{\cdot j}$), and bisection on S satisfies the budget.

This closed-form LP (CF-LP) is applicable when the only friction is proportional loss, e.g., bid-ask spreads with $c_{ij} = 1 - m_{ij}$. In such settings CF-LP achieves 10–4,500× speedup over generic LP at $N = 500$. However, most applications involve an independent cost matrix c (e.g., Euclidean distance), so CF-LP is restricted to a narrow special case. We therefore present it here for completeness rather than as a primary contribution.

B.5 CLOSENESS TO THE EXACT UNREGULARIZED LP

We quantify how close the entropic NC-SINKHORN plan is to the exact *unregularized* NC-OT optimum on the portfolio problem (separable bid–ask m, capital-gains tax cost $c \neq 1 - m$; the regime where NC-OT is genuinely needed). Let $\pi^\star$ be the exact LP optimum (scipy/HiGHS, no entropic regularization) and π_ε the NC-SINKHORN plan. We report two quantities:

- the **objective gap** $\Delta_{\text{obj}} = \langle c, \pi_\varepsilon \rangle - \langle c, \pi^\star \rangle$ (how much more expensive the regularized plan is);

- the **tracking error** $\mathrm{TE} = \max_j |\hat{\nu}_j - \nu_j|$ with $\hat{\nu}_j = \frac{1}{S}\sum_i m_{ij}(\pi_\varepsilon)_{ij}$, i.e. the worst-case deviation of the realized arrived allocation from the target shape (the decision a portfolio manager acts on).

Table 9: Distance of NC-SINKHORN to the exact *unregularized* LP optimum on the portfolio problem ($c \neq 1 - m$, Dirichlet marginals, seed 42). The tracking error (arrived-allocation decision) is at machine precision for *all* N and *all* ε, because the target-shape constraint is enforced exactly; it is ε-independent. The objective gap to the unregularized LP shrinks by roughly an order of magnitude per ε halving, to relative $<1\%$ at ε=0.002.

		$\Delta_{\mathrm{obj}} = \langle c, \pi_\varepsilon\rangle - \langle c, \pi^\star\rangle$			
N	$\langle c, \pi^\star\rangle$	ε=0.01	ε=0.005	ε=0.002	max TE
100	1.7×10^{-2}	3.9×10^{-3}	4.9×10^{-4}	3.5×10^{-5}	1×10^{-15}
500	2.0×10^{-2}	7.2×10^{-3}	8.2×10^{-4}	1.0×10^{-4}	1×10^{-15}
1000	2.0×10^{-2}	9.8×10^{-3}	1.2×10^{-3}	1.4×10^{-4}	1×10^{-14}

Two points follow. First, the two error notions are *decoupled*: the arrived-allocation decision is exact regardless of ε (hard constraint), while the objective gap is controlled separately by ε. Second, the objective values $\langle c, \pi^\star\rangle$ are nearly constant in N because the problem is intensive (marginals normalize to 1 and per-edge costs are N-independent). Taken with the scaling data (Table 7), this delineates the operating regime: at small N the exact LP is affordable and NC-SINKHORN matches it closely; at large N the LP is intractable while NC-SINKHORN retains an exact arrived allocation and an objective within $<1\%$ of optimal at small ε.

C MIXTURE-OF-EXPERTS ROUTING IN DEPTH

We now turn from the portfolio study to the learned-routing application. The thread here is *why* quality-weighted balance helps: we first certify with an exact LP that the quality gap between NC-OT and greedy routers is a *necessary* consequence of the balance constraint rather than solver slack (Appendix C.1); we then isolate the mechanism on a controlled synthetic routing problem (Appendix C.2); give the full language-modeling protocol, baselines, and tuning budgets (Appendix C.3); and finally trace how the routing shapes expert specialization over training (Appendix C.4).

C.1 AN EXACT-LP FEASIBILITY CERTIFICATE

To verify that the quality gap between NC-OT and greedy methods is a necessary consequence of load balance rather than algorithmic suboptimality, we solve the NC-OT LP *without* entropic regularization (CVXPY/Clarabel, exact optimum) over 5 seeds. The LP cost provides a hard floor: no QW-feasible plan can achieve less. See Figure 5 in the main text for the results.

C.2 A CONTROLLED SYNTHETIC ROUTING STUDY

The language modeling experiment in Section 4.2 uses NC-OT as a gradient-detached routing step with a dynamically learned preservation map. To isolate the routing mechanism in a controlled setting, we additionally study a synthetic problem with known structure.

Setup. We route 128 token types to 16 experts arranged in 4 specialization blocks. Within-block quality is high ($m \approx 0.95$–1.0) while cross-block quality drops to $m \approx 0.70$. We compare NC-SINKHORN (Algorithm 1) against six baselines: unconstrained top-k (k=1, 2); capacity-constrained top-1 with capacity factor F=1.0 and F=1.25 (Switch Transformer (Fedus et al., 2022)); expert-choice routing (Zhou et al., 2022); and balanced Sinkhorn (Lewis et al., 2021). All numbers are averaged over 20 seeds. We report *worst-expert accuracy* $\min_j \hat{a}_j$, where $\hat{a}_j = \sum_i m_{ij}^\beta \pi_{ij} / \sum_i \pi_{ij}$ (β=3), and *relative QW-imbalance* $\max_j |\hat{\nu}_j - \nu_j|/\nu_j$.

Results. NC-OT is the *only* method achieving exactly zero imbalance with the highest worst-expert accuracy (0.983); the full seven-method comparison is in Table 10. Top-1 reaches 0.978 but at 91% imbalance; Balanced OT has 3.4% imbalance but degrades its worst expert to 0.896 because it

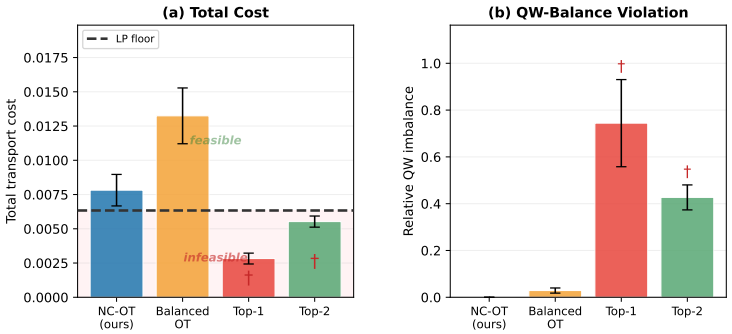


Figure 5: LP feasibility certificate (5 seeds, mean ± std). **(a)** Total cost: the dashed line is the LP floor (minimum over all QW-balanced plans). NC-OT and Balanced OT lie above the floor (feasible); Top-k (†) falls below, proving their plans violate QW-balance. **(b)** Relative QW imbalance $\max_j |\hat{\nu}_j - \nu_j|/\nu_j$: NC-OT achieves zero violation; Top-k exceeds 50%. Per-expert breakdowns, routing heatmaps, and throughput curves are in Figures 6 and 7.

balances raw mass rather than quality-weighted load. NC-OT achieves 100% effective throughput at the ideal per-expert budget vs. 99.6% for Balanced OT and 81.4% for Top-1 (Figure 6).

Table 10: MoE routing comparison (128 token types, 16 experts, averaged over 20 seeds). NC-OT is the only method achieving exactly zero imbalance while routing all tokens. Worst-expert accuracy $\min_j \hat{a}_j$ with $\hat{a}_j = \sum_i m_{ij}^{\beta} \pi_{ij} / \sum_i \pi_{ij}$ (β=3).

Method	**Imbalance**	**Worst-Acc**	**Notes**
NC-SINKHORN (ours)	**0.000**	**0.983**	hard balance, no drops
Top-1	0.912	0.978	no balance
Switch (C=1.0)	0.584	0.978	drops 21% of tokens
Switch (C=1.25)	0.631	0.978	drops 11% of tokens
Top-2	0.430	0.974	moderate imbalance
Expert Choice	0.084	0.600	77% tokens orphaned
Balanced OT	0.034	0.896	residual QW-imbalance

C.3 THE LANGUAGE-MODELING EXPERIMENT: PROTOCOL, BASELINES, AND TUNING

This is the full protocol behind the WikiText-103 results in Section 4.2 (Tables 3 to 5): the architecture and training recipe, every routing baseline, and the single-knob tuning budget each method received.

