Towards Anytime-Valid Statistical Watermarking

Baihe Huang* UC Berkeley
Eric Xu UC Berkeley
Kannan Ramchandran UC Berkeley
Jiantao Jiao UC Berkeley, NVIDIA

Michael I. Jordan
UC Berkeley, ENS, & Inria

Abstract

The proliferation of Large Language Models (LLMs) necessitates efficient mechanisms to distinguish machine-generated content from human text. While statistical watermarking has emerged as a promising solution, existing methods suffer from two critical limitations: the lack of a principled approach for selecting sampling distributions and the reliance on fixed-horizon hypothesis testing, which precludes valid early stopping. In this paper, we bridge this gap by developing an e-value-based watermarking framework, *Anchored E-Watermarking*, that unifies optimal sampling with anytime-valid inference. Unlike traditional approaches where optional stopping invalidates Type-I error guarantees, our framework enables valid, anytime-inference by constructing a test supermartingale for the detection process. By leveraging an anchor distribution to approximate the target model, we characterize an optimal e-value with respect to the *worst-case log-growth rate* and derive the optimal expected stopping time. Our theoretical claims are substantiated by simulations and evaluations on established benchmarks, showing that our framework can significantly enhance sample efficiency and reduce the average token budget required for detection.

1 Introduction

The revolutionary success of Large Language Models (LLMs) at generating human-like texts [6–8] has raised several societal concerns regarding the misuse of LLM outputs. Unregulated LLM outputs pose risks ranging from the contamination of future training corpora [43, 13] to the propagation of disinformation [55, 49] and academic misconduct [24, 34]. Consequently, there is an urgent need for reliable detection mechanisms capable of distinguishing LLM-generated text from human-authored content.

To address this challenge, Aaronson [1] and Kirchenbauer et al. [28] propose *statistical watermarking* as a theoretically-grounded method to distinguish machine-generated content from a target distribution. These methods operate by injecting statistical signals into the generated texts in the decoding phase of LLMs. Mechanistically, the watermark generator couples the output tokens by a sequence of pseudorandom seeds. During detection, the detector reconstructs the seeds and performs a hypothesis test to check the dependence between the seeds and the observed tokens. This formulation allows the detection problem to be treated as a rigorous statistical hypothesis test [22] where the detection power and sample complexity are governed by the randomness in the target distribution. Various designs for the underlying seed distributions, such as binary [10], exponential [29], and Gaussian [4] have been explored. Recently, Huang et al. [23] advanced the concept of using an auxiliary open-source model as an *anchor* to generate random seeds and coupling them with tokens via speculative decoding [30]. This approach is motivated by the theoretical insight that watermarking achieves the optimal power

*Correspondence: baihe_huang@berkeley.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

when the seed distribution approximates the target distribution [22]. By exploiting the distributional proximity between modern LLMs, anchor-based methods have shown promise in improving both detection efficiency and robustness to removing-attacks [35]. Empirically, watermarked texts can be detected within short token horizons (< 100 tokens) and high success rate (nearly 50% under paraphrasing) under various attacks.

Despite these successes, state-of-the-art watermarking remains constrained by two fundamental limitations. First, the design of generation and detection schemes lacks a unifying guiding principle. Huang et al. [23] utilize a hashed green-red-list seed adapted from Kirchenbauer et al. [28], but this heuristic lacks rigorous justification. What is needed is a characterization of optimal generation and detection schemes under given anchor distributions. Second, there exists a methodological disconnect between generation and detection. While text generation is inherently autoregressive and variable in length, current detection paradigms are *fixed-horizon* and *ad hoc*. In practice, for example, a detector may wish to process a stream of tokens and stop as soon as confidence is high. However, with the standard statistical paradigm based on p-values, it is not allowed to continuously monitor the test statistic and decide when to stop based on the current value—this is known as "p-hacking" and it invalidates the Type-I error guarantee. This inability to perform *post-hoc* sequential analysis severely limits detection efficiency. These gaps highlight a critical research challenge:

How can we design provably efficient watermarking schemes leveraging anchor distributions that allow for valid early stopping?

In this work, we provide a formal resolution to this question. We formulate the problem of *Anchored E-Watermarking*, where both the generator and detector share access to an anchor distribution p_0, aiming to watermark a family of distributions in the δ-proximity of p_0. Departing from classical p-value analysis, we adopt *e-values* as the central detection paradigm. E-values [51, 41, 50] are nonnegative random variables E satisfying $\mathbb{E}_{\mathbf{H}_0}[E] \leq 1$ under the null. Unlike p-values, e-values arise from supermartingales, and it is therefore possible to preserve Type-I error guarantees under optional stopping based on ongoing analysis of data. We analyze an optimal e-value for watermark detection with respect to the worst-case log-growth rate [27] and the expected stopping time [18, 53], thus fully characterize the average per-step growth rate of evidence and sample efficiency. Our main theoretical contributions are summarized informally as follows:

Theorem 1.1 (Informal version of Theorem 4.1 and Theorem 4.3)**.** *The optimal worst-case log-growth rate $\mathbb{E}_{\mathbf{H}_1}[\log E]$ under the alternative $\mathbf{H}_1$ is given by:*

$$J^* = h + \left(1 - \frac{\delta}{2}\right) \log\left(1 - \frac{\delta}{2}\right) + \frac{\delta}{2} \log\left(\frac{\delta}{2(n-1)}\right),$$

where $h = H(p_0)$ is the Shannon entropy of the anchor distribution p_0 and $\delta > 0$ is a robustness tolerance parameter. Furthermore, the optimal expected stopping time to achieve a Type-I error α scales as $\frac{\log(1/\alpha)}{J^*}$.

By enabling valid sequential testing, our framework significantly improves detection efficiency, allowing the system to flag machine-generated text with fewer tokens than fixed-horizon counterparts. With the ability to early-stop, this sequential paradigm enhances robustness against adaptive attacks: because the detector can terminate immediately upon accumulating sufficient evidence, the watermark remains effective even if the attacker perturbs the text heavily in later segments (post-stopping). Consequently, our approach offers a theoretically rigorous and practically superior alternative to existing heuristic detection methods. Through experiments on real watermarking benchmark [35], we show that the theoretical scheme implied by Theorem 1.1 achieves consistently higher sample efficiency than state-of-the-art methods while reducing the token consumption.

1.1 Related works

Statistical watermarking. Watermarking is a powerful white-box method for detecting LLM-generated texts [45]. Comparing with edit-based watermarking [38, 3, 54, 25], the idea of statistical watermarking, dating back to the work of Venugopal et al. [48], injects watermarks into text generation (generative watermarks). Aaronson [1] and Kirchenbauer et al. [28] introduce statistical watermarking as a provable method of embedding statistical signals into language model generation. Huang et al. [22], Kuditipudi et al. [29], Zhao et al. [56], Li et al. [31], Block et al. [4] develop formal guarantees

and use them to design provably efficient watermarking schemes. To address vulnerability to attacks, several works focus on the design of robust pseudo-random codes as the basis for constructing robust watermarks [9, 16]. Following Kirchenbauer et al. [28], several recent works [37, 32, 20, 19, 15, 23] advance the efficiency of watermarking methods by leveraging semantic information. These works exploit the fact that the target distribution (i.e., human language) has certain benign structure. In particular, Huang et al. [23] encode semantic information using a proposal language model, thus introducing speculative decoding [30] into statistical watermarking.

E-values and sequential hypothesis testing. The field of sequential hypothesis testing aims to design statistical methods that can accommodate streams of data and data-dependent stopping times [52, 5, 39]. The earliest work in the field relied on likelihood ratios and p-values. These tools are appropriate for a limited range of problems but in general fail to control Type-I error under data-dependent stopping rules and composite hypotheses. These deficiencies led to the development of tests based on nonnegative supermartingales, which yield e-values at arbitrary stopping times under composite null hypotheses [50, 40, 18, 36]. Moreover, this framework also provides leverage on the problem of statistical power (i.e., the control of Type-II error). For example, Grünwald et al. [18] introduced worst-case growth rate optimality (GROW) as a criterion to assess power for a composite alternative. More recently, Waudby-Smith et al. [53] established optimal rates for a broad class of betting-based e-processes (a generalization of nonnegative supermartingales), and also obtained matching upper and lower bounds on mean rejection times based on these processes. A concurrent work by Su et al. [44] also studies anytime-valid detection from watermark statistics, while focusing on sequential detection from null-uniform pivots, especially for Gumbel-max.

2 Preliminaries

2.1 Statistical watermarking

Statistical watermarking for Large Language Models (LLMs) provides a probabilistic framework for distinguishing machine-generated text from human-written text. Unlike post-hoc detection methods that rely on classifiers trained on model artifacts, watermarking actively embeds a statistical signal into the generation process. This signal is designed to be imperceptible to humans yet statistically significant to an algorithmic detector possessing a secret key.

Statistical watermarking as a hypothesis testing problem. Formally, statistical watermarking is cast as a hypothesis testing problem. Let q denote the target distribution over the sample space $\mathcal{V}$, and let $\mathcal{S}$ represent the space of pseudorandom seeds. The watermarking protocol modifies the standard decoding mechanism of the LLM such that the output tokens $V \in \mathcal{V}$ and the pseudorandom seeds $S \in \mathcal{S}$ are sampled jointly from a watermarked distribution $\mathcal{P}_W$.

The detection task seeks to determine whether an observed sequence $v^{1:T} = (v^1, \ldots, v^T)$ was generated by the watermarked model. We write the sequential detector as $\mathcal{D}_T : \mathcal{V}^T \times \mathcal{S}^T \to \{\text{Watermarked}, \text{Unwatermarked}\}$. At each time t, the detector reconstructs the seed s^t from the secret key and computes a one-step score; the final decision is based on the accumulated evidence across the sequence. This reduces to testing for independence between the token sequence $v^{1:T}$ and the reconstructed seed sequence $s^{1:T}$. We define the null and alternative hypotheses as follows [22]:

- **Null hypothesis $\mathbf{H}_0$:** The sequence $v^{1:T}$ is generated independently of the seeds $s^{1:T}$ (e.g., by a human or an unwatermarked model).
- **Alternative hypothesis $\mathbf{H}_1$:** The token-seed pairs $(v^t, s^t)_{t=1}^T$ are sampled from the joint watermarked mechanism, implying they come from the watermarked model.

It is pertinent to distinguish this *distributional* watermarking approach from classical instance-level watermarking techniques applied to static media, such as images or audio [12]. While classical methods embed a signature into a specific, fixed outcome (post-generation), statistical watermarking modifies the stochastic sampling process itself, ensuring that any realization from the model carries the statistical evidence of its origin without requiring rigid alterations to the final output.

Statistical guarantees. A robust watermarking scheme is characterized by its ability to provide three fundamental guarantees.

First, to preserve generation quality and ensure watermarked content remains indistinguishable from ordinary samples, the distance (e.g., KL-divergence) between the watermarked distribution $\mathcal{P}_W$ and the original target distribution q must be minimized. Ideally, we seek a scheme that introduces zero distributional shift [21].

Definition 2.1 (Distortion-free)**.** A watermark is considered *distortion-free* (or unbiased) if the outcome marginal of the watermarked distribution $\mathcal{P}_W$ matches the target distribution q. Formally, a distortion-free watermark satisfies:

$$\mathcal{P}_W(A \times \mathcal{S}) = q(A), \qquad \forall A \subseteq \mathcal{V}.$$

Second, it is often required in practice that the detector operates without access to the watermarked distribution since the underlying model's parameters are generally unknown to the detector. This means that the detection scheme needs to be designed for a *family of target distributions* in a model-agnostic way [29], leading to the next concept.

Definition 2.2 (Model-agnosticity)**.** A watermark is *model-agnostic* if the seed-marginal of $\mathcal{P}_W$ is chosen independently of the target distribution q. That is, the distribution of seeds S does not depend on the model's logits.

Third, the reliability of the watermark is quantified by standard statistical error rates. The *Type I error* (false positive rate) measures the probability of incorrectly identifying non-watermarked text (e.g., human-written) as watermarked. Given the high stakes of false accusations (e.g., in academic integrity), this error must be bounded by a small significance level α:

$$\mathbb{P}_{H_0}\big(\mathcal{D}_T(V^{1:T}, S^{1:T}) = \text{Watermarked}\big) \leq \alpha.$$

The *Type II error* (false negative rate) measures the probability that watermarked text fails to be detected. Minimizing this error is equivalent to maximizing the statistical power of the test. For a target error rate β, we require:

$$\mathbb{P}_{H_1}\big(\mathcal{D}_T(V^{1:T}, S^{1:T}) = \text{Unwatermarked}\big) \leq \beta.$$

2.2 Sequential hypothesis testing and e-values

Consider a measurable space $(\Omega, \mathcal{F})$ and a family of probability distributions $\mathcal{P}$. We are interested in testing a null hypothesis $\mathcal{H}_0 : P \in \mathcal{P}_0 \subset \mathcal{P}$ against an alternative $\mathcal{H}_1 : P \in \mathcal{P}_1 = \mathcal{P} \setminus \mathcal{P}_0$. An e-value is a random variable whose expectation is bounded by unity under the null hypothesis [50].

Definition 2.3 (E-value)**.** A positive random variable E is an e-value for $\mathcal{H}_0$ if for all $P \in \mathcal{P}_0$,

$$\mathbb{E}_P[E] \leq 1.$$

Intuitively, an e-value represents the amount of wealth a gambler would have after betting against the null hypothesis, in a game where the game is fair or unfavorable under $\mathcal{H}_0$, and starting with an initial wealth of one. On the other hand, if E becomes large, this suggests that the null hypothesis is unlikely to be true.

Sequential evidence and stopping time guarantees. The primary advantage of e-values lies in their behavior regarding stopping times. In a sequential setting, we are given a filtration $(\mathcal{F}_t)_{t\geq 0}$ and construct a sequence of e-values $(E_t)_{t\geq 0}$. If $(E_t)_{t\geq 0}$ is a nonnegative supermartingale with respect to the null distributions (often called a *test martingale*), it allows for *anytime-valid* inference. This property derives from Ville's inequality, a time-uniform generalization of Markov's inequality.

Theorem 2.4 (Ville's inequality)**.** *Let $(E_t)_{t\geq 0}$ be a nonnegative supermartingale with respect to $P \in \mathcal{P}_0$ such that $E_0 \leq 1$. Then for any $\alpha \in (0, 1)$,*

$$P\left(\exists t \geq 0 : E_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

This result guarantees that a researcher can track the e-value process continuously and stop at any data-dependent time τ (a stopping time) while maintaining Type-I error control. Specifically, $\mathbb{E}_P[E_\tau] \leq 1$ holds for any stopping time τ (possibly unbounded), a kind of guarantee which is not available for p-values.

3 Problem Formulation

In this section, we formulate the problem of *Anchored E-Watermarking*. Similar to statistical watermarking, the goal is to embed statistical signals into samples from a target distribution to enable detection. However, in Anchored E-Watermarking, the generator and the detector share access to an anchor distribution that serves as a robust *a priori* estimate of the target distribution. This setting is common in practice; for example, when watermarking Large Language Models or image generators, the target distribution is known to be human language or natural images, respectively. Consequently, watermarking efficacy can be improved by leveraging this structure. Besides, Anchored E-Watermarking uses e-values for detection, thus enjoying improved efficiency in sequential testing. In the following, we introduce the specific components of Anchored E-Watermarking.

Anchor distribution. In the Anchored E-Watermarking framework, we assume that both the generator and the detector share access to an *anchor distribution* $p_0 \in \Delta(\mathcal{V})$, which is in the same probability simplex as the target distribution q. This p_0 serves as the best *a priori* estimate of the target distribution. For instance, in the context of watermarking Large Language Models (LLMs), p_0 could be an open-source, smaller-scale LLM such as Qwen3-8B [46]. Leveraging its role as a known reference, we designate p_0 as the marginal distribution for the watermark signal (or random seed), denoted as s. Therefore, the seed space $\mathcal{S}$ is equal to $\mathcal{V}$ in our framework. Throughout the paper, we assume $0 < \delta \leq \min_{v\in\mathcal{V}} p_0(v)$.[2] It gives $q(v) \geq p_0(v) - \delta/2 \geq \delta/2 > 0$ throughout $\mathcal{Q}(p_0, \delta)$.

Target distribution. The target distribution $q \in \Delta(\mathcal{V})$ represents the desired distribution from which (watermarked) output samples v are generated. Consistent with the definition of the anchor, we premise that q remains sufficiently proximal to p_0. Formally, we assume q resides within the δ-neighborhood of the anchor:

$$\mathcal{Q}(p_0, \delta) = \{q \in \Delta(\mathcal{V}) : \|q - p_0\|_1 \leq \delta\}.$$

Here $\|\cdot\|_1$ is the ℓ_1 distance and the radius $\delta > 0$ represents a robustness tolerance parameter that controls how the true target distribution q can deviate from the anchor. The neighborhood condition depends on the chosen anchor, target, and radius. The measurements in Appendix D.3 assess its practicality on the partition alphabet used in the implementation.

Generator. The objective of the generator is to sample an output v from the target distribution q while embedding a statistical signal s. To achieve this, the generator constructs a joint distribution (or coupling) w over v and s, subject to the constraint that the marginals recover the target and anchor distributions, respectively. The feasible set of valid couplings is defined as:

$$\mathcal{P}(p_0, q) = \left\{ w \in \Delta(\mathcal{V} \times \mathcal{V}) : \sum_{s\in\mathcal{V}} w(v, s) = q(v), \quad \sum_{v\in\mathcal{V}} w(v, s) = p_0(s) \right\}.$$

Note that the generator has access to q, as is common in practice (e.g., when the generator is a service provider). During generation, the system samples the outcome v and the signal s jointly from w. This procedure ensures that the marginal distribution of the outcome v is unbiased (satisfying Definition 2.1), while the watermark signal is embedded within the statistical dependency between v and s.

E-value. In the detection phase, the detector sequentially receives an outcome-signal pair (v, s) and computes an e-value $e(v, s) > 0$. The objective is to distinguish between the following one-step hypotheses:

$$\mathbf{H}_0 : v \sim q \text{ and } s \sim p_0 \text{ are sampled independently,}$$
$$\mathbf{H}_1 : (v, s) \text{ are sampled from the joint coupling } w.$$

The null distribution q is arbitrary; in particular, human text need not lie in $\mathcal{Q}(p_0, \delta)$. To guarantee Type-I error at most α, the e-value must have expectation at most one under every such null:

$$\sup_{q\in\Delta(\mathcal{V})} \mathbb{E}_{v\sim q,\, s\sim p_0}[e(v, s)] \leq 1. \tag{1}$$

[2]Here $\mathcal{V}$ is the finite alphabet on which the watermark is applied. In the partition-based implementation, p_0 and q are distributions over partition cells, and the condition is assessed at that level.

The set $\mathcal{Q}(p_0, \delta)$ is used below for alternative-side power analysis, not for null validity. We let $\mathcal{E}$ denote the set of all nonnegative functions $e : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ satisfying Eq. (1).

The process above describes a single conditional generation step. In an autoregressive model, all distributions are conditional on the history $\mathcal{F}_{t-1}$: the anchor and target become $p_0^t(\cdot) = p_0(\cdot \mid \mathcal{F}_{t-1})$ and $q^t(\cdot) = q(\cdot \mid \mathcal{F}_{t-1})$, and the detector uses the conditional e-value $e_t(v^t, s^t)$ induced by p_0^t. Since $\sum_s p_0^t(s) e_t(v, s) \leq 1$ holds conditionally for every v, the product of one-step e-values is a test supermartingale under $\mathbf{H}_0$.

3.1 Robust log-optimality

While the definition of an e-value ensures validity (safety) under the null, it does not guarantee power (growth) under the alternative. To reject $\mathbf{H}_0$ effectively, we desire the e-value to be large when the data is generated from the alternative distribution $\mathbf{H}_1$. The standard criterion for selecting an e-value is *log-optimality*, often referred to as the "Kelly criterion" in the finance literature [27]. This approach seeks to maximize the expected logarithmic growth rate of evidence against the null, under the alternative. For simple hypotheses where $\mathcal{P}_0 = \{P\}$ and $\mathcal{P}_1 = \{Q\}$, the likelihood ratio $e^* = \frac{dQ}{dP}$ is log-optimal, such that

$$\mathbb{E}_Q[\log e^*] \geq \mathbb{E}_Q[\log e]$$

for any valid e-value e.

In Anchored E-Watermarking, the alternative is defined by the joint coupling w constructed by the generator. However, since the target distribution q lies in a set $\mathcal{Q}(p_0, \delta)$ that is unknown to the detector (i.e., the designer of the e-value), we must optimize the worst-case log-growth rate over arbitrary $q \in \mathcal{Q}(p_0, \delta)$. This gives rise to the following robust log-optimality problem.

Problem 3.1 (Robust log-optimality)**.** An e-value e is *robust log-optimal* in Anchored E-Watermarking if it optimizes the following objective:

$$\sup_{e \in \mathcal{E}} \inf_{q \in \mathcal{Q}(p_0, \delta)} \sup_{w \in \mathcal{P}(p_0, q)} \sum_{v,s \in \mathcal{V}} w(v, s) \cdot \log e(v, s). \tag{2}$$

In this formulation, the objective $\sum w(v, s) \log e(v, s)$ represents the log-growth rate under the alternative. The inner maximization $\sup_{w \in \mathcal{P}(p_0, q)}$ reflects the generator's optimization of the coupling w given knowledge of q; the minimization $\inf_{q \in \mathcal{Q}(p_0, \delta)}$ enforces robustness against the worst-case target distribution in the family; and the outer maximization $\sup_{e \in \mathcal{E}}$ seeks the optimal detector without access to q subject to model-agnosticity (Definition 2.2).

Remark 3.2 (Relationship to growth rate optimality in the worst case (GROW) [18])**.** *GROW studies the worst-case optimal expected capital growth rate under a composite null $\mathcal{P}_0$ and alternative $\mathcal{P}_1$:*

$$\mathrm{GROW}(\mathcal{P}_1) = \sup_{E \in \mathcal{E}(\mathcal{P}_0)} \inf_{\theta \in \mathcal{P}_1} \mathbb{E}_{P_\theta}[\log E],$$

where $\mathcal{E}(\mathcal{P}_0)$ is the set of all valid e-values for the null. If the generator were fixed as a map $q \mapsto w_q$, then Eq. (2) *would reduce to a GROW problem over the induced alternative family $\{w_q : q \in \mathcal{Q}(p_0, \delta)\}$. Our formulation generalizes GROW because the generator is optimized actively through the inner maximization over couplings.*

3.2 Stopping time

While Problem 3.1 addresses the one-step growth rate, it remains to analyze the sample complexity of the framework. In the context of sequential hypothesis testing, sample efficiency is characterized by the stopping time $\tau_\alpha = \inf\{n \in \mathbb{Z}_+ : W_n \geq \log(1/\alpha)\}$ under the alternative, where

$$W_n = \sum_{t=1}^{n} \log e(v^t, s^t),$$

is the accumulated wealth process. This quantity represents the number of samples required to reject the null hypothesis. We consider a stochastic process $\mu(\mathcal{A}, \mathcal{G})$ governed by the interaction between an adversary $\mathcal{A}$ and a generator $\mathcal{G}$. At each step $t \in \mathbb{Z}_+$:

- The adversary $\mathcal{A}$ selects a conditional target distribution $q^t \in \mathcal{Q}(p_0^t, \delta)$ based on the e-value and the history $(v^1, s^1), \ldots, (v^{t-1}, s^{t-1})$;
- The generator $\mathcal{G}$ specifies the joint coupling $w^t \in \mathcal{P}(p_0^t, q^t)$ and draws $(v^t, s^t) \sim w^t$.

Due to the adaptive nature of the adversary, the outcome sequence $v^1, v^2, \ldots$ may be generated autoregressively, conditional on previous outcomes. This formulation extends the i.i.d. setting of Huang et al. [22] to arbitrary dependent distributions. In this setting, we write the stopping time as $\tau_\alpha(e)$ to highlight the dependence on the e-value e. In the context of sequential testing [5, 11, 42, 26, 53], the expected stopping time $\mathbb{E}[\tau_\alpha]$ typically scales linearly with $\log(1/\alpha)$, modulated by an information-theoretic quantity capturing the complexity of the testing problem. Therefore, we are explicitly interested in the following question:

Problem 3.3 (Sample efficiency)**.** For any e-value e satisfying the constraint in Eq. (1), define its *sample complexity* by

$$\mathrm{SC}(e) = \sup_{\mathcal{A}} \inf_{\mathcal{G}} \lim_{\alpha \to 0} \frac{\mathbb{E}_{\mu(\mathcal{A},\mathcal{G})}[\tau_\alpha(e)]}{\log(1/\alpha)}.$$

We are interested in identifying the e-value that minimizes this sample complexity.

This formulation is stronger than the classical expected stopping-time setting because the adversary may alter the conditional target distribution at each time step. Thus samples need not be i.i.d., which is essential for autoregressive language-model watermarking.

4 Theoretical Results

In this section, we answer the questions posed in the previous section. We first derive an optimal detector for the single-step decision problem and then extend this analysis to the sequential setting to characterize the fundamental limit of sample efficiency. Let n denote the cardinality of the sample space $\mathcal{V}$ and $\mathbf{e}_v$ denote the one-hot vector (i.e., Dirac measure) that assigns mass 1 to $v \in \mathcal{V}$.

4.1 Optimal log-growth rate

We begin by solving the robust log-optimality problem in Problem 3.1. The following theorem gives an optimal e-value, its worst-case log-growth rate, and an optimal coupling at each extreme target distribution.

Theorem 4.1 (Log-growth rate)**.** *Let $n = |\mathcal{V}| \geq 2$ and $0 < \delta \leq \min_{v \in \mathcal{V}} p_0(v)$. An optimal e-value for Eq.* (2) *is*

$$e^*(v, s) = \begin{cases} \frac{1-\delta/2}{p_0(s)}, & s = v, \\ \frac{\delta}{2(n-1)p_0(s)}, & s \neq v. \end{cases} \tag{3}$$

For this e-value, the middle infimum is attained precisely at the $q \in \mathcal{Q}(p_0, \delta)$ satisfying $\|q - p_0\|_1 = \delta$. In particular, for every $a \neq b$, the target $q_{a,b} = p_0 + \frac{\delta}{2}(\mathbf{e}_a - \mathbf{e}_b)$ is worst-case, and an optimal coupling for the inner maximization is given by

$$w^*_{a,b} = \mathrm{diag}(p_0) + \frac{\delta}{2}\left(\mathbf{e}_a \mathbf{e}_b^\top - \mathbf{e}_b \mathbf{e}_b^\top\right). \tag{4}$$

The optimal robust log-growth rate is

$$J^* = H(p_0) + \left(1 - \frac{\delta}{2}\right) \log\left(1 - \frac{\delta}{2}\right) + \frac{\delta}{2} \log\left(\frac{\delta}{2(n-1)}\right). \tag{5}$$

The result in Equation (5) fundamentally expresses the optimal worst-case growth rate of e-values in Anchored E-Watermarking. The first term is the entropy of the anchor distribution, which formalizes the intuition that watermarking is easier for higher-entropy anchors. The remaining terms equal $-H(\nu_\delta)$, where $\nu_\delta = (1 - \delta/2, \delta/(2(n-1)), \ldots, \delta/(2(n-1))) \in \Delta([n])$. Thus the rate can be written as $H(p_0) - H(\nu_\delta)$, an entropy-gap analogue of the mutual information of a symmetric additive-noise channel.

Remark 4.2 (Relationship to SEAL [23])**.** *SEAL uses a smaller language model as p_0 and speculative decoding as the generator. For the detector e^*, the generator-side inner problem is to maximize the diagonal mass of a coupling between q and p_0, which is exactly the maximal-coupling principle implemented by speculative decoding, confirming SEAL's choice of generator. On worst-case boundary points of $\mathcal{Q}(p_0, \delta)$ this reduces to the explicit coupling in Eq.* (4)*. Our contribution relative to SEAL is therefore an optimal e-value detector and its anytime-valid sequential analysis.*

4.2 Sample efficiency

Having characterized the one-step optimal growth rate, we now analyze the long-term performance of the watermark with history-dependent anchors. The following theorem connects the average conditional log-growth rate to the sample complexity required to reject the null hypothesis against an adaptive adversary.

Theorem 4.3 (Expected stopping time)**.** *Let $n = |\mathcal{V}| \geq 2$ and let $\mu(\mathcal{A}, \mathcal{G})$ be the stochastic process in Problem 3.3, with $\mathcal{F}_{t-1}$-measurable anchors p_0^t satisfying $p_0^t(v) \geq \delta > 0$ for every $v \in \mathcal{V}$ and $t \geq 1$ almost surely. A predictable e-value strategy $e = (e_t)_{t\geq 1}$ consists of nonnegative $\mathcal{F}_{t-1}$-measurable scores satisfying*

$$\sum_{s\in\mathcal{V}} p_0^t(s) e_t(v, s) \leq 1 \qquad \text{for every } v \in \mathcal{V}.$$

The adversary observes p_0^t and e_t before choosing q^t. Write

$$\tau_\alpha(e) = \inf\left\{ T \geq 1 : \sum_{t=1}^{T} \log e_t(v^t, s^t) \geq \log(1/\alpha) \right\}.$$

Suppose there is a deterministic h such that, under every admissible detector, adversary, and generator,

$$\frac{1}{T}\sum_{t=1}^{T} H(p_0^t) \longrightarrow h \qquad \text{a.s. as } T \to \infty. \tag{6}$$

The value h is common to all strategies, which are fixed as $\alpha \downarrow 0$. Define

$$J_\infty^* = h - H(\nu_\delta) > 0. \tag{7}$$

Then:

- ***Lower bound (converse):*** *For any predictable e-value strategy e,*

$$\sup_{\mathcal{A}} \inf_{\mathcal{G}} \liminf_{\alpha\downarrow 0} \frac{\mathbb{E}_{\mu(\mathcal{A},\mathcal{G})}[\tau_\alpha(e)]}{\log(1/\alpha)} \geq \frac{1}{J_\infty^*}.$$

- ***Upper bound (achievability):*** *Let $e^* = (e_t^*)_{t\geq 1}$, where e_t^* is Eq.* (3) *evaluated at p_0^t. Then*

$$\sup_{\mathcal{A}} \inf_{\mathcal{G}} \limsup_{\alpha\downarrow 0} \frac{\mathbb{E}_{\mu(\mathcal{A},\mathcal{G})}[\tau_\alpha(e^*)]}{\log(1/\alpha)} \leq \frac{1}{J_\infty^*}.$$

Consequently, e^* achieves the optimal first-order sample complexity

$$\inf_{e} \sup_{\mathcal{A}} \inf_{\mathcal{G}} \limsup_{\alpha\downarrow 0} \frac{\mathbb{E}_{\mu(\mathcal{A},\mathcal{G})}[\tau_\alpha(e)]}{\log(1/\alpha)} = \frac{1}{J_\infty^*}.$$

Theorem 4.3 establishes $1/J_\infty^*$ as the fundamental information-theoretic limit of first-order sample complexity for Anchored E-Watermarking under Eq. (6). The anchors may depend on the observed history, and the detector applies the optimal one-step e-value at the current anchor. The entropy condition holds for deterministic anchor sequences whose average entropy converges to h, and allows history-dependent anchors with the same limiting average entropy. For a fixed anchor, $h = H(p_0)$ and $J_\infty^* = J^*$ in Eq. (5).

Remark 4.4 (Relationship with the rates in Huang et al. [22])**.** *Huang et al. [22] shows that the minimum number of samples required to watermark with Type I error α scales as $\log(1/\alpha) \cdot \frac{\log(1/h)}{h}$, where h is the average entropy per token. Our rate $\frac{\log(1/\alpha)}{h - H(\nu_\delta)}$ has the same scaling in terms of α; here h is the average conditional anchor entropy in Eq.* (6)*. The condition $p_0^t(v) \geq \delta > 0$ uniformly lower bounds the anchor entropy, whereas Huang et al. [22] study an asymptotic regime where $h \to 0$.*

5 Experiments

We evaluate the performance of our watermarking scheme in Theorem 4.1 by comparing it with several baselines on a real watermarking task[3].