Architecture. 4-layer Transformer with hidden dimension 256, 4 attention heads. Layers 1 and 3 use a dense FFN of dimension 128; layers 2 and 4 use sparse MoE with top-2 gating and the same per-expert FFN dimension. We vary the number of experts $E \in \{8, 32, 64, 128\}$ (and E=256 for the fixed-α control, Table 4). The token embedding and output head share weights. Exact parameter counts are 15.2M (E=8), 18.4M (E=32), 22.6M (E=64), 31.1M (E=128), and 48.0M for the E=256 control. Tokenization uses the GPT-2 BPE vocabulary (50,257 tokens).

Training. AdamW optimizer with learning rate 3×10^{-4}, weight decay 0.01, cosine schedule with 1,000 linear warmup steps. Batch size 32, sequence length 256, trained for 3,000 steps on WikiText-103 (Merity et al., 2017) (119M train tokens). All experiments run on NVIDIA A100-SXM4-40GB GPUs (CUDA); each configuration is repeated over 3 random seeds.

NC-OT routing. The preservation map is $m(i,j) = \alpha + (1-\alpha)\,\sigma(\mathbf{x}_i^\top \mathbf{w}_j)$ where σ is softmax over experts, giving $m \in [\alpha, 1]$. The floor parameter is set as $\alpha = 1 - 2/E$, a simple heuristic that scales the non-conservative flexibility with the number of experts (e.g., $\alpha = 0.75$ at E=8; $\alpha \approx 0.98$ at E=128). Transport cost is $c(i,j) = -\text{logit}(i,j) + \min_j \text{logit}(i,j)$ (shifted so $\min_j c(i,j) = 0$ per token). The NC-Sinkhorn solver uses ε=0.1, 10 outer iterations, and 5 coupled Newton steps per

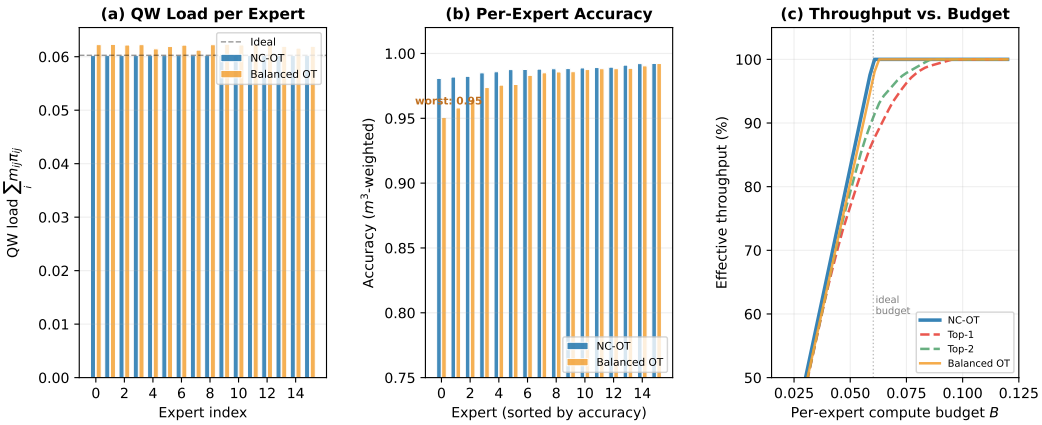


Figure 6: MoE routing quality (128 token types, 16 experts; means ± std over 20 seeds). **(a)** Per-expert QW load: NC-OT (blue) is perfectly flat; Balanced OT (orange) shows visible variation. **(b)** Per-expert accuracy (m^3-weighted, sorted): NC-OT worst expert 0.983 vs. Balanced OT worst 0.896. **(c)** Effective throughput vs. per-expert budget: NC-OT reaches 100% at the ideal budget; others require over-provisioning.

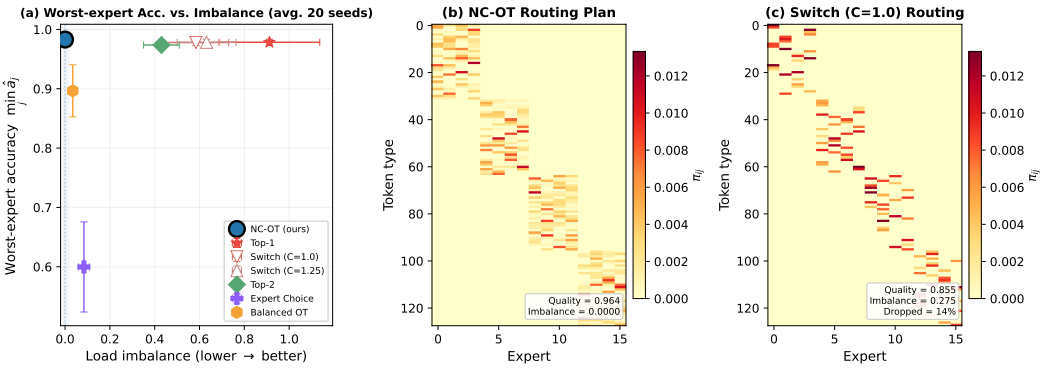


Figure 7: MoE routing (128 token types, 16 experts; means ± std over 20 seeds). **(a)** Worst-expert accuracy ($\min_j \hat{a}_j$) vs. load imbalance: NC-OT (blue circle, top-left corner) is the only method combining zero imbalance with high worst-expert accuracy; error bars show ±1 std. **(b)** NC-OT routing plan (representative seed): tokens are concentrated on specialist experts while quality-weighted load is distributed uniformly. **(c)** Switch Transformer (C=1.0) routing: drops 21% of tokens yet still suffers 58% imbalance.

iteration. The solver is detached (no gradients flow through it); gradients pass through the softmax weights at the chosen expert indices.

Balanced OT routing. Standard entropic OT (Sinkhorn with m=1) using the same solver parameters (ε=0.1, 10 iterations, 5 Newton steps). This is the conservative ($m \equiv 1$) special case of NC-OT and serves as the primary ablation to isolate the effect of quality-weighted routing. For the budget-matched comparison (Table 5) we additionally sweep its regularization $\varepsilon \in \{0.05, 0.10, 0.20\}$, matching the single-knob search afforded to NC-OT; the best value (ε=0.05, PPL 150.9) is reported alongside the paper setting (ε=0.1, PPL 152.0).

GShard top-2 + auxiliary-loss routing. A GShard-style top-2 router (Lepikhin et al., 2021) with the standard load-balancing auxiliary loss $\mathcal{L}_{\text{aux}} = \lambda E \sum_j f_j p_j$, where f_j is the fraction of tokens routed to expert j and p_j the mean gate probability. We report $\lambda \in \{0.01, 0.1\}$ (PPL 154.2 / 158.3 at E=128). This is the standard strong balanced-routing baseline; its balance is *soft* (enforced by the added loss term) rather than the hard constraint of the OT methods.

Table 11: Tuning budget for the E=128 MoE comparison (Table 5). Each OT method searches a single scalar over the listed grid; the aux-loss baselines search the loss weight λ. All other hyperparameters (optimizer, schedule, steps, seeds) are shared.

Method	**Search grid**	**Selected**
NC-OT (default)	n/a (parameter-free)	α=$1-2/E$
NC-OT (tuned)	$\alpha \in \{0.70, \ldots, 0.99\}$, warmup on/off	α=0.70 + warmup
Balanced OT	$\varepsilon \in \{0.05, 0.10, 0.20\}$	ε=0.05
GShard top-2	$\lambda \in \{0.01, 0.10\}$	λ=0.01
Switch (top-1)	$\lambda \in \{0.01, 0.10\}$, $F \in \{1.0, 1.25\}$	λ=0.01, F=1.25

Top-2 routing. Standard top-2 gating with no load-balancing loss or capacity constraints. This baseline highlights the cost of expert collapse when no balancing mechanism is employed.

Expert utilization threshold. An expert is counted as "alive" if it receives more than a $0.1/E$ fraction of tokens on the validation set. This adaptive threshold (rather than a fixed percentage) prevents false negatives at large E: with E=128 experts, a perfectly balanced router assigns $\sim$0.78% per expert, so a fixed 1% threshold would incorrectly flag balanced experts as dead.