Setup. We select Llama2-7B-chat [47] with temperature 0.7 as target distribution (i.e., the model to be watermarked) and Phi-3-mini-128k-instruct [33] as the anchor distribution. We evaluate our watermark using the MARKMYWORDS benchmark [35], which is an open-source benchmark designed to evaluate symmetric key watermarking schemes.

MARKMYWORDS generates 300 outputs spanning three tasks—book summarization, creative writing, and news article generation—which mimic potential misuse scenarios. Then, watermarked outputs undergo a set of transformations mimicking realistic user perturbation or attacks: (1) character-level perturbations (contractions, expansions, misspellings and typos); (2) word-level perturbations (random removal, addition, and swap of words in each sentence, replacing words with synonyms); (3) text-level perturbations (paraphrasing, translating the output to another language and back). Detection is run in a sequential test setting with early stopping. For baseline methods, we apply Bonferroni corrections to preserve *anytime valid* Type-I error guarantees: let p_k denote the p-value at time step $k \in \mathbb{Z}_+$, reject at the first time $p_k < \frac{\alpha}{k(k+1)}$.

Results. In this section, we present a comprehensive comparison of our method against these baselines in terms of quality and size. Quality measures the utility of the watermarked text. It is computed using Meta-Llama-3-8B-Instruct [14] with greedy decoding as a judge model, and ranges from zero to one. Size represents the median number of tokens required to detect the watermark at significance level α. All experiments enforce a Type-I error constraint of $\alpha = 0.02$. Lower values indicate higher efficiency. We compare against five state-of-the-art watermarking schemes: "Distribution Shift" [28], "Exponential" [2], "Binary" [10], "Inverse Transform" [29], and "SEAL" [23].

As shown in Table 1, our Anchored E-Watermarking framework significantly outperforms baseline methods in terms of detection efficiency while maintaining high generation quality. Specifically, our method improves over SEAL, confirming the superiority of an optimal e-value. Furthermore, we achieve nearly $2\times$ speed improvement (from 145 to 72.0 tokens) compared to the best non-anchored baseline. Crucially, this efficiency gain does not come at the cost of text quality: the quality score of our method remains competitive with the baselines, demonstrating that E-value-based watermarking offers a superior trade-off between detectability and utility. Due to space constraints, we defer additional experiment results and hyperparameter setups to Appendix D.

Scheme	**Quality** (↑)	**Size** (↓)
Exponential	0.907	∞
Inverse Transform	0.917	734.0
Binary	**0.919**	∞
Distribution Shift	0.912	145.0
SEAL	0.901	84.5
Theorem 4.1	**0.919**	**72.0**

Table 1: Comparison of our e-value-based watermarking scheme in Theorem 4.1 with baselines across quality and size. We report the median over different task prompts and private keys. The best result in each category is highlighted in bold. ∞ size suggests over half of the watermarked generations fail to be detected. E-value-based watermarking demonstrates higher efficiency comparing with all baselines.

6 Discussions

We introduced *Anchored E-Watermarking*, an e-value-based watermarking framework that unifies optimal sampling with anytime-valid inference. Theorem 4.1 characterizes the optimal e-value e^* and the worst-case log-growth rate J^* in closed form, and Theorem 4.3 establishes a matching converse identifying $\log(1/\alpha)/J^*$ as the fundamental sample-complexity limit, even against an adaptive adversary that varies the target distribution within $\mathcal{Q}(p_0, \delta)$. On MARKMYWORDS, the resulting detector reduces the median detection size compared to the strongest baseline while preserving generation quality and Type-I error control.

[3]Code and data are available at `https://github.com/realgourmet/e-watermark`

Broader Impact. Anytime-valid watermark detection strengthens defenses against the misuse of large language models—including misinformation, academic misconduct, and contamination of future training corpora—and lowers the cost of auditing third-party generation services in budget-constrained pipelines. The worst-case Type-I guarantee inherited from Ville's inequality further bounds false accusations against human-written text, providing a principled basis for deployment in high-stakes provenance decisions. The same primitives could nevertheless be repurposed for surveillance or content control if deployed without disclosure or appropriate safeguards.

Limitations and Future Works. Our framework has three principal limitations, each suggesting a direction for future work. First, the log-growth optimality of e^* requires the target to lie in the ℓ_1 ball $\mathcal{Q}(p_0, \delta)$ around a shared anchor; relaxing this to richer divergence balls (KL or Wasserstein) and to online-adapted anchors would broaden applicability. Second, the theory operates on a coarse partition of the vocabulary, leaving an open gap to the full-token regime in which $\inf_v p_0(v)$ may vanish. Third, our analysis targets sample efficiency rather than robustness against semantic attacks; coupling anytime-valid testing with explicit robustness guarantees, and extending the framework beyond text to image and code generation, remain natural next steps.

Acknowledgments

This work was supported in part by the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941. Part of this work was funded by NSF fund CIF-2507523 and ARO fund 062982. The authors would like to thank Ian Waudby-Smith for helpful discussions.

References

[1] Scott Aaronson. My AI safety lecture for UT effective altruism. Shtetl-Optimized, 2022. URL `https://scottaaronson.blog/?p=6823`. Blog post, accessed 2023-09-11.

[2] Scott Aaronson. Watermarking GPT outputs. *Scott Aaronson*, 2022. URL `https://www.scottaaronson.com/talks/watermark.ppt`.

[3] Sahar Abdelnabi and Mario Fritz. Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 121–140. IEEE, 2021.

[4] Adam Block, Ayush Sekhari, and Alexander Rakhlin. Gaussmark: A practical approach for structural watermarking of language models. *arXiv preprint arXiv:2501.13941*, 2025.

[5] Leo Breiman. Optimal gambling systems for favorable games. *Fourth Berkeley Symposium on Probability and Statistics*, pages 65–78, 1961.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

[9] Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. In *Annual International Cryptology Conference*, pages 325–347. Springer, 2024.

[10] Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. *arXiv preprint arXiv:2306.09194*, 2023.

[11] Ben Chugg, Santiago Cortes-Gomez, Bryan Wilder, and Aaditya Ramdas. Auditing fairness by betting. *Advances in Neural Information Processing Systems*, 36:6070–6091, 2023.

[12] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann, Burlington, MA, 2 edition, 2007. ISBN 978-0-12-372585-1.

[13] Debarati Das, Karin De Langis, Anna Martin, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*, 2024.

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[15] Yu Fu, Deyi Xiong, and Yue Dong. Watermarking conditional text generation for ai detection: Unveiling challenges and a semantic-aware watermark remedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18003–18011, 2024.

[16] Noah Golowich and Ankur Moitra. Edit distance robust watermarks for language models. *arXiv preprint arXiv:2406.02633*, 2024.

[17] Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives, 2024.

[18] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

[19] Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation, 2024. URL `https://arxiv.org/abs/2310.03991`.

[20] Abe Bohan Hou, Jingyu Zhang, Yichen Wang, Daniel Khashabi, and Tianxing He. k-semstamp: A clustering-based semantic watermark for detection of machine-generated text. *arXiv preprint arXiv:2402.11399*, 2024.

[21] Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. *arXiv preprint arXiv:2310.10669*, 2023.

[22] Baihe Huang, Hanlin Zhu, Banghua Zhu, Kannan Ramchandran, Michael I Jordan, Jason D Lee, and Jiantao Jiao. Towards optimal statistical watermarking. *arXiv preprint arXiv:2312.07930*, 2023.

[23] Baihe Huang, Hanlin Zhu, Julien Piet, Banghua Zhu, Jason D Lee, Kannan Ramchandran, Michael Jordan, and Jiantao Jiao. Watermarking using semantic-aware speculative sampling: from theory to practice, 2025. URL `https://openreview.net/pdf?id=LdIlnsePNt`.

[24] Adeeb M Jarrah, Yousef Wardat, and Patricia Fidalgo. Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say. *Online Journal of Communication and Media Technologies*, 13(4):e202346, 2023.

[25] Nurul Shamimi Kamaruddin, Amirrudin Kamsin, Lip Yee Por, and Hameedur Rahman. A review of text watermarking: theory, methods, and applications. *IEEE Access*, 6:8011–8028, 2018.

[26] Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

[27] John L Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35 (4):917–926, 1956.

[28] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

[29] Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

[30] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

[31] Xiang Li, Feng Ruan, Huiyuan Wang, Qi Long, and Weijie J Su. A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *arXiv preprint arXiv:2404.01245*, 2024.

[32] Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*, 2024.

[33] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

[34] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023.

[35] Julien Piet, Chawin Sitawarin, Vivian Fang, Norman Mu, and David Wagner. Mark my words: Analyzing and evaluating language model watermarks. *arXiv preprint arXiv:2312.00273*, 2023.

[36] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-sts894.

[37] Jie Ren, Han Xu, Yiding Liu, Yingqian Cui, Shuaiqiang Wang, Dawei Yin, and Jiliang Tang. A robust semantics-based watermark for large language model against paraphrasing, 2024. URL https://arxiv.org/abs/2311.08721.

[38] Stefano Giovanni Rizzo, Flavio Bertini, and Danilo Montesi. Fine-grain watermarking for intellectual property protection. *EURASIP Journal on Information Security*, 2019:1–20, 2019.

[39] Herbert E. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952. URL https://api.semanticscholar.org/CorpusID:15556973.

[40] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021.

[41] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84 – 101, 2011. doi: 10.1214/10-STS347. URL https://doi.org/10.1214/10-STS347.

[42] Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2023.

[43] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[44] Weijie Su, Ruodu Wang, and Zinan Zhao. Online llm watermark detection via e-processes, 2026. URL https://arxiv.org/abs/2602.14286.

[45] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting LLM-generated texts. *arXiv preprint arXiv:2303.07205*, 2023.

[46] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[48] Ashish Venugopal, Jakob Uszkoreit, David Talbot, Franz Och, and Juri Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK., July 2011.

[49] James Vincent. AI-generated answers temporarily banned on coding Q&A site Stack Overflow. *The Verge*, 5, 2022.

[50] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

[51] Vladimir G Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(2):317–341, 1993.

[52] Abraham Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.

[53] Ian Waudby-Smith, Ricardo Sandoval, and Michael I Jordan. Universal log-optimality for general classes of e-processes and sequential hypothesis tests. *arXiv preprint arXiv:2504.02818*, 2025.

[54] Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11613–11621, 2022.

[55] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. *arXiv preprint arXiv:2306.17439*, 2023.

A Proof of Theorem 4.1

Notation. All logarithms are natural. We identify $\mathcal{V}$ with $[n] = \{1, \dots, n\}$ and write $\Delta(\mathcal{V}) = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$; $\mathbf{1}$ is the all-ones vector. For a probability vector x, $H(x) = -\sum_i x_i \log x_i$, with $0 \log 0 = 0$. The binary entropy is $h_2(a) = -a \log a - (1-a)\log(1-a)$ for $0 \le a \le 1$, and the total variation distance is $\mathtt{TV}(x, y) = \frac{1}{2}\sum_i |x_i - y_i|$. We write $x_+ = \max(x, 0)$. A function on $\mathcal{V} \times \mathcal{V}$ is identified with its $n \times n$ matrix of values: $e_{ij} = e(i, j)$, $R_{ij} = r(i, j)$, and $W_{ij} = w(i, j)$. For matrices X, Y we write $\langle X, Y \rangle = \sum_{i,j} X_{ij} Y_{ij}$. In sums such as $\sum_{i,j} W_{ij} \log e_{ij}$ or $\langle W, \log R \rangle$, a term with zero mass contributes zero, whereas positive mass on a zero entry of e or R contributes $-\infty$.

Proposition A.1 (Minimax value and optimal coupling)**.** *Let $n \ge 2$ and $p_0 \in \Delta(\mathcal{V})$. Suppose*

$$0 < \delta \le \min_i p_0(i). \tag{8}$$

Write $p = p_0$, so $p_i = p_0(i)$, and let $t = \delta/2$. Define

$$\begin{aligned} \mathcal{Q} := \mathcal{Q}(p_0, \delta) &= \{q \in \Delta(\mathcal{V}) : \mathtt{TV}(q, p) \le t\}, \\ \mathcal{P}(p_0, q) &= \{W \in \mathbb{R}^{n \times n}_+ : W\mathbf{1} = q,\ W^\top \mathbf{1} = p\}, \\ \mathcal{E} &= \left\{ e \in \mathbb{R}^{n \times n}_+ : \sum_j p_j e_{ij} \le 1 \text{ for every } i \in [n] \right\}. \end{aligned}$$

Rows index the observed symbol and columns index the seed; elements of $\mathcal{E}$ are called detectors. The robust log-growth value is

$$\begin{aligned} J(p, t) &:= \sup_{e \in \mathcal{E}} \inf_{q \in \mathcal{Q}} \sup_{W \in \mathcal{P}(p_0, q)} \sum_{i,j} W_{ij} \log e_{ij} \\ &= H(p) + (1-t)\log(1-t) + t \log \frac{t}{n-1}. \end{aligned} \tag{9}$$

The detector

$$e^*_{ij} = \begin{cases} (1-t)/p_i, & i = j, \\ t/((n-1)p_j), & i \ne j \end{cases} \tag{10}$$

attains the outer supremum. For every $q \in \mathcal{Q}$, the inner supremum for e^* is attained by the coupling $W^{\mathrm{M}}(q)$ below, that is, by the generator that outputs i given the seed j with probability $W^{\mathrm{M}}_{ij}(q)/p_j$; here $\tau = \mathtt{TV}(q, p)$ and

$$W^{\mathrm{M}}_{ij}(q) = \begin{cases} \min(q_i, p_i), & i = j, \\ (q_i - p_i)_+ (p_j - q_j)_+ / \tau, & i \ne j,\ \tau > 0, \\ 0, & i \ne j,\ \tau = 0. \end{cases} \tag{11}$$

For $q = p$ this gives $W^{\mathrm{M}}(p) = \mathrm{diag}(p)$. For e^*, the infimum over $q \in \mathcal{Q}$ in Eq. (9) *is attained exactly at those q with $\mathtt{TV}(q, p) = t$.*

Proof sketch

This proof is computer-assisted and developed through extensive use of an AI assistant (GPT-6 Astra).

Overview. Proposition A.1 computes the value Eq. (9) of a game with three moves: the detector picks e; the adversary picks an alternative q within total variation t of p; and the generator, who wants to be detected, picks a coupling W with row marginal q and column marginal p. Measure probability in units of t, so that the total mass is $L = 1/t$. Since $g(L) = L\, h_2(1/L)$ satisfies $t\, g(1/t) = h_2(t)$, the claimed value reads

$$J(p, t) = H(p) - h_2(t) - t \log(n-1) = H(p) - t\big[g(L) + \log(n-1)\big].$$

Achievability, that e^* attains this value, is a short computation (Step 10). The *converse*, that no detector does better, is the hard part.

The main insight is that the adversary's best move shifts exactly one of the L units of probability to another symbol, and every detector pays twice for this unit. Then $g(L)$ is the least total log-loss of predicting, for each of the L units, whether its seed differs from its symbol, when exactly one of them does; and $\log(n-1)$ is the log-loss of guessing which of the other $n-1$ seeds the shifted unit came from. Step 3 makes this precise.

Steps 1–4 reduce the converse to a lower bound on optimal-transport costs, show that this bound is tight at e^*, and isolate the main difficulty: the optimal transports depend on the detector. Steps 5–6 overcome it by bounding an average of the costs by $\log \kappa$ plus one explicit term per row, whatever the optimal transports are. Steps 7–8 use the fact that every column of a transport has a fixed total mass to show that the row terms add up to at least $g(L)$, first when the largest entries of the rows lie in distinct columns and then in general. The cases are covered as follows:

$n = 2$:	by hand (Step 9),
$n \geq 53,$ every $L > 2n$:	Step 8(a),
$5 \leq n \leq 52,\ L \geq 2048$:	Steps 7 and 8(b),
$n = 3, 4,$ every $L > 2n$:	computer (Step 9),
$5 \leq n \leq 52,\ L \leq 2048$:	computer (Step 9),
boundary cases:	limits (Step 10).

Step 1: Normalizing the detector. A detector is valid when every row has mean at most one, $\sum_j p_j e_{ij} \leq 1$. A row with mean zero is identically zero, and since every alternative gives each symbol probability $q_i \geq p_i - t > 0$, every coupling then has payoff $-\infty$. Otherwise, dividing each row by its mean keeps the detector valid and can only increase the payoff. We may therefore write

$$e_{ij} = \frac{R_{ij}}{p_j}, \qquad R \text{ row stochastic},$$

and call R the *normalized detector*. Since every coupling has column marginal p, the payoff becomes Eq. (40)

$$\langle W, \log e \rangle = H(p) - \langle W, -\log R \rangle .$$

In words, row i of R is the detector's prediction of the seed after it sees the symbol i, and the log-growth is the entropy of the seed minus the log-loss of this prediction Eq. (41). The optimal detector e^* corresponds to the matrix R^* with diagonal entries $1-t$ and off-diagonal entries $t/(n-1)$ Eq. (153).

Step 2: The worst alternative moves mass t between two symbols. Fix R. The payoff is linear in the coupling, and couplings for different alternatives can be mixed: if $W_k \in \mathcal{P}(p_0, q_k)$ and the weights $a_k \geq 0$ sum to one, then $\sum_k a_k W_k \in \mathcal{P}(p_0, \sum_k a_k q_k)$. Since the generator moves after the adversary, its best payoff is therefore a concave function of q, and its minimum over the polytope $\mathcal{Q}$ is attained at an extreme point (Lemma A.7):

$$\mathcal{Q} = \operatorname{conv}\{q_{uv} : u \neq v\}, \qquad q_{uv} = p + t(\mathbf{e}_u - \mathbf{e}_v),$$

where $\mathbf{e}_u$ is the uth coordinate vector. Note that q_{uv} moves probability t from the symbol v to the symbol u, and that it is a distribution because $p_v \geq 2t$. So the adversary has only $n(n-1)$ moves.

Step 3: An optimal-transport inequality. The trick is to count mass in units of t: put $P = p/t$, so that $\sum_i P_i = L = 1/t$. The hypothesis $\delta \leq \min_i p_i$ becomes $P_i \geq 2$; until Step 10 we assume $P_i > 2$, so that $L > 2n$. Dividing a coupling for q_{uv} by t gives a *transport*, that is, an element of

$$\mathcal{X}^{uv}(P) = \{X \geq 0 :\ X\mathbf{1} = P + \mathbf{e}_u - \mathbf{e}_v,\ X^\top \mathbf{1} = P\}.$$

As before, rows index symbols and columns index seeds. Column j holds exactly P_j units in every transport; we call P_j its *capacity*. Row i holds P_i units, except that the *source* u holds one extra unit and the *sink* v one unit less. A unit in cell (i, j) costs the log-loss $-\log R_{ij} \in [0, \infty]$. The generator picks an *optimal* transport, and the adversary picks the most expensive move:

$$C_{uv}(R, P) = \min_{X \in \mathcal{X}^{uv}(P)} \langle X, -\log R \rangle , \qquad K_R(P) = \max_{u \neq v} C_{uv}(R, P).$$

Then $J(p, t) = H(p) - t \inf_R K_R(P)$ Eq. (45). We may assume that all these costs are finite, that is, that every move has a transport avoiding the zero entries of R; otherwise $K_R(P) = \infty$. With

$$g(L) = L \log L - (L-1)\log(L-1) = L\, h_2(1/L), \qquad B_n(L) = g(L) + \log(n-1),$$

we have $t\, B_n(1/t) = h_2(t) + t \log(n-1)$ Eq. (47), so the converse becomes

$$K_R(P) \ \geq \ B_n(L) \qquad \text{for all row-stochastic } R \text{ and all } P \text{ with } P_i > 2.$$

Step 4: From transport costs to entropies. For a nonnegative matrix M with rows M_i and row sums r_i, define, as in Eq. (48),

$$\mathcal{H}(M) = \sum_i r_i \log r_i - \sum_{i,j} M_{ij} \log M_{ij} = \sum_i r_i \, H(M_i/r_i),$$

the entropy of each row, weighted by its mass. Gibbs' inequality, applied row by row, gives $\langle M, -\log R\rangle \geq \mathcal{H}(M)$, with equality when $M = \operatorname{diag}(r)R$. A maximum is at least any average, and the cost is linear in the transport. So every probability distribution (λ_{uv}) on the moves gives, for optimal transports X^{uv} Eq. (50),

$$\begin{aligned} K_R(P) \;&\geq\; \sum_{u\neq v} \lambda_{uv} C_{uv}(R,P) = \Big\langle \sum_{u\neq v} \lambda_{uv} X^{uv}, -\log R \Big\rangle \\ &\geq\; \mathcal{H}\Big(\sum_{u\neq v} \lambda_{uv} X^{uv}\Big). \end{aligned}$$

Draw the source u with probability

$$\omega_u = \frac{P_u - \nu}{L - 1 - \nu}, \qquad \nu = \frac{1}{n-1}$$

(these sum to one because $n\nu = 1 + \nu$), and then the sink uniformly among the other $n-1$ symbols, so that $\lambda_{uv} = \nu\,\omega_u$. Row i gains a unit with probability ω_i and loses one with probability $\nu(1-\omega_i)$, so the mean transport $M = \sum_{u\neq v} \lambda_{uv} X^{uv}$ has row masses Eq. (84)

$$P_i + \omega_i - \nu(1-\omega_i) = L\,\omega_i.$$

Note that each row's mean mass is proportional to its probability of being the source; this is what lets Step 7 compare every row with a single row of mass L. At R^*, the formula for X^{uv} in Step 3 gives $M = \operatorname{diag}(L\omega)R^*$, so both inequalities above are equalities.

The main technical difficulty is that R enters this bound through the optimal transports, which change abruptly as R varies. At R^* the source puts its extra unit directly in the sink's column, so the source row visibly pays $\log(n-1)$. For a general R, the unit may be relayed through several rows, along a path that depends on R and on the sink, and it is unclear which row pays.

Step 5: Optimal transports share a common part. Optimal transports need not be unique, and an arbitrary choice could leave the transports for different sinks unrelated. The trick is to break ties consistently. Fix an order of the entries where $R > 0$. Among the cheapest transports, take the one whose first entry is smallest, among those the one whose second entry is smallest, and so on Eq. (52). This rule chooses every X^{uv}, and also the optimal transport X^0 with both marginals equal to P (no source and no sink), which we call the *central transport*.

Take two chosen transports X and Y whose row sums differ by $X\mathbf{1} - Y\mathbf{1} = \mathbf{e}_a - \mathbf{e}_b$. Then X arises from Y by moving one unit of mass out of row b and into row a along chains that alternate between rows and columns: row b gives up mass in some column, another row takes it there and gives up as much in a second column, and so on until row a (Lemma A.8). No mass moves around a cycle: if $X - Y$ contained a cycle C, then $X - \varepsilon C$ and $Y + \varepsilon C$ would both be feasible, and one of them would improve on a chosen transport, in cost or in the tie-breaking order. Consequently, entrywise, row a only grows and row b only shrinks from Y to X.

Now fix the source u and a row $i \neq u$. Since $X^{uv}\mathbf{1} - X^{ui}\mathbf{1} = \mathbf{e}_i - \mathbf{e}_v$, row i only grows from X^{ui}, the transport in which i itself is the sink, to every other X^{uv} Eq. (54):

$$X_i^{ui} \leq X_i^{uv} \quad \text{entrywise, for every sink } v, \qquad |X_i^{ui}| = P_i - 1,$$

where X_i denotes row i of X and $|\cdot|$ its total mass. We call X_i^{ui} the *common part* of row i; for the source row, the common part is X_u^0, since $X^{uv}\mathbf{1} - X^0\mathbf{1} = \mathbf{e}_u - \mathbf{e}_v$. Whatever a row holds beyond its common part is its *excess*: exactly one unit, except in the sink's row, which has none Eq. (58). In words, the transports with source u differ only in where each row puts its single excess unit. At R^*, every excess unit lies on the diagonal, except that of the source, which lies in the sink's column.

Step 6: Averaging over the sinks. Fix a source u and a set V of κ sinks, and average the costs uniformly over $v \in V$. We want a lower bound with one explicit term per row that holds whatever the optimal transports look like.

Setup. Select a set I of rows and give each $i \in I$ a *main column* $f(i)$, different rows getting different columns, with *main entry* $z_i = R_{i,f(i)} \in (0,1)$; in all applications, the main entry is a largest entry of its row. Averaged over the sinks, let r_i be the mass of row i and m_i its mass outside column $f(i)$. For $i \in I \cap V$, let η_i be the average part of its excess that lies outside column $f(i)$; it is zero when $v = i$, so $\eta_i \le 1 - 1/\kappa$.

The bound. The fixed-source inequality (Lemma A.12) holds for every I. It gives each selected sink a share $(\log\kappa)/\kappa$, and each unselected row a *remainder* Eq. (72), which is nonnegative unless the row is a sink Eq. (75). When every sink is selected, $V \subseteq I$, the shares add up to $\log\kappa$, and dropping the remainders gives

$$\frac{1}{\kappa}\sum_{v\in V} C_{uv}(R,P) \;\ge\; \log\kappa + \sum_{i\in I}\Big\{ r_i \log\frac{1}{z_i} + m_i \log\frac{z_i}{1-z_i}\Big\} - \sum_{i\in V}\chi_\kappa(\eta_i).$$

Step 8 uses the remainders to make the unselected rows pay. The brace of row i is what the row would cost if all its mass outside column $f(i)$ sat in a single entry of R equal to $1 - z_i$, the largest value such an entry can have.

Where does $\log\kappa$ *come from?* It is the entropy of the uniformly random sink, and the transports must pay for it. By the log-sum inequality, the true cost of a row exceeds its brace by at least the entropy of how its mass outside the main column is spread over the columns. To see that these extra entropies add up to $\log\kappa$, let $\mathrm{out}_i(v)$ be the mass of row i outside column $f(i)$ when the sink is v, and $\mathrm{in}_i(v)$ the mass of the other rows in column $f(i)$. Row i holds $P_i + \mathbb{1}_{\{i=u\}} - \mathbb{1}_{\{v=i\}}$ units and column $f(i)$ holds $P_{f(i)}$, both including the main entry, so Eq. (77)

$$\mathrm{in}_i(v) = \mathrm{out}_i(v) - \big(P_i - P_{f(i)} + \mathbb{1}_{\{i=u\}}\big) + \mathbb{1}_{\{v=i\}}.$$

So, up to a constant, the inflow of a row equals its outflow plus a one-unit spike when the row is the sink. How much a quantity $F(v)$ varies with the random sink is measured by the potential Eq. (67)

$$\Phi[F] = \int_0^\infty h\big(\Pr[F(v) > s]\big)\, ds, \qquad h(y) = -y\log y, \qquad v \text{ uniform on } V.$$

It vanishes on constants, is unchanged when a constant is added, and takes the value $h(1/\kappa) = (\log\kappa)/\kappa$ on the spike $\mathbb{1}_{\{v=i\}}$ Eq. (68). The splitting inequality (Lemma A.11) shows that the extra entropies are at least $\sum_{i\in I}\{\Phi[\mathrm{in}_i] - \Phi[\mathrm{out}_i]\}$ Eq. (79), whatever paths the units take. By the identity above, each of the κ sinks contributes its spike, $(\log\kappa)/\kappa$, which gives $\log\kappa$ in total.

The correction χ_κ. A sink's spike can be partly cancelled by the variation of its own outflow. This happens only when part of the row's excess lies outside its main column, and it costs at most $\chi_\kappa(\eta_i)$ Eq. (71), where χ_κ is concave and piecewise linear with

$$\chi_\kappa(j/\kappa) = \frac{g(j+1)}{\kappa} \quad (0 \le j < \kappa), \qquad \chi_\kappa(0) = 0, \qquad \chi_\kappa(1-1/\kappa) = h_2(1/\kappa).$$

Note that the same mass is counted in m_i, at the price $\log\frac{z_i}{1-z_i}$ per unit, which pays for the correction wherever this price exceeds the slope of χ_κ.

From braces to numbers. Two facts are used throughout. First, for $z_i \ge 1/2$ the price $\log\frac{z_i}{1-z_i}$ is nonnegative, so lower bounds on m_i may be inserted, and minimizing over the unknown z_i gives the *binary payment* Eq. (80)

$$\mathcal{A}(r,m) = \inf_{1/2\le z<1}\Big\{ r\log\frac{1}{z} + m\log\frac{z}{1-z}\Big\} = r\, h_2(m/r) \qquad (m \le r/2).$$

It is concave, homogeneous and nondecreasing, and $\mathcal{A}(L,1) = g(L)$: a row of mass L with one unit off its main entry pays at least $g(L)$, the first term of $B_n(L)$, with equality when the main entry is $1 - 1/L$. Second, capacity forces mass out of the main column. Put $d_i = P_i - P_{f(i)}$. Column $f(i)$ holds only $P_{f(i)}$ units, while row i holds $P_i + 1$ units as the source, $P_i - 1$ as the sink and P_i otherwise, so at least

$$(1+d_i)_+ \text{ (source)}, \qquad (d_i - 1)_+ \text{ (sink)}, \qquad (d_i)_+ \text{ (otherwise)}$$

of its units lie outside column $f(i)$.

Step 7: The model case: distinct main columns. Suppose that f is a permutation of $[n]$ and every main entry is at least $1/2$. This case contains R^*, with f the identity, so no estimate may lose anything. Apply Step 6 to every source u, with $I = [n]$ and $V = [n] \setminus \{u\}$, and average over the source drawn from ω (Lemma A.13). Each source gives $\log(n-1)$. Since z_i does not depend on the source, the brace is linear in (r_i, m_i), so averaging replaces r_i by its mean $L\omega_i$ and m_i by its mean. This gives Eq. (82)

$$\begin{aligned} K_R(P) \;&\geq\; \log(n-1) + \sum_i U_i, \\ U_i &= \min_{\eta}\Big\{\mathcal{A}\big(L\omega_i, \underline{m}_i(\eta)\big) - (1-\omega_i)\chi_{n-1}(\eta)\Big\}, \\ \underline{m}_i(\eta) &= \omega_i(1+d_i)_+ + (1-\omega_i)\big[(d_i-1)_+ + \eta\big]. \end{aligned}$$

Here $\underline{m}_i(\eta)$ is a lower bound on the mean mass of row i outside its main column. When i is the source, this mass is at least $(1+d_i)_+$. Otherwise it consists of at least $(d_i-1)_+$ in the common part plus the mean excess η outside the main column, and $\eta \geq (1-\nu)\min\{(d_i)_+, 1\}$ by the capacity bounds of Step 6 Eq. (81).

The trick is homogeneity. Row i is an ω_i-scaled copy of a row of mass L, and concavity of $\mathcal{A}$ gives the chord bound

$$\mathcal{A}(L\omega_i, \omega_i x) = \omega_i\, \mathcal{A}(L, x) \;\geq\; \omega_i\, x\, g(L) \qquad (0 \leq x \leq 1).$$

With probability ω_i the row is the source and must put $x = (1+d_i)_+$ units outside its main column. This suggests the *row bound* Eq. (128)

$$U_i \;\geq\; g(L)\min\{1,\; \omega_i(1+d_i)_+\}.$$

The cap at one is needed because $\mathcal{A}(L,x) \leq x\, g(L)$ for $x \geq 1$, so a row cannot be charged $g(L)$ for each of several units. It is harmless because the coefficients only need to add up to one, and they do. Summing the identity

$$2(P_i - \nu)d_i = (P_i - \nu)^2 - (P_{f(i)} - \nu)^2 + d_i^2$$

over i, the squares cancel because f is a permutation; dividing by $2(L-1-\nu)$ and using $\sum_i \omega_i = 1$ gives

$$\sum_i \omega_i\big(1 + P_i - P_{f(i)}\big) = 1 + \frac{\sum_i (P_i - P_{f(i)})^2}{2(L-1-\nu)} \;\geq\; 1.$$

Since $(1+d_i)_+ \geq 1+d_i$ and $\sum_i \min\{1, y_i\} \geq \min\{1, \sum_i y_i\}$ for $y_i \geq 0$, the row bounds, wherever they hold, add up to at least $g(L)$, and then $K_R(P) \geq B_n(L)$.