Wall-clock overhead. NC-OT's routing overhead is dominated by the Sinkhorn solve (10 outer iterations $\times$ $O(B \cdot E)$ matrix-vector products per routing call). At larger model scale, expert FFN computation ($O(B \cdot k \cdot d_{\text{ffn}})$) dominates the $O(B \cdot E)$ routing solve, amortizing the overhead. In our setting (d_{ffn}=128, E up to 128), routing adds $\sim$25–35% wall-clock overhead over Top-2; this fraction would shrink with larger expert FFNs typical of production MoE models.

Sensitivity to α and tuning protocol. At E=128, we swept $\alpha \in \{0.70, 0.80, 0.85, 0.90, 0.95, 0.97, 0.98, 0.99\}$ (8 values $\times$ 3 seeds). High-α configurations ($\alpha \geq 0.80$) yield PPL in the narrow range 151.5–151.7 with all 128 experts alive. In this plateau, the preservation map has little dynamic range, and the parameter-free default α=$1-2/E \approx 0.984$ also lands there. Lowering α to 0.7 (with a warmup schedule, below) breaks out of this plateau and achieves PPL 148.5 ± 0.2. We keep the two protocols strictly separate to avoid the tuned-vs-untuned confound: the *default* row uses the parameter-free α=$1-2/E$ heuristic (no warmup, PPL $\sim$151.5 at E=128), while the *tuned* row uses α=0.7 with a linear warmup (0.95 $\rightarrow$ 0.7 over 1,000 steps, preventing early expert collapse while gate logits are still random). Read like-for-like: at the default setting NC-OT (151.5) and Balanced OT (152.0) are within noise; under matched single-knob tuning, NC-OT (148.5) leads tuned Balanced OT (150.9) by 2.4 PPL (Table 5). The earlier framing of a "3.5-point improvement" compared the *tuned* NC-OT number against the *default* Balanced OT number and is not like-for-like; we report the matched comparisons instead. Table 11 lists the search grid and selected value for every method.

Fixed-α controls across expert count. Because the default heuristic ties α to E through α=$1-2/E$, it entangles the effect of the expert count with the effect of the floor. To disentangle them we rerun the comparison with a *single* α=0.8 held constant across all E (same ε=0.1, 3 seeds), reported in Table 4. With α fixed, Balanced OT is slightly ahead at $E \leq 64$ (a statistical match), and NC-OT's advantage then grows monotonically to +0.5 at E=128 and +2.6 PPL at E=256; since α is constant this is a property of the NC-OT mechanism rather than the α=$1-2/E$ schedule. NC-OT's seed variance also shrinks with E (0.82 $\rightarrow$ 0.11), so the largest-E gap is also the tightest.

Target distribution ν. All experiments use the uniform target $\nu_j = 1/E$. We also tested quality-weighted demand ($\nu_j \propto \sum_i m_{ij}\pi_{ij}$) and blends thereof at E=8; non-uniform variants had only a marginal effect on PPL (0.5–3 points), with the uniform prior remaining near-optimal.

C.4 How the routing shapes expert specialization over training

The controlled study above is a single snapshot. A natural follow-up is whether the quality-starvation effect *compounds* over training: if an expert consistently receives mismatched tokens, does its specialization degrade?

We simulate this with a simple model. Each expert j maintains a specialization vector $w_j \in \Delta^{S-1}$ (probability over $S{=}4$ specializations), initialized uniformly. At each training step: (i) the affinity matrix is modulated by expert specialization, $m_{ij} \propto m_{\text{base},ij} \cdot (S \cdot w_j[\text{spec}(i)])^{0.5}$; (ii) tokens are routed by the chosen method; (iii) expert weights are updated via a quality-weighted gradient toward the specializations of received tokens.

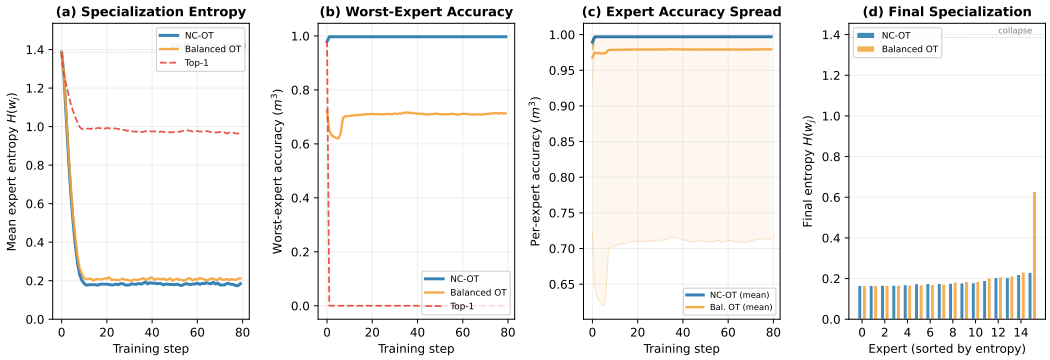


Figure 8: Expert learning dynamics over 80 training steps. **(a)** Mean specialization entropy: NC-OT (blue) drops rapidly to 0.18; Top-1 (red) barely decreases due to collapsed experts. **(b)** Worst-expert accuracy: NC-OT maintains 0.997 uniformly; Balanced OT's worst expert stalls at 0.71; Top-1 drops to 0. **(c)** Per-expert accuracy trajectories (thin lines) and means (thick): NC-OT has tight spread; Balanced OT shows persistent outliers. **(d)** Final per-expert entropy: NC-OT uniformly low; Balanced OT has a clear worst-expert tail.

Results (Figure 8). Under NC-OT, all 16 experts converge to sharp specializations (mean entropy 0.18, worst 0.23) and uniform accuracy (worst expert 0.997). Balanced OT concentrates most experts well, but its worst expert retains entropy 0.63 (nearly 3× higher than NC-OT's worst) and accuracy 0.713, because that expert consistently receives low-quality tokens (equal raw mass but mismatched specializations), receiving noisy gradients that prevent full specialization. Top-1 routing causes catastrophic expert collapse: 12 out of 16 experts become dead (zero routed mass), the remaining 4 monopolize all tokens, mean entropy remains near the uniform ceiling (0.97), and worst-expert accuracy is 0.000.

These dynamics illustrate the mechanism behind expert collapse in practice: load imbalance → poor training signal → further degradation. NC-OT's hard quality-weighted balance ensures that *every* expert receives a strong, well-matched gradient at every step, preventing this positive feedback loop.

D MULTI-STAGE NC-OT

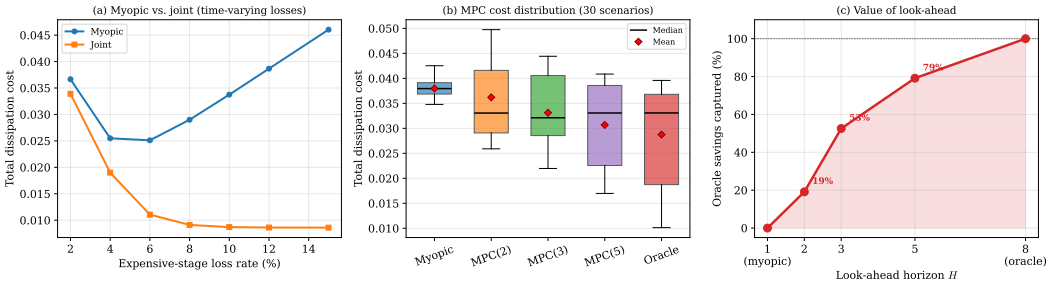


Figure 9: Multi-stage NC-OT. **(a)** Myopic vs. joint cost under time-varying loss rates ($K{=}4$, $N{=}5$); joint optimization avoids high-loss stages, achieving up to 81% cost reduction. **(b)** MPC cost distributions over 30 Monte Carlo scenarios ($K{=}8$, stochastic Markov losses). **(c)** Fraction of oracle savings captured as a function of MPC look-ahead horizon H; MPC(5) recovers 79% of the savings.

Real-world lossy transport often occurs over multiple stages: sequential rebalancing periods with time-varying preservation maps, shipments through multi-hop logistics chains, or staged chemical processes. We discretize a continuous dynamic formulation into K stages, evolving intermediate distributions $\mu_0 \to \mu_1 \to \cdots \to \mu_K = \tilde{\nu}$. The joint optimization over intermediate targets is reminiscent of Wasserstein barycenter problems (Cuturi & Doucet, 2014), but here each stage has a different lossy kernel.

Myopic vs. joint optimization. The myopic approach solves each stage independently, passing the output distribution forward. The joint approach treats intermediate targets $\mu_1, \ldots, \mu_{K-1}$ as decision variables and minimizes total cost over the simplex via SLSQP with multi-start initialization:

$$\min_{\mu_1,\ldots,\mu_{K-1}} \sum_{k=0}^{K-1} \mathcal{C}^{(k)}(\mu_k, \mu_{k+1}) \quad \text{s.t.} \quad \mu_0 \text{ fixed}, \ \mu_K = \tilde{\nu}. \tag{7}$$

Deterministic experiments. We first consider two deterministic settings where all future loss rates are known.

Experiment 1: Time-varying loss rates (Figure 9a). K=4 stages, N=5, with loss pattern [low, high, high, low] where low $= 0.3\%$ and the high rate varies from 2% to 15%. The joint optimizer re-routes mass through cheap stages, achieving 7.5% cost reduction at 2% high loss, growing to 81% at 15%. Intuitively, the optimizer front-loads trades in the initial cheap stage and back-loads in the final cheap stage, nearly skipping the expensive middle stages entirely.

Experiment 2: Location-specific loss windows. K=4, N=4. Locations $\{0, 1\}$ have low loss in stages 0–1 but high in 2–3; locations $\{2, 3\}$ the reverse. Joint optimization front-loads transport through each location's cheap window, reaching 19.4% improvement at 8% cross-window loss.

Stochastic setting: Model Predictive Control. In practice, future loss rates are uncertain. We bridge the gap between deterministic joint optimization and stochastic reality via *model predictive control* (MPC): at each stage k, solve the joint problem over a look-ahead window of H stages using realized loss rates, execute only the first stage, then re-solve at $k{+}1$ with the updated distribution.

Experiment 3: MPC under stochastic loss rates (Figure 9b,c). K=8, N=5, losses drawn from a 2-state Markov chain (low $= 0.3\%$, high $= 8\%$, persistence p=0.7), evaluated over 30 Monte Carlo paths. MPC(H=2) captures 19% of oracle savings, MPC(3) captures 53%, and MPC(5) captures 79%. Panel (c) shows the value of look-ahead: even a modest horizon of H=3 recovers more than half of the full-information oracle improvement, with diminishing returns beyond H=5.

D.1 Implementation details for the multi-stage experiments

We close with the hyperparameters behind the multi-stage results above.

Entropic regularization. All multi-stage experiments use $\varepsilon = 0.02$. At this value, the spread cost $c_{ij} \approx 2 \cdot \text{spread}$ dominates the entropy term; at $\varepsilon = 0.5$, joint improvement vanishes ($< 0.01\%$).

Joint optimization. The outer optimization over $\mu_1, \ldots, \mu_{K-1}$ uses SLSQP with simplex constraints, multi-start initialization (front-load, back-load, linear interpolation), and typically converges in 50–500 iterations. For $K = 4$, $N = 5$, each evaluation requires 4 Sinkhorn solves; total wall-clock per sweep point is 20–35 s.

MPC protocol. At each stage k, solve joint optimization over a look-ahead window of H stages using realized loss rates, execute only the first stage, then re-solve at $k{+}1$. Loss rates follow a 2-state Markov chain (low $= 0.3\%$, high $= 8\%$, persistence $p = 0.7$) over 30 Monte Carlo paths.

E Proofs and Theoretical Details

E.1 Detailed Derivation of the NC-OT Plan

We derive the optimal plan form (3) from first principles, showing how the chain rule through the endogenous variable $S(\pi)$ produces the correction term G.

Adding entropic regularization to (2), the Lagrangian with source multipliers f_i and target multipliers g_j is

$$\begin{aligned}\mathcal{L} = &\textstyle\sum_{i=1}^{N}\sum_{j=0}^{M}\big[c_{ij}\pi_{ij} + \varepsilon\,\pi_{ij}(\log\pi_{ij}-1)\big] \\ &- \sum_{i=1}^{N} f_i\Big(\sum_{j=0}^{M}\pi_{ij} - \mu_i\Big) - \sum_{j=1}^{M} g_j\Big(\sum_{i=1}^{N} m_{ij}\pi_{ij} - \nu_j S\Big),\end{aligned} \tag{8}$$

where $j = 0$ denotes the sink node, $\sum_i$ and $\sum_j$ without limits run over sources $\{1,\dots,N\}$ and destinations $\{1,\dots,M\}$ respectively. Because $S = \sum_{ij} m_{ij}\pi_{ij}$ depends on the plan, S appears implicitly in every target constraint through $\nu_j S$. Differentiating the target-constraint terms with respect to π_{ij} requires the chain rule. Define $G := \sum_j g_j\nu_j$. The total contribution of π_{ij} through all M target constraints is

$$-\sum_{j'} g_{j'}\Big(m_{ij}\,\delta_{j'j} - \nu_{j'}\,m_{ij}\Big) = -g_j\,m_{ij} + m_{ij}\,G = -m_{ij}(g_j - G). \tag{9}$$

The first term is the direct contribution to the j-th column constraint; the second arises because π_{ij} changes S, which feeds back into *every* column constraint. This coupling through G is the fundamental difference from balanced OT, where $m \equiv 1$ makes G absorbable into the multipliers.

Setting $\partial\mathcal{L}/\partial\pi_{ij} = 0$ gives

$$c_{ij} + \varepsilon\,\log\pi_{ij} - f_i - m_{ij}(g_j - G) = 0 \implies \pi_{ij} = \exp\Big(\tfrac{f_i - c_{ij} + m_{ij}(g_j - G)}{\varepsilon}\Big). \tag{10}$$

Defining $u_i := e^{f_i/\varepsilon}$, $K_{ij} := e^{-c_{ij}/\varepsilon}$, and $v_j := e^{(g_j-G)/\varepsilon}$:

$$\boxed{\pi_{ij} = u_i\,K_{ij}\,v_j^{\,m_{ij}}.} \tag{11}$$

The exponent m_{ij} on v_j is a direct consequence of the chain rule (9); in balanced OT it reduces to $u_iK_{ij}v_j$. For the sink node ($j = 0$), no target constraint applies and $\pi_{i0} = u_iK_{i0}$.

Finally, the definition $v_j = e^{(g_j-G)/\varepsilon}$ implies

$$\sum_j \nu_j \log v_j = \sum_j \nu_j \frac{g_j - G}{\varepsilon} = \frac{G - G}{\varepsilon} = 0, \tag{12}$$

using $\sum_j \nu_j = 1$. This is *not* a gauge choice as in balanced OT (see Lemma 3 and Remark 4); it is a necessary consequence of how G is defined. Any iterative algorithm that maintains v_j must enforce this constraint; violating it drives convergence to an incorrect fixed point.

E.2 Derivation of the Sherman–Morrison Newton Step

We derive the $O(M)$ Newton formula (5) for the coupled column system.