The main technical difficulty is the row bound when $d_i > 0$. The row must then pay more than its share $\omega_i g(L)$, after a correction of up to $(1-\omega_i)h_2(\nu)$ for the sink entropy, and more excess increases both the payment and the correction. Lemma A.19 proves the row bound for $n \geq 5$, $P_i \geq 2$ and $L \geq 2048$ when every main entry is at least $3/5$, so that $\mathcal{A}$ may be minimized over $z \geq 3/5$ only.

Why $3/5$*?* A unit off such an entry costs at least $\log \frac{3/5}{2/5} = \log \frac{3}{2}$ more than a unit on it, which covers the slopes of χ_{n-1} from its third piece on Eq. (91). On the first two pieces, of slopes $\log 4$ and less than $\log 2$, the row is nearly pure (less than a fifth, respectively a third, of its mass lies off its main entry), so an extra unit off the main entry costs more than $\log 4$, respectively $\log 2$. Hence more excess never lowers the bound, the least excess allowed by capacity is the worst case, and for $d_i \leq 0$ the chord bound then gives the row bound. With main entries of only $1/2$, the extra price $\log \frac{z_i}{1-z_i}$ could vanish.

Why 2048*?* The remaining one-variable inequalities, in particular at the point $\omega_i(1+d_i) = 1$ where the cap becomes active Eq. (130), are checked with numerical bounds that hold for $L \geq 2048$ Eq. (129). The lemma is used only for $5 \leq n \leq 52$; for $n \geq 53$, Step 8(a) covers permutations with the threshold $5/8$ and no condition on L beyond $L > 2n$.

Step 8: When main columns collide. In general, let the main column of a row be the column of its largest entry. Several rows may share a main column, and some rows have no large entry at all. This is the main difficulty: Step 6 can select only one row per main column, because its flow argument

relabels the columns so that the main entries lie on the diagonal, and a row that is left out may pay almost nothing if its largest entry is close to one. The trick is capacity again: rows that share a main column also share its capacity, so their overflow must lie outside their main entries, where it is paid for.

(a) Dimensions $n \geq 53$ (Appendix A.7). Call a row *high* if its largest entry exceeds $5/8$. In each main column, select the high row with the largest marginal P_i; let I be the set of selected rows, and call all other rows *omitted*. Average Step 6 over all sources, drawn from ω, and all $n-1$ sinks. The rows then pay as follows (Lemma A.14).

- Omitted rows are sinks that the flow argument does not count, so their shares $(\log(n-1))/(n-1)$ of $\log(n-1)$ are missing. We add these shares anyway and charge them to the omitted rows, which costs row i at most $(1-\omega_i)h_2(\nu)$.
- A row that is not high has all entries at most $5/8$, so each unit costs at least $\log\frac{8}{5}$; the row pays this on all but at most one unit of its marginal Eq. (89).
- A unit off a main entry above $5/8$ costs at least $\log\frac{5/8}{3/8} = \log\frac{5}{3} \approx 0.51$ more than a unit on it. We keep $\frac{11}{25} = 0.44$ of this extra price as a credit: omitted high rows pay $\frac{11}{25}$ per unit off their main entry Eq. (90).
- In a selected row, the rest of the extra price, $\log\frac{5}{3} - \frac{11}{25} > 0.07$ per unit, still exceeds the slopes of χ_{n-1} beyond its fourteenth piece, and on the first fourteen pieces the row is nearly pure, as in Step 7. So the analogue of the row bound for the threshold $5/8$, which uses $L > 2n \geq 106$ and $\nu \leq 1/52$, shows that a selected row pays at least Eq. (92)

  $$g(L)\,c_i + \tfrac{11}{25}(\bar{m}_i - c_i), \qquad c_i = \min\{1, \omega_i(1+d_i)_+\},$$

  where $\bar{m}_i \geq c_i$ is its mean mass outside column $f(i)$.

The selected rows thus supply $g(L)(1-\Delta)$, where $\Delta = (1 - \sum_{i\in I} c_i)_+$ is the *missing coefficient*, and the other payments must make up for $g(L)\Delta$. Summed over I only, the identity of Step 7 no longer cancels completely; it keeps the squares of the omitted rows and of the columns outside $f(I)$ Eq. (123):

$$(L-1-\nu)\Big(1 - \sum_{i\in I}\omega_i(1+d_i)\Big) = \sum_{i\notin I}(P_i-\nu) + \frac{1}{2}\sum_{i\notin I}(P_i-\nu)^2 - \frac{1}{2}\sum_{j\notin f(I)}(P_j-\nu)^2 - \frac{1}{2}\sum_{i\in I} d_i^2.$$

Distinct main columns. If no two high rows share a main column, every omitted row is not high, and the omitted rows can be paired with the columns outside $f(I)$. In the identity, an omitted row adds $(P_i-\nu) + \frac{1}{2}(P_i-\nu)^2$, and its column, of marginal above two, subtracts more than $\frac{1}{2}(2-\nu)^2$. Hence Eq. (124)

$$\Delta \;\leq\; \min\Big\{1,\; \sum_{i\notin I} \frac{(P_i-1)^2/2 + (2-\nu)(P_i-1) - 1/2}{L-1-\nu}\Big\},$$

and the price $\log\frac{8}{5}$ per unit lets each omitted row pay $g(L)$ times its term in this sum, capped at one Eq. (97). Together, the omitted rows pay at least $g(L)\Delta$.

Repeated main columns. Let Z be the mass that high rows place outside their main entries in the mean transport M, plus the marginals of the rows that are not high:

$$Z = \sum_{i \text{ high}} \sum_{j\neq f(i)} M_{ij} + \sum_{i \text{ not high}} P_i.$$

Capacity forces Z to be large: each omitted row leaves a column that is the main column of no high row, and this column needs more than two units, which essentially only mass counted in Z can supply. Hence $\sum_{i\notin I}(1-\omega_i) \leq Z/2$, so the charges for omitted sinks cost at most $\frac{1}{2}h_2(\nu)$ per unit of Z. Since each unit of Z earns $\frac{11}{25}$, up to bounded corrections, collecting all payments gives Eq. (127)

$$K_R(P) \;\geq\; \log(n-1) + g(L)(1-\Delta) + aZ - \tfrac{11}{25} - k\big(\tfrac{22}{25} - \log\tfrac{8}{5}\big),$$

with $a = \frac{11}{25} - \frac{1}{2}h_2(\nu) > 0.39$ and k the number of rows that are not high. It remains to show that the loss $g(L)\Delta$ is smaller than the gain $aZ - \frac{11}{25} - k(\frac{22}{25} - \log\frac{8}{5})$.

Why is Δ small? Consider the graph with an edge $i \to f(i)$ from each high row to its main column, identifying column j with row j. Each component is a tree or a cycle with trees leading into it. Since $L - 1 - \nu = \sum_j (P_j - \nu)$, the quantity $(L - 1 - \nu)\big(1 - \sum_{i\in I}\omega_i(1 + d_i)_+\big)$, which bounds $(L - 1 - \nu)\Delta$, splits into one term per column. For a main column j with selected row i, this term is at most the increase of $x^2/2 + (1 - \nu)x$ from $x = P_i$ to $x = P_j$ Eq. (101), which is smaller when P_i is larger. This is why the row with the largest marginal is selected: on a cycle, P_i is then at least the marginal of the predecessor of j on the cycle, and the increases telescope to at most zero around the cycle. On the trees, the terms add up to at most a quadratic in the *spare capacity*, the total amount by which columns exceed the marginals of the high rows whose main column they are; capacity bounds this by Z plus less than one unit. Together with a cruder linear count, this gives (Lemma A.16)

$$\Delta \;\le\; \min\Big\{1,\; \frac{Z^2/2 + (2 - \nu)Z - 1/2}{L - 1 - \nu},\; \frac{L - 2n + (2 - \nu)Z}{L - 1 - \nu}\Big\}.$$

Each of the three bounds is needed: the quadratic fraction controls small Z, since $g(L)/(L-1-\nu) < 0.054$; the linear one controls L close to $2n$; and the cap controls large Z. Maximized over L, the resulting bound on the loss $g(L)\Delta$ grows more slowly in Z than the gain aZ, so it suffices to check the smallest admissible Z, where the margins are thin (Lemmas A.17 and A.18).

Why $n \ge 53$? The constants are evaluated at $n = 53$: $\nu \le 1/52$ keeps the charges $h_2(\nu)$ small and the first fourteen pieces of χ_{n-1} short, and $L > 2n \ge 106$ keeps $g(L)/(L - 1 - \nu)$ small. No computer and no condition on L beyond $L > 2n$ are needed.

(b) Dimensions $5 \le n \le 52$ with $L \ge 2048$ (Appendix A.9). Now call a row *high* if its largest entry exceeds $3/5$, and *low* otherwise. If all rows are high with distinct main columns, Step 7 applies, and if there is at most one main column, a direct argument works (Lemma A.24). Otherwise, let $\kappa + 1$ be the number of main columns, and keep one high row per main column, called the *representatives*. They are chosen by a rule on the graph of (a): on a cycle, the predecessor on the cycle; off the cycles, the row whose tree carries the most mass, that is, its marginal plus the overflow of the columns in its tree (Lemma A.21). Sources and sinks are both taken among the representatives (Lemma A.20), so every sink is selected, but Step 6 now yields only $\log\kappa$, a loss of $\log\frac{n-1}{\kappa}$. Each source gives its own lower bound on $K_R(P)$, and the largest one is used.

On the set I of representatives, the sum-of-squares identity survives up to one negative term: for every shift β Eq. (139),

$$\sum_{i\in I}(P_i - \beta)(P_i - P_{f(i)}) = \frac{1}{2}\sum_{i\in I}(P_i - P_{f(i)})^2 + \frac{1}{2}\sum_{j\notin f(I)}(P_j - \beta)^2 - E,$$
$$E = \frac{1}{2}\sum_{i\notin I}(P_i - \beta)^2.$$

Hence $\sum_{i\in I}(P_i - \beta)(1 + d_i)_+ \ge \sum_{i\in I}(P_i - \beta) - E$ Eq. (138). In words, with the representatives weighted by $P_i - \beta$, the number of units that a representative must put outside its main column when it is the source is at least one on average, up to the error E divided by the total weight. The shift is $\beta = \frac{1}{\kappa} + \min\{\frac{3}{2}, \frac{10}{\kappa}\} \le 2$ for $\kappa \ge 2$; its second term absorbs the correction χ_κ of each sink.

The rows outside I are paid for by *credits*, costs not used so far: $\frac{31}{100}$ per unit that high rows place outside their main entries (when main columns repeat), and $\log\frac{5}{3}$ per unit of a low row, except one unit per row. Capacity forces credited mass to exist: each column that is the main column of no high row needs more than two units, which only low rows or mass outside main entries can supply, and an overloaded main column pushes mass out of the main entries Eq. (133). This bounds both E and the total marginal b of the rows outside I in terms of the credited mass (Lemma A.21). Finally, Lemma A.22, which extends $\mathcal{A}(L, 1) = g(L)$ to representatives with different main entries, gives Eq. (143)

$$K_R(P) - B_n(L) \;\ge\; \text{credits} - \log\frac{n-1}{\kappa} - \xi - \frac{b + 17 + E\,g(L)}{L - b - 18}.$$

Here $\xi = \frac{31}{100}$ when main columns repeat and $\xi = 0$ otherwise; it is lost because the source's forced overflow is already among the credits. The constant 17 bounds the total shift $(\kappa + 1)\beta$. Note that the

credits grow linearly in the credited mass, while the loss $\log \frac{n-1}{\kappa}$ grows only logarithmically and the last term is of order $(b + E \log L)/L$.

Why $L \geq 2048$? The last term must be small. The comparison is checked at $L = 2048$ in six cases, some with thin margins, and extends to larger L by monotonicity (Lemma A.23).

Step 9: The remaining cases. *Dimension two, by hand* (Lemma A.10). Here $B_2(L) = g(L)$, and the difficulty of Step 4 disappears. A 2×2 transport with given marginals is determined by its $(1,1)$ entry, and its cost is linear in that entry, with slope $\log \frac{R_{12}R_{21}}{R_{11}R_{22}}$. So the optimal transports are the most or the least diagonal ones, depending only on the sign of $R_{11}R_{22} - R_{12}R_{21}$.

If $R_{11}R_{22} \geq R_{12}R_{21}$, they are the most diagonal ones,

$$X^{12} = \begin{pmatrix} P_1 & 1 \\ 0 & P_2 - 1 \end{pmatrix}, \qquad X^{21} = \begin{pmatrix} P_1 - 1 & 0 \\ 1 & P_2 \end{pmatrix},$$

and we mix them with the source probabilities of Step 4 for $n = 2$, namely $\nu = 1$ and $\omega_u = \frac{P_u - 1}{L-2}$. Each row of the mixture N then has exactly the fraction $1/L$ of its mass off the diagonal, as for R^*, so Step 4 gives $K_R(P) \geq \mathcal{H}(N) = L\, h_2(1/L) = g(L)$.

If $R_{11}R_{22} < R_{12}R_{21}$, the optimal transports are the least diagonal ones, but capacity still keeps mass on the diagonal. After swapping both rows and both columns if necessary, which preserves this case and $K_R(P)$, we have $d = P_1 - P_2 \geq 0$. With source 1, row 1 holds $P_2 + d + 1$ units and column 2 only P_2, so at least $1 + d$ units stay in cell $(1,1)$. For $d \geq 1$, row 1 alone already costs more than $g(L)$. The delicate case is $0 \leq d < 1$: at $d = 0$ the detector R^* with its two columns swapped is also optimal (see the remark on nonuniqueness at the end of Appendix A.11), so there is no slack. Here a mixture of the two optimal transports with weights $\frac{1+d}{2}$ and $\frac{1-d}{2}$ closes the case, through monotonicity in the smaller marginal and convexity in d Eq. (65).

The computer-assisted cases (Proposition A.25) are $n = 3, 4$ for every L, including $L \to \infty$, and $5 \leq n \leq 52$ with $L \leq 2048$. Their proof is outlined in Appendix A.10. It works at a minimizer of $K_R(P)$ (in some dimensions, of an average of costs that is at most $K_R(P)$), and four ideas reduce a continuum of cases to a finite check.

1. *Stationarity.* At a minimizer there is a distribution λ on the moves for which $N = \sum_{u \neq v} \lambda_{uv} X^{uv}$ has rows proportional to those of R (Lemma A.8). So Gibbs' inequality in Step 4 is an equality, and the minimal value is $\mathcal{H}(N)$ Eq. (57).
2. *Finitely many shapes.* The chosen transports are vertices of transportation polytopes, so their supports are forests, and tX^{uv} is affine in $(P/L, 1/L)$ on each of finitely many regions.
3. *Compactness.* In the variable $t = 1/L$ the domain is compact, with $L = \infty$ at $t = 0$ Eq. (145).
4. *Exact arithmetic.* Concavity of $\mathcal{H}$, with an explicit margin for the curvature of the target $t\, B_n(1/t)$ Eq. (146), leaves finitely many comparisons per region. They are done in exact rational arithmetic, with logarithms enclosed by Lemma A.9.

For $5 \leq n \leq 52$, a crude lower bound on $K_R(P)$ limits how many rows can be low or share a main column Eq. (147), and the permutation case, which contains R^*, uses the exact cancellation of Step 7 instead of slack. Dimension six uses dimension five, which is completed first for all L (for $L \geq 2048$ by Steps 7 and 8(b)).

Step 10: Boundary cases and achievability. Steps 3–9 assume $P_i > 2$, that is, $\delta < \min_i p_i$, and parts of them assume $R > 0$; limits remove both restrictions (Lemma A.26). For positive R, adding ε to each diagonal entry of an optimal transport gives Eq. (150)

$$K_R(P + \varepsilon \mathbf{1}) \;\leq\; K_R(P) + \varepsilon \sum_i (-\log R_{ii}),$$

so the bound passes from $P + \varepsilon\mathbf{1}$ to P as $\varepsilon \downarrow 0$; this covers $P_i = 2$. Zero entries are removed by smoothing: for a positive row-stochastic Q, the matrix $R_\varepsilon = (1 - \varepsilon)R + \varepsilon Q$ is at least $(1 - \varepsilon)R$ entrywise, and every transport has mass L, so Eq. (63)

$$K_R(P) \;\geq\; K_{R_\varepsilon}(P) + L \log(1 - \varepsilon),$$

and a lower bound that does not depend on R passes to all R as $\varepsilon \downarrow 0$.

Achievability is easy because R^* takes only two values, so the payoff of a coupling depends only on its diagonal mass Eq. (154):

$$\langle W, \log R^* \rangle = \Big(\sum_i W_{ii}\Big) \log(1-t) + \Big(1 - \sum_i W_{ii}\Big) \log \frac{t}{n-1}.$$

This increases with the diagonal mass, since $t \le 1/(2n)$ gives $1 - t > t/(n-1)$. As $W_{ii} \le \min(q_i, p_i)$, the diagonal mass is at most $1 - \mathtt{TV}(q,p)$, with equality for the maximal coupling Eq. (11), which is therefore the generator's best response to every q. The resulting payoff decreases strictly in $\mathtt{TV}(q,p)$, so the worst alternatives are exactly those with $\mathtt{TV}(q,p) = t$, and there the value equals the converse bound Eq. (157).

Roadmap. Appendix A.1 proves Theorem 4.1 under the extra assumption of cycle-diagonal dominance: for every cycle of distinct symbols, the product of the diagonal detector values dominates the product of the values shifted along the cycle. Appendix A.2 gives the reduction; Appendix A.3 the entropy bounds, the lexicographic choice, common parts, stationarity and smoothing; Appendix A.4 rational bounds for logarithms, used for numerical constants; Appendix A.5 the case $n = 2$; Appendix A.6 the averaging inequalities used in the next three subsections, which treat $n \ge 53$ (Appendix A.7), complete matchings with $L \ge 2048$ (Appendix A.8), and $5 \le n \le 52$ with $L \ge 2048$ (Appendix A.9); Appendix A.10 the computer-assisted cases; and Appendix A.11 the assembly, the boundary cases, the proof of Proposition A.1 and Theorem 4.1, and nonuniqueness.

A.1 Proof under the cycle-diagonal-dominance assumption

We first give a proof of Theorem 4.1 under the cycle-diagonal-dominance assumption that $v_0, \dots, v_{k-1}$ with $v_k = v_0$: $\prod_{i=0}^{k-1} e(v_i, v_i) \ge \prod_{i=0}^{k-1} e(v_i, v_{i+1})$ for every directed cycle of distinct vertices $v_0, \dots, v_{k-1}$, where $2 \le k \le n$ and $v_k = v_0$. Equivalently, around any cycle, the detector should collectively favor seed tokens that match the actual tokens over the corresponding cyclic mismatch. Throughout this subsection, every detector e is positive, $e(v,s) > 0$, and satisfies this assumption; both properties are invariant under rescaling the rows of e. Accordingly, every supremum over r below, including the one over $\mathcal{R}$ and the one in Proposition A.5, ranges over matrices $r(v,s) = p_0(s)e(v,s)$ with such e; these satisfy $r(v,s) > 0$ and $\prod_{i=0}^{k-1} r(v_i, v_i) \ge \prod_{i=0}^{k-1} r(v_i, v_{i+1})$ for every directed cycle, since the factors $p_0(v_i)$ cancel.

Proof. Let

$$\mathcal{R} = \left\{ r \in \mathbb{R}^{n \times n} : \sum_{s \in \mathcal{V}} r(v,s) = 1,\ r(v,s) \ge 0,\ \forall v, s \in \mathcal{V} \right\}.$$

By Lemma A.6, the original problem is equivalent to

$$J' : \sup_{r \in \mathcal{R}} \inf_{q \in \mathcal{Q}(p_0, \delta)} \sup_{w \in \mathcal{P}(p_0, q)} \sum_{v,s \in \mathcal{V}} w(v,s) \cdot (\log r(v,s) - \log p_0(s)),$$

with $r(v,s) = p_0(s)e(v,s) \in \mathcal{R}$. Note that

$$\sum_{v,s \in \mathcal{V}} w(v,s) \cdot (-\log p_0(s)) = \sum_{s \in \mathcal{V}} p_0(s) \cdot (-\log p_0(s)) = H(p_0)$$

where $H(p_0)$ is the entropy of p_0, for every $w \in \mathcal{P}(p_0, q)$. Hence the optimum of J' is $H(p_0)$ plus the supremum below. By Proposition A.5,

$$\sup_{r \in \mathcal{R}} \inf_{q \in \mathcal{Q}(p_0, \delta)} \sup_{w \in \mathcal{P}(p_0, q)} \sum_{v,s \in \mathcal{V}} w(v,s) \cdot \log r(v,s) = (1 - \tfrac{\delta}{2}) \log\left(1 - \frac{\delta}{2}\right) + \tfrac{\delta}{2} \log\left(\frac{\delta}{2(n-1)}\right),$$

achieved at

$$r^*(v,s) = \begin{cases} 1 - \delta/2, & s = v, \\ \delta/(2n-2), & s \ne v, \end{cases}$$

and for any $q \in \mathcal{Q}(p_0, \delta)$ written as $q = p_0 + \frac{\delta}{2} \cdot (\mathbf{e}_{v_i} - \mathbf{e}_{s_i})$, the optimizer of the inner problem is given by $\frac{\delta}{2} \cdot (\mathbf{e}_{v_i}\mathbf{e}_{s_i}^\top - \mathbf{e}_{s_i}\mathbf{e}_{s_i}^\top) + \mathrm{diag}(p_0)$. Therefore, the optimum value of the original problem is equal to

$$(1 - \tfrac{\delta}{2}) \log\left(1 - \frac{\delta}{2}\right) + \tfrac{\delta}{2} \log\left(\frac{\delta}{2(n-1)}\right) + H(p_0).$$

In particular, this is achieved at

$$e^*(v, s) = \begin{cases} \frac{1-\delta/2}{p_0(s)}, & s = v, \\ \frac{\delta}{2(n-1)p_0(s)}, & s \neq v, \end{cases}$$

For this e-value, the boundary couplings above attain the worst-case value, and interior targets are handled by maximal couplings as stated in Theorem 4.1. This completes the proof. □

A.1.1 Supporting lemma

Lemma A.2 (Optimizer of Inner Problem)**.** *Let $\mathcal{V}$ be the set $\{1, \ldots, n\}$ and $p_0 \in \Delta(\mathcal{V})$ be a distribution over $\mathcal{V}$ such that $\inf_{v \in \mathcal{V}} p_0(v) \geq \delta$. Let $M : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ be a matrix that satisfies*

$$\sum_{i=0}^{\kappa_v - 1} M(\sigma^i(v), \sigma^i(v)) \geq \sum_{i=0}^{\kappa_v - 1} M(\sigma^i(v), \sigma^{i+1}(v)) \tag{12}$$

for any $v \in \mathcal{V}$, $\sigma \in S_n$, where $\kappa_v = \inf\{k \geq 1 : \sigma^k(v) = v\}$. Fix distinct $a, b \in \mathcal{V}$ and define

$$q = p_0 + \frac{\delta}{2} \cdot (\mathbf{e}_a - \mathbf{e}_b).$$

Consider the optimization problem

$$\sup_{w \in \mathcal{P}(p_0, q)} J(w), \qquad J(w) := \sum_{v,s \in \mathcal{V}} w(v, s) M(v, s),$$

where

$$\mathcal{P}(p_0, q) = \left\{ w : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+ \Big| \sum_{s \in \mathcal{V}} w(v, s) = q(v) \; \forall v \in \mathcal{V}, \; \sum_{v \in \mathcal{V}} w(v, s) = p_0(s) \; \forall s \in \mathcal{V} \right\}.$$

Then there exists a permutation $\sigma \in S_n$ such that

$$w^* = \frac{\delta}{2} \sum_{i=0}^{\kappa - 2} \left(\mathbf{e}_{\sigma^i(a)} \mathbf{e}_{\sigma^{i+1}(a)}^\top - \mathbf{e}_{\sigma^{i+1}(a)} \mathbf{e}_{\sigma^{i+1}(a)}^\top \right) + \mathrm{diag}(p_0) \tag{13}$$

is an optimizer of J, where κ is the minimum positive integer such that $\sigma^\kappa(a) = a$.

Proof. First, we establish existence of an optimizer. Define w^{feas} by

$$\begin{aligned} w^{\mathrm{feas}}(a, b) &= \frac{\delta}{2}, \\ w^{\mathrm{feas}}(v, v) &= \begin{cases} p_0(b) - \frac{\delta}{2}, & v = b, \\ p_0(v), & v \neq b, \end{cases} \\ w^{\mathrm{feas}}(v, s) &= 0, \quad (v, s) \notin \{(a, b)\} \cup \{(v, v) : v \in \mathcal{V}\}. \end{aligned}$$

Because $\inf_v p_0(v) \geq \delta > \delta/2$, we have $w^{\mathrm{feas}} \geq 0$. Its row sums satisfy

$$\begin{aligned} \sum_s w^{\mathrm{feas}}(a, s) &= p_0(a) + \frac{\delta}{2} = q(a), \\ \sum_s w^{\mathrm{feas}}(b, s) &= p_0(b) - \frac{\delta}{2} = q(b), \\ \sum_s w^{\mathrm{feas}}(v, s) &= p_0(v) = q(v), \quad v \notin \{a, b\}, \end{aligned}$$

and its column sums satisfy

$$\sum_v w^{\mathrm{feas}}(v,b) = p_0(b) - \frac{\delta}{2} + \frac{\delta}{2} = p_0(b),$$
$$\sum_v w^{\mathrm{feas}}(v,s) = p_0(s), \quad s \neq b.$$

Thus $w^{\mathrm{feas}} \in \mathcal{P}(p_0, q)$. The set $\mathcal{P}(p_0, q)$ is a nonempty bounded polytope, and J is linear, so there exists

$$w^{(0)} \in \arg \sup_{p \in \mathcal{P}(p_0,q)} J(w).$$

For a feasible w, let G_w be the directed graph on vertex set $\mathcal{V}$ with an edge (v, s) whenever $v \neq s$ and $w(v, s) > 0$. A directed cycle in G_w is a k-tuple $(v_0, \ldots, v_{k-1})$ of distinct vertices, $k \geq 2$, such that

$$w(v_i, v_{i+1}) > 0, \quad i = 0, \ldots, k-1,$$

with indices understood modulo k (so $v_k = v_0$).

Next, we show that for an optimal distribution $\bar{w}$, the graph $G_{\bar{w}}$ has no directed cycles. Fix any feasible w and such a cycle $(v_0, \ldots, v_{k-1})$. For $\varepsilon > 0$ define $\tilde{w}$ by

$$\tilde{w}(v_i, v_{i+1}) = w(v_i, v_{i+1}) - \varepsilon, \qquad i = 0, \ldots, k-1, \tag{14}$$
$$\tilde{w}(v_i, v_i) = w(v_i, v_i) + \varepsilon, \qquad i = 0, \ldots, k-1, \tag{15}$$
$$\tilde{w}(x, y) = w(x, y), \qquad (x, y) \notin \{(v_i, v_{i+1}), (v_i, v_i) : 0 \leq i \leq k-1\}.$$

Choose

$$0 < \varepsilon \leq \inf_{0 \leq i \leq k-1} w(v_i, v_{i+1}),$$

so that $\tilde{w} \geq 0$.

Row sums. For each i,

$$\begin{aligned}\sum_{s \in \mathcal{V}} \tilde{w}(v_i, s) &= (w(v_i, v_{i+1}) - \varepsilon) + (w(v_i, v_i) + \varepsilon) + \sum_{s \notin \{v_i, v_{i+1}\}} w(v_i, s) \\ &= \sum_{s \in \mathcal{V}} w(v_i, s).\end{aligned}$$

All other rows are unchanged, so

$$\sum_s \tilde{w}(v, s) = \sum_s w(v, s) = q(v), \qquad \forall v \in \mathcal{V}. \tag{16}$$

Column sums. For each i,

$$\begin{aligned}\sum_{x \in \mathcal{V}} \tilde{w}(x, v_i) &= (w(v_{i-1}, v_i) - \varepsilon) + (w(v_i, v_i) + \varepsilon) + \sum_{x \notin \{v_{i-1}, v_i\}} w(x, v_i) \\ &= \sum_{x \in \mathcal{V}} w(x, v_i),\end{aligned}$$

where again indices are taken modulo k. All other columns are unchanged, so

$$\sum_v \tilde{w}(v, s) = \sum_v w(v, s) = p_0(s), \qquad \forall s \in \mathcal{V}. \tag{17}$$

Thus $\tilde{w} \in \mathcal{P}(p_0, q)$.

Objective value. Only entries on the cycle and the corresponding diagonals change, hence

$$\begin{aligned}J(\tilde{w}) - J(w) = \sum_{i=0}^{k-1} \Big[&\tilde{w}(v_i, v_i) M(v_i, v_i) + \tilde{w}(v_i, v_{i+1}) M(v_i, v_{i+1}) \\ &- w(v_i, v_i) M(v_i, v_i) - w(v_i, v_{i+1}) M(v_i, v_{i+1}) \Big]\end{aligned} \tag{18}$$

$$= \varepsilon \sum_{i=0}^{k-1} \left(M(v_i, v_i) - M(v_i, v_{i+1}) \right). \tag{19}$$

Let $\sigma \in S_n$ be the permutation whose cycle on the set $\{v_0, \dots, v_{k-1}\}$ is $(v_0 v_1 \dots v_{k-1})$ and which fixes all other vertices. For $v = v_0$ we have $\kappa_v = k$ and

$$\sigma^i(v_0) = v_i, \quad \sigma^{i+1}(v_0) = v_{i+1}, \quad i = 0, \dots, k-1,$$

so by Eq. (12),

$$\begin{aligned} \sum_{i=0}^{k-1} M(v_i, v_i) &= \sum_{i=0}^{\kappa_v - 1} M(\sigma^i(v_0), \sigma^i(v_0)) \\ &\geq \sum_{i=0}^{\kappa_v - 1} M(\sigma^i(v_0), \sigma^{i+1}(v_0)) = \sum_{i=0}^{k-1} M(v_i, v_{i+1}). \end{aligned} \tag{20}$$

Combining Eq. (19) and Eq. (20) yields

$$J(\tilde{w}) - J(w) \geq 0. \tag{21}$$

Now start from the optimizer $w^{(0)}$. If $G_{w^{(0)}}$ has no directed cycles, set $\bar{w} = w^{(0)}$. Otherwise, choose a directed cycle in $G_{w^{(0)}}$, apply the above transformation with maximal ε as chosen above, and obtain $w^{(1)} \in \mathcal{P}(p_0, q)$ with $J(w^{(1)}) \geq J(w^{(0)})$. Since $w^{(0)}$ is optimal, $J(w^{(1)}) = J(w^{(0)})$, so $w^{(1)}$ is also optimal. Moreover, at least one edge of the chosen cycle has $w^{(1)}(v_i, v_{i+1}) = 0$.

Iterating this construction, we obtain a sequence of optimal plans $w^{(m)}$ in $\mathcal{P}(p_0, q)$ in which the set of off-diagonal edges with positive mass strictly decreases whenever there is a directed cycle. Since there are only finitely many off-diagonal entries, this procedure must terminate. We thus obtain an optimal plan $\bar{w}$ such that $G_{\bar{w}}$ contains no directed cycle.

Define the off-diagonal flow

$$F(v, s) := \begin{cases} \bar{w}(v, s), & v \neq s, \\ 0, & v = s. \end{cases}$$

For $v \in \mathcal{V}$ define

$$\text{out}(v) := \sum_{s \neq v} F(v, s), \qquad \text{in}(v) := \sum_{u \neq v} F(u, v).$$

Using the marginal constraints of $\bar{w}$,

$$\begin{aligned} q(v) &= \sum_s \bar{w}(v, s) = \bar{w}(v, v) + \text{out}(v), \\ p_0(v) &= \sum_u \bar{w}(u, v) = \bar{w}(v, v) + \text{in}(v), \end{aligned}$$

so

$$\text{out}(v) - \text{in}(v) = q(v) - p_0(v) = \begin{cases} \frac{\delta}{2}, & v = a, \\ -\frac{\delta}{2}, & v = b, \\ 0, & v \notin \{a, b\}. \end{cases} \tag{22}$$

Thus F is a nonnegative flow of value $\delta/2$ from a to b on the directed acyclic graph $G_{\bar{w}}$.