Given current row scalings u (held fixed), we solve M equations in M unknowns $\ell_j := \log v_j$:

$$\begin{aligned} F_j(\boldsymbol{\ell}) &:= C_j(\boldsymbol{\ell}) - \nu_j\,S(\boldsymbol{\ell}) = 0, \quad j = 1,\dots,M, \\ \text{where}\quad C_j &:= \sum_i m_{ij}\,u_iK_{ij}\,e^{m_{ij}\,\ell_j}, \quad S := \sum_j C_j. \end{aligned} \tag{13}$$

Since $\sum_j F_j = S - S\cdot 1 = 0$ identically, only $M-1$ of these equations are independent. We replace the last equation with the dual consistency constraint

$$\mathcal{G}(\boldsymbol{\ell}) := \sum_j \nu_j\,\ell_j = 0. \tag{14}$$

Each C_j depends only on ℓ_j (not on other ℓ_k), so the Jacobian is sparse:

$$\frac{\partial C_j}{\partial \ell_j} = \phi_j' := \sum_i m_{ij}^2\,u_iK_{ij}\,e^{m_{ij}\ell_j}, \qquad \frac{\partial C_j}{\partial \ell_k} = 0 \quad (k \neq j). \tag{15}$$

Therefore $\partial S/\partial \ell_j = \phi'_j$ and

$$\frac{\partial F_j}{\partial \ell_k} = \begin{cases} \phi'_j - \nu_j \, \phi'_j = \phi'_j(1-\nu_j) & \text{if } k = j, \\ -\nu_j \, \phi'_k & \text{if } k \neq j. \end{cases} \tag{16}$$

Writing this in matrix form, the Jacobian for the first $M-1$ rows plus the $\mathcal{G}$-constraint row is:

$$J = \begin{pmatrix} D - \boldsymbol{\nu}\, \boldsymbol{\phi}'^{\top} \\ \boldsymbol{\nu}^{\top} \end{pmatrix} \quad \text{with} \quad D = \operatorname{diag}(\phi'_1, \ldots, \phi'_M), \tag{17}$$

where the last row of $D - \boldsymbol{\nu}\, \boldsymbol{\phi}'^{\top}$ is *replaced* by $\boldsymbol{\nu}^{\top}$.

We seek $\Delta\boldsymbol{\ell}$ satisfying

$$J \, \Delta\boldsymbol{\ell} = -\mathbf{r}, \qquad \mathbf{r} = \begin{pmatrix} F_1 \\ \vdots \\ F_{M-1} \\ \mathcal{G} \end{pmatrix}. \tag{18}$$

For the first $M-1$ rows, $J_{jk} = \phi'_j \, \delta_{jk} - \nu_j \, \phi'_k = [D - \boldsymbol{\nu}\, \boldsymbol{\phi}'^{\top}]_{jk}$. Since D is diagonal and invertible (each $\phi'_j > 0$), we recognize this as a rank-1 perturbation of D. By Sherman–Morrison:

$$(D - \boldsymbol{\nu}\, \boldsymbol{\phi}'^{\top})^{-1} = D^{-1} + \frac{D^{-1}\, \boldsymbol{\nu}\, \boldsymbol{\phi}'^{\top}\, D^{-1}}{1 - \boldsymbol{\phi}'^{\top} D^{-1} \boldsymbol{\nu}}.$$

Note that $\boldsymbol{\phi}'^{\top} D^{-1} = \mathbf{1}^{\top}$ (since $D^{-1}_{jj} = 1/\phi'_j$), so the denominator is $1 - \sum_j \nu_j = 0$, and the naive Sherman–Morrison formula is singular. This confirms that $\sum_j F_j \equiv 0$ makes the column-constraint system rank-deficient.

Replacing the M-th row by $\boldsymbol{\nu}^{\top}$ resolves this. We solve the augmented system by introducing a Lagrange-like multiplier Λ.

From rows $j = 1, \ldots, M$:

$$\phi'_j \, \Delta\ell_j - \nu_j \sum_k \phi'_k \, \Delta\ell_k = -F_j. \tag{19}$$

Let $\Sigma := \sum_k \phi'_k \, \Delta\ell_k$. Then $\Delta\ell_j = (-F_j + \nu_j \, \Sigma)/\phi'_j$. Substituting into $\Sigma = \sum_j \phi'_j \, \Delta\ell_j = \sum_j (-F_j + \nu_j \, \Sigma) = -\underbrace{\sum_j F_j}_{=0} + \Sigma$ gives the identity $\Sigma = \Sigma$, confirming the rank deficiency.

Now use the $\mathcal{G}$-constraint row: $\sum_j \nu_j \, \Delta\ell_j = -\mathcal{G}$. Substituting $\Delta\ell_j = (-F_j + \nu_j \, \Sigma)/\phi'_j$:

$$\sum_j \nu_j \frac{-F_j + \nu_j \, \Sigma}{\phi'_j} = -\mathcal{G} \quad \Longrightarrow \quad -\sum_j \frac{\nu_j F_j}{\phi'_j} + \Sigma \sum_j \frac{\nu_j^2}{\phi'_j} = -\mathcal{G}. \tag{20}$$

Solving for Σ:

$$\Sigma = \Lambda := \frac{\sum_j \nu_j F_j / \phi'_j - \mathcal{G}}{\sum_j \nu_j^2 / \phi'_j}, \tag{21}$$

and the Newton step is

$$\boxed{\Delta\ell_j = \frac{-F_j + \nu_j \, \Lambda}{\phi'_j}, \qquad j = 1, \ldots, M.} \tag{22}$$

Evaluating C_j and ϕ'_j requires $O(N)$ per column (sum over sources), giving $O(NM)$ total, the same asymptotic cost as M independent column updates. The coupled solve adds only $O(M)$ scalar operations for Λ.

E.3 Dual Consistency as a Necessary Constraint

The following lemma formalizes why the constraint $\sum_j \nu_j \log v_j = 0$ is a binding structural requirement in NC-OT.

Lemma 3 (Dual consistency is binding for non-constant m)**.** *Let $\pi_{ij} = u_i K_{ij} v_j^{m_{ij}}$ be the optimal plan form derived from entropic regularization of NC-OT with column preservation constraints $\sum_i m_{ij}\pi_{ij} = \nu_j S$ (Appendix E.1), and let $v_j := e^{(g_j - G)/\varepsilon}$ with $G := \sum_l g_l \nu_l$.*

(i) The constraint $\sum_j \nu_j \log v_j = 0$ is a necessary algebraic identity: any dual (f, g) satisfying the first-order conditions must satisfy it, since $\sum_j \nu_j \log v_j = (G - G)/\varepsilon = 0$.

(ii) Suppose m is not identically constant across all (i, j). If $\sum_j \nu_j \log v_j = \delta \neq 0$, the perturbed plan $\pi'_{ij} = u_i K_{ij} (v_j e^{\delta})^{m_{ij}}$ changes column ratios by a factor $e^{(m_{ij} - m_{ij'})\delta} \neq 1$, and no source-only reparametrization $u_i \to u_i e^{\alpha}$ can restore the original plan. The violated constraint thus produces a different objective value and incorrect preserved mass.

Proof. **(i)** By definition, $\sum_j \nu_j \log v_j = \sum_j \nu_j (g_j - G)/\varepsilon = (G - G)/\varepsilon = 0$.

(ii) Under $v_j \to v_j e^{\delta}$, the plan becomes $\pi'_{ij} = u_i K_{ij} v_j^{m_{ij}} e^{m_{ij}\delta}$. The ratio between two columns j, j' at source i changes as

$$\frac{\pi'_{ij}}{\pi'_{ij'}} = \frac{\pi_{ij}}{\pi_{ij'}} \cdot e^{(m_{ij} - m_{ij'})\delta}.$$

Since $m_{ij} \neq m_{ij'}$ for some (i, j, j') by assumption, the relative allocation changes by a factor $\neq 1$. No source-only reparametrization $u_i \to u_i e^{\alpha}$ (which scales all columns of row i equally) can restore the original ratios. Moreover, row constraints force $\sum_{j=0}^{M} \pi'_{ij} = \mu_i$, so the new row scaling $u'_i = \mu_i / \sum_j K_{ij} v_j^{m_{ij}} e^{m_{ij}\delta}$ yields a plan with different column proportions and hence different objective value. □

Remark 4 (Balanced OT contrast). When $m \equiv 1$, the shift $v_j \to v_j e^{\delta}$ is absorbed by $u_i \to u_i e^{-\delta}$, leaving π unchanged. The constraint $\sum_j \nu_j \log v_j = 0$ is then automatically satisfiable by gauge choice and imposes no restriction.

Consequence for iterative algorithms. An iteration that updates columns independently (without enforcing $\sum_j \nu_j \log v_j = 0$) is free to drift along the gauge direction δ. This does not violate column residuals $F_j = C_j - \nu_j S = 0$ (since $\sum_j F_j \equiv 0$ regardless), but it produces a plan with suboptimal cost and incorrect preserved mass (Figure 3). The coupled Newton step (5) in NC-Sinkhorn prevents this drift.