We now decompose F into simple a–b paths. Define $F^{(1)} := F$, $\beta_1 := \delta/2$, and proceed inductively. Assume $F^{(m)} \not\equiv 0$ is a nonnegative flow on $G_{\bar{w}}$ satisfying Eq. (22) with $\delta/2$ replaced by some $\beta_m \geq 0$, where out and in now refer to $F^{(m)}$. As $G_{\bar{w}}$ is acyclic, the first vertex u of a longest directed path with positive mass has $\text{in}(u) = 0 < \text{out}(u)$, so $u = a$, the only vertex with positive balance. Hence $\text{in}(a) = 0$, $\text{out}(a) = \beta_m > 0$, and there exists $s_1 \neq a$ with $F^{(m)}(a, s_1) > 0$. Set $v_0^{(m)} := a$ and $v_1^{(m)} := s_1$.

Suppose $v_0^{(m)} = a, \dots, v_j^{(m)} \neq b$ are already recursively constructed such that we use Eq. (22) and nonnegativity to obtain for $j \geq 1$,

$$\text{in}(v_j^{(m)}) \geq F^{(m)}(v_{j-1}^{(m)}, v_j^{(m)}) > 0.$$

For $v_j^{(m)} \notin \{a, b\}$, Eq. (22) gives $\text{out}(v_j^{(m)}) = \text{in}(v_j^{(m)}) > 0$, so there exists $v_{j+1}^{(m)} \neq v_j^{(m)}$ with $F^{(m)}(v_j^{(m)}, v_{j+1}^{(m)}) > 0$. Since $G_{\bar{w}}$ is acyclic and finite, the sequence $v_0^{(m)}, v_1^{(m)}, \dots$ cannot visit a vertex twice; hence the process must terminate at a vertex with no outgoing edges. By (22), such a vertex must satisfy $\text{out}(v) - \text{in}(v) \leq 0$. Because all intermediate vertices have balance 0, the only possible terminal vertex is b. Thus we obtain a simple path

$$P^m : a = v_0^{(m)} \to v_1^{(m)} \to \cdots \to v_{k_m}^{(m)} = b.$$

Let

$$\alpha_m := \inf_{0 \leq i \leq k_m - 1} F^{(m)}(v_i^{(m)}, v_{i+1}^{(m)}) > 0,$$

and define

$$F^{(m+1)}(v, s) := F^{(m)}(v, s) - \alpha_m \cdot \mathbf{1}\{(v, s) = (v_i^{(m)}, v_{i+1}^{(m)}) \text{ for some } i = 0, \dots, k_m - 1\}.$$

Then $F^{(m+1)} \geq 0$. Removing α_m along a path from a to b changes out $-$ in only at a and b, by $-\alpha_m$ and $+\alpha_m$, so $F^{(m+1)}$ satisfies Eq. (22) with $\delta/2$ replaced by $\beta_{m+1} := \beta_m - \alpha_m$, which is nonnegative because $\alpha_m \leq F^{(m)}(a, v_1^{(m)}) \leq \text{out}(a) = \beta_m$. Moreover, $F^{(m+1)}$ vanishes on an edge of P^m where $F^{(m)}$ is positive, and it has no new positive edges.

Since $G_{\bar{w}}$ has finitely many edges, this procedure terminates after some L steps with $F^{(L+1)} \equiv 0$, and hence $\beta_{L+1} = 0$. For $\ell = 1, \dots, L$, we obtain simple paths

$$P^\ell : a = v_0^\ell \to v_1^\ell \to \cdots \to v_{k_\ell}^\ell = b, \qquad \ell = 1, \dots, L,$$

and coefficients $\alpha_\ell > 0$ such that

$$F(v, s) = \sum_{\ell=1}^{L} \alpha_\ell \mathbf{1}\{(v, s) = (v_i^\ell, v_{i+1}^\ell) \text{ for some } i = 0, \dots, k_\ell - 1\}, \qquad v \neq s, \tag{23}$$

$$\sum_{\ell=1}^{L} \alpha_\ell = \frac{\delta}{2}. \tag{24}$$

Finally, we prove that the optimal G_{w^*} contains only a single path and derive the closed form solution w^*.

For a general feasible $w \in \mathcal{P}(p_0, q)$, the column constraints give

$$w(v, v) = p_0(v) - \sum_{u \neq v} w(u, v), \qquad v \in \mathcal{V}. \tag{25}$$

Thus

$$\begin{aligned} J(w) &= \sum_v w(v, v) M(v, v) + \sum_{u \neq v} w(u, v) M(u, v) \\ &= \sum_v \left(p_0(v) - \sum_{u \neq v} w(u, v) \right) M(v, v) + \sum_{u \neq v} w(u, v) M(u, v) \\ &= \underbrace{\sum_v p_0(v) M(v, v)}_{=: J_0} + \sum_{u \neq v} w(u, v) \left(M(u, v) - M(v, v) \right). \end{aligned} \tag{26}$$

For the optimal plan $\bar{w}$, the off-diagonal part is F, so

$$J(\bar{w}) = J_0 + \sum_{u \neq v} F(u, v) \left(M(u, v) - M(v, v) \right). \tag{27}$$

For each path P^ℓ, define its *gain*

$$W(P^\ell) := \sum_{i=0}^{k_\ell - 1} \left(M(v_i^\ell, v_{i+1}^\ell) - M(v_{i+1}^\ell, v_{i+1}^\ell) \right). \tag{28}$$

Using the decomposition Eq. (23), we obtain from Eq. (27)

$$\begin{aligned} J(\bar{w}) &= J_0 + \sum_{u \neq v} \left(\sum_{\ell=1}^{L} \alpha_\ell \mathbf{1}\{(u,v) = (v_i^\ell, v_{i+1}^\ell) \text{ for some } i = 0, ..., k_\ell - 1\} \right) (M(u,v) - M(v,v)) \\ &= J_0 + \sum_{\ell=1}^{L} \alpha_\ell \sum_{i=0}^{k_\ell - 1} \left(M(v_i^\ell, v_{i+1}^\ell) - M(v_{i+1}^\ell, v_{i+1}^\ell) \right) \\ &= J_0 + \sum_{\ell=1}^{L} \alpha_\ell W(P^\ell). \end{aligned} \tag{29}$$

By Eq. (24),

$$J(\bar{w}) = J_0 + \frac{\delta}{2} \sum_{\ell=1}^{L} \theta_\ell W(P^\ell), \qquad \theta_\ell := \frac{\alpha_\ell}{\delta/2}, \quad \theta_\ell \geq 0, \ \sum_\ell \theta_\ell = 1. \tag{30}$$

Let

$$W^* := \sup\{W(P) : P \text{ is a simple directed path from } a \text{ to } b\}. \tag{31}$$

Since the graph on $\mathcal{V}$ is finite, the maximum exists. From Eq. (30) we deduce

$$J(\bar{w}) \leq J_0 + \frac{\delta}{2} W^*. \tag{32}$$

Now fix an arbitrary simple directed path

$$P : a = u_0 \to u_1 \to \cdots \to u_K = b$$

with distinct vertices $u_0, \ldots, u_K$. Define w^P by

$$w^P := \frac{\delta}{2} \sum_{i=0}^{K-1} \left(\mathbf{e}_{u_i} \mathbf{e}_{u_{i+1}}^\top - \mathbf{e}_{u_{i+1}} \mathbf{e}_{u_{i+1}}^\top \right) + \operatorname{diag}(p_0). \tag{33}$$

We first verify $w^P \in \mathcal{P}(p_0, q)$.

The diagonal entries of w^P are

$$\begin{aligned} w^P(u_0, u_0) &= p_0(u_0), \\ w^P(u_j, u_j) &= p_0(u_j) - \frac{\delta}{2}, \quad 1 \leq j \leq K, \\ w^P(v, v) &= p_0(v), \quad v \notin \{u_0, \ldots, u_K\}. \end{aligned} \tag{34}$$

Since $\inf_v p_0(v) \geq \delta$, we have $p_0(u_j) - \delta/2 \geq \delta/2 > 0$ for $1 \leq j \leq K$; thus all diagonals are nonnegative. Off-diagonal entries are either 0 or $\delta/2$, so $w^P \geq 0$.

For $v = u_0 = a$,

$$\sum_s w^P(a, s) = w^P(a, a) + w^P(a, u_1) = p_0(a) + \frac{\delta}{2} = q(a).$$

For an internal vertex u_j with $1 \leq j \leq K - 1$,

$$\begin{aligned} \sum_s w^P(u_j, s) &= w^P(u_j, u_j) + w^P(u_j, u_{j+1}) \\ &= \left(p_0(u_j) - \frac{\delta}{2} \right) + \frac{\delta}{2} = p_0(u_j) = q(u_j). \end{aligned}$$

For $v = u_K = b$,

$$\sum_s w^P(b, s) = w^P(b, b) = p_0(b) - \frac{\delta}{2} = q(b).$$

For $v \notin \{u_0, \dots, u_K\}$, the only nonzero entry in row v is the diagonal, so

$$\sum_s w^P(v,s) = p_0(v) = q(v).$$

Thus the row constraints are satisfied.

For a vertex u_{i+1} on the path (with $0 \le i \le K-1$),

$$\begin{aligned}\sum_v w^P(v,u_{i+1}) &= w^P(u_{i+1},u_{i+1}) + w^P(u_i,u_{i+1}) \\ &= \left(p_0(u_{i+1}) - \frac{\delta}{2}\right) + \frac{\delta}{2} = p_0(u_{i+1}).\end{aligned}$$

For $s \notin \{u_1, \dots, u_K\}$, the only nonzero entry in column s is $w^P(s,s) = p_0(s)$, so

$$\sum_v w^P(v,s) = p_0(s).$$

Hence $w^P \in \mathcal{P}(p_0, q)$.

From Eq. (33) and linearity of J,

$$J(w^P) - J_0 = \frac{\delta}{2} \sum_{i=0}^{K-1} \left(M(u_i,u_{i+1}) - M(u_{i+1},u_{i+1})\right) = \frac{\delta}{2} W(P), \tag{35}$$

where $W(P)$ is defined as in Eq. (28) for this path P.

Now choose a path P^* attaining the maximum gain $W(P^*) = W^*$ in Eq. (31), and set $w^* := w^{P^*}$. Then

$$\begin{aligned}J(w^*) &= J_0 + \frac{\delta}{2} W^* \ge J_0 + \frac{\delta}{2} W(P^\ell) \quad \forall \ell, \\ &\ge J_0 + \frac{\delta}{2} \sum_{\ell=1}^{L} \theta_\ell W(P^\ell) = J(\bar{w}) \ge J(w), \quad \forall w \in \mathcal{P}(p_0, q),\end{aligned} \tag{36}$$

where we used Eq. (30) and Eq. (32) in the last line. Thus, w^* is an optimizer of J over $\mathcal{P}(p_0, q)$ with only a single path.

Write the maximizing path as

$$P^* : a = u_0 \to u_1 \to \cdots \to u_K = b.$$

Define a permutation $\sigma \in S_n$ by

$$\begin{aligned}\sigma(u_i) &= u_{i+1}, \quad i = 0, \dots, K-1, \\ \sigma(u_K) &= u_0, \\ \sigma(v) &= v, \quad v \notin \{u_0, \dots, u_K\}.\end{aligned}$$

Then the orbit of a under σ is

$$a, \sigma(a), \dots, \sigma^K(a) = u_0, u_1, \dots, u_K,$$

and $\sigma^{K+1}(a) = a$, so the minimal positive integer κ with $\sigma^\kappa(a) = a$ is $\kappa = K+1$. In particular,

$$\sigma^i(a) = u_i, \quad i = 0, \dots, K.$$

Therefore, the definition Eq. (33) of w^* can be rewritten as

$$\begin{aligned}w^* &= \frac{\delta}{2} \sum_{i=0}^{K-1} \left(\mathbf{e}_{u_i} \mathbf{e}_{u_{i+1}}^\top - \mathbf{e}_{u_{i+1}} \mathbf{e}_{u_{i+1}}^\top\right) + \mathrm{diag}(p_0) \\ &= \frac{\delta}{2} \sum_{i=0}^{\kappa-2} \left(\mathbf{e}_{\sigma^i(a)} \mathbf{e}_{\sigma^{i+1}(a)}^\top - \mathbf{e}_{\sigma^{i+1}(a)} \mathbf{e}_{\sigma^{i+1}(a)}^\top\right) + \mathrm{diag}(p_0),\end{aligned}$$

which is exactly Eq. (13). This completes the proof. □

Lemma A.3 (Optimizer of Middle Problem)**.** *Let $\mathcal{V}$ be $\{1,\dots,n\}$, $p_0 \in \Delta(\mathcal{V})$ be a distribution over $\mathcal{V}$ such that $\inf_{v\in\mathcal{V}} p_0(v) \geq \delta$, and $M : \mathcal{V}\times\mathcal{V} \mapsto \mathbb{R}$ be a matrix. Consider the optimization problem*

$$J : \inf_{q\in\mathcal{Q}(p_0,\delta)} \sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot M(v,s)$$

where

$$\mathcal{Q}(p_0,\delta) = \{q\in\Delta(\mathcal{V}) : \|q-p_0\|_1 \leq \delta\}$$
$$\mathcal{P}(p_0,q) = \left\{w(v,s) : \sum_s w(v,s) = q(v), \sum_v w(v,s) = p_0(s)\right\}$$

Then there exists an optimal solution q^* of J coming from the set

$$\mathcal{Q}_{\rm ext}(p_0,\delta) := \left\{p_0 + \frac{\delta}{2}\cdot(\mathbf{e}_a - \mathbf{e}_b) : a\neq b\in\mathcal{V}\right\}.$$

Proof. By Lemma C.1, $\mathcal{Q}_{\rm ext}(p_0,\delta)$ is the set of extreme points of the convex polytope $\mathcal{Q}(p_0,\delta)$. Define $J(q) = \sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot M(v,s)$. We show that $J(q)$ is a concave function. Indeed for all $q_1\neq q_2$, we let

$$w_1 := \arg\sup_{w\in\mathcal{P}(p_0,q_1)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot M(v,s)$$
$$w_2 := \arg\sup_{w\in\mathcal{P}(p_0,q_2)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot M(v,s).$$

We have

$$\begin{aligned} J((q_1+q_2)/2) &= \sup_{w\in\mathcal{P}(p_0,(q_1+q_2)/2)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot M(v,s) \\ &\geq \sum_{v,s\in\mathcal{V}} (w_1(v,s)+w_2(v,s))/2\cdot M(v,s) \\ &= (J(q_1)+J(q_2))/2, \end{aligned}$$

where the second step is because the feasibility of $(w_1(v,s)+w_2(v,s))/2$:

$$\sum_s (w_1(v,s)+w_2(v,s))/2 = (q_1(v)+q_2(v))/2.$$

Since $J(q)$ is concave over a convex feasible set, its optimizer must be found on the extreme set $\mathcal{Q}_{\rm ext}(p_0,\delta)$. □

Lemma A.4 (Optimal Subpaths)**.** *Let $\mathcal{V}$ be $\{1,\dots,n\}$, $p_0\in\Delta(\mathcal{V})$ be a distribution over $\mathcal{V}$ such that $\inf_{v\in\mathcal{V}} p_0(v)\geq\delta$. Let $M : \mathcal{V}\times\mathcal{V}\mapsto\mathbb{R}$ be a matrix that satisfies*

$$\sum_{i=0}^{\kappa_v-1} M(\sigma^i(v),\sigma^i(v)) \geq \sum_{i=0}^{\kappa_v-1} M(\sigma^i(v),\sigma^{i+1}(v))$$

for any $v\in\mathcal{V}$, $\sigma\in S_n$, where $\kappa_v = \inf\{k\geq 1 : \sigma^k(v) = v\}$. Consider the optimization problem

$$J(q) = \sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot M(v,s)$$

where

$$\mathcal{P}(p_0,q) = \left\{w(v,s) : \sum_s w(v,s) = q(v), \sum_v w(v,s) = p_0(s)\right\}.$$

If there exists $q = p_0 + \frac{\delta}{2}\cdot(\mathbf{e}_a - \mathbf{e}_b)$ for some $a\neq b\in\mathcal{V}$ and w_q^* such that w_q^* is the optimizer of $J(q)$ and for a permutation σ we have that

$$w^*(q) = \frac{\delta}{2}\cdot\sum_{i=0}^{\kappa-2}(\mathbf{e}_{\sigma^i(a)}\mathbf{e}_{\sigma^{i+1}(a)}^\top - \mathbf{e}_{\sigma^{i+1}(a)}\mathbf{e}_{\sigma^{i+1}(a)}^\top) + \mathrm{diag}(p_0),$$

where κ is the minimum positive integer such that $\sigma^\kappa(a) = a$. Then define $c := \sigma^i(a)$ and $d := \sigma^{i+1}(a)$ with $i < \kappa - 1$, and let $q' = p_0 + \frac{\delta}{2} \cdot (\mathbf{e}_c - \mathbf{e}_d)$. We have that the $w^*(q')$ that optimizes $J(q')$ is given by

$$w^*(q') = \frac{\delta}{2} \cdot (\mathbf{e}_c \mathbf{e}_d^\top - \mathbf{e}_d \mathbf{e}_d^\top) + \operatorname{diag}(p_0).$$

Proof. Let $c = \sigma^i(a)$ and $d = \sigma^{i+1}(a)$. Define the local transport term corresponding to the edge (c, d) as:

$$T_{c,d} := \frac{\delta}{2} \cdot (\mathbf{e}_c \mathbf{e}_d^\top - \mathbf{e}_d \mathbf{e}_d^\top).$$

Note that the proposed optimizer for the subproblem $J(q')$ is given by $\hat{w} = \operatorname{diag}(p_0) + T_{c,d}$.

We proceed by contradiction. Assume that $\hat{w}$ is not the optimizer for $J(q')$. Then there exists a feasible transport plan $\tilde{w} \in \mathcal{P}(p_0, q')$ such that:

$$\sum_{v,s \in \mathcal{V}} \tilde{w}(v, s) M(v, s) > \sum_{v,s \in \mathcal{V}} \hat{w}(v, s) M(v, s).$$

By Lemma A.2, $\tilde{w}$ can be chosen to take the form of

$$\tilde{w} = \frac{\delta}{2} \sum_{i=0}^{\bar{\kappa}-2} \left(\mathbf{e}_{\bar{\sigma}^i(c)} \mathbf{e}_{\sigma^{i+1}(c)}^\top - \mathbf{e}_{\bar{\sigma}^{i+1}(c)} \mathbf{e}_{\bar{\sigma}^{i+1}(c)}^\top \right) + \operatorname{diag}(p_0)$$

for some $\bar{\sigma} \in S_n$ and $\bar{\kappa} = \inf\{k : \bar{\sigma}^k(c) = c\}$. It follows that $\inf_{v \in \mathcal{V}} (\tilde{w} - \operatorname{diag}(p_0))(v, v) \geq -\delta/2$ and $\inf_{v \neq s \in \mathcal{V}} (\tilde{w} - \operatorname{diag}(p_0))(v, s) \geq 0$.

Substituting $\hat{w} = \operatorname{diag}(p_0) + T_{c,d}$ into the inequality, we have:

$$\sum_{v,s \in \mathcal{V}} (\tilde{w}(v, s) - \operatorname{diag}(p_0)_{v,s}) M(v, s) > \sum_{v,s \in \mathcal{V}} (T_{c,d})_{v,s} M(v, s). \tag{37}$$

Now, consider the global optimizer $w^*(q)$. By the hypothesis, $w^*(q)$ decomposes into a sum of path segments. We can separate the specific term $T_{c,d}$ from the rest of the path:

$$w^*(q) = \underbrace{\left(\operatorname{diag}(p_0) + \frac{\delta}{2} \sum_{j \neq i}^{\kappa-2} (\mathbf{e}_{\sigma^j(a)} \mathbf{e}_{\sigma^{j+1}(a)}^\top - \mathbf{e}_{\sigma^{j+1}(a)} \mathbf{e}_{\sigma^{j+1}(a)}^\top) \right)}_{R} + T_{c,d},$$

where R represents the flow on the path excluding the step from c to d. We construct a new global transport plan w_{new} by replacing the local step $T_{c,d}$ in $w^*(q)$ with the "better" local flow derived from $\tilde{w}$:

$$w_{\text{new}} := R + (\tilde{w} - \operatorname{diag}(p_0)).$$

Substituting $R = w^*(q) - T_{c,d}$, we get:

$$w_{\text{new}} = w^*(q) - T_{c,d} + \tilde{w} - \operatorname{diag}(p_0).$$

We verify that w_{new} is feasible for the original problem $J(q)$:

Since $\tilde{w} \in \mathcal{P}(p_0, q')$, its row sum is $q' = p_0 + \frac{\delta}{2}(\mathbf{e}_c - \mathbf{e}_d)$. The term $T_{c,d}$ also corresponds to a row marginal shift of $\frac{\delta}{2}(\mathbf{e}_c - \mathbf{e}_d)$. Thus, w_{new} preserves the row sums of $w^*(q)$, which equal q.

Both $\tilde{w}$ and $\operatorname{diag}(p_0) + T_{c,d}$ maintain column sums equal to p_0. Thus, w_{new} maintains the column sums of $w^*(q)$, which equal p_0.

Since $\inf_v p_0(v) \geq \delta$, $\inf_v T_{c,d}(v, v) \geq -\delta/2$, and $\inf_v (\tilde{w} - \operatorname{diag}(p_0))(v, v) \geq -\delta/2$, we have $\inf_v w_{\text{new}}(v, v) \geq 0$. Furthermore, all off-diagonal entries of $w^*(q) - T_{c,d}$ and $\tilde{w} - \operatorname{diag}(p_0)$ are non-negative, thus we conclude that w_{new} is non-negative.

Finally, we compare the objective value of w_{new} to $w^*(q)$:

$$\begin{aligned} J(w_{\text{new}}) &= \sum_{v,s}(R_{v,s} + \tilde{w}(v,s) - \operatorname{diag}(p_0)_{v,s})M(v,s) \\ &= \sum_{v,s} R_{v,s}M(v,s) + \sum_{v,s}(\tilde{w}(v,s) - \operatorname{diag}(p_0)_{v,s})M(v,s) \\ &> \sum_{v,s} R_{v,s}M(v,s) + \sum_{v,s}(T_{c,d})_{v,s}M(v,s) \quad \text{(by Ineq. 37)} \\ &= J(w^*(q)). \end{aligned}$$

We have constructed a feasible solution $w_{\text{new}} \in \mathcal{P}(p_0, q)$ with a strictly higher objective value than $w^*(q)$. This contradicts the optimality of $w^*(q)$. Therefore, $\hat{w}$ must be the optimizer for $J(q')$. □

Proposition A.5 (Reformulated Problem)**.** *Let $\mathcal{V}$ be $\{1, \dots, n\}$ and $p_0 \in \Delta(\mathcal{V})$ be a distribution over $\mathcal{V}$ such that $\inf_{v\in\mathcal{V}} p_0(v) \geq \delta$. Consider the following problem*

$$\sup_{r} \inf_{q\in\mathcal{Q}(p_0,\delta)} \sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot \log r(v,s) \tag{38}$$

$$s.t. \sum_{s\in\mathcal{V}} r(v,s) = 1,\ \forall v \in \mathcal{V} \tag{39}$$

$$r(v,s) \geq 0,\ \forall v,s \in \mathcal{V}$$

where p_0 is a fixed distribution over the sample space $\mathcal{V}$ such that $\inf_{v\in\mathcal{V}} p_0(v) \geq \delta$, and

$$\begin{aligned} \mathcal{Q}(p_0,\delta) &= \{q \in \Delta(\mathcal{V}) : \|q - p_0\|_1 \leq \delta\} \\ \mathcal{P}(p_0,q) &= \left\{ w(v,s) : \sum_{s\in\mathcal{V}} w(v,s) = q(v), \sum_{v\in\mathcal{V}} w(v,s) = p_0(s) \right\}. \end{aligned}$$

Then the optimum value is

$$J^* = (1-\tfrac{\delta}{2})\log\left(1 - \frac{\delta}{2}\right) + \tfrac{\delta}{2}\log\left(\frac{\delta}{2(n-1)}\right).$$

In particular, this is achieved at

$$r^*_{a,b}(v,s) = \begin{cases} 1-\delta/2, & s = v, \\ \delta/(2n-2), & s \neq v, \end{cases}$$

and for every boundary point $q_{a,b} = p_0 + \frac{\delta}{2}(\mathbf{e}_a - \mathbf{e}_b)$, $a \neq b$, the inner problem is optimized by

$$w^*_{a,b} = \operatorname{diag}(p_0) + \frac{\delta}{2}\left(\mathbf{e}_a\mathbf{e}_b^\top - \mathbf{e}_b\mathbf{e}_b^\top\right).$$

For a general $q \in \mathcal{Q}(p_0,\delta)$, the inner optimizer for the proposed r^* is any maximal coupling of q and p_0. Since the objective places larger weight on diagonal entries than off-diagonal entries, its value is minimized over $\mathcal{Q}(p_0,\delta)$ by boundary points with $\|q - p_0\|_1 = \delta$.

Proof. By Lemma A.3, it suffices to restrict the middle infimum to boundary targets

$$q_{a,b} = p_0 + \frac{\delta}{2}(\mathbf{e}_a - \mathbf{e}_b), \qquad a \neq b.$$

For every directed cycle of distinct vertices $v_0, \dots, v_{k-1}$ where $2 \leq k \leq n$ and $v_k = v_0$ since $\prod_{i=0}^{k-1} e(v_i, v_i) \geq \prod_{i=0}^{k-1} e(v_i, v_{i+1})$, we also have $\prod_{i=0}^{k-1} r(v_i, v_i) \geq \prod_{i=0}^{k-1} r(v_i, v_{i+1})$. Thus for each $q_{a,b}$, applying Lemma A.2 with $M = \log r$, which is real-valued as $r > 0$, shows that the inner supremum admits an optimizer of path-flow form

$$\frac{\delta}{2}\sum_{i=0}^{\kappa-2}\left(\mathbf{e}_{\sigma^i(a)}\mathbf{e}_{\sigma^{i+1}(a)}^\top - \mathbf{e}_{\sigma^{i+1}(a)}\mathbf{e}_{\sigma^{i+1}(a)}^\top\right) + \operatorname{diag}(p_0),$$

where $\sigma \in S_n$, $\kappa = \inf\{k \in \mathbb{Z}_+ : \sigma^k(a) = a\}$, and the path endpoint satisfies $\sigma^{\kappa-1}(a) = b$. Define

$$\begin{aligned} J(r, w, q) &= \sum_{v,s\in\mathcal{V}} w(v,s) \cdot \log r(v,s) \\ J^* &= \sup_r \inf_{q\in\mathcal{Q}_{\text{ext}}(p_0,\delta)} \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,q)} J(r,w,q) \\ R^* &= \{r : \inf_{q\in\mathcal{Q}_{\text{ext}}(p_0,\delta)} \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,q)} J(r,w,q) = J^*\} \\ r^* &= \arg\sup\{\text{tr}[r] : r \in R^*\}. \end{aligned}$$

We say a solution $(\bar{w}, \bar{q})$ is active if

$$\begin{aligned} \bar{w} &= \arg \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,\bar{q})} J(r^*, w, \bar{q}) \\ \bar{q} &= \arg \inf_{q\in\mathcal{Q}_{\text{ext}}(p_0,\delta)} \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,q)} J(r^*, w, q) \\ J(\bar{w}, \bar{q}) &= J^*. \end{aligned}$$

Fix $v \neq s \in \mathcal{V}$ and consider the perturbation for sufficiently small $\epsilon > 0$, which keeps $\bar{r}$ positive and preserves the cycle condition (a diagonal entry increases, an off-diagonal entry decreases)

$$\begin{aligned} \bar{r}(v,v) &\leftarrow r^*(v,v) + \epsilon \\ \bar{r}(v,s) &\leftarrow r^*(v,s) - \epsilon \\ \bar{r}(v',s') &\leftarrow r^*(v',s'),\ \forall (v',s') \notin \{(v,v),(v,s)\}. \end{aligned}$$

Then $\bar{r}$ cannot be a valid solution since it has higher trace than r^*. It follows that

$$\exists \text{ an active pair } (\bar{w},\bar{q}) \text{ such that } \quad \frac{dJ(\bar{r},\bar{w},\bar{q})}{d\epsilon} = \frac{\bar{w}(v,v)}{r^*(v,v)} - \frac{\bar{w}(v,s)}{r^*(v,s)} \leq 0.$$

Notice that the derivative $\frac{dJ(\bar{r},\bar{w},\bar{q})}{d\epsilon}$ must be either

- No-transport: $\bar{w}(v,v) = p_0(v), \bar{w}(v,s) = 0$ so $\frac{dJ(\bar{r},\bar{w},\bar{q})}{d\epsilon} = \frac{p_0(v)}{r^*(v,v)} > 0$.
- Middle-way: $\bar{w}(v,v) = p_0(v) - \delta/2, \bar{w}(v,s) = \delta/2$ so $\frac{dJ(\bar{r},\bar{w},\bar{q})}{d\epsilon} = \frac{p_0(v)-\delta/2}{r^*(v,v)} - \frac{\delta/2}{r^*(v,s)}$.
- Transport-start: $\bar{w}(v,v) = p_0(v), \bar{w}(v,s) = \delta/2$ so $\frac{dJ(\bar{r},\bar{w},\bar{q})}{d\epsilon} = \frac{p_0(v)}{r^*(v,v)} - \frac{\delta/2}{r^*(v,s)}$.
- Transport-end: $\bar{w}(v,v) = p_0(v) - \delta/2, \bar{w}(v,s) = 0$ so $\frac{dJ(\bar{r},\bar{w},\bar{q})}{d\epsilon} = \frac{p_0(v)-\delta/2}{r^*(v,v)} > 0$.

Since $\frac{dJ(\bar{r},\bar{w},q)}{d\epsilon} \leq 0$, we can rule out the first and last cases. Now we can summarize that for any $v \neq s \in \mathcal{V}$ we have either Middle-way: $\bar{w}(v,v) = p_0(v) - \delta/2, \bar{w}(v,s) = \delta/2$ or Transport-start: $\bar{w}(v,v) = p_0(v), \bar{w}(v,s) = \delta/2$, and in any case $\frac{p_0(v)-\delta/2}{r^*(v,v)} - \frac{\delta/2}{r^*(v,s)} \leq 0$ holds.

Since $p_0(v) \geq \delta$, in either Middle-way or Transport-start case we have

$$0 \geq \frac{p_0(v) - \delta/2}{r^*(v,v)} - \frac{\delta/2}{r^*(v,s)} \geq \frac{\delta/2}{r^*(v,v)} - \frac{\delta/2}{r^*(v,s)}.$$

It follows that $r^*(v,v) \geq r^*(v,s)$. Since this argument holds for all $v \neq s \in \mathcal{V}$, r^* must satisfy $r^*(v,v) \geq r^*(v,s)$ for all $v \neq s \in \mathcal{V}$.

Next, we show that all active $\bar{q}, \bar{w}$ can be written as $\bar{q} = p_0 + \frac{\delta}{2} \cdot (\mathbf{e}_v - \mathbf{e}_s), \bar{w} = \frac{\delta}{2} \cdot (\mathbf{e}_v\mathbf{e}_s^\top - \mathbf{e}_s\mathbf{e}_s^\top) + \text{diag}(p_0)$ for some $v \neq s \in \mathcal{V}$.

In either Middle-way or Transport-start case, there exists $a \neq b \in \mathcal{V}$ and $\sigma \in S_n$ and $i \in \mathbb{Z}$ such that $\bar{q} = p_0 + \frac{\delta}{2} \cdot (\mathbf{e}_a - \mathbf{e}_b)$ and $v = \sigma^i(a), s = \sigma^{i+1}(a)$, due to Lemma A.2. By Lemma A.4, with respect to $\bar{q} = p_0 + \frac{\delta}{2} \cdot (\mathbf{e}_v - \mathbf{e}_s)$, the optimizer of the inner problem must be written as

$$\frac{\delta}{2} \cdot (\mathbf{e}_v\mathbf{e}_s^\top - \mathbf{e}_s\mathbf{e}_s^\top) + \text{diag}(p_0) = \arg \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,\bar{q})} J(r^*, w, \bar{q}).$$

Repeating this perturbation separately for each ordered pair (v, s) with $v \neq s$ yields, for each such pair, at least one active boundary coupling that witnesses the nonpositive directional derivative. The subsequent case analysis is applied to that witnessing active pair.

We can now explicitly write:

$$\inf_{q\in\mathcal{Q}_{\text{ext}}(p_0,\delta)} \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,q)} J(r^*, w, q) = \inf_{q\in\mathcal{Q}_{\text{ext}}(p_0,\delta)} \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,q)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot \log r^*(v,s)$$

$$= \inf_{v\neq s\in\mathcal{V}} \sum_{x\in\mathcal{V}} p_0(x)\log r^*(x,x) - \frac{\delta}{2}\cdot(\log r^*(s,s) - \log r^*(v,s))$$

Define $J(r, v, s) = \sum_{x\in\mathcal{V}} p_0(x)\log r(x,x) - \frac{\delta}{2}\cdot(\log r(s,s) - \log r(v,s))$, we claim that $J(r^*, v, s)$ must be the same for all $v \neq s \in \mathcal{V}$. Otherwise suppose

$$J(r^*, \bar{v}, \bar{s}) = \sup_{v,s} J(r^*, v, s) > J^*.$$

Set

$$\bar{r}(\bar{v}, \bar{v}) \leftarrow r^*(\bar{v}, \bar{v}) + \epsilon$$
$$\bar{r}(\bar{v}, \bar{s}) \leftarrow r^*(\bar{v}, \bar{s}) - \epsilon$$
$$\bar{r}(v', s') \leftarrow r^*(v', s'),\ \forall (v', s') \notin \{(\bar{v}, \bar{v}), (\bar{v}, \bar{s})\}$$

for sufficiently small $\epsilon > 0$. Then for any ordered pair (v, s) with $v \neq s$ and $(v, s) \neq (\bar{v}, \bar{s})$, the off-diagonal entry $r(v, s)$ is unchanged, so

$$\frac{dJ(\bar{r}, v, s)}{d\epsilon} \geq \frac{p_0(\bar{v}) - \delta/2}{r^*(\bar{v}, \bar{v})} \geq 0,$$

and thus

$$\inf_{q\in\mathcal{Q}_{\text{ext}}(p_0,\delta)} \sup_{w\in\mathcal{P}_{\text{flow}}(p_0,q)} J(\bar{r}, w, q) \geq \inf_{v\neq s\in\mathcal{V}, (v,s)\neq(\bar{v},\bar{s})} J(\bar{r}, v, s) \geq J^*.$$

But $\text{tr}[\bar{r}] > \text{tr}[r^*]$, this is a contradiction.

From the last argument, we know that the optimal solution r^* is of the form

$$r^*_{a,b}(v,s) = \begin{cases} a, & s = v, \\ b, & s \neq v, \end{cases} \qquad a \geq 0,\ b \geq 0, \quad a + (n-1)b = 1,$$

with $a \geq b$ (diagonal dominance). We now optimize over the two parameters (a, b) subject to this constraint:

$$\sup_{a,b}\ \Phi(a,b) \quad \text{s.t. } a + (n-1)b = 1,\ a > 0,\ b > 0,\ a \geq b.$$

straightforward algebra shows the optimal objective value:

$$\Phi(a^*, b^*) = (1 - \tfrac{\delta}{2})\log a^* + \tfrac{\delta}{2}\log b^*$$
$$= (1 - \tfrac{\delta}{2})\log\left(1 - \frac{\delta}{2}\right) + \tfrac{\delta}{2}\log\left(\frac{\delta}{2(n-1)}\right).$$

attained at $a^* = 1 - \frac{\delta}{2}$ and $b^* = \frac{\delta}{2(n-1)}$. Thus, the optimal value is

$$J^* = (1 - \tfrac{\delta}{2})\log\left(1 - \frac{\delta}{2}\right) + \tfrac{\delta}{2}\log\left(\frac{\delta}{2(n-1)}\right).$$

This completes the proof.

□

Lemma A.6 (Row Normalization)**.** *Let $\mathcal{V}$ be $\{1, \ldots, n\}$ and $p_0 \in \Delta(\mathcal{V})$ be a distribution over $\mathcal{V}$ such that $\inf_{v\in\mathcal{V}} p_0(v) \geq \delta$. Consider the problem*

$$\sup_e \inf_{q\in\mathcal{Q}(p_0,\delta)} \sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s\in\mathcal{V}} w(v,s)\cdot\log e(v,s)$$
$$s.t. \sum_{v,s\in\mathcal{V}} q(v)p_0(s)e(v,s) \leq 1,\ \forall q \in \Delta(\mathcal{V})$$
$$e(v,s) \geq 0,\ \forall v, s \in \mathcal{V}$$

where

$$\mathcal{Q}(p_0, \delta) = \{q \in \Delta(\mathcal{V}) : \|q - p_0\|_1 \leq \delta\}$$
$$\mathcal{P}(p_0, q) = \left\{ w(v,s) : \sum_s w(v,s) = q(v), \sum_v w(v,s) = p_0(s) \right\}$$

Now define the kernel matrix $r : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ such that each of its entries are defined as

$$r(v,s) = \frac{p_0(s)e(v,s)}{A(v)}, \quad \forall v, s \in \mathcal{V},$$

where $A(v) := \sum_s p_0(s)e(v,s)$. Then $A^*(v) = \sum_s p_0(s)e^*(v,s) = 1$ at the optimizer e^*.

Proof. Let $e(v,s)$ be any feasible solution to the optimization problem. We define the scaling factor of e at node v as:

$$A(v) := \sum_{s \in \mathcal{V}} p_0(s)e(v,s).$$

Since e is positive (the standing assumption of this subsection), $A(v) > 0$, and all objective values below are finite.

We can decompose the matrix $e(v,s)$ into a scale-independent "shape" matrix $\bar{e}(v,s)$ and the scaling factors $A(v)$ as follows:

$$e(v,s) = A(v) \cdot \bar{e}(v,s), \quad \text{where } \bar{e}(v,s) = \frac{e(v,s)}{A(v)}.$$

By construction, the normalized matrix $\bar{e}$ satisfies the normalization property:

$$\sum_{s \in \mathcal{V}} p_0(s)\bar{e}(v,s) = \sum_{s \in \mathcal{V}} p_0(s)\frac{e(v,s)}{A(v)} = \frac{1}{A(v)} \sum_{s \in \mathcal{V}} p_0(s)e(v,s) = 1.$$

Now, let us analyze the constraint given in the problem statement. The condition is:

$$\sum_{v,s \in \mathcal{V}} q(v)p_0(s)e(v,s) \leq 1, \quad \forall q \in \Delta(\mathcal{V}).$$

Substituting the decomposition of $e(v,s)$:

$$\sum_{v \in \mathcal{V}} q(v) \left(\sum_{s \in \mathcal{V}} p_0(s)e(v,s) \right) = \sum_{v \in \mathcal{V}} q(v)A(v) \leq 1, \quad \forall q \in \Delta(\mathcal{V}).$$

Therefore $A(v) \leq 1$ for any $v \in \mathcal{V}$.

Next, we substitute the decomposition into the objective function. Using the property that $w \in \mathcal{P}(p_0, q)$ implies $\sum_s w(v,s) = q(v)$, we have:

$$\begin{aligned}
\sum_{v,s \in \mathcal{V}} w(v,s) \log e(v,s) &= \sum_{v,s \in \mathcal{V}} w(v,s) \log \left(A(v) \cdot \bar{e}(v,s) \right) \\
&= \sum_{v,s \in \mathcal{V}} w(v,s) \log \bar{e}(v,s) + \sum_{v,s \in \mathcal{V}} w(v,s) \log A(v) \\
&= \sum_{v,s \in \mathcal{V}} w(v,s) \log \bar{e}(v,s) + \sum_{v \in \mathcal{V}} q(v) \log A(v).
\end{aligned}$$

The first term depends only on the normalized shape $\bar{e}$, while the second term depends only on the scaling factors $A(v)$. To maximize the total objective, we must maximize the second term subject to the feasibility constraint derived above.

Consider the term $\sum_{v \in \mathcal{V}} q(v) \log A(v)$. Since the logarithm is a concave function, we can apply Jensen's inequality:

$$\sum_{v \in \mathcal{V}} q(v) \log A(v) \leq \log \left(\sum_{v \in \mathcal{V}} q(v)A(v) \right).$$

From the feasibility constraint, we know that $\sum_{v\in\mathcal{V}} q(v)A(v) \le 1$. Therefore:

$$\sum_{v\in\mathcal{V}} q(v)\log A(v) \le \log(1) = 0.$$

Thus for every q,

$$\sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s} w(v,s)\log e(v,s) \le \sup_{w\in\mathcal{P}(p_0,q)} \sum_{v,s} w(v,s)\log \bar{e}(v,s).$$

And the inequality is strict unless $\sum_v q(v)A(v) = 1$ and $A(v)$ is constant on the support of q.

Taking the minimum over q, we conclude:

$$\inf_q \sup_w \sum_{v,s} w(v,s)\log e(v,s) \le \inf_q \sup_w \sum_{v,s} w(v,s)\log \bar{e}(v,s).$$

Equality is achieved if and only if $A(v) = 1$ for all $v \in \mathcal{V}$. Thus, for any optimal solution e^*, the scaling factors must be set to 1. Consequently:

$$A^*(v) = \sum_{s\in\mathcal{V}} p_0(s)e^*(v,s) = 1.$$

□

A.2 Normalization and reduction to extreme points

We first prove the converse inequality under $\delta < \min_i p_i$. The boundary case follows from the limiting argument in Lemma A.26.

For $e \in \mathcal{E}$, put

$$A_i = \sum_j p_j e_{ij}.$$

Every feasible q satisfies $q_i \ge p_i - t > 0$. If $A_i = 0$ for some i, then row i of e vanishes, since every p_j is positive. Every coupling then has objective value $-\infty$, so such an e cannot improve Eq. (9).

Otherwise, $0 < A_i \le 1$ for every i. Define

$$\bar{e}_{ij} = e_{ij}/A_i, \qquad R_{ij} = p_j\bar{e}_{ij}.$$

The matrix R, which we call the normalized detector, is nonnegative and row stochastic, and

$$\sum_i q_i \log A_i \le 0.$$

Hence, for every $W \in \mathcal{P}(p_0, q)$,

$$\langle W, \log e\rangle \le \langle W, \log \bar{e}\rangle = H(p) + \langle W, \log R\rangle . \tag{40}$$

The equality uses the fixed column marginal p. Conversely, for every row-stochastic R, the matrix $e_{ij} = R_{ij}/p_j$ is a valid detector: its expectation under $q \otimes p$ is one for every q. Therefore, with

$$T_R(q,p) = \sup_{W\in\mathcal{P}(p_0,q)} \langle W, \log R\rangle ,$$
$$F_{p,t}(R) = \inf_{q\in\mathcal{Q}} T_R(q,p),$$

we obtain

$$J(p,t) = H(p) + \sup_R F_{p,t}(R). \tag{41}$$

Here and below the supremum is over the set $\mathcal{R} = \{R \in \mathbb{R}_+^{n\times n} : R\mathbf{1} = \mathbf{1}\}$ of row-stochastic matrices. The value $J(p_0, \delta/2)$ is the optimal value of Eq. (2), that is, J^* in Theorem 4.1.

Lemma A.7 (Reduction to extreme points)**.** *Let $\mathbf{e}_i$ denote the ith coordinate vector, and for $u \neq v$ call $q_{uv} = p + t(\mathbf{e}_u - \mathbf{e}_v)$ the extreme points of $\mathcal{Q}$. Then*

$$\begin{aligned} \mathcal{Q} &= \operatorname{conv}\{p + t(\mathbf{e}_u - \mathbf{e}_v) : u \neq v\}, \\ F_{p,t}(R) &= \min_{u \neq v} T_R(p + t(\mathbf{e}_u - \mathbf{e}_v), p). \end{aligned} \tag{42}$$

Proof. For $d = q - p$, let d^+, d^- be its positive and negative parts, and let $\tau = \sum_i d_i^+ = \sum_i d_i^- \leq t$. When $\tau > 0$,

$$d = \sum_{u \neq v} \frac{d_u^+ d_v^-}{\tau} (\mathbf{e}_u - \mathbf{e}_v).$$

The coefficients sum to τ. Hence q is a convex combination of p and the extreme points. The point p is itself the midpoint of two extreme points q_{uv} and q_{vu}. Every extreme point lies in the simplex because $p_i \geq 2t > t$, and it is at total-variation distance t from p.

If $T_R(q_{uv}, p)$ is finite for all $u \neq v$, take optimal couplings at the extreme points and mix them with the same convex coefficients. The mixture has row marginal q and column marginal p, and its objective value is at least the smallest of the values at the extreme points. If one of these values is $-\infty$, the infimum over $\mathcal{Q}$ is already $-\infty$. This proves the second identity, including the case where R has zero entries. □

We rescale by setting

$$P_i = p_i/t > 2, \qquad L = \sum_i P_i = 1/t. \tag{43}$$

The vector P is the column marginal of every matrix in $\mathcal{X}^{uv}(P)$ below; we call P_i the marginal of i and L the total mass. Conversely, every vector P with all entries $P_i > 2$ arises in this way, from $t = 1/L$ and $p_i = P_i/L$. For $u \neq v$, define

$$\begin{aligned} \mathcal{X}^{uv}(P) &= \{X \geq 0 : X\mathbf{1} = P + \mathbf{e}_u - \mathbf{e}_v,\ X^\top \mathbf{1} = P\}, \\ C_{uv}(R, P) &= \min_{X \in \mathcal{X}^{uv}(P)} \langle X, -\log R \rangle, \\ K_R(P) &= \max_{u \neq v} C_{uv}(R, P). \end{aligned} \tag{44}$$

We call the elements of $\mathcal{X}^{uv}(P)$ transports, and a minimizer in the definition of $C_{uv}(R, P)$ an optimal transport X^{uv} at q_{uv}. For q_{uv} we call u the source and v the sink: the row marginal of X^{uv} exceeds P by one in row u and falls short of it by one in row v. A positive entry of X at a zero entry of R contributes $+\infty$, and a zero entry of X contributes 0. Lemma A.7 and the substitution $W = tX$ give

$$F_{p,t}(R) = -tK_R(P). \tag{45}$$

Define

$$\begin{aligned} g(x) &= x \log x - (x-1)\log(x-1) \quad (x \geq 1), \\ B_n(L) &= g(L) + \log(n-1), \end{aligned} \tag{46}$$

where $g(1) = 0$ by the convention $0 \log 0 = 0$. Substituting $L = 1/t$ gives

$$tB_n(1/t) = -(1-t)\log(1-t) - t \log \frac{t}{n-1}. \tag{47}$$

It therefore remains to prove, for every $n \geq 2$, $K_R(P) \geq B_n(L)$ for every P with all $P_i > 2$ and every row-stochastic R. We record the derivatives

$$g'(x) = \log \frac{x}{x-1} > 0, \qquad g''(x) = -\frac{1}{x(x-1)} < 0 \quad (x > 1).$$

Hence g is increasing and concave on $[1, \infty)$. Moreover, since $\log(1+y) < y$ for $y > 0$, $g(x) = \log x + (x-1)\log(1 + 1/(x-1)) < 1 + \log x$ for $x > 1$.

A.3 Optimal transports and entropy

For a nonnegative matrix M, with row masses $r_i = \sum_j M_{ij}$, put

$$\mathcal{H}(M) = \sum_i r_i \log r_i - \sum_{i,j} M_{ij} \log M_{ij}. \tag{48}$$

This is the total mass of M times the conditional entropy of the column given the row; it is positively homogeneous and concave. For a row-stochastic R, Gibbs' inequality gives

$$\langle M, -\log R \rangle \geq \mathcal{H}(M). \tag{49}$$

Hence, for any probability distribution ω on ordered pairs of distinct indices and any optimal transports at the extreme points,

$$K_R(P) \geq \sum_{u \neq v} \omega_{uv} C_{uv}(R, P) \geq \mathcal{H}\left(\sum_{u \neq v} \omega_{uv} X^{uv} \right). \tag{50}$$

We also use the averaged objectives

$$T_s(R, P) = \max_{\substack{u,\ V \subset [n] \setminus \{u\} \\ |V| = s-1}} \frac{1}{s-1} \sum_{v \in V} C_{uv}(R, P), \qquad 2 \leq s \leq n. \tag{51}$$

Here $T_2 = K$ and $T_s \leq K$, so a lower bound for T_s is also a lower bound for K. For a fixed source, the average cost over any $\kappa \geq s-1$ sinks is at most T_s: it is the average of the means over all $(s-1)$-element subsets of these sinks. Since the costs are nonnegative and every extreme point occurs in such an average, finiteness of T_s implies that the costs at all extreme points are finite.

For fixed R with finite costs at all extreme points, write $E = \{(i,j) : R_{ij} > 0\}$ and $c_{ij} = -\log R_{ij}$ on E. We call an optimal transport with both marginals equal to P (the case $q = p$) a central transport. We choose the central transport X^0 and the optimal transports X^{uv} at the extreme points by the same lexicographic minimization of

$$(\langle X, c \rangle, X_{e_1}, \ldots, X_{e_{|E|}}), \tag{52}$$

where $e_1, \ldots, e_{|E|}$ is a fixed ordering of the support. The central transport has finite cost, since the average of the optimal transports at the opposite extreme points q_{uv} and q_{vu} has both marginals equal to P and finite cost; successive minimization over compact faces makes every choice unique.

Lemma A.8 (Optimal transports and stationarity)**.** *Fix P and let G be K or one of the T_s. Then G attains a finite minimum over R, and the following properties hold.*

1. *Two transports chosen by* (52)*, with the same column marginals and row marginals differing by $\mathbf{e}_a - \mathbf{e}_b$, differ by an acyclic unit flow from row a to row b. Passing from the second transport to the first, row a increases entrywise and row b decreases entrywise. For $i \neq u$, the common part of row i for the source u is the row*

$$A_i^u = X_i^{ui}, \qquad |A_i^u| = P_i - 1, \tag{53}$$

where $|\cdot|$ denotes the total mass of a row, and

$$A_i^u \leq X_i^0, \qquad A_i^u \leq X_i^{uv} \quad (v \neq u), \qquad X_u^0 \leq X_u^{uv}. \tag{54}$$

All row inequalities are entrywise.

2. *Let R be a local minimum of G with finite value on the set of matrices with the same support as R. Then every choice of optimal transports at the extreme points admits a probability distribution λ such that*

$$\begin{aligned} N &= \sum_{u \neq v} \lambda_{uv} X^{uv} = \operatorname{diag}(v) R, \qquad N^\top \mathbf{1} = P, \\ v &= P + \alpha - \zeta, \qquad \alpha_i = \sum_{j \neq i} \lambda_{ij}, \qquad \zeta_i = \sum_{j \neq i} \lambda_{ji}. \end{aligned} \tag{55}$$

Here α_i and ζ_i are the probabilities under λ that i is the source and the sink, respectively. They satisfy

$$\sum_i \alpha_i = \sum_i \zeta_i = 1, \qquad \alpha_i + \zeta_i \le 1, \qquad v_i \ge P_i - 1 > 0, \tag{56}$$

and

$$G(R,P) = \mathcal{H}(N) = \sum_i v_i H(R_i). \tag{57}$$

For K, λ is supported on the extreme points of maximal cost. For T_s, it is a convex combination of the uniform distributions on the extreme points of the source averages that attain the maximum in (51)*.*

Proof. The cost function $x(-\log r)$, extended to $r = 0$ with values in $[0,\infty]$, is nonnegative and lower semicontinuous. Along any convergent sequence of matrices R, compactness of the transportation polytopes gives a convergent subsequence of optimal transports. Hence every C_{uv}, and thus G, is lower semicontinuous on the product of row simplices. This product is compact and a positive matrix has finite cost, so G attains a finite minimum.

For part 1, orient the positive entries of the difference of the two transports from rows to columns and the negative entries from columns to rows. A directed cycle would permit opposite feasible perturbations of the two transports. Its lexicographic cost is nonzero, so one of those perturbations would improve the chosen transport. The difference is therefore acyclic and decomposes into simple a-to-b paths of total weight one, and no path enters a or leaves b. Applying this to (X^{ui}, X^0), (X^{ui}, X^{uv}), and (X^0, X^{uv}) gives (54).

For part 2, fix optimal transports X^{uv} at R and replace $C_{uv}(Q,P)$ in G by $\langle X^{uv}, -\log Q\rangle$. The resulting convex function majorizes G and agrees with it at R, so it too is minimized there on the matrices with the support of R. Hence a convex combination of the gradients of its maximal terms vanishes on the perturbations of each row that preserve its sum. If λ is the corresponding distribution on extreme points, this says that $-N_{ij}/R_{ij}$ is constant on the positive entries of each row, and summing over the row shows that the constant is $-v_i$. The marginals of the transports X^{uv} give (55); since the source and sink of an extreme point differ, the events that i is the source and that i is the sink are disjoint. Finally all maximal terms have value G, whence $G = \langle N, -\log R\rangle = \mathcal{H}(N)$. □

The common parts give a decomposition of each X^{uv}. Let B^u have rows A^u_i for $i \ne u$ and X^0_u for $i = u$. Then

$$Z^{uv} = X^{uv} - B^u \ge 0, \qquad Z^{uv}\mathbf{1} = \mathbf{1} - \mathbf{e}_v, \tag{58}$$

and we call Z^{uv} the excess of X^{uv} over B^u. For fixed u, all Z^{uv} with $v \ne u$ have the same column marginals, and each row of Z^{uv} has mass zero or one. At a minimum as in part 2, averaging over λ the bounds that part 1 gives for the unit flows from X^0 to X^{uv} yields

$$\sum_j (N_{ij} - X^0_{ij})_+ \le 1 - \zeta_i, \qquad \sum_j (X^0_{ij} - N_{ij})_+ \le 1 - \alpha_i. \tag{59}$$

We also use two elementary entropy bounds. If x is a probability row and $z = \max_j x_j$, then

$$H(x) \ge 2(\log 2)(1 - z). \tag{60}$$

For $z \ge 1/2$, merging the other entries reduces the entropy to a binary one, and the bound follows from the chord of h_2; for $z \le 1/2$, $H(x) \ge -\log z \ge 2(\log 2)(1-z)$. At a minimum as in part 2, let $I_{2/3} = \{i : \max_j R_{ij} > 2/3\}$ (the selected rows), and let b count the distinct columns containing their maxima. Write V for the sum of v_i over the other rows and T for the mass of N in selected rows outside their maximal entries. With $h_0 = h_2(1/3)$, the entropy of a row is at least h_0 for an unselected row and at least $3ah_0$ for a selected row whose entries outside the maximum sum to $a \le 1/3$. The other $n - b$ columns have total mass at least $2(n-b)$, all of it in unselected rows or outside the selected maxima, so

$$G(R,P) \ge h_0(V + 3T) \ge 2(n-b)h_0. \tag{61}$$

Column j receives total mass P_j in every transport; we call P_j its capacity. If the maximum of row i exceeds $1/2$ and occurs in column j, then

$$P_j \geq v_i R_{ij} > \frac{P_i - 1}{2}. \tag{62}$$

For rows with the same maximizing column, the sum of v_i is at least the sum of their marginals minus one, since the sink probabilities ζ_i sum to one in Eq. (56).

Finally, positive smoothing $R_\varepsilon = (1-\varepsilon)R + \varepsilon Q$, with $Q > 0$ row stochastic, gives

$$K_R(P) \geq K_{R_\varepsilon}(P) + L \log(1-\varepsilon). \tag{63}$$

The same inequality holds for every transport cost, because every transport has total mass L, and hence for every C_{uv} and T_s. Therefore a lower bound that does not depend on R and holds for all matrices with positive entries also holds for their limits.

A.4 Rational bounds for logarithms

Lemma A.9 (Logarithm bounds)**.** *Let $y \geq 1$, put $z = (y-1)/(y+1)$, and let $N \geq 1$ be an integer. Then*

$$\begin{aligned} 2 \sum_{j=0}^{N-1} \frac{z^{2j+1}}{2j+1} &\leq \log y \\ &\leq 2 \sum_{j=0}^{N-1} \frac{z^{2j+1}}{2j+1} + \frac{2z^{2N+1}}{(2N+1)(1-z^2)}. \end{aligned} \tag{64}$$

Consequently, if $x > 0$ is rational and $x = 2^k y$ with $k \in \mathbb{Z}$ and $1 \leq y \leq 2$, then these bounds for y and for 2, together with $\log x = k \log 2 + \log y$, give rational lower and upper bounds for $\log x$.

Proof. Integrating the geometric series for $1/(1-z^2)$ gives $\log y = 2\sum_{j\geq 0} z^{2j+1}/(2j+1)$, a series of nonnegative terms. In the tail beginning at $j = N$, every denominator is at least $2N+1$; bounding it by this value and summing the resulting geometric series proves (64). All quantities in the finite sum and its remainder bound are rational when y is rational. For $k < 0$, multiplication by k interchanges the lower and upper bounds on $\log 2$; in either case we obtain rational lower and upper bounds on the logarithm of every rational $x > 0$. □

A.5 The bound in dimension two

In two dimensions a transport with fixed marginals has only one free entry, and its cost is linear in that entry, so an optimal transport places this entry at one end of its feasible interval. For one ordering of the entries of R, a mixture of the optimal transports at the two extreme points has off-diagonal probability $1/L$ in each normalized row, and its entropy is $g(L)$. For the opposite ordering we use the difference between the two marginals: when it is at least one, a single extreme-point cost exceeds $g(L)$, and otherwise an explicit mixture still has entropy at least $g(L)$.

Lemma A.10 (Dimension two)**.** *Let $P = (A, B)$ with $A, B > 2$, and let $L = A + B$. Every row-stochastic matrix $R \in \mathbb{R}^{2\times 2}$ satisfies*

$$K_R(P) \geq g(L) = B_2(L).$$

Proof. First suppose all entries of R are positive, and write

$$R = \begin{pmatrix} r & 1-r \\ s & 1-s \end{pmatrix}.$$

For nonnegative x, y, write

$$\mathcal{J}(x,y) = (x+y)\log(x+y) - x\log x - y\log y, \qquad g(L) = \mathcal{J}(1, L-1),$$

with $0 \log 0 = 0$. The free entry of a transport with fixed marginals is its $(1,1)$ entry, and the coefficient of this entry in the cost is

$$\log \frac{s(1-r)}{r(1-s)}.$$

Hence an optimal transport takes this entry as large as possible when $r \geq s$, and as small as possible when $r \leq s$.

Case $r \geq s$. Optimal transports for the row marginals $(A+1, B-1)$ and $(A-1, B+1)$ of the two extreme points are, respectively,

$$X^+ = \begin{pmatrix} A & 1 \\ 0 & B-1 \end{pmatrix}, \qquad X^- = \begin{pmatrix} A-1 & 0 \\ 1 & B \end{pmatrix}.$$

Let $\lambda = (A-1)/(L-2)$ and $N = \lambda X^+ + (1-\lambda)X^-$. Its row masses are $L\lambda$ and $L(1-\lambda)$, while its off-diagonal entries are λ and $1-\lambda$. Each normalized row therefore has off-diagonal probability $1/L$. By (50) and Gibbs' inequality (49), applied row by row,

$$K_R(P) \geq \langle N, -\log R \rangle \geq \mathcal{H}(N) = Lh_2(1/L) = g(L).$$

Case $r \leq s$. Interchanging the two rows and, simultaneously, the two columns if necessary, we may assume $A \geq B$; put $d = A - B \geq 0$. This relabeling maps (r, s) to $(1-s, 1-r)$, so it preserves $r \leq s$; it also exchanges the two extreme points, and hence leaves $K_R(P)$ unchanged.

If $d \geq 1$, optimal transports at the two extreme points are

$$X^+ = \begin{pmatrix} d+1 & B \\ B-1 & 0 \end{pmatrix}, \qquad X^- = \begin{pmatrix} d-1 & B \\ B+1 & 0 \end{pmatrix}.$$

Their costs $C_\pm = \langle X^\pm, -\log R \rangle$ satisfy $C_+ - C_- = 2\log(s/r) \geq 0$. Dropping the nonnegative term $-(B-1)\log s$ and minimizing the remaining cost $-(d+1)\log r - B\log(1-r)$ over $r \in (0,1)$ gives

$$K_R(P) = C_+ \geq \mathcal{J}(d+1, B).$$

The function $Q(d, B) = \mathcal{J}(d+1, B) - g(d+2B)$ satisfies

$$\partial_d Q = \log\left(1 + \frac{B}{d+1}\right) - \log\left(1 + \frac{1}{d+2B-1}\right) > 0.$$

Also

$$\frac{d}{dB} Q(1, B) = \log\left(1 + \frac{2}{B}\right) - 2\log\left(1 + \frac{1}{2B}\right) > 0 \qquad (B \geq 2).$$

Both inequalities follow by comparing the arguments of the logarithms, after exponentiating in the second. Therefore

$$\mathcal{J}(d+1, B) - g(d+2B) \geq Q(1,2) = 12\log 2 - 5\log 5 = \log\frac{4096}{3125} > 0.$$

It remains to treat $0 \leq d \leq 1$. The optimal transports are now

$$X^+ = \begin{pmatrix} d+1 & B \\ B-1 & 0 \end{pmatrix}, \qquad X^- = \begin{pmatrix} 0 & A-1 \\ A & 1-d \end{pmatrix}.$$

Let $\lambda = (1+d)/2$. The mixture $N = \lambda X^+ + (1-\lambda)X^-$ has the following rows, listed with its two columns interchanged (which does not change its entropy):

$$\left(B - \frac{1}{2} + d - \frac{d^2}{2}, \frac{(1+d)^2}{2}\right), \qquad \left(\frac{(1-d)^2}{2}, B - \frac{1}{2} - \frac{d^2}{2}\right).$$

Again $K_R(P) \geq \mathcal{H}(N)$ by Eq. (50). Define

$$\begin{aligned} \mathcal{E}_2(B, d) = \mathcal{J}&\left(B - \frac{1}{2} + d - \frac{d^2}{2}, \frac{(1+d)^2}{2}\right) \\ &+ \mathcal{J}\left(B - \frac{1}{2} - \frac{d^2}{2}, \frac{(1-d)^2}{2}\right) - g(2B+d). \end{aligned}$$

We show that this entropy difference is nonnegative in two steps: it is nondecreasing in the smaller marginal B, and at $B = 2$ it is a convex function of d that vanishes together with its derivative at $d = 0$.

First, $\mathcal{E}_2$ is nondecreasing in $B \geq 2$ for fixed $d \in [0,1]$. With $L = 2B + d$, differentiation gives

$$\partial_B \mathcal{E}_2 = \log \frac{L^2 - 9d^2}{(L-1-d^2)^2 - d^2} - 2\log \frac{L}{L-1}.$$

All arguments of these logarithms are positive, so after cross-multiplication the derivative has the sign of

$$\begin{aligned}
&(L^2 - 9d^2)(L-1)^2 - L^2\big((L-1-d^2)^2 - d^2\big) \\
&\quad = d^2\left\{L^2(2L - 1 - d^2) - 9(L-1)^2\right\} \\
&\quad \geq d^2(L-1)(2L^2 - 9L + 9) \geq 0,
\end{aligned}$$

because $L \geq 4$ and $d^2 \leq 1$. It therefore suffices to show $\varphi(d) := \mathcal{E}_2(2, d) \geq 0$.

A direct computation gives $\varphi(0) = \varphi'(0) = 0$ and, for $0 \leq d < 1$,

$$\varphi''(d) = \log \frac{(3-d)(3-d^2)}{(1+d)(1-d)^2} - \frac{2}{3-d} + \frac{1}{2-d} - \frac{6}{3-d^2} + \frac{1}{(d+3)(d+4)}. \tag{65}$$

At the origin,

$$\varphi''(0) = \log 9 - \frac{25}{12} > 0.$$

Here the lower bound in (64) with $y = 3$ and $N = 2$ (two terms of the series) gives $\log 3 > 13/12$ and therefore $\log 9 > 25/12$.