E.4 Comparison with Unbalanced OT

Unbalanced OT softens both marginal constraints:

$$\min_{\pi \geq 0} \langle C, \pi \rangle + \varepsilon \, \mathrm{KL}(\pi \| K) + \lambda_1 \, D(\pi \mathbf{1} \| \mu) + \lambda_2 \, D(\pi^{\top} \mathbf{1} \| \nu).$$

Scaling algorithms for this problem (Chizat et al., 2018a) alternate three-way soft projections, converging to plans where mass is created or destroyed *uniformly* across all edges (controlled by penalty weights λ_1, λ_2).

NC-OT differs structurally:

1. Mass loss is *edge-dependent*: m_{ij} determines how much survives on each edge, not a global penalty.
2. The source marginal is matched *exactly* ($\pi \mathbf{1} = \mu$), not softly.
3. The target is matched *in shape*: $\sum_i m_{ij}\pi_{ij} = \nu_j S$, with S endogenous.
4. Two plans with identical marginal violations in the unbalanced sense can produce very different preserved masses S.

One cannot recover NC-OT solutions by tuning unbalanced penalty parameters λ_1, λ_2, because the unbalanced formulation does not encode which edges cause mass loss.

E.5 Cost vs. Preservation: Why $c \neq 1 - m$ in Practice

When $c_{ij} = 1 - m_{ij}$, the NC-OT objective reduces to

$$\textstyle\sum_{ij}(1 - m_{ij})\,\pi_{ij} = \sum_i \mu_i - S = 1 - S,$$

so minimizing cost is *equivalent* to maximizing preserved mass S. This degenerate case admits an $O(N + M)$ closed-form solution (Appendix B.4), but it obscures the fundamental trade-off that makes NC-OT interesting: in general, the cost-optimal plan does *not* maximize S.

E.5.1 Example 1: Portfolio rebalancing with tax lots

Consider rebalancing a portfolio of N assets. Selling one unit of asset i at bid price P_i^b and buying asset j at ask price P_j^a preserves a fraction $m_{ij} = (P_i^b/q_i)(q_j/P_j^a)$, where q_i is the reference price. The preservation map captures *only the bid-ask spread*, which determines the mass loss on each edge.

But selling asset i also triggers *capital gains tax* on each tax lot k held. With K_i lots of sizes n_{ik} shares at cost bases B_{ik}, selling u_i dollars of asset i requires choosing a lot-liquidation schedule s_{ik} subject to $\sum_k s_{ik} = u_i$, $0 \le s_{ik} \le n_{ik}q_i$. The full transport cost is

$$c_{ij} = \underbrace{(1 - m_{ij})}_{\text{spread loss}} + \underbrace{L_i\big(\textstyle\sum_{j' \neq i} \pi_{ij'}\big)}_{\text{tax cost (source-only)}}, \tag{23}$$

where

$$L_i(u) = \min\Big\{\sum_k T_{ik}\, s_{ik} : \sum_k s_{ik} = u,\ 0 \le s_{ik} \le n_{ik}q_i\Big\}$$

is a piecewise-linear convex function (optimal lot selection orders lots by T_{ik} ascending: loss lots first, then zero-basis, then gains), with $T_{ik} = \tau \cdot \max(0, P_i^b - B_{ik})$ the per-share tax.

Key structural points.

1. The tax term is *source-dependent but destination-independent*: it depends on *which* asset you sell (and its embedded gains), not where proceeds are routed.
2. This creates a genuine separation: edge (i, j) with high m but large embedded gains has *high cost, high preservation*, while edge (i', j) with lower m but small gains has *lower cost, lower preservation*.
3. The cost-minimizing plan will sacrifice S to avoid high-tax trades. A plan that maximizes S ignores tax entirely and incurs unnecessarily high cost.
4. Tax payments further reduce post-tax wealth: $S_{\text{post}} = \sum_{ij} m_{ij}\pi_{ij} - \sum_{ik} T_{ik}s_{ik}$, so the true preserved wealth is *lower* than S. Loss lots ($T_{ik} < 0$, i.e. cost basis above current price) create a beneficial nonlinearity: realizing losses increases S_{post} and can shift the target constraints in a way that reduces total spread cost.

E.5.2 Example 2: MoE token routing

In the MoE application (Section 4.2), m_{ij} is the quality retention when token i is processed by expert j, and $S = \sum_{ij} m_{ij}\pi_{ij}$ is the total *routing quality*, the analogue of preserved mass. Both routers use the same shifted logit cost $c_{ij} = -z_{ij} + \min_j z_{ij}$; they differ in the preservation map used by the constraints. Balanced OT sets $m \equiv 1$ and balances raw token mass, whereas NC-OT uses the learned quality-retention map and balances quality-weighted load. Thus c is not proportional to $1 - m$ in either formulation.

In the controlled synthetic study, NC-OT achieves exact QW balance and worst-expert accuracy 0.983. Balanced OT has QW imbalance 0.034 and worst-expert accuracy 0.896 (Table 10 and Figure 6b). These values show the effect of enforcing quality-weighted rather than raw-mass balance.

E.5.3 Connection to the convergence diagnostic

The convergence plot (Figure 3b) exhibits the same phenomenon. The *independent* solver converges to $S = 0.914$ (higher than the optimal $S^\star = 0.861$) but at *higher cost*. Without dual consistency, the

iteration drifts toward a plan that preserves more mass (routing through high-m edges) but incurs more cost from edges whose c_{ij} exceeds $1 - m_{ij}$. The coupled Newton correctly balances preservation against the full cost, accepting lower S when doing so reduces the objective.

In each application, the pattern is the same: when $c \neq 1 - m$, maximizing S is *not* the same as minimizing cost, and any solver that conflates the two, whether by ignoring dual consistency or by using balanced OT, converges to a feasible but suboptimal plan.

Remark 5 (When does $c = 1 - m$ hold?). Only when the *sole* friction is proportional mass loss (pure bid-ask spreads with no taxes, or attenuation in a channel with no routing cost. In logistics (distance $\neq$ spoilage), chemistry (energy input $\neq$ yield), and MoE routing (computational cost $\neq$ quality), c and m are generically independent.

E.6 Convergence of the NC-Sinkhorn Iteration

We prove two results for the NC-Sinkhorn iteration (Algorithm 1): *qualitative convergence* to the unique entropic NC-OT optimum (Theorem 1), and an *iteration complexity bound* of $O(\delta^{-2})$ for a sequential variant (Theorem 2). The main argument recognizes NC-Sinkhorn as alternating I-projections under KL divergence, so convergence follows from the classical theorem of Bregman (Bregman, 1967) and Csiszár (Csiszár, 1975). We close with a complementary dual-ascent perspective that also covers the practically important case of inexact (finite) Newton inner iterations.

E.6.1 Entropic NC-OT as KL Minimization

Define the Gibbs reference measure $\xi_{ij} := e^{-c_{ij}/\varepsilon}$ for $j \geq 1$ and $\xi_{i0} := e^{-c_{i0}/\varepsilon}$ for the sink. With this choice, the entropic NC-OT objective (2) plus regularization becomes

$$\min_{\pi \geq 0} \ \varepsilon \cdot \mathrm{KL}(\pi \| \xi) \quad \text{s.t.} \quad \pi \in \mathcal{R} \cap \mathcal{C}, \tag{24}$$

where $\mathrm{KL}(\pi\|\xi) = \sum_{i,j} \pi_{ij} \log(\pi_{ij}/\xi_{ij}) - \pi_{ij} + \xi_{ij}$ is the generalized KL divergence, and the two constraint families are

$$\begin{aligned} \mathcal{R} &:= \{\pi \geq 0 : \textstyle\sum_{j=0}^{M} \pi_{ij} = \mu_i, \ \forall\, i\}, \\ \mathcal{C} &:= \{\pi \geq 0 : \textstyle\sum_{i} m_{ij}\pi_{ij} = \nu_j S(\pi), \ \forall\, j\}, \end{aligned} \tag{25}$$

with $S(\pi) = \sum_{ij} m_{ij}\pi_{ij}$. Both families are defined by linear equalities in π. The source balance constraints are manifestly linear. The target balance constraints rearrange to $\sum_i m_{ij}\pi_{ij} - \nu_j \sum_{i',j'} m_{i'j'}\pi_{i'j'} = 0$, which is a homogeneous linear equation in the entries of π (the endogenous S does not introduce nonlinearity because it is itself linear in π).