We split the logarithm in Eq. (65) as

$$\log \frac{(3-d)(3-d^2)}{(1+d)(1-d)^2} = \log 9 + \log \frac{1 - d/3}{1-d} + \log \frac{1 - d^2/3}{1-d^2}.$$

Since $\log[(1-x/3)/(1-x)] \geq 2x/3$ for $0 \leq x < 1$ (the left side vanishes at $x = 0$, and its derivative $1/(1-x) - 1/(3-x) = 2/((1-x)(3-x))$ is at least $2/3$ on $[0,1)$), and

$$-\frac{2}{3-d} + \frac{2}{3} \geq -\frac{d}{3}, \qquad -\frac{6}{3-d^2} + 2 \geq -d^2,$$
$$\frac{1}{2-d} - \frac{1}{2} \geq \frac{d}{4}, \qquad \frac{1}{(d+3)(d+4)} - \frac{1}{12} \geq -\frac{d}{18},$$

we obtain

$$\varphi''(d) \geq \log 9 - \frac{25}{12} + \frac{19d}{36} - \frac{d^2}{3} \geq \log 9 - \frac{25}{12} + \frac{7d}{36} > 0.$$

Thus $\varphi''(d) > 0$ on $[0,1)$, so $\varphi(0) = \varphi'(0) = 0$ implies $\varphi(d) \geq 0$ there; by continuity (recall $0 \log 0 = 0$), this extends to $d = 1$. Hence $\mathcal{E}_2(B, d) \geq 0$, and the bound is proved when all entries of R are positive.

For an arbitrary row-stochastic R, apply the bound to the positive smoothings in Eq. (63) and let the smoothing parameter tend to zero. This gives $K_R(P) \geq g(L)$. Equality holds for the matrix with diagonal entries $1 - 1/L$ and off-diagonal entries $1/L$: its costs at both extreme points equal $g(L)$. □

A.6 Averaging over extreme points

For a fixed source, the optimal transports at the extreme points differ only in their excess over the common part (Lemma A.8). We use this to split an average of their costs into two contributions: an entropy term coming from the choice of the sink, and the cost of the mass that each row places outside a selected entry. The first contribution equals $\log \kappa$ when all κ possible sinks are selected rows.

For a nonnegative vector a, write

$$\mathcal{J}(a) = |a| \log |a| - \sum_j a_j \log a_j, \qquad |a| = \sum_j a_j. \tag{66}$$

Like $\mathcal{H}$, this function is concave and homogeneous. We also write $\mathcal{J}(a_1, \dots, a_m)$ for its value at the vector with these entries, and put $h(x) = -x \log x$.

Let λ be a probability distribution with positive weights on a finite set of labels (labels of probability zero can be discarded). For a vector $x = (\lambda_v a_v)_v$ with ratios $a_v \geq 0$, define

$$\Phi_\lambda(x) = \int_0^\infty h\left(\sum_{v : a_v > z} \lambda_v\right) dz. \tag{67}$$

This quantity measures how unequal the ratios are: at each level z, the integrand is the entropy summand of the labels whose ratio exceeds z. If all a_v are equal, it is zero. It is unchanged when the same amount is added to every ratio (translation invariance), and its value at a point mass is given by the second identity:

$$\Phi_\lambda(x + c\lambda) = \Phi_\lambda(x), \qquad \Phi_\lambda(\lambda_i \mathbf{e}_i) = h(\lambda_i). \tag{68}$$

The first identity holds whenever both vectors are nonnegative.

Lemma A.11 (Splitting inequality)**.** *The function Φ_λ is convex, homogeneous and subadditive. For nonnegative vectors $x^1, \dots, x^m$,*

$$\sum_a \Phi_\lambda(x^a) - \Phi_\lambda\left(\sum_a x^a\right) \leq \mathcal{J}((|x^a|)_a). \tag{69}$$

Proof. We first write the integral as a maximum of linear functions. Let

$$\mathcal{B}_\lambda = \left\{ b : \sum_v \lambda_v b_v = 0, \quad \sum_{v \in A} \lambda_v b_v \leq h(\lambda(A)) \text{ for every } A \right\}. \tag{70}$$

Then

$$\Phi_\lambda(x) = \max_{b \in \mathcal{B}_\lambda} b \cdot x.$$

To see this, order the ratios $a_v = x_v/\lambda_v$ decreasingly. Summation by parts bounds $b \cdot x$ by the integral in (67). Equality is attained by assigning the successive differences of h along that ordering:

$$\lambda_{\pi(j)} b_{\pi(j)} = h\left(\sum_{l \leq j} \lambda_{\pi(l)}\right) - h\left(\sum_{l < j} \lambda_{\pi(l)}\right).$$

Since h is concave, these differences decrease as the preceding mass increases, which gives every subset inequality in (70). Convexity, homogeneity and subadditivity follow from the representation as a maximum.

For the splitting bound, discard the zero vectors, choose maximizers b^a in this representation, and set

$$w_a = \frac{|x^a|}{\sum_b |x^b|}, \qquad z_v = \max_a \{b_v^a + \log w_a\}.$$

For any subset A, partition its labels according to an index a attaining this maximum. Write $\mu_a = \lambda(A_a)$ and $\mu = \lambda(A)$. The subset inequalities in Eq. (70) and the log-sum inequality give

$$\sum_{v \in A} \lambda_v z_v \leq \sum_a \{h(\mu_a) + \mu_a \log w_a\} \leq h(\mu).$$

Therefore z can be increased coordinatewise to some $\widetilde{b} \in \mathcal{B}_\lambda$. At a maximal vector satisfying these inequalities, every coordinate belongs to a subset with equality. By the decreasing differences used above, the union of two such subsets again has equality, so the whole label set has equality, and its bound is $h(1) = 0$. Hence

$$\widetilde{b} \cdot x^a \geq \Phi_\lambda(x^a) + |x^a| \log w_a.$$

Summing this inequality and using $\widetilde{b} \cdot \sum_a x^a \leq \Phi_\lambda(\sum_a x^a)$ leaves the term $-\sum_a |x^a| \log w_a = \mathcal{J}((|x^a|)_a)$. □

For an integer $\kappa \geq 2$, put $\nu = 1/\kappa$ and define the continuous, piecewise affine function χ_κ on $[0, 1-\nu]$ by

$$\chi_\kappa(j/\kappa) = \frac{g(j+1)}{\kappa}, \qquad 0 \leq j \leq \kappa - 1. \tag{71}$$

This function is increasing and concave. For $\kappa = 1$ its domain is $\{0\}$ and $\chi_1(0) = 0$. In Lemma A.12 it bounds the loss in the entropy term when a selected row is also a sink.

Lemma A.12 (Fixed-source inequality)**.** *Fix a source u and a nonempty set of sinks $V \subseteq [n] \setminus \{u\}$, with $|V| = \kappa$ and $\nu = 1/\kappa$. Let $I \subseteq [n]$ be any set of rows, called selected, and let $f : I \to [n]$ be an injective map that assigns to each selected row the column of its selected entry. Assume all extreme-point costs are finite and choose the optimal transports at the extreme points by the rule* (52)*. For each $i \in I$, assume $z_i := R_{i,f(i)} \in (0,1)$ and set*

$$a_i = -\log z_i, \qquad \mu_i = \log \frac{z_i}{1 - z_i},$$
$$r_i = P_i + \mathbb{1}_{\{i=u\}} - \nu \mathbb{1}_{\{i \in V\}}, \qquad m_i = \nu \sum_{v \in V} \sum_{j \neq f(i)} X_{ij}^{uv}.$$

For a selected row that is also a sink, $i \in I \cap V$, write

$$b_i = \sum_{j \neq f(i)} X_{ij}^{ui}, \qquad d_i^v = \sum_{j \neq f(i)} (X_{ij}^{uv} - X_{ij}^{ui}), \qquad \eta_i = \nu \sum_{v \in V} d_i^v.$$

Then $d_i^i = 0$, $0 \leq d_i^v \leq 1$, $m_i = b_i + \eta_i$, and $0 \leq \eta_i \leq 1 - \nu$.

Let λ be the uniform distribution on V. For each unselected row $i \notin I$, define its remainder

$$\mathcal{Q}_i^{u,V} = \nu \sum_{v \in V} \sum_j X_{ij}^{uv} (-\log R_{ij}) - \sum_j \Phi_\lambda((\nu X_{ij}^{uv})_{v \in V}). \tag{72}$$

Then

$$\begin{aligned} \nu \sum_{v \in V} C_{uv}(R, P) \geq |I \cap V| h(\nu) + \sum_{i \in I} (r_i a_i + \mu_i m_i) - \sum_{i \in I \cap V} \chi_\kappa(\eta_i) \\ + \sum_{i \notin I} \mathcal{Q}_i^{u,V}. \end{aligned} \tag{73}$$

The first term is $\log \kappa$ when $V \subseteq I$. For an unselected row, put $c_i = P_i + \mathbb{1}_{\{i=u\}}$ and $\epsilon_i = \mathbb{1}_{\{i \in V\}}$. Then

$$\mathcal{Q}_i^{u,V} \geq (c_i - 1)(-\log \max_j R_{ij}) + \min_{\substack{k_j \in \mathbb{Z}_{\geq 0} \\ \sum_j k_j = \kappa - \epsilon_i}} \sum_j \frac{k_j}{\kappa} \log \frac{k_j/\kappa}{R_{ij}}. \tag{74}$$

In this minimum, positive counts at zero entries of R are not allowed. With $s_i = 1 - \nu\epsilon_i$ and $0 \log 0 = 0$, this implies

$$\mathcal{Q}_i^{u,V} \geq (c_i - 1)(-\log \max_j R_{ij}) + s_i \log s_i. \tag{75}$$

In particular $\mathcal{Q}_i^{u,V} \geq 0$ for $i \notin V$.

Proof. Extend f to a column permutation σ and put

$$S_i = P_{\sigma(i)}. \tag{76}$$

Relabel the columns accordingly, so that the selected entries lie on the diagonal, and remove the diagonal entry from each selected row of X^{uv}, obtaining T^v. Then

$$\sum_j T_{ij}^v - \sum_j T_{ji}^v = P_i - S_i + \mathbb{1}_{\{i=u\}} - \mathbb{1}_{\{i=v\}}. \tag{77}$$

Put $x_{ij} = (\nu T_{ij}^v)_{v \in V}$, $O_i = \sum_j x_{ij}$ and $I_i' = \sum_j x_{ji}$. By translation invariance, $\Phi_\lambda(I_i') - \Phi_\lambda(O_i)$ is zero off V and equals $\Phi_\lambda(O_i + \nu \mathbf{e}_i) - \Phi_\lambda(O_i)$ on V. For an unselected i, column i contains no removed entry, so $I_i' = S_i \lambda$ and $\Phi_\lambda(I_i') = 0$.

For a selected sink, Lemma A.8 gives $0 \le d_i^v \le 1$ and $d_i^i = 0$. Translation invariance removes the common part outside the selected entry, and Eq. (67), with $N_i(z) = \#\{v \ne i : d_i^v > z\}$, gives

$$\Phi_\lambda(I_i') - \Phi_\lambda(O_i) = h(\nu) - \nu \int_0^1 g(N_i(z) + 1)\, dz \ge h(\nu) - \chi_\kappa(\eta_i). \tag{78}$$

The inequality follows from concavity of the interpolation (71), since $\int_0^1 N_i(z)\, dz = \kappa\eta_i$. For each selected row, the log-sum inequality applied to the entries other than the selected one bounds the mean cost of the row from below by $r_i a_i + \mu_i m_i + \mathcal{J}((\nu \sum_v T_{ij}^v)_{j \ne i})$. Applying Lemma A.11 in each row and subadditivity in each column gives

$$\nu \sum_v C_{uv}(R, P) \ge \sum_{i \in I} \{r_i a_i + \mu_i m_i + \Phi_\lambda(I_i') - \Phi_\lambda(O_i)\} + \sum_{i \notin I} \mathcal{Q}_i^{u,V}. \tag{79}$$

Substituting Eq. (78) proves (73) for every I.

An unselected row has a common part of mass $c_i - 1$ and, in addition, $\kappa - \epsilon_i$ excess units. Translation invariance removes the common part from each term of the second sum in (72); what remains is concave in the excess. Its minimum is therefore attained when each excess unit lies in a single column with a positive entry of the row. If k_j units lie in column j, the remainder is $\sum_j (k_j/\kappa) \log((k_j/\kappa)/R_{ij})$. The cost of the common part is at least $(c_i - 1)(-\log \max_j R_{ij})$. This proves Eq. (74). The log-sum inequality, applied to counts of total mass s_i, gives the lower bound $s_i \log s_i$ in Eq. (75). □

If a selected entry equals one, finiteness of the extreme-point costs forces every other entry of that row to be zero in each of the optimal transports. Its excess outside the selected entry and its terms in the entropy differences then vanish, and so does its contribution as a selected row. With this contribution set to zero, the fixed-source inequality still follows from (79).

For $r > 0$ and $m \ge 0$, define

$$\begin{aligned} \mathcal{A}(r, m) &= \inf_{1/2 \le z < 1} \left\{ -r \log z + m \log \frac{z}{1 - z} \right\} \\ &= \begin{cases} r \log r - (r - m) \log(r - m) - m \log m, & 0 \le m \le r/2, \\ r \log 2, & m \ge r/2. \end{cases} \end{aligned} \tag{80}$$

Differentiating in z shows that the infimum is attained at $z = 1 - m/r$ when $0 < m \le r/2$, is approached as $z \to 1$ when $m = 0$, and is attained at $z = 1/2$ when $m \ge r/2$. As an infimum of affine functions of (r, m) with nonnegative coefficients, $\mathcal{A}$ is concave, homogeneous, and nondecreasing in both arguments.

Lemma A.13 (Averaging sources for a permutation)**.** *Let f be a permutation with $R_{i,f(i)} \ge 1/2$ for every i. Put*

$$\kappa = n - 1, \quad \nu = 1/\kappa, \quad D = L - 1 - \nu, \quad \omega_i = \frac{P_i - \nu}{D}.$$

Choose the source u with probability ω_u and then its sink uniformly among the other rows. The mean row masses are

$$P_i + \omega_i - \nu(1 - \omega_i) = L\omega_i.$$

For $P = P_i$, $T = P_{f(i)}$, $d = P - T$ and $\omega = \omega_i$, define

$$\begin{aligned} U_\kappa(P, T, \omega) &= \min_{s \le \eta \le 1 - \nu} \{\mathcal{A}(L\omega, m_d(\eta)) - (1 - \omega)\chi_\kappa(\eta)\}, \\ s &= (1 - \nu) \min\{d_+, 1\}, \\ m_d(\eta) &= \omega(1 + d)_+ + (1 - \omega)\{(d - 1)_+ + \eta\}. \end{aligned} \tag{81}$$

Then every $2 \le s' \le n$ satisfies

$$T_{s'}(R, P) \ge \log(n - 1) + \sum_i U_{n-1}(P_i, P_{f(i)}, \omega_i). \tag{82}$$

If $R_{i,f(i)} \ge \theta \ge 1/2$, the same lower bound holds with $\mathcal{A}$ replaced by the infimum over $z \ge \theta$.

Proof. If an extreme-point cost is infinite, every $T_{s'}$ is infinite and the conclusion follows. Otherwise choose the finite optimal transports by the rule Eq. (52). Apply Lemma A.12 with every row selected and average with probabilities ω_u. The entries of R do not depend on the source, so after averaging, the linear term $r_i a_i + \mu_i m_i$ of each row has r_i replaced by its mean $L\omega_i$ and m_i by the mean mass outside the selected entry. Concavity of χ_κ bounds the mean entropy correction by $(1-\omega_i)\chi_\kappa(\eta_i)$, where η_i is the mean excess over the sources other than the row itself.

It remains to bound from below the mass outside the selected column. The column of the selected entry receives total mass equal to its marginal in every transport (its capacity). Hence, when the row is the source, at least $(1+d)_+$ units lie outside that column. When the row is the sink, the corresponding bound is $(d-1)_+$; at the other extreme points it is d_+. By Lemma A.8, the mean over the sources other than the row itself is therefore at least

$$(d-1)_+ + \max\{\eta_i, (1-\nu)\min(d_+, 1)\}.$$

On $0 \le \eta_i \le s$ this bound is constant while χ_κ increases, so replacing η_i by s can only weaken the lower bound. On $[s, 1-\nu]$ it equals $m_d(\eta_i)$ in Eq. (81). Minimizing over the possible values of the selected entry gives the bound for each row. Finally, for each source the average over all its sinks is at most $T_{s'}$, as observed after Eq. (51). □

A.7 Dimensions at least 53

Let $n \ge 53$, $P_i > 2$, and $L = \sum_i P_i$. We prove $K_R(P) \ge g(L) + \log(n-1)$ for every positive row-stochastic matrix R, using the optimal transports X^{uv} chosen by (52). Matrices with zero entries then follow from the limit (63).

The argument has two steps. First, we average Lemma A.12 over all sources, with source probabilities chosen so that the mean mass of row i is proportional to its source probability. This bounds $K_R(P)$ below by $\log(n-1)$ plus a sum of row contributions. Second, we consider the columns in which the rows attain their maxima. If these columns are distinct, the row contributions satisfy a quadratic identity in which the losses cancel. If several row maxima lie in one column, the column marginal limits the mass that column can receive, so part of the mass of these rows lies outside their maximal entries; the resulting contribution compensates for the loss in that identity.

A.7.1 Source weights and row estimates

Define

$$\kappa = n-1, \qquad \nu = \kappa^{-1}, \qquad \tau_\kappa = 1-\nu, \qquad D = L - 1 - \nu, \qquad \omega_i = \frac{P_i - \nu}{D}. \tag{83}$$

Then $\omega_i > 0$, $\sum_i \omega_i = 1$, and $L > 2n \ge 106$. Draw the source u with probability ω_u and the sink uniformly from the other $n-1$ rows. The mean transport is

$$\begin{aligned} M &= \sum_u \omega_u \nu \sum_{v \ne u} X^{uv}, \\ M\mathbf{1} &= L\omega, \qquad M^\top \mathbf{1} = P, \\ P_i - \nu + (1+\nu)\omega_i &= L\omega_i. \end{aligned} \tag{84}$$

The last identity explains the choice of ω: row i gains one unit with probability ω_i and loses one with probability $\nu(1-\omega_i)$, so its mean mass is $L\omega_i$. Since M is an average of the optimal transports X^{uv}, $K_R(P) \ge \langle M, -\log R \rangle$.

Set

$$\theta = \frac{5}{8}, \qquad \gamma_* = \frac{11}{25}, \qquad \ell_* = \log\frac{8}{5}, \qquad \mu_0 = \log\frac{5}{3}.$$

When the selected entry of a row is at least θ, we use the following restricted minimization:

$$\begin{aligned} \mathcal{A}_\theta(r,m) &= \inf_{\theta \le z < 1} \left\{ -r\log z + m \log\frac{z}{1-z} \right\} \\ &= \begin{cases} r\log r - (r-m)\log(r-m) - m\log m, & 0 \le m \le 3r/8, \\ r\ell_* + m\mu_0, & m \ge 3r/8. \end{cases} \end{aligned} \tag{85}$$

Here r is the row mass and m is the mass outside the selected entry. The entropy expression applies when the minimizer $1 - m/r$ is at least θ; otherwise the infimum is attained at θ and is linear. The linear expression also holds for $m > r$. As an infimum of linear functions of (r, m), $\mathcal{A}_\theta$ is concave and homogeneous. It is nondecreasing in each argument, and $\partial_m \mathcal{A}_\theta \geq \mu_0 > \gamma_*$. The function χ_κ is the one defined in (71); it is increasing and concave, and $\chi_\kappa(\tau_\kappa) = h_2(\nu)$.

The following rational bounds follow from (64) with $N = 12$, after reducing the arguments of the logarithms to $[1, 2]$ by powers of two:

$$\ell_* > .47, \qquad .5108 < \mu_0 < .511, \qquad h_2(1/52) < .0951, \qquad g(106) < 5.65871,$$
$$\frac{g(L)}{L - 1 - \nu} < .054 \quad (L \geq 106,\ 0 < \nu \leq 1/52).$$

The last bound follows from evaluation at the ends of these ranges and the strict decrease of $g(L)/(L-c)$ for $L \geq 106$ and $1 \leq c \leq 53/52$: its derivative has numerator $(L-c)g'(L) - g(L) < 1 - g(L) < 0$.

In the next lemma the selected rows have distinct selected columns, and every unselected row contributes its remainder $\mathcal{Q}_i^{u,V_u}$ from (72). These remainders are needed below when several row maxima lie in one column.

Lemma A.14 (Averaging over sources and sinks)**.** *Let ω be the source distribution and M the mean matrix in* (84)*, and let $V_u = [n] \setminus \{u\}$. Let $f : I \to [n]$ be an injection on a set of rows I, with $z_i = R_{i,f(i)} \geq \theta$ for $i \in I$. For a selected row define*

$$a_i = -\log z_i, \qquad \mu_i = \log \frac{z_i}{1 - z_i},$$
$$r_i^{(u)} = P_i - \nu + (1+\nu)\mathbb{1}_{\{i=u\}}, \qquad m_i^{(u)} = \nu \sum_{v \neq u} \sum_{j \neq f(i)} X_{ij}^{uv},$$
$$\eta_i^{(u)} = \nu \sum_{v \neq u} \sum_{j \neq f(i)} (X_{ij}^{uv} - X_{ij}^{ui}) \quad (u \neq i), \qquad \eta_i^{(i)} = 0.$$

The row contributions are

$$\begin{aligned} \mathcal{P}_i &= \sum_u \omega_u \left[r_i^{(u)} a_i + \mu_i m_i^{(u)} - \mathbb{1}_{\{i \neq u\}} \chi_\kappa(\eta_i^{(u)}) \right] \quad (i \in I), \\ \mathcal{Q}_i &= \sum_u \omega_u \left[\mathcal{Q}_i^{u,V_u} - \mathbb{1}_{\{i \neq u\}} h(\nu) \right] \qquad\qquad (i \notin I), \end{aligned} \tag{86}$$

where $\mathcal{Q}_i^{u,V_u}$ is the remainder in (72) *and $h(\nu) = -\nu \log \nu$. These contributions satisfy*

$$K_R(P) \geq \log \kappa + \sum_{i \in I} \mathcal{P}_i + \sum_{i \notin I} \mathcal{Q}_i. \tag{87}$$

For a selected row put

$$T_i = P_{f(i)}, \qquad d_i = P_i - T_i, \qquad m_i = \sum_{j \neq f(i)} M_{ij}, \qquad s_i = \tau_\kappa \min\{(d_i)_+, 1\}.$$

Then there is $\eta_i \in [s_i, \tau_\kappa]$ such that

$$\begin{aligned} &\mathcal{P}_i \geq \mathcal{A}_\theta(L\omega_i, m_i) - (1 - \omega_i)\chi_\kappa(\eta_i), \\ &m_i \geq \omega_i(1 + d_i)_+ + (1 - \omega_i)\big(((d_i)_+ - 1)_+ + \eta_i\big). \end{aligned} \tag{88}$$

For an unselected row let $a_i = -\log \max_j R_{ij}$. Then

$$\mathcal{Q}_i \geq (P_i - 1 + \omega_i)a_i - (1 - \omega_i)h_2(\nu). \tag{89}$$

If the largest entry of an unselected row exceeds $5/8$, let m_i be the mean mass of the row outside that entry. Then

$$\mathcal{Q}_i \geq \gamma_* m_i - (1 - \omega_i)h_2(\nu). \tag{90}$$

Proof. For each source u we apply Lemma A.12. Each selected row in V_u contributes $h(\nu)$. We add the same quantity for every unselected row in V_u and subtract it from that row's remainder. Since there are κ rows in V_u, the resulting constant is $\kappa h(\nu) = \log \kappa$. Averaging over u gives Eq. (87).

For an unselected row $i \neq u$, the common part has mass $P_i - 1$ and the excess has mean mass τ_κ. Hence (75) gives

$$\mathcal{Q}_i^{u,V_u} - h(\nu) \geq (P_i - 1)a_i + \tau_\kappa \log \tau_\kappa - h(\nu) = (P_i - 1)a_i - h_2(\nu).$$

For the source row, the common part has mass P_i and the excess has mean mass one, which gives $P_i a_i$. Averaging these bounds proves (89).

For an unselected row with maximum $z > 5/8$, we also need a bound in terms of the mass outside that entry. We subtract γ_* times this mass before minimizing over the counts in (74). Each unit of the common part still contributes at least $a = -\log z$, since $\log(z/(1-z)) > \gamma_*$. For the excess, the subtraction replaces each entry outside the maximizing entry by $e^{\gamma_*} R_{ij}$. The modified entries have sum $z + e^{\gamma_*}(1 - z)$. By the log-sum inequality, the excess part of the remainder of a row $i \neq u$ is at least

$$\tau_\kappa \log \tau_\kappa - \tau_\kappa \log\{z + e^{\gamma_*}(1 - z)\} \geq \tau_\kappa \log \tau_\kappa - \tau_\kappa a.$$

The inequality holds because $e^{\gamma_*} < 2$ and $z\{z + e^{\gamma_*}(1 - z)\} \leq 2z - z^2 \leq 1$. Since the common part has mass $P_i - 1 > 1 \geq \tau_\kappa$, it compensates for the subtracted term $\tau_\kappa a$. For the source row, the excess has mass one and the common part has mass $P_i > 2$. Adding back γ_* times the mass outside the maximizing entry and averaging gives Eq. (90).

Finally, the selected column of row i has total mass T_i in every transport, whereas the row has mass $P_i + 1$ when it is the source, $P_i - 1$ when it is the sink, and P_i otherwise. This and Lemma A.8 imply

$$m_i^{(i)} \geq (1 + d_i)_+, \qquad m_i^{(u)} \geq ((d_i)_+ - 1)_+ + \max\{\eta_i^{(u)}, s_i\} \quad (u \neq i).$$

Set

$$\bar{\eta}_i = \frac{1}{1 - \omega_i} \sum_{u \neq i} \omega_u \eta_i^{(u)}, \qquad \eta_i = \max\{\bar{\eta}_i, s_i\}.$$

Convexity of the maximum gives the claimed lower bound on m_i. Concavity and monotonicity of χ_κ give

$$\sum_{u \neq i} \omega_u \chi_\kappa(\eta_i^{(u)}) \leq (1 - \omega_i)\chi_\kappa(\eta_i).$$

The remaining terms are $L\omega_i(-\log z_i) + m_i \log(z_i/(1 - z_i))$. Taking the infimum over $z_i \geq \theta$ gives (88). □

For $0 \leq j \leq \kappa - 2$, the slope of χ_κ on $[j/\kappa, (j+1)/\kappa]$ is $g(j+2) - g(j+1)$. Consequently

$$\chi'_\kappa \leq \log 4, \qquad \chi'_\kappa \leq \log \frac{j+1}{j} \text{ on } \Big[\frac{j}{\kappa}, \frac{j+1}{\kappa}\Big] \quad (j \geq 1). \tag{91}$$

The next lemma attaches a coefficient c_i to each selected row and bounds its contribution below by $g(L)c_i$ plus a nonnegative linear term in the mass outside its selected entry beyond c_i. This linear term is used when several rows have the same maximizing column.

For the row estimates below, fix a selected row, let P be its marginal and T the marginal of its selected column. Define

$$d = P - T, \qquad \omega = \frac{P - \nu}{D}, \qquad r = L\omega, \qquad G = g(L), \qquad c = \min\{1, \omega(1 + d)_+\},$$
$$s = \tau_\kappa \min\{d_+, 1\}, \qquad m_d(\eta) = \omega(1 + d)_+ + (1 - \omega)\big((d_+ - 1)_+ + \eta\big), \qquad s \leq \eta \leq \tau_\kappa.$$

Here P and T are entries of the same marginal vector $(P_1, \ldots, P_n)$; when $d \neq 0$, they are distinct entries.

Lemma A.15 (Lower bounds for the individual rows)**.** *For a selected row, let $\overline{m}_i$ denote its mean mass outside its selected column, and put $c_i = \min\{1, \omega_i(1 + d_i)_+\}$. Then*

$$\mathcal{P}_i \geq \gamma_* \overline{m}_i + (G - \gamma_*)c_i. \tag{92}$$

For an unselected row with $\max_j R_{ij} \leq 5/8$, put $z = P_i - 1$ and $f_\nu(z) = z^2/2 + (2 - \nu)z - 1/2$. Then

$$\mathcal{Q}_i \geq G \min\{1, f_\nu(z)/D\}.$$

Proof. For the first assertion we use the notation defined before the lemma, applied to this row, and write $\overline{m} = \overline{m}_i$. By Eq. (88), there is $\eta \in [s, \tau_\kappa]$ with $\bar{m} \ge m_d(\eta)$. Since $\partial_m \mathcal{A}_\theta \ge \mu_0 > \gamma_*$, it suffices to prove

$$\mathcal{A}_\theta(r, m_d(\eta)) - \gamma_* m_d(\eta) - (1-\omega)\chi_\kappa(\eta) \ge (G - \gamma_*)c. \tag{93}$$

Minimization over the excess. When $d < 1$, the left side of (93) is nondecreasing in the excess, so it suffices to take the smallest value allowed by the column capacities, namely $\eta = s$. The derivative of this left side is $(1-\omega)(\partial_m \mathcal{A}_\theta - \gamma_* - \chi'_\kappa)$. On the intervals with index $j \ge 14$, (91) gives $\chi'_\kappa \le \log(15/14) < \mu_0 - \gamma_*$. On the earlier intervals, $L\omega > 2 - \nu$ and $1 + d \le 1 + \eta/\tau_\kappa$ when $d \ge 0$ (with a stronger bound when $d < 0$) give

$$\frac{m_d(\eta)}{r} \le \frac{1 + \eta/\tau_\kappa}{L} + \frac{\eta}{2-\nu} \le \begin{cases} (1 + 1/51)/106 + 1/103 < 3/100, & j = 0, \\ (1 + 14/51)/106 + 14/103 < 3/20, & 1 \le j \le 13. \end{cases}$$

Hence the first case of Eq. (85) applies, and $\partial_m \mathcal{A}_\theta$ exceeds $\log(97/3) > \gamma_* + \log 4$ and $\log(17/3) > \gamma_* + \log 2$ in the two cases, respectively. By continuity the monotonicity extends across the ends of the intervals. Adding $\gamma_* m_d$ shows that it also holds for the expression without the subtracted linear term.

An estimate at $\omega(1+d) = 1$. For $d \ge 1$, the proof below minimizes a function that is concave in ω on each side of this point. We first show that this function is positive at the point itself. Set

$$z_d = d^2 + (1-\nu)d + 1, \qquad L_0(d) = d^2 + (3-\nu)d + 3,$$

and

$$\Theta(L, d, \nu) = \mathcal{A}_\theta(L, z_d) - (d+1)g(L) - \gamma_* d(d+1) - d h_2(\nu).$$

Then

$$\Theta(L, d, \nu) > 0 \quad \text{if } L \ge 106,\ L \ge L_0(d),\ 0 < \nu \le 1/52. \tag{94}$$

For fixed d, ν, Θ increases with L. In the first case of (85), convexity of $-\log(1-x)$ gives $\partial_L \Theta \ge (z_d - d - 1)g'(L) > 0$; in the second case,

$$\partial_L \Theta > \ell_* - (d+1)/(L-1) \ge \ell_* - 1/(d + 2 - \nu) > 0.$$

It decreases with ν, since z_d decreases and $h_2(\nu)$ increases. It therefore suffices to take $\nu_0 = 1/52$ and $L = \max\{106, L_0(d)\}$; the positive solution of $L_0(d_*) = 106$ satisfies $8 < d_* < 44/5$.