The primal objective $\varepsilon \cdot \mathrm{KL}(\pi\|\xi)$ is strictly convex on $\mathbb{R}_{\geq 0}^{N\times(M+1)}$, so the minimizer $\pi^\star$ over the (polyhedral) feasible set $\mathcal{R} \cap \mathcal{C}$ is unique.

E.6.2 Proof of Theorem 1: Alternating I-Projections

Given a reference measure $\xi > 0$ and a closed convex set $\mathcal{A}$, the *I-projection* of ξ onto $\mathcal{A}$ is $P_{\mathcal{A}}(\xi) := \arg\min_{\pi\in\mathcal{A}} \mathrm{KL}(\pi\|\xi)$. When $\mathcal{A}$ is defined by linear equalities, the I-projection takes the exponential-family form $\pi^\star_{ij} = \xi_{ij} \exp(\text{linear function of dual variables})$. The classical alternating-projection result (Bregman, 1967; Csiszár, 1975) states that if $\mathcal{A}_1, \ldots, \mathcal{A}_K$ are closed convex sets defined by linear equalities with $\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_K \neq \emptyset$, then the cyclic projection sequence $\pi^{(t+1)} = P_{\mathcal{A}_{(t \bmod K)+1}}(\pi^{(t)})$ converges to $P_{\mathcal{A}_1\cap\cdots\cap\mathcal{A}_K}(\pi^{(0)})$, the I-projection of the starting point onto the intersection.

The NC-Sinkhorn iteration alternates two half-steps, each of which is an I-projection:

- The *row update* computes the I-projection of the current plan onto $\mathcal{R}$. Since $\mathcal{R}$ consists of N independent row-sum constraints, the I-projection has the closed form $\pi_{ij}^{(t+1/2)} = \pi_{ij}^{(t)} \cdot \mu_i / \sum_{j'} \pi_{ij'}^{(t)}$, which is exactly the row scaling $u_i \leftarrow \mu_i/[\cdots]$.
- The *column update* computes the I-projection of $\pi^{(t+1/2)}$ onto $\mathcal{C}$. In NC-OT, the m_{ij}-dependent exponents make this projection nonlinear in $\ell_j = \log v_j$, so it cannot be computed in closed form. The coupled Newton step (22) finds the dual variables (g_j) (subject to dual consistency

$\sum_j \nu_j \log v_j = 0$) that satisfy the stationarity conditions of $\min_{\pi \in \mathcal{C}} \mathrm{KL}(\pi \| \pi^{(t+1/2)})$. When the Newton iteration converges, the result is the exact I-projection onto $\mathcal{C}$.

The dual consistency constraint $\sum_j \nu_j \log v_j = 0$ is essential: it selects the unique point within $\mathcal{C}$ that corresponds to a valid dual, breaking the gauge degeneracy that would otherwise leave a one-dimensional family of spurious fixed points (cf. Figure 3).

Theorem 1 (Convergence of NC-SINKHORN, restated)**.** *Suppose the coupled Newton column update is run to convergence at each outer step (i.e., each I-projection onto $\mathcal{C}$ is exact). Then the* NC-SINKHORN *iterates $\pi^{(t)}$ converge to $\pi^\star$, the unique solution of the entropically regularized NC-OT problem.*

Proof of Theorem 1. The iteration alternates exact I-projections onto $\mathcal{R}$ and $\mathcal{C}$, both of which are closed convex sets defined by linear equalities with $\mathcal{R} \cap \mathcal{C} \neq \emptyset$ (the feasible set is non-empty by assumption). By the Csiszár–Bregman alternating projection theorem, $\pi^{(t)}$ converges to the I-projection of the starting point onto $\mathcal{R} \cap \mathcal{C}$, which is the minimizer of $\mathrm{KL}(\pi \| \xi)$ subject to $\pi \in \mathcal{R} \cap \mathcal{C}$. This is precisely the entropic NC-OT optimum $\pi^\star$. Uniqueness follows from the strict convexity of $\mathrm{KL}(\cdot \| \xi)$. □

Remark 6 (Role of dual consistency). Without the constraint $\sum_j \nu_j \ell_j = 0$, the column update solves M independent equations $C_j = \nu_j S$, which are necessary but not sufficient for dual optimality. They leave a one-dimensional family of solutions parametrized by G. The unconstrained iteration converges to a fixed point in this family that generically does *not* satisfy $\sum_j \nu_j \ell_j = 0$, yielding a suboptimal plan with incorrect S (cf. Figure 3). The coupled Newton step (22) simultaneously enforces dual consistency and column balance, collapsing the spurious family to the unique optimum.

E.6.3 ITERATION COMPLEXITY VIA SEQUENTIAL I-PROJECTIONS

The preceding proof establishes convergence but does not give a rate. The difficulty is that NC-SINKHORN uses coupled Newton iterations for the column update, making the per-step progress non-trivial to quantify directly. To obtain a worst-case bound, we instead analyze a conservative "constraint-by-constraint Sinkhorn" variant that performs a single I-projection onto the most-violated constraint at each step. This simpler algorithm serves as the analytical backbone: it admits a clean $O(\delta^{-2})$ iteration bound, and our Newton-based implementation can only converge faster, since each Newton step simultaneously satisfies multiple constraints.

The sequential I-projection algorithm. The sequential I-projection algorithm proceeds as follows: at each step, identify the constraint row $a_{i^\star}^\top \pi = b_{i^\star}$ with the largest violation, and project onto it by solving

$$\pi^{(k+1)} = \arg\min_{\pi' \geq 0} \mathrm{KL}(\pi' \| \pi^{(k)}) \quad \text{s.t.} \quad a_{i^\star}^\top \pi' = b_{i^\star}.$$

The solution is a multiplicative reweighting: $\pi_{ij}^{(k+1)} \propto \pi_{ij}^{(k)} \exp(-\eta_k a_{i^\star,(i,j)})$, which is exactly the same exponential update used by each row/column scaling in Sinkhorn. The only difference from standard Sinkhorn is that we update one constraint at a time (the worst one) rather than cycling through blocks.

Writing the full NC-OT problem in standard form

$$\min_{\pi \geq 0} \langle C, \pi \rangle \qquad \text{s.t.} \qquad A\pi = b, \tag{26}$$

with $A \in \mathbb{R}^{(N+M) \times N(M+1)}$ and $b = (\mu_1, \ldots, \mu_N, 0, \ldots, 0)^\top$, adding the entropic term $\varepsilon \, \mathrm{KL}(\pi \| \bar{\pi})$ makes (26) strictly convex with purely linear constraints, which is the setting of the following theorem.

Theorem 7 (Theorem 2 restated: iteration complexity of sequential I-projections)**.** *Consider the entropically regularized NC-OT problem*

$$\min_{\pi \in \Delta^d} \left\{ \textstyle\sum_{i=1}^N \sum_{j=0}^M c_{ij} \, \pi_{ij} + \varepsilon \, \mathrm{KL}(\pi \| \bar{\pi}) \right\} \qquad \textit{s.t.} \quad A\pi = b,$$

where $A \in \mathbb{R}^{r \times d}$ encodes the source-marginal and target-balance constraints of (2), $d = N(M+1)$, and $\bar{\pi} \in \Delta^d_{>0}$. Let $K_{ij} := \bar{\pi}_{ij} \, e^{-c_{ij}/\varepsilon}$, define $\tilde{\pi} := K / \|K\|_1$, and let $\pi^\star$ be the optimum.

Then the sequential I-projection algorithm initialized at $\tilde{\pi}$, projecting onto the most-violated constraint with step size $\eta_k = u_k/\|A\|_\infty^2$, returns $\pi^{(K)}$ with $\|A\pi^{(K)} - b\|_\infty \le \delta$ in at most

$$K \;\le\; \frac{2\,\|A\|_\infty^2\,\mathrm{KL}(\pi^\star\|\tilde{\pi})}{\delta^2}$$

iterations.