At $L = 106$, it suffices to consider the two ends of the interval $[1, d_*]$. Put $a = 1 - \nu_0$, $b_0 = 4 - a^2 > 0$, and $z = d^2 + ad + 1$. In the first case of (85), $z \le 3L/8$, and

$$\frac{\partial^3 \Theta}{\partial d^3} = -\frac{L(2d+a)\{2z(L+z) + b_0(L-2z)\}}{z^2(L-z)^2} < 0.$$

Hence $\partial_d \Theta$ is concave, and it is positive at the left end: $\log(103/3) > 7/2$ gives $\partial_d \Theta(106, 1, \nu_0) > 3.3$. The set where it is positive is therefore an interval, so Θ first increases and then possibly decreases in this case. In the second case,

$$\partial_d \Theta = (\mu_0 - \gamma_*)(2d+1) - \mu_0 \nu_0 - g(106) - h_2(\nu_0) < 0 \qquad (d \le d_* < 9),$$

using $\mu_0 < .511$ and $g(106) > 5$. Hence the minimum over $[1, d_*]$ is attained at an end of the interval. At $d = 1$, with $z = 155/52$,

$$\mathcal{A}_\theta(106, z) \ge z \log(106/z) + z - z^2/106$$

and $\log 106 > 4.6$, $\log 3 < 1.1$ give $\Theta > .9$. At the other end the second case of (85) applies, and

$$\Theta(106, d_*, \nu_0) > .0708 d_*^2 - 5.695 d_* + 44.67 > \frac{2297}{62500} > 0,$$

where the quadratic decreases on $[1, 44/5]$ and its value at $44/5$ is $2297/62500$. For $d \ge d_*$ we have $\max\{106, L_0(d)\} = L_0(d)$, and at $L = L_0(d)$ the second case still applies. Writing the value there as $\Theta_0(d) = a_2 d^2 + a_1 d + a_0 - (d+1)g(L_0(d))$, one has $a_2 > .5408$ and $a_1 > 1.36$. Since $L_0(d) < (d+2)^2$ and $(d+1)L_0'(d)/(L_0(d) - 1) < 2$, this function has derivative

$> 1.08d - 1.64 - 2\log(d+2) > 0$ for $d \geq 8$. The last inequality holds at 8, because $\log 10 < 2.31$, and its left side increases thereafter. The value at d_* is positive by the preceding bound.

Rows with $P > T$. For $d > 0$ we prove the following bound:

$$\begin{aligned} U_\theta(P,T,\omega) &:= \min_{s \leq \eta \leq \tau_\kappa} \{\mathcal{A}_\theta(L\omega, m_d(\eta)) - (1-\omega)\chi_\kappa(\eta)\} \\ &\geq G\min\{1, \omega(1+d)\} + \gamma_* d. \end{aligned} \tag{95}$$

For $d \geq 1$, $s = \tau_\kappa$ and $m_d = d - \nu + (1+\nu)\omega$. The difference between the two sides of (95) is concave in ω on each side of $\omega_* = (1+d)^{-1}$: both arguments of the first term on the left are affine, the subtracted term is affine, and $\min\{1, \omega(1+d)\}$ is affine on each side. We prove the bound for all ω in $[\omega_{\min}, 1]$, where $\omega_{\min} = (d+2-\nu)/D$; this interval contains every feasible value of ω. The point ω_* belongs to this interval if and only if $L \geq L_0(d)$; there $m_d/\omega_* = z_d$, so by homogeneity (94) gives the bound at this point. At $\omega = 1$, concavity of $\mathcal{A}_\theta(L, \cdot)$ on the interval between the masses 1 and $L-1$ gives

$$\mathcal{A}_\theta(L, 1+d) - G \geq d\frac{L\log 2 - G}{L-2} > \gamma_* d,$$

because $d < L-2$, $G/L < .054$, and $\log 2 > .69$; here $\mathcal{A}_\theta(L, L-1) = L\log(8/3) - \log(5/3) \geq L\log 2$. At $\omega_{\min}$ one has $m_d = L\omega_{\min} - 2$. While $\omega_{\min}(1+d) \leq 1$, both arguments of the first term are affine in d, and the term $G(d+2-\nu)(d+1)/D$ subtracted from it is convex. The difference is therefore concave in d, and its minimum occurs at $d = 1$ or where $\omega_{\min}(1+d) = 1$. At the latter point the bound follows from (94), as above; at $d = 1$,

$$\mathcal{A}_\theta(L\omega_{\min}, m_d) \geq 2(1-\nu)\log 2, \qquad \text{difference} > 2(51/52)(.69) - .324 - .1 - .44 > .489.$$

Once $\omega_{\min}(1+d)$ reaches one, the first term on the right side of Eq. (95) is constant. Here $d > 8$, the second case of (85) applies to $\mathcal{A}_\theta$, and the difference has positive derivative $\{L(\ell_* + \mu_0) + h_2(\nu)\}/D - \gamma_*$.

For $0 < d < 1$, the monotonicity in the excess proved above gives $\eta = \tau_\kappa d$, and hence $m_d = \tau_\kappa d + (1+\nu d)\omega$. Since P and $T = P - d$ are distinct coordinates of the vector of marginals, $L > P + T + 2(n-2)$ gives $\omega < 1/2$. Concavity in ω therefore reduces the estimate to the two ends $\omega_{\min} = (2+d-\nu)/D$ and $1/2$. At $1/2$, we use $\chi_\kappa(\tau_\kappa d) \leq 2\tau_\kappa d\log 2$ and concavity in d. The difference between the two sides is at least

$$\frac{d}{2}\{\mathcal{A}_\theta(L, 3-\nu) - 2g(L) - 2\gamma_* - 2(1-\nu)\log 2\} > 0. \tag{96}$$

The point $(L, 3-\nu)$ lies in the first case of (85). There the bracket increases in L because $\partial_L \mathcal{A}_\theta \geq (3-\nu)g'(L)$, and decreases in ν because $\partial_m \mathcal{A}_\theta > 2\log 2$. At $(106, 1/52)$ it exceeds .0237, as one sees by taking $q = 149/50$ and using

$$\mathcal{A}_\theta(106, q) \geq q\log(106/q) + q - \frac{q^2}{212} - \frac{q^3}{6\cdot 106^2(1-q/106)},$$

with $\log(5300/149) > 3.5715$, $g(106) < 5.65871$, and $\log 2 < .693148$.

At the other end, we use the smaller mass $\tau_\kappa d + \omega_{\min}$ and the upper bound $\chi_\kappa(\tau_\kappa d) \leq \min\{7d/5, 1/10\}$. The resulting lower bound for the difference is concave on $[0, 1/14]$ and on $[1/14, 1]$: both arguments of the first term are affine, and the right side of (95), which is subtracted, is convex. It vanishes at zero and exceeds .489 at one. At $d = 1/14$, write $r = L\omega_{\min}$ and $m = \tau_\kappa d + \omega_{\min}$. Then $r \geq 1494/728$ and $m \geq 51/728$. Monotonicity in r, m gives $\mathcal{A}_\theta(r, m) > .30$, by the lower bound $m\log(r/m) + m - m^2/r$, valid in the first case of (85), and $\log 29 > 3.36$. The terms $G\omega_{\min}(1+d)$, $\gamma_* d$ and $(1-\omega_{\min})\chi_\kappa(\tau_\kappa d)$ are less than .120, .032 and .100, respectively, and their sum is less than .30.

Conclusion for a selected row. For $d \leq 0$, the minimum over the excess is at $\eta = 0$. Writing $x = (1+d)_+ \in [0,1]$, homogeneity and the chord inequality in x give

$$\mathcal{A}_\theta(r, \omega x) - \gamma_*\omega x = \omega\{\mathcal{A}_\theta(L, x) - \gamma_* x\} \geq \omega x(G - \gamma_*) = (G - \gamma_*)c,$$

because $\mathcal{A}_\theta(L, 0) = 0$, $\mathcal{A}_\theta(L, 1) = G$, and $\mathcal{A}_\theta$ is concave. For $0 < d < 1$ the minimum is at s, where $m_d - c = \tau_\kappa d(1-\omega) \leq d$. For $d \geq 1$ we have $s = \tau_\kappa$, so $\eta = \tau_\kappa$, and again $m_d - c \leq d$.

Subtracting $\gamma_* m_d$ from (95) gives (93), and monotonicity in the mass outside the selected entry then gives the bound for $\bar{m}$.

Unselected rows with maximum at most $5/8$. By Eq. (89), the contribution of such a row is at least $z\ell_* - h_2(\nu)$, where $z = P_i - 1 > 1$ and $L > 105 + z$. The condition $L > 105 + z$ holds because the other $n-1$ marginals sum to more than $2(n-1) \geq 104$. It therefore suffices to show that, for $z \geq 1$ and $L \geq 105 + z$,

$$z \log(8/5) - h_2(\nu) \geq g(L) \min\left\{1, \frac{f_\nu(z)}{D}\right\}. \tag{97}$$

We have $f_\nu(z) \leq f_0(z) = z^2/2 + 2z - 1/2$ and $D \geq L - 53/52$. For $1 \leq z \leq 14$, the ratio $g(L)/(L - 53/52)$ decreases in L, while $g(105 + z) \leq g(119) < 5.78$. It is therefore enough to prove

$$(.47z - .0951)(z + 104 - 1/52) - 5.78 f_0(z) > 0.$$

The left side is a concave quadratic, with values greater than 27 and 39 at $z = 1$ and $z = 14$, respectively. For $z \geq 14$, the maximum of the right side over $L \geq 106$ is attained at $L = f_0(z) + 53/52$, and there it is less than $1 + 2\log(z+1)$ because $f_0(z) + 53/52 < (z+1)^2$. At $z = 14$ this is less than 6.418, while $.47z - .0951 = 6.4849$, and thereafter the latter has the larger derivative. □

A.7.2 The maximizing columns and their capacities

We now choose the rows to which Lemma A.14 is applied. Let

$$\mathcal{U} = \{i : \max_j R_{ij} > 5/8\}, \qquad f(i) = \arg\max_j R_{ij} \quad (i \in \mathcal{U}).$$

The maximizing entry is unique because it exceeds $1/2$ and each row sums to one. If several rows have the same maximizing column, we select one whose marginal P_i is largest. Let $I \subseteq \mathcal{U}$ be the set of selected rows. Then f is injective on I, as required by Lemma A.14. We call the rows that are not selected *omitted* and write

$$O = [n] \setminus I, \qquad k = n - |\mathcal{U}|, \qquad q = |\mathcal{U}| - |I|, \qquad b_{\mathrm{om}} = |O| = k + q. \tag{98}$$

Of the omitted rows, k have largest entry at most $5/8$, and the other q share their maximizing column with a selected row. For a set of indices A, write $\Omega_A = \sum_{i \in A} \omega_i$.

For $i \in I$, let $d_i = P_i - P_{f(i)}$ and $c_i = \min\{1, \omega_i(1 + d_i)_+\}$. If $\sum_{i \in I} c_i \geq 1$, the selected rows alone contribute the required $g(L)$. Otherwise we bound the *missing coefficient*, the amount by which this sum falls short of one:

$$\Delta = \left(1 - \sum_{i \in I} c_i\right)_+.$$

We use the quantities

$$\begin{aligned}
B &= \sum_{j=1}^{n} \left(P_j - \sum_{\substack{i \in \mathcal{U} \\ f(i) = j}} P_i \right)_+, \\
T &= \sum_{i \in \mathcal{U}} \sum_{j \neq f(i)} M_{ij}, \qquad P_{\mathrm{rem}} = \sum_{i \notin \mathcal{U}} P_i, \qquad Z = T + P_{\mathrm{rem}}, \\
\Psi_\nu(x) &= \frac{x^2}{2} + (1 - \nu)x, \qquad c_\nu = 2 - \nu, \\
\mathcal{D} &= D - \sum_{i \in I} (P_i - \nu)(1 + d_i)_+, \qquad \mathcal{N} = n - \sum_{i \in I} (1 + d_i)_+.
\end{aligned} \tag{99}$$

Here B is the sum, over all columns, of the amount by which the marginal of the column exceeds the total marginal of the rows that maximize in it. The quantity T is the mass that the mean transport places outside the maximizing entries. We first bound $\mathcal{D}$ by a quadratic function of B, and then use the column capacities to bound B by Z plus a small correction.

Lemma A.16 (Bounds on the missing coefficient)**.** *Let*

$$\varepsilon_i = (1+\nu)\omega_i - \nu, \qquad \varepsilon_A = \sum_{i\in A} \varepsilon_i, \qquad e = \sum_i (-\varepsilon_i)_+.$$

Then

$$0 \le e \le \frac{L-2n}{D} < 1, \qquad B \le Z + e, \qquad b_{\mathrm{om}} - \Omega_O \le Z/2, \qquad 0 \le \Delta \le (\mathcal{D}/D)_+ \le 1.$$

If $b_{\mathrm{om}} = 0$*, then* $\Delta = 0$*. Otherwise* $Z > 2k + q(2-\nu)$ *and*

$$\Delta \le \max\left\{0, \min\left\{1, \frac{Z^2/2 + (2-\nu)Z - 1/2}{D}, \frac{L - 2n + (2-\nu)Z}{D}\right\}\right\}. \tag{100}$$

For $k \ge 1$*, the sharper bound* (114)*, proved below, also holds.*

Proof. Contributions of single columns. Consider a column with marginal p that is the maximizing column of at least one row, and let a be the largest marginal of such a row. The contributions of this column to $\mathcal{D}$ and $\mathcal{N}$ are

$$\delta(p,a) = p - \nu - (a-\nu)(1+a-p)_+, \qquad \delta_0(p,a) = 1 - (1+a-p)_+.$$

If no row maximizes in the column, its contributions are $p - \nu$ and 1. Both displayed functions are nonincreasing in a; this is why we select the row with the largest marginal. Moreover,

$$\delta(p,a) \le \Psi_\nu(p) - \Psi_\nu(a), \qquad \delta_0(p,a) \le p - a. \tag{101}$$

For the first inequality, set $x = p - a$. The difference between its right and left sides is $x^2/2$ if $x \le 1$, and $(x-1)(a-\nu) + x^2/2$ otherwise. Where the positive part is nonzero, $\partial_a \delta = p + \nu - 1 - 2a < 0$; elsewhere the first function is constant in a. The second inequality and the monotonicity of the second function follow from its formula.

The graph of maximizing columns. Consider the directed graph whose vertices are the indices, each standing for both a row and a column, with edges $i \to f(i)$ for $i \in \mathcal{U}$. Every vertex in $\mathcal{U}$ has one outgoing edge, and every other vertex has none. Each component therefore consists either of a directed cycle together with trees whose edges point toward it, or of a tree whose edges point toward its one vertex outside $\mathcal{U}$. For a vertex on a cycle, the largest marginal of the rows that maximize in its column is at least the marginal of the vertex preceding it on the cycle. By monotonicity and (101), the contributions to $\mathcal{D}$ and $\mathcal{N}$ are bounded by the differences of $\Psi_\nu(P_j)$ and P_j between consecutive vertices. These bounds telescope, so each cycle contributes a nonpositive total.

It remains to treat the trees. For a vertex off the cycles, consider the subtree formed by this vertex (its root) and all vertices with a directed path to it, and write p for the marginal of the root. Let s be the sum of the summands of B indexed by the vertices of the subtree. Without the positive parts, these summands telescope to p; hence $s \ge p > 2$. We claim that, over the columns of the subtree, the contributions described by δ and δ_0 sum to at most

$$\Psi_\nu(s) - c_\nu \qquad \text{and} \qquad s - 1. \tag{102}$$

We prove this by induction on the number of vertices.

For a single vertex, $s = p$ and the claim reduces to

$$\Psi_\nu(p) - (p-\nu) = p^2/2 - \nu p + \nu \ge 2 - \nu, \qquad 1 \le p - 1.$$

If the root has one child, with marginal a and subtree sum $s_1 \ge a$, then $s = s_1 + (p-a)_+$. When $p \le a$, both contributions of the root are nonpositive. When $p > a$, convexity gives

$$\delta(p,a) \le \Psi_\nu(p) - \Psi_\nu(a) \le \Psi_\nu(s_1 + p - a) - \Psi_\nu(s_1), \qquad \delta_0(p,a) \le p - a,$$

and in both cases adding the bounds for the subtree of the child proves the claim.

Suppose the root has $j \ge 2$ children, with marginals $p_l > 2$ and subtree sums $s_l \ge p_l$. Let $A = \sum_l p_l$ and $b = (p - A)_+$, so that $s = \sum_l s_l + b$. Then

$$\Psi_\nu(s) - \sum_l \Psi_\nu(s_l) = \sum_{l<l'} s_l s_{l'} + b\sum_l s_l + b^2/2 + (1-\nu)b \ge p.$$

This holds because $p_l p_{l'} \geq p_l + p_{l'}$, so the first term is at least A, and, if $b > 0$, the second is at least b. The first contribution of the root is at most $p - \nu$, and the bounds for the subtrees of the children together subtract jc_ν; this gives the first bound in Eq. (102). The second contribution of the root is at most one, and the j bounds for the children together subtract j; this gives the second bound.

Deleting the cycle vertices leaves a forest. Its trees are the trees attached to the cycles and the components ending outside $\mathcal{U}$. Their number t is at least k, and is positive whenever $b_{\mathrm{om}} > 0$. Their sums satisfy $s_l > 2$ and $\sum_l s_l \leq B$. The quadratic is increasing and superadditive on nonnegative arguments, so summing the bounds for the trees and the nonpositive cycle contributions gives

$$\mathcal{D} \leq \Psi_\nu(B) - c_\nu, \qquad \mathcal{N} \leq B - 1 \qquad (b_{\mathrm{om}} > 0). \tag{103}$$

Keeping track of the number of trees gives a sharper bound. By convexity, for a total at most B and each $s_l \geq 2$, the sum of the quadratic terms is largest when all excess above $2t$ lies in one summand. Since $\Psi_\nu(2) = 2c_\nu$,

$$\sum_l \Psi_\nu(s_l) - tc_\nu \leq \Psi_\nu(B - 2(t-1)) + (t-2)c_\nu.$$

The right-hand side decreases as t increases: one more tree lowers it by $\Psi_\nu(x) - \Psi_\nu(x-2) - c_\nu = 2x - 2 - \nu > 0$, evaluated at the argument of the quadratic before the increase. Since $t \geq k$, we obtain

$$\mathcal{D} \leq \Psi_\nu(B - 2(k-1)) + (k-2)c_\nu, \qquad \mathcal{N} \leq B - k \qquad (k \geq 1). \tag{104}$$

If $b_{\mathrm{om}} = 0$, the map defining the edges is a permutation, and the graph consists of cycles only. Hence

$$\mathcal{D} \leq 0, \qquad \mathcal{N} \leq 0. \tag{105}$$

Column capacity. By Eq. (84), the marginals of the mean transport are

$$M^\top \mathbf{1} = P, \qquad M\mathbf{1} = P + \varepsilon = L\omega, \qquad \varepsilon_i = \frac{(1+\nu)P_i - \nu L}{D}. \tag{106}$$

Write $P_i = 2 + x_i$ and $s = \sum_i x_i = L - 2n$. Since $1 + \nu = n\nu$,

$$e = \frac{1+\nu}{D} \sum_i (s/n - x_i)_+.$$

The arguments of the positive parts average to zero, so at most $n - 1$ summands are positive, and each is at most s/n because the marginals exceed two. This proves the first chain of inequalities in

$$0 \leq e \leq \frac{L - 2n}{D} < 1, \qquad B \leq Z + e. \tag{107}$$

For the last inequality, let $b_j = \sum_{i \in \mathcal{U}: f(i) = j} P_i - P_j$. Column j has total mass P_j, which bounds the sum of its maximizing entries; each of them equals the mass of its row minus the mass outside the maximizing entry. Hence

$$b_j \leq \sum_{i \in \mathcal{U}: f(i) = j} \left(\sum_{j' \neq f(i)} M_{ij'} - \varepsilon_i \right).$$

Summing over the columns with $b_j > 0$ and bounding the last term of each summand by its positive part gives $\sum_j (b_j)_+ \leq T + e$. Since $\sum_j b_j = -P_{\mathrm{rem}}$, we obtain $B = P_{\mathrm{rem}} + \sum_j (b_j)_+ \leq Z + e$.

Because the selected rows have distinct maximizing columns, there are b_{om} columns outside $f(I)$, each of mass $P_j > 2$. Every row in $\mathcal{U}$ shares its maximizing column with a selected row, so its entries in these columns lie outside its maximizing entry. The other rows have total mass $P_{\mathrm{rem}} + \varepsilon_{[n] \setminus \mathcal{U}}$. Hence

$$2b_{\mathrm{om}} < Z + \varepsilon_{[n] \setminus \mathcal{U}} \qquad (b_{\mathrm{om}} > 0). \tag{108}$$

All rows outside $\mathcal{U}$ belong to O, and

$$\varepsilon_{[n] \setminus \mathcal{U}} = (1+\nu)\Omega_{[n] \setminus \mathcal{U}} - k\nu \leq 2\Omega_O.$$

Therefore

$$b_{\mathrm{om}} - \Omega_O \leq Z/2. \tag{109}$$

Inequality Eq. (108) also gives the lower bound on Z stated in the lemma. Since $P_{\mathrm{rem}} \le Z$,

$$\varepsilon_{[n]\setminus\mathcal{U}} \le \frac{(1+\nu)Z - k\nu L}{D}.$$

Substituting this in Eq. (108) and using $L > 2n$ gives

$$Z > 2b_{\mathrm{om}} - \frac{2b_{\mathrm{om}}(1+\nu)}{L} + k\nu > 2k + q(2-\nu). \tag{110}$$

In particular,

$$Y := Z - 2(k-1) > 4 - \nu \qquad (q \ge 1). \tag{111}$$

Bounding the missing coefficient. For nonnegative numbers a_i, $\sum_i \min\{1, a_i\} \ge \min\{1, \sum_i a_i\}$. Applying this to $a_i = \omega_i(1+d_i)_+$ gives

$$0 \le \Delta \le (\mathcal{D}/D)_+ \le 1. \tag{112}$$

Besides Eq. (103) and (104), the inequality $P_i - \nu \ge c_\nu$ gives

$$\mathcal{D} \le D - c_\nu \sum_{i\in I}(1+d_i)_+ = L - 2n + c_\nu \mathcal{N}.$$

Combining this with Eq. (103) and $B \le Z + e$ yields

$$\mathcal{D} \le \min\{D, \Psi_\nu(Z+e) - c_\nu, L - 2n + c_\nu(Z+e-1)\} \qquad (b_{\mathrm{om}} > 0). \tag{113}$$

For $k \ge 1$, write $Y = Z - 2(k-1)$. Combining (104) with the same bound and $B \le Z + e$ gives

$$\mathcal{D} \le \min\{D, \Psi_\nu(Y+e) + (k-2)c_\nu, L - 2n + c_\nu(Y+e+k-2)\}. \tag{114}$$

In both displays the quadratic is evaluated at nonnegative arguments, where it is increasing, because $Z + e \ge B > 2k$. Finally, using $e \le 1$ in (113) gives

$$\mathcal{D} \le \min\{D, Z^2/2 + (2-\nu)Z - 1/2, L - 2n + (2-\nu)Z\}, \tag{115}$$

which, together with Eq. (112), proves Eq. (100). When $b_{\mathrm{om}} = 0$, Eq. (105) gives $\Delta = 0$. □

A.7.3 A scalar inequality for repeated maximizing columns

When maximizing columns repeat, the proof of Lemma A.18 below reduces the remaining inequality to one between the missing coefficient Δ, the mass Z, and the number k of rows with maximum at most $5/8$. The contribution of the selected rows falls short of its largest possible value by at most $g(L)\Delta$. The next lemma bounds this shortfall by the linear terms contributed by the omitted rows.

We use the following constants and functions, which correspond to the smallest dimension $n = 53$:

$$\nu_* = \frac{1}{52}, \qquad \sigma_* = 1 + \nu_* = \frac{53}{52}, \qquad b_* = 2 - \nu_* = \frac{103}{52}, \qquad \eta_* = h_2(\nu_*),$$

$$a_* = \gamma_* - \frac{\eta_*}{2}, \qquad \xi_* = \frac{27}{26000}, \qquad D_*(L) = L - \sigma_*,$$

$$\Psi_*(x) = \Psi_{\nu_*}(x), \qquad f_*(z) = \Psi_*(z+1) - b_* = \frac{z^2}{2} + b_* z - \frac{1}{2},$$

$$\rho_0(z) = a_* z - \gamma_*, \qquad \rho_1(z) = a_* z - 3\gamma_* + \ell_*.$$

Lemma A.17 (Scalar bound for the missing coefficient)**.** *Let $n \ge 53$, $\nu = 1/(n-1)$, $L \ge 2n$, $D = L - 1 - \nu$, $a_\nu = \gamma_* - h_2(\nu)/2$, and $0 \le \Delta \le 1$. For $k = 0$, suppose $Z \ge 2$ and*

$$\Delta \le \min\left\{1, \frac{Z^2/2 + (2-\nu)Z - 1/2}{D}, \frac{L - 2n + (2-\nu)Z}{D}\right\}. \tag{116}$$

For an integer $k \ge 1$, let $Y = Z - 2(k-1)$ and suppose instead that $Y \ge 4 - \nu$,

$$0 \le e \le \min\left\{1, \frac{L-2n}{D}\right\}$$

and

$$\Delta \le \min\left\{1, \frac{\Psi_\nu(Y+e) + (k-2)(2-\nu)}{D}, \frac{L - 2n + (2-\nu)(Y+e+k-2)}{D}\right\}. \tag{117}$$

Under either set of assumptions,

$$g(L)\Delta < a_\nu Z - \gamma_* + k(\ell_* - 2\gamma_*). \tag{118}$$

Proof. Define

$$d_*(L,z) = \min\{1, f_*(z)/D_*(L), (L-106+b_*z)/D_*(L)\}.$$

Maximization over total mass. We first prove that, for every $L \geq 106$,

$$\rho_0(z) > g(L)d_*(L,z) + \xi_* z \qquad (z \geq 2), \tag{119}$$
$$\rho_1(z) > g(L)d_*(L,z) + \xi_* z \qquad (z \geq 257/52). \tag{120}$$

The rational bounds of Lemma A.9 and the monotonicity of $g(L)/(L-\sigma_*)$ give

$$\eta_* < .1, \quad a_* > .39, \quad \ell_* > .47, \quad \frac{g(L)}{D_*(L)} < \begin{cases} .054, & L \geq 106, \\ .042, & L \geq 144. \end{cases} \tag{121}$$

The ratio decreases because its derivative has numerator $(L-\sigma_*)g'(L) - g(L) < 1 - g(L) < 0$; the two bounds then follow from $g(106) < 5.66$ and $g(144) < 5.97$.

For $2 \leq z \leq 53$, the maximum of $g(L)d_*(L,z)$ over $L \geq 106$ occurs at $L_z = (z^2+211)/2$, where the last two fractions in the minimum coincide. Before this point the minimum equals $1 - b_*(53-z)/D_*(L)$, which increases; after it the product $f_*(z)g(L)/D_*(L)$ decreases. For $z \geq 53$ the last fraction is at least one, and the maximum is $g(f_*(z)+\sigma_*)$.

Monotonicity of the differences. To compare the two linear left-hand sides with this maximum, write

$$d_0 = 5433/52, \qquad w(z) = 1 - \frac{b_*(53-z)}{z^2/2 + d_0}, \qquad B(z) = w(z)g(L_z).$$

Concavity of $\log$ gives a bound valid on the whole interval $2 \leq z \leq 53$: from $g(x) < 1 + \log x$ and the tangent line at 160,

$$g(L_z) < \log 160 + \frac{L_z}{160} < \frac{z^2+1837}{320}.$$

Since $104 < d_0 < 105$, $b_* < 2$, and $0 \leq w \leq 1$, it follows that

$$B'(z) < \frac{(210+106z-z^2)(z^2+1837) + 160z(z^2+208)}{80(z^2+208)^2} < \frac{31}{80} < a_* - \xi_*.$$

The middle inequality follows from the identity

$$\begin{aligned} &31(z^2+208)^2 - (210+106z-z^2)(z^2+1837) - 160z(z^2+208) \\ &\quad = 32z^2\left(z - \frac{133}{32}\right)^2 + \frac{2247}{32}z^2 + 13900\left(z - \frac{33}{4}\right)^2 + 1348z + \frac{37381}{4} > 0. \end{aligned}$$

Hence the differences between the two sides of (119) and of (120) increase on this interval. At $z = 2$, $g(107.5) < 5.68$ shows that the first difference exceeds $3353481/71981000$; at $z = 257/52$, $g(L_z) < 5.78$ shows that the second difference exceeds $60537567/8283704000$. For $z \geq 53$, the derivative of $g(f_*(z)+\sigma_*)$ satisfies

$$(z+b_*)g'(f_*(z)+\sigma_*) < \frac{z+b_*}{f_*(z)+\sigma_*-1} < \frac{2}{z} < \frac{31}{80}.$$

Hence both differences also increase for $z \geq 53$. This proves Eq. (119) and (120).

Small values of Y. For $k \geq 1$, we use the second bound in the hypothesis on e, which is sharper when L is near 106. Let

$$e_*(L) = \frac{L-106}{D_*(L)} < 1.$$

Since $D \geq D_*(L)$ and $L - 2n \leq L - 106$, the hypotheses give $e \leq e_*(L)$. Define

$$\widetilde{d}_*(L,Y) = \min\left\{1, \frac{\Psi_*(Y+e_*) - b_*}{D_*(L)}, \frac{L-106+b_*(Y+e_*-1)}{D_*(L)}\right\}.$$

We claim that, for every $L \geq 106$,

$$\rho_1(Y) > g(L)\widetilde{d}_*(L,Y) + \xi_* Y \qquad (Y \geq 207/52). \tag{122}$$

For $Y \geq 257/52$, the claim follows from $e_* \leq 1$ and (120). On $[207/52, 257/52]$, if $L \leq 144$ then $e_* \leq 1976/7435 < 34/125$ and $g(L)\widetilde{d}_* \leq .054\{\Psi_*(Y + 34/125) - b_*\}$. If $L \geq 144$, the same product is at most $.042 f_*(Y)$. Subtracting these bounds from $.39Y - .85 - \xi_* Y$ gives concave quadratics, so it suffices that they are positive at both ends of the interval. The smaller of the two end values is $3877333877/42250000000 > .09$ for the first quadratic and $29971/540800 > .05$ for the second. This proves the claim.

General n and k. We have $D \geq D_*(L)$, $a_\nu \geq a_*$, and $0 \leq \nu_* - \nu \leq \nu_*$. For $k = 0$, each numerator in (116) is at most the corresponding numerator of $d_*(L, Z)$ plus $\nu_* Z$. The numerators are nonnegative, so

$$g(L)\Delta \leq g(L) d_*(L, Z) + \xi_* Z < \rho_0(Z),$$

since $.054\nu_* = \xi_*$. For $k \geq 1$, $Y \geq 4 - \nu \geq 207/52$ and $e \leq e_*(L)$. Writing $Q_* = \Psi_*(Y + e_*) - b_*$ and $N_* = L - 106 + b_*(Y + e_* - 1)$ for the numerators of $\widetilde{d}_*$, we obtain by expanding

$$\begin{aligned} \Psi_\nu(Y + e) + (k - 2)(2 - \nu) &\leq Q_* + \nu_* Y + 2(k - 1), \\ L - 2n + (2 - \nu)(Y + e + k - 2) &\leq N_* + \nu_* Y + 2(k - 1). \end{aligned}$$

Thus $g(L)\Delta \leq g(L)\widetilde{d}_* + \xi_* Y + .108(k - 1)$. On the other hand, since $Z = Y + 2(k - 1)$,

$$a_\nu Z - \gamma_* + k(\ell_* - 2\gamma_*) = a_\nu Y - 3\gamma_* + \ell_* + (k - 1)(\ell_* - h_2(\nu)) \geq \rho_1(Y) + (k - 1)(\ell_* - \eta_*).$$

By Eq. (122), this lower bound exceeds the preceding upper bound, because $\ell_* - \eta_* - .108 > .262$. □

A.7.4 Completion of the case $n \geq 53$

Lemma A.18 (The bound in dimensions at least 53)**.** *For every $n \geq 53$, every vector of marginals with $P_i > 2$, and every row-stochastic matrix R,*

$$K_R(P) \geq B_n(L) = g(L) + \log(n - 1).$$

Proof. Distinct maximizing columns. Suppose first that $q = 0$. Then every row in $\mathcal{U}$ is selected, and every omitted row has maximum at most $5/8$. If some $c_i = 1$, then $\Delta = 0$. Otherwise, taking positive parts can only increase the sum of the coefficients of the selected rows, and the following identity bounds the missing coefficient:

$$\begin{aligned} D\left(1 - \sum_{i \in I} \omega_i(1 + d_i)\right) = \sum_{i \in O} (P_i - \nu) \\ + \frac{1}{2}\left\{\sum_{i \in O} (P_i - \nu)^2 - \sum_{j \notin f(I)} (P_j - \nu)^2 - \sum_{i \in I} d_i^2\right\}. \end{aligned} \tag{123}$$

To verify it, sum $2(P_i - \nu)d_i = (P_i - \nu)^2 - (P_{f(i)} - \nu)^2 + d_i^2$ over the selected rows; since f is injective, the terms of the selected columns cancel. There are $|O|$ columns outside $f(I)$, each with marginal greater than two. Using this and dropping the terms $-d_i^2$ gives

$$\Delta \leq \min\left\{1, \sum_{i \in O} \frac{f_\nu(P_i - 1)}{D}\right\}, \qquad f_\nu(z) = z^2/2 + (2 - \nu)z - 1/2. \tag{124}$$

For an omitted row, Lemma A.15 gives $\mathcal{Q}_i \geq G \min\{1, f_\nu(P_i - 1)/D\}$. For a selected row, the extreme points with this row as source give $m_i \geq \omega_i(1 + d_i)_+ \geq c_i$, so Lemma A.15 gives $\mathcal{P}_i \geq G c_i$. Finally, $\sum_i \min\{1, x_i\} \geq \min\{1, \sum_i x_i\}$ for nonnegative x_i. The omitted rows therefore contribute at least $G\Delta$, and the selected rows at least $G(1 - \Delta)$. Substitution in Eq. (87) proves $K_R(P) \geq \log \kappa + G$. The argument also covers $I = \varnothing$ and the case where f is a permutation of all rows.

Repeated maximizing columns. Now suppose $q \geq 1$. Let $\Omega_{\mathrm{rem}} = \Omega_{[n] \setminus \mathcal{U}}$ and $a_\nu = \gamma_* - h_2(\nu)/2$. Each selected row contributes the bound of Lemma A.15, which consists of its coefficient term and a term linear in its mass outside the selected entry. By Eq. (90), the omitted rows in $\mathcal{U}$ contribute the

same linear term, up to an entropy term. The other k rows contribute at least $\ell_*(P_{\rm rem} - k + \Omega_{\rm rem})$, again up to entropy terms. Summing gives

$$\begin{aligned} K_R(P) &\geq \log\kappa + (G-\gamma_*)(1-\Delta) + \gamma_* T + \ell_*(P_{\rm rem} - k + \Omega_{\rm rem}) \\ &\quad - h_2(\nu)(b_{\rm om} - \Omega_O) && (125)\\ &\geq \log\kappa + G(1-\Delta) + a_\nu Z - \gamma_* + (\ell_* - \gamma_*)P_{\rm rem} - \ell_* k + \ell_*\Omega_{\rm rem} && (126)\\ &\geq \log\kappa + G(1-\Delta) + a_\nu Z - \gamma_* + k(\ell_* - 2\gamma_*). && (127) \end{aligned}$$

The second inequality uses $b_{\rm om} - \Omega_O \leq Z/2$ and drops $\gamma_*\Delta \geq 0$. The last inequality uses $P_{\rm rem} \geq 2k$, $\ell_* > \gamma_*$, and $\Omega_{\rm rem} \geq 0$.

If $k = 0$, the q columns outside $f(I)$ receive only mass outside maximizing entries. Their total mass exceeds $2q$, so $Z > 2q \geq 2$. The bound Eq. (100) therefore gives the hypotheses of Lemma A.17 with $k = 0$. If $k \geq 1$, Eq. (111) gives $Y = Z - 2(k-1) > 4 - \nu$, and (107) gives the required bound on e. We divide Eq. (114) by D and use $\Delta \leq (\mathcal{D}/D)_+$. Both numerators are positive: the quadratic one is at least $\Psi_\nu(4-\nu) - (2-\nu) > 0$, and the linear one is positive because $L - 2n \geq 0$ and $Y + e + k - 2 \geq Y - 1 > 0$. Hence Eq. (117) holds.

In either case, Lemma A.17 bounds $G\Delta$ strictly by the right-hand side of (118), which consists of the last three terms in (127). This proves $K_R(P) > G + \log\kappa$ when maximizing columns repeat. Hence the bound holds for every positive R. Approximation by positive matrices and Eq. (63) give the stated nonstrict inequality for every row-stochastic R. □

A.8 A complete matching when the total mass is large

We first treat the case in which every row has an entry of at least $3/5$ and these entries lie in different columns. The proof chooses sources with probabilities proportional to $P_i - \nu$, as in Lemma A.13; with these weights, the differences $P_i - P_{f(i)}$ add up to a nonnegative sum of squares.

Lemma A.19 (Complete matching)**.** *Let $n \geq 5$, $P_i \geq 2$, and $L = \sum_i P_i \geq 2048$. Suppose a permutation f satisfies $R_{i,f(i)} \geq 3/5$ for every i. Then*

$$T_s(R,P) \geq g(L) + \log(n-1) \qquad (2 \leq s \leq n).$$

Proof. Assume first that $R > 0$; matrices with zero entries are treated at the end of the proof. Write

$$\kappa = n-1, \quad \nu = \frac{1}{\kappa}, \quad \tau_\kappa = 1-\nu, \quad D = L - 1 - \nu, \quad \omega_i = \frac{P_i - \nu}{D}, \quad G = g(L).$$

The probabilities ω_i sum to one. Since the selected entries are at least $3/5$, the restriction $z \geq 3/5$ can be kept in the minimization over z that defines $\mathcal{A}$, which gives

$$\mathcal{A}_{3/5}(r,m) = \begin{cases} \mathcal{J}(r-m, m), & m \leq 2r/5, \\ r\log(5/3) + m\log(3/2), & m \geq 2r/5. \end{cases}$$

We call these the entropy case and the linear case. Fix a row and abbreviate $P = P_i$, $T = P_{f(i)}$, $d = P - T$ and $\omega = \omega_i$. By Lemma A.13, with $\mathcal{A}_{3/5}$ in place of $\mathcal{A}$, the contribution of this row is at least

$$\begin{aligned} U(P,T) &= \min_{\tau_\kappa \min(d_+,1) \leq \eta \leq \tau_\kappa} \left\{ \mathcal{A}_{3/5}(L\omega, m_d(\eta)) - (1-\omega)\chi_{n-1}(\eta) \right\}, \\ m_d(\eta) &= \omega(1+d)_+ + (1-\omega)\{(d-1)_+ + \eta\}. \end{aligned}$$

We will prove the row bound

$$U(P,T) \geq G\min\{1, \omega(1+d)_+\}. \tag{128}$$

We use the numerical bounds

$$\nu \leq \tfrac{1}{4}, \qquad h_2(\nu) < \tfrac{9}{16}, \qquad g(2048) < \tfrac{87}{10}, \qquad \frac{G}{D} < \frac{1}{200}. \tag{129}$$

The last ratio decreases with L; all numerical bounds on logarithms follow from Lemma A.9.

The minimum over the excess. If $d < 1$, the minimum defining U is attained at the left end of the interval for η. By (91), the slope of χ_{n-1} is at most $\log 4$ on the first interval, at most $\log 2$ on the second, and at most $\log(3/2)$ on every later interval. On the first two intervals, respectively,

$$\frac{m_d(\eta)}{L\omega} \le \frac{1+\eta/\tau_\kappa}{L} + \frac{\eta}{2-\nu} \le \frac{1}{1536} + \frac{1}{7} < \frac{1}{5}, \qquad \frac{m_d(\eta)}{L\omega} \le \frac{5}{6144} + \frac{2}{7} < \frac{1}{3}.$$

Hence, in the entropy case, the derivative $\log[(r-m)/m]$ of $\mathcal{A}_{3/5}$ in m is at least the corresponding slope of χ_{n-1}. On the remaining intervals the derivative of $\mathcal{A}_{3/5}$ in m is at least $\log(3/2)$ by its definition. For $d \le 0$, evaluation at $\eta = 0$ then proves (128), by homogeneity and the chord between $\mathcal{A}_{3/5}(L,0) = 0$ and $\mathcal{A}_{3/5}(L,1) = G$.

An estimate at $\omega(1+d) = 1$. For $d \ge 1$, set

$$z_d = d^2 + (1-\nu)d + 1, \qquad L_0(d) = d^2 + (3-\nu)d + 3.$$

We claim that

$$\mathcal{A}_{3/5}(L, z_d) - (d+1)g(L) - dh_2(\nu) > 0, \qquad L \ge \max\{2048, L_0(d)\}. \tag{130}$$

The left side increases with L and decreases with ν. In the entropy case its derivative in L is at least $(z_d - d - 1)g'(L) > 0$; in the linear case it is at least $\log(5/3) - 1/(d+2-\nu) > 0$. It therefore suffices to take $\nu = 1/4$ and $L = \max\{2048, L_0(d)\}$.

For $L = 2048$, let $F(d)$ denote the left side, and put $a = 3/4$ and $z = d^2 + ad + 1$. In the entropy case,

$$F'''(d) = -\frac{L(2d+a)}{z^2(L-z)^2}\left\{2z(L+z) + (4-a^2)(L-2z)\right\} < 0.$$

Thus F' is concave. At $d = 1$, the inequality $\mathcal{J}(L-z, z) \ge z\{\log(L/z) + 1 - z/L\}$ gives $F(1) > 5/2$, and

$$F'(1) > \tfrac{11}{4}\tfrac{13}{2} - \tfrac{87}{10} - \tfrac{9}{16} > 0.$$

At the other end of the entropy case, $z = 2L/5$ and $d > 28$, so

$$F'(d) > (56 + \tfrac{3}{4})\tfrac{2}{5} - \tfrac{87}{10} - \tfrac{9}{16} > 0.$$

By concavity, F' is positive between these two points. In the linear case $F'' = 2\log(3/2) > 0$, so F' remains positive there. Hence $F(d) > 0$ up to the point d_0 with $L_0(d_0) = 2048$.

One has $43 < d_0 < 44$. For $d \ge d_0$ the linear case applies at $L = L_0(d)$. Let $F_0(d)$ denote the left side of (130) at this value of L and at $\nu = 1/4$. Before $(d+1)g(L_0(d))$ is subtracted, the coefficient of d in F_0 is

$$\frac{11}{4}\log\frac{5}{3} + \frac{3}{4}\log\frac{3}{2} - h_2(1/4) > 1.$$

Differentiation therefore gives

$$F_0'(d) > \frac{9}{5}d - 2 - 2\log(d+2) > 0.$$

The first inequality uses $L_0(d) < (d+2)^2$, $(d+1)L_0'(d)/(L_0(d)-1) < 2$, and $\log(5/2) > 9/10$. The second holds at $d = 43$, using $\log 45 < 4$, and its left side increases thereafter. Since $L_0(d_0) = 2048$, we have $F_0(d_0) = F(d_0) > 0$, and hence $F_0(d) > 0$ for every $d \ge d_0$. This proves Eq. (130).

The case $d > 0$. Suppose first $d \ge 1$. Then the interval for η reduces to the point $\eta = \tau_\kappa$, and $m_d = d - \nu + (1+\nu)\omega$. The difference between the two sides of (128) is concave in ω on either side of $\omega_* = (1+d)^{-1}$. Every possible value of ω lies in the interval

$$\frac{d+2-\nu}{D} \le \omega \le 1.$$

At $\omega = 1$, monotonicity of $\mathcal{A}_{3/5}(L, \cdot)$ proves Eq. (128). When ω_* belongs to this interval, (128) is (130) after division by ω_*. At the lower end, $m_d = L\omega - 2$. While $\omega(1+d) \le 1$, the difference is concave in d, so it suffices to consider $d = 1$ and the point $\omega(1+d) = 1$. At that point the bound was proved above. At $d = 1$, $\mathcal{A}_{3/5}(L\omega, m_d)$ is at least $2(1-\nu)\log 2$, whereas the subtracted terms are

at most $9/16 + 6G/D < 9/16 + 3/100$. Once $\omega(1+d) \ge 1$, we have $d > 43$ and the linear case applies; the derivative of the difference in d is $\{L\log(5/2) + h_2(\nu)\}/D > 0$.

For $0 < d < 1$, the minimum over η is attained at $\tau_\kappa d$, and $m_d = \tau_\kappa d + (1+\nu d)\omega$. Since P and T are distinct coordinates of the vector of marginals, whose coordinates sum to L, we have

$$\frac{2 + d - \nu}{D} \le \omega < \frac{1}{2}.$$

Concavity in ω again reduces (128) to the two ends of this interval. At $\omega = 1/2$, concavity in d and $\chi_{n-1}(\tau_\kappa d) \le 2\tau_\kappa d \log 2$ show that the difference is at least

$$\frac{d}{2}\left\{\mathcal{A}_{3/5}(L, 3-\nu) - 2g(L) - 2(1-\nu)\log 2\right\} > 0.$$

The expression in braces increases with L and decreases with ν. At $(L,\nu) = (2048, 1/4)$ the inequality $\mathcal{J}(L-z,z) \ge z\{\log(L/z) + 1 - z/L\}$ with $z = 3-\nu$ shows that it is greater than 2.

At the lower end, we replace m_d by the smaller value $\tau_\kappa d + \omega$, using monotonicity of $\mathcal{A}_{3/5}$, and use

$$\chi_{n-1}(\tau_\kappa d) \le \min\{3d/2, 3/5\}.$$

The resulting lower bound for the difference is concave on $[0, 2/5]$ and $[2/5, 1]$: the arguments of $\mathcal{A}_{3/5}$ are affine in d and the right side of Eq. (128) is convex. Its value at zero is zero. At one it is at least $\frac{3}{2}\log 2 - 3/5 - 3/100 > 0$. At $d = 2/5$, put $r = L\omega$ and $m = \tau_\kappa d + \omega$. Then $r - m \ge 37/20$ and $m \ge 3/10$, so, since $\mathcal{A}_{3/5}$ is nondecreasing in both arguments,

$$\mathcal{A}_{3/5}(L\omega, \tau_\kappa d + \omega) \ge \mathcal{J}(37/20, 3/10) > \frac{3}{4}.$$

For the last inequality, use $\log(37/6) > 7/4$ and $\log(1+x) \ge x/(1+x)$. The right side of (128) is less than $1/50$, so the difference exceeds $3/4 - 3/5 - 1/50 > 0$. This proves Eq. (128) in all cases.

Summing the rows. Since f is a permutation, the choice of source probabilities gives

$$\sum_i \omega_i(1 + P_i - P_{f(i)}) = 1 + \frac{1}{2D}\sum_i (P_i - P_{f(i)})^2 \ge 1.$$

Taking positive parts and then replacing each summand by its minimum with one leaves a sum at least one. Hence Lemma A.13 and (128) give an average over sources of at least $G + \log(n-1)$. This average is at most every T_s. Finally, if R has zero entries, smooth it toward a positive matrix whose selected entries equal $3/5$; the selected entries stay at least $3/5$. Since Eq. (63) also holds with K_R replaced by any extreme-point cost C_{uv}, and hence by any T_s, the bound passes to the limit. □

A.9 Large total mass

We prove $K_R(P) \ge B_n(L)$ for $5 \le n \le 52$ and $L \ge 2048$. We select one row for each distinct maximizing column; the selected rows form a partial permutation of the columns. Two quantities measure how far it is from a full permutation: the total marginal mass of the rows that are not selected, and a quadratic moment of that mass. Using the directed graph of the maximizing columns, we bound both in terms of the mass placed outside the maximizing entries.

Throughout this subsection put

$$a_0 = \log(5/3), \qquad \xi_0 = 31/100, \qquad \mu(x) = -\log(e^x - 1) \quad (0 < x \le a_0).$$

Let

$$H = \{i : \max_j R_{ij} > 3/5\}, \qquad J = f(H), \qquad m = |J|, \quad k = n - |H|, \quad r = |H| - m,$$

where $f(i)$ is the unique maximizing column of row $i \in H$. We call the rows in H high rows and the remaining k rows low rows. Write $\sum_{i \notin H} P_i = 2k + e$, so $e \ge 0$. Assume $2 \le m < n$, and choose one row from each set $\{i \in H : f(i) = j\}$, $j \in J$. We call the chosen rows representatives, denote their set by G, and call the rows of $H \setminus G$ omitted. For a source $u \in G$ we average the costs

$C_{uv}(R,P)$ uniformly over the sinks $v \in G \setminus \{u\}$, using the optimal transports chosen by the rule (52). Set

$$\begin{aligned} \kappa &= m-1, & \nu &= 1/\kappa, \\ M_u &= \frac{1}{\kappa} \sum_{v \in G\setminus\{u\}} \sum_{i \in H} \sum_{j \neq f(i)} X_{ij}^{uv}, & M &= \min_{u \in G} M_u, \\ a_i &= -\log R_{i,f(i)}, & \theta_i &= (1+P_i-P_{f(i)})_+. \end{aligned} \tag{131}$$

In words, M_u is the mean mass that the high rows place outside their maximizing entries. Throughout this subsection we assume that all extreme-point costs $C_{uv}(R,P)$ are finite. If $a_i = 0$, this forces $\theta_i = 0$, and we read every product involving $\theta_i \mu(a_i)$ as zero.

Lemma A.20 (Averaging over sinks in a set of representatives)**.** *For $\xi \in \{0, \xi_0\}$ put*

$$\rho_1 = 0, \qquad \rho_\kappa = \min\{3/2, 10/\kappa\} \quad (\kappa \geq 2), \qquad \beta_\kappa = \nu + \rho_\kappa, \quad A_\kappa = 1+\beta_\kappa, \quad w_i = P_i - \beta_\kappa.$$

Then $0 < \beta_\kappa \leq 2$ and $w_i > 0$. For each $u \in G$,

$$\begin{aligned} \frac{1}{\kappa} \sum_{v \in G\setminus\{u\}} C_{uv}(R,P) \geq\ & \log\kappa + \sum_{i \in G} w_i a_i + A_\kappa a_u + \theta_u\{\mu(a_u) - \xi\} + \xi M_u \\ & + a_0\big(k + e + k\mathbb{1}_{\{\kappa=1\}}\big). \end{aligned} \tag{132}$$

Moreover, let

$$S = \sum_{j \in J} \left(\sum_{i \in H : f(i) = j} P_i - P_j \right)_+,$$

the total overflow of the maximizing columns. Then

$$M + e \geq 2r, \qquad S \leq M + 1. \tag{133}$$

Proof. For $\kappa = 1$ the correction $\rho_\kappa a$ is zero. For $\kappa \geq 2$ it compensates for the variation of a selected row when that row is itself a sink: for $0 < a \leq a_0$,

$$D_{\kappa,\xi}(a) := \frac{1}{\kappa} \max_{0 \leq j < \kappa} \{g(j+1) - j[\mu(a) - \xi]\} \leq \rho_\kappa a.$$

For the bound $D \leq 3a/2$, it suffices to show $g(j+1) - j[\mu(a) - \xi] \leq 3(j+1)a/2$. For $j \geq 2$ the left side minus the right side is increasing in a and negative at a_0, since $g(j+1) \leq (j+1)\log 2$. For $j = 1$ its maximum over a is $\xi - \log(27/16) < 0$. For the other bound, maximizing over real $j \geq 0$ gives

$$D_{\kappa,\xi}(a) \leq -\frac{1}{\kappa}\log\{1 - e^\xi(e^a - 1)\} \leq \frac{a \log 15}{\kappa \log(5/3)} < \frac{10a}{\kappa}.$$

The middle inequality holds because the function of a on its left is convex and therefore lies below its chord from 0 to a_0; here $e^{\xi_0} < 7/5$.

Apply Lemma A.12 with $I = G$ and $V = G \setminus \{u\}$. For a selected sink, its mean mass outside the maximizing entry is at least its mean excess. After ξ times that mass is subtracted, the preceding bound leaves $w_i a_i$. For the source, column $f(u)$ receives total mass $P_{f(u)}$ in every transport, so at least θ_u of the mass of row u lies outside its maximizing entry; this leaves $w_u a_u + A_\kappa a_u + \theta_u\{\mu(a_u) - \xi\}$.

For an omitted row $i \in H \setminus G$, put $z = R_{i,f(i)}$. When ξ is subtracted per unit outside column $f(i)$, its common part has mass $P_i - 1 > 1$ and unit cost at least $-\log z$, and the excess part of (74) is at least $-\log\{z + e^\xi(1-z)\}$. Their sum is nonnegative, because

$$z\{z + e^\xi(1-z)\} \leq 2z - z^2 \leq 1.$$

Hence the remaining cost of row i is at least ξ times its mean mass outside the maximizing entry. Each low row contributes at least $(P_i - 1)a_0$. When $\kappa = 1$, its single excess unit contributes another a_0, giving $P_i a_0$. Summing these bounds proves (132).

The $k + r$ columns outside J receive mass at least $2(k+r)$. Low rows contribute at most $2k + e$ of it, and high rows reach these columns only outside their maximizing entries, so $M_u + e \geq 2r$. For

the second assertion write $D_j = \sum_{i \in H : f(i) = j} P_i - P_j$. In the mean transport with source u, the row marginal of the rows $\{i \in H : f(i) = j\}$ differs from their marginal mass by 1 for $j = f(u)$ and by $-\nu$ for every other $j \in J$. The sum of these decreases is one. Since column j receives only P_j,

$$\sum_{j \in J} (D_j)_+ \le 1 + \sum_{j \in J} (D_j + \mathbb{1}_{\{j = f(u)\}} - \nu \mathbb{1}_{\{j \ne f(u)\}})_+ \le 1 + M_u.$$

Taking the minimum over $u \in G$ in both bounds gives (133). □

Define the increasing functions

$$\mathsf{B}(y) = (2y+1)\left(1 + \frac{\log(y+3)}{\log 4}\right), \qquad \mathsf{Q}(y) = \frac{3}{4}y^2 + \frac{7}{2}y + \frac{7}{4}. \tag{134}$$

Lemma A.21 (Omitted marginal mass and a sum-of-squares identity)**.** *The set of representatives G can be chosen so that, with $B = [n] \setminus G$, $b = \sum_{i \in B} P_i$, and $p_{\max} = \max_{i \notin H} P_i$ (zero when $H = [n]$),*

$$\sum_{i \in H \setminus G} P_i^2 \le \frac{1}{2} \sum_{i \notin H} P_i^2 + p_{\max} S + \frac{3}{2} S^2, \tag{135}$$

$$P_i \le \max\{S, (p_{\max} + S)/2\} \quad (i \in H \setminus G). \tag{136}$$

Let M be defined by Eq. (131) *for this set G, and let $X = M + e$ and $Y = k + X$. For every $0 \le \beta \le 2$,*

$$\begin{aligned} b &\le (2k + X + 1)\left(1 + \frac{\log(X+3)}{\log 4}\right) \le \mathsf{B}(Y), \\ E_\beta &:= \frac{1}{2} \sum_{i \in B} (P_i - \beta)^2 \le \mathsf{Q}(Y). \end{aligned} \tag{137}$$

When $\beta = 1$, the last bound improves to $E_1 \le \mathsf{Q}(Y) - 3|B|/2$. When $k = 0$,

$$b \le S\left(1 + \frac{\log S}{\log 4}\right), \qquad \sum_{i \in B} P_i^2 \le \frac{3}{2} S^2.$$

Finally, for any set of representatives, with $w_i = P_i - \beta$ and $W = \sum_{i \in G} w_i$,

$$\sum_{i \in G} w_i \theta_i \ge W - E_\beta. \tag{138}$$

Proof. Consider the directed graph on $[n]$ with an edge $i \to f(i)$ for each $i \in H$. Every high vertex has out-degree one and every low vertex has out-degree zero, so every component consists of trees directed toward either a low vertex or a directed cycle. Put

$$d_j = P_j - \sum_{i \in H : f(i) = j} P_i, \qquad S = \sum_j (-d_j)_+.$$

For a high vertex i outside the cycles, let $\mathcal{T}_i$ be the tree of vertices with a directed path to i, including i, and set

$$A_i = \sum_{j \in \mathcal{T}_i} (d_j)_+.$$

Since the contributions of the internal edges cancel,

$$A_i = P_i + \sum_{j \in \mathcal{T}_i} (-d_j)_+ \ge P_i, \qquad A_i \ge \sum_{j : f(j) = i} A_j.$$

At each column in J off the cycles choose a predecessor with largest A_i; at each column on a cycle choose its predecessor on that cycle. The chosen rows form a set of representatives G.

The bounds follow by applying two elementary inequalities repeatedly along these trees and telescoping. If $x_* \geq x_i \geq 0$ for every i, so that x_* is a largest member of a finite list $(x_i)_i$, then

$$\sum_{i \neq *} x_i^2 \leq \frac{1}{2}\left\{\left(\sum_i x_i\right)^2 - \sum_i x_i^2\right\}, \qquad \sum_{i \neq *} x_i \leq \frac{\mathcal{J}((x_i)_i)}{\log 4}.$$

The first inequality follows from $x_i^2 \leq x_* x_i$ for $i \neq *$. For the second, normalize the list to a probability vector with largest coordinate p_*. If $p_* \geq 1/2$, then $H(p) \geq h_2(p_*) \geq (1-p_*)\log 4$, since the binary entropy lies above its chord on $[1/2, 1]$. If $p_* \leq 1/2$, then $H(p) \geq -\log p_* \geq (1-p_*)\log 4$.

Consider first a component whose tree ends at a low vertex j. Let $s_j = \sum_{i \text{ in this component}} (-d_i)_+$. The sum of the positive parts $(d_i)_+$ over this component is $P_j + s_j$. Applying the first inequality at every vertex shows that the sum of P_i^2 over the omitted rows in this component is at most $(P_j + s_j)^2/2$. The second inequality shows that the total marginal mass of these omitted rows is at most $(P_j + s_j)\log(P_j + s_j)/\log 4$. In both cases the terms $A_i^2/2$ and $A_i \log A_i / \log 4$, respectively, cancel between a child and its parent; the terms that remain at vertices without predecessors are nonnegative, since $A_i \geq P_i > 2$.

In a component containing a cycle, let s_D be the sum of the negative parts $(-d_i)_+$ over the component. The total A_i of the trees entering the cycle is at most s_D. The last vertices of these trees are all omitted, because the predecessors on the cycle were selected. Together with the omitted rows inside these trees, they contribute at most $3s_D^2/2$ to $\sum_{i \in H \setminus G} P_i^2$ and at most $s_D(1 + \log s_D/\log 4)$ to the omitted marginal mass. A component that is a bare cycle contains no omitted row. Each vertex lies in exactly one component, so the numbers s_j, s_D sum to S. Hence

$$\sum_{i \in H \setminus G} P_i^2 \leq \frac{1}{2}\sum_{j \notin H}(P_j + s_j)^2 + \frac{3}{2}\sum_D s_D^2 \leq \frac{1}{2}\sum_{j \notin H} P_j^2 + p_{\max} S + \frac{3}{2}S^2.$$

An omitted row in a component ending at a low vertex has a selected sibling with at least the same value of A, so $P_i \leq (P_j + s_j)/2$. An omitted row in a component containing a cycle has $P_i \leq s_D$. This proves Eq. (136).

The values $P_j + s_j$ and s_D in the preceding argument total at most $2k + e + S$, and each is at most $2 + e + S$. Adding the low rows themselves gives

$$b \leq (2k + e + S)\left(1 + \frac{\log(2 + e + S)}{\log 4}\right).$$

By Eq. (133), $S \leq M + 1$, which gives the bound on b. For E_β, we insert $p_{\max} \leq 2 + e$ and $\sum_{i \notin H} P_i^2 \leq 4k + 4e + e^2$ into (135). Since $(P_i - \beta)^2 \leq P_i^2$,

$$E_\beta \leq 3k + \frac{7}{2}e + \frac{3}{4}e^2 + \frac{1}{2}eM + \frac{5}{2}M + \frac{3}{4}M^2 + \frac{7}{4} \leq \mathsf{Q}(k + M + e).$$

For $\beta = 1$, we subtract $b - |B|/2 \geq 3|B|/2$ from this bound. When $k = 0$, only components containing a cycle occur, which gives the sharper estimates.

For Eq. (138) put $C = [n] \setminus f(G)$. The identity $2x(x-y) = (x-y)^2 + x^2 - y^2$ gives

$$\sum_{i \in G}(P_i - \beta)(P_i - P_{f(i)}) = \frac{1}{2}\sum_{i \in G}(P_i - P_{f(i)})^2 + \frac{1}{2}\sum_{j \in C}(P_j - \beta)^2 - E_\beta. \tag{139}$$

The first two sums on the right are nonnegative, and $\theta_i \geq 1 + P_i - P_{f(i)}$. Since the weights w_i are nonnegative, (138) follows. □

Lemma A.22 (A scalar inequality at large total mass)**.** *Let $1 \leq A \leq 12$, $S \geq 1027$, and $w_i \geq 0$ with $\sum_i w_i = S - A$. Suppose $0 \leq x_i \leq a_0$, $v_i \geq 0$, $v_i = 0$ when $x_i = 0$, and $\sum_i w_i v_i \geq \sum_i w_i$. For $\xi \in \{0, \xi_0\}$,*

$$\sum_i w_i x_i + \max_i\{A x_i + v_i[\mu(x_i) - \xi]\} \geq g(S) - \xi. \tag{140}$$

Proof. Write $q(x) = \mu(x) - \xi > 0$ and $F_D(x) = (D - Ax)/q(x)$, with $F_D(0) = 0$. We first construct a linear upper bound for F_D. On the interval where the numerator of this quotient is nonnegative,

$$F_D''(x) = \frac{-2Aq(1 - e^{-x}) + (D - Ax)(2 - e^{-x}q)}{(1 - e^{-x})^2 q^3}.$$

Where $2 - e^{-x}q \le 0$ this is negative. On the rest of the interval its sign is the sign of $D/A - \Lambda(x)$, where

$$\Lambda(x) = x + \frac{2q(1 - e^{-x})}{2 - e^{-x}q}, \qquad \Lambda'(x) = -\frac{e^{-x}q^2(2 - e^{-x})}{(2 - e^{-x}q)^2} < 0.$$

Hence F_D changes at most once from concave to convex.

For $h \ge 1024$, let

$$t = \log(1 + 1/h), \qquad q_t = \log h - \xi, \qquad D = At + q_t.$$

The tangent to F_D at t is

$$\ell_t(x) = 1 + s_t(x - t), \qquad s_t = \frac{h + 1 - A}{q_t} > 0.$$

Its intercept is positive because $s_t t < 1/q_t < 1$. The function F_D is concave near t, since $e^{-t}q_t > 2$ makes both terms in the numerator of $F_D''(t)$ negative. The tangent also lies above F_D at a_0: the difference between the two there increases with A and is bounded below by

$$1 + \frac{h(a_0 - t)}{\log h - \xi} - \frac{\log h - \xi}{q(a_0)} > 0.$$

The positivity of this bound follows from

$$a_0 - t > \frac{509}{1000}, \quad q(a_0) > \begin{cases} 2/5, & \xi = 0, \\ 19/200, & \xi = \xi_0, \end{cases} \quad \frac{173}{25} < \log 1024 < \frac{347}{50}.$$

At $h = 1024$ it suffices that

$$\frac{509}{1000}\,1024 > \frac{5}{2}\left(\frac{347}{50}\right)^2, \qquad \frac{509}{1000}\,1024 > \frac{200}{19}\left(\frac{663}{100}\right)^2.$$

Both remain true for $h > 1024$ because $h/(\log h - \xi)^2$ increases. Since the curvature changes at most once, $F_D \le \ell_t$ on the interval where the numerator is nonnegative. Where the numerator is negative, the function itself