Proof. Absorbing the linear cost into the reference reduces the problem to $\min \mathrm{KL}(\pi\|\tilde{\pi})$ s.t. $A\pi = b$ (the additive constant $\varepsilon \log \|K\|_1$ is immaterial). The dual is $\max_{\alpha\ge 0}\ \omega(\alpha) := -\log Z(\alpha) - \langle \alpha, b^\pm\rangle$ with $b^\pm = (b^\top, -b^\top)^\top$ and $Z(\alpha) = \sum_{i,j} \tilde{\pi}_{ij} \exp(-((A^\pm)^\top\alpha)_{(i,j)})$, where $A^\pm := (A^\top, -A^\top)^\top$. At $\alpha = 0$: $\omega(0) = 0$ and $\pi(\cdot;0) = \tilde{\pi}$; strong duality gives $\omega(\alpha^\star) = \mathrm{KL}(\pi^\star\|\tilde{\pi})$.

Each I-projection step on the most-violated constraint $i^\star$ with violation $u_k > \delta$ corresponds to dual coordinate ascent $\alpha \mapsto \alpha + \eta_k e_{i^\star}$. A standard cumulant-generating-function bound (Popoviciu) gives $\psi''(\eta) \le \|A\|_\infty^2$, yielding per-step dual progress $\ge u_k^2/(2\|A\|_\infty^2) \ge \delta^2/(2\|A\|_\infty^2)$. Summing K steps and bounding by the total available dual gap $\mathrm{KL}(\pi^\star\|\tilde{\pi})$ gives the result. □

Remark 8 (Sinkhorn as a special case). When $m \equiv 1$ (balanced OT), the constraint matrix A encodes only row and column marginals, and the cyclic Sinkhorn algorithm of Cuturi (2013) is precisely this sequential I-projection scheme, cycling through constraint rows rather than selecting the most-violated one. Our NC-SINKHORN algorithm (Section 3) extends this to NC-OT via coupled Newton steps for the target-balance constraints, log-domain stabilization, and the Sherman–Morrison acceleration; Theorem 7 provides the worst-case $O(\delta^{-2})$ guarantee for the underlying sequential backbone that these enhancements accelerate in practice.

Remark 9 (Convergence rate of NC-SINKHORN). Within the alternating I-projection framework, the row update is a closed-form maximization (linear convergence in f), while the column update converges quadratically (Newton on a smooth, strictly concave function). Each outer iteration therefore makes at least linear progress in the dual gap. A tighter bound analogous to the $\widetilde{O}(n^2/\delta^3)$ result for δ-approximate balanced Sinkhorn (Altschuler et al., 2017) would require bounding the condition number of the Jacobian (17) uniformly in t; this is left for future work.

E.6.4 DUAL-ASCENT PERSPECTIVE AND INEXACT NEWTON

The I-projection proof above assumes exact column projections (Newton run to convergence). In practice, we use only 3–5 inner Newton steps per outer iteration, so the column I-projection is slightly inexact. A dual-ascent viewpoint handles this directly.

The Lagrangian (8) yields the dual problem

$$\max_{f,g}\ \mathcal{D}(f,g) := \sum_i f_i\mu_i + \sum_j g_j\nu_j S(f,g) - \varepsilon \sum_{ij} \pi_{ij}(f,g), \tag{27}$$

where $\pi_{ij}(f,g) = \exp\!\big(\frac{f_i - c_{ij} + m_{ij}(g_j - G)}{\varepsilon}\big)$. Since the entropic primal is strictly convex on a polyhedral feasible set, Fenchel–Rockafellar duality gives strong duality, a unique primal optimum $\pi^\star$, and a unique dual optimum $(f^\star, g^\star)$ (up to the sink constant; there is no gauge freedom for g in NC-OT, cf. Lemma 3 and Remark 4).

Proposition 10 (Dual ascent)**.** *Each step of Algorithm 1 does not decrease the dual objective $\mathcal{D}$.*

Proof sketch. **Row update.** Fixing g and optimizing over f, the stationarity condition $\partial\mathcal{D}/\partial f_i = 0$ yields the row update $u_i \leftarrow \mu_i/[\cdots]$. Since $\mathcal{D}$ is concave in f for fixed g, this closed-form maximization strictly increases $\mathcal{D}$ unless f was already optimal. **Column update.** The coupled Newton step (5) solves $F_j(\boldsymbol{\ell}) = 0$ for all j, which are the stationarity conditions $\partial\mathcal{D}/\partial g_j = 0$ restricted to the consistency manifold. Newton on a strictly concave function with a non-degenerate Hessian converges quadratically. Even with only 3–5 inner iterations, each partial Newton step moves uphill on $\mathcal{D}$. □

Since each outer step does not decrease $\mathcal{D}$ (Proposition 10) and $\mathcal{D}$ is bounded above (the primal is feasible and bounded below), the dual sequence $\{\mathcal{D}^{(t)}\}$ converges. At any limit point, row stationarity ($\partial\mathcal{D}/\partial f_i = 0$) and column residuals ($F_j = 0$) together with dual consistency yield the full KKT

conditions, so $\pi^{(t)} \to \pi^\star$ by uniqueness. This argument requires only that each Newton step increases $\mathcal{D}$; exact convergence is unnecessary. It therefore justifies the finite-Newton implementation used in practice.

E.7 The Column Coupling Is Not an Entropic Artifact: Quadratic Regularization and the Semi-Dual View

A natural question is whether the dual-consistency coupling that defines NC-OT is an artifact of the *entropic* regularizer, or an intrinsic feature of edge-dependent dissipation. We show it is intrinsic by rederiving the optimality structure under a *quadratic* regularizer, and we relate NC-Sinkhorn to the semi-dual paradigm (Cuturi & Peyré, 2018; Blondel et al., 2018).

Quadratic (non-entropic) regularization. Replace the entropic penalty in (2) by the quadratic penalty $\frac{\varepsilon}{2}\sum_{ij}\pi_{ij}^2$. With source multipliers f_i and column multipliers $w_j := g_j - G$ (the same ν-centered combination as in the entropic case), the KKT stationarity condition $\partial\mathcal{L}/\partial\pi_{ij} = 0$ with the nonnegativity constraint gives a *sparse, thresholded* plan

$$\pi_{ij} = \tfrac{1}{\varepsilon}\big[\, f_i - c_{ij} + m_{ij}\,\varepsilon\, w_j \,\big]_+, \tag{28}$$

where $[\cdot]_+ = \max(\cdot, 0)$. Zero entries mark source–destination pairs where the preservation m_{ij} is too low (or the cost c_{ij} too high) to be worth using, so the quadratic regularizer yields genuinely sparse plans, unlike the dense entropic optimum. Crucially, the same edge-dependent factor m_{ij} multiplies w_j, and the same dual-consistency condition

$$\textstyle\sum_j \nu_j\, w_j = 0 \tag{29}$$

still couples the column variables (it follows, exactly as in (12), from $w_j = g_j - G$ and $G = \sum_j \nu_j g_j$). Thus edge-dependent dissipation induces the column coupling *regardless* of the regularizer. The only thing that changes is the per-column solve: under entropy the exponent $v_j^{m_{ij}}$ makes the balance equation $\sum_i m_{ij}\pi_{ij} = \nu_j S$ transcendental (hence the inner Newton loop), whereas under quadratic regularization the $[\cdot]_+$ threshold makes it piecewise-linear, so after fixing active sets the row and coupled-column updates become piecewise closed-form (a threshold/breakpoint sweep). The main solver in this paper remains the entropic NC-Sinkhorn; (28) is included to show the coupling is structural.

Semi-dual view. NC-Sinkhorn also admits a natural semi-dual reading (Cuturi & Peyré, 2018). The row potentials admit a closed-form c-transform (exactly our row rescaling $u_i \leftarrow \mu_i/[\sum_j K_{ij} v_j^{m_{ij}} + K_{i0}]$), so they can be eliminated, leaving an optimization purely over the column potentials, which we solve with the coupled Newton step. This is precisely the semi-dual paradigm: eliminate one dual block in closed form and optimize the other with a (quasi-)Newton solver. Two features are specific to NC-OT. First, the preservation map m_{ij} breaks the closed-form *column* c-transform that standard balanced Sinkhorn enjoys, so the Newton solve is necessary rather than optional. Second, the dual-consistency condition (29) supplies the extra scalar coupling with no analogue in the balanced semi-dual. The same coupling therefore appears under both the entropic and quadratic regularizers, confirming that the inner Newton step is only the numerical means to the exact column projection that Theorem 1 assumes, not a feature of the entropic choice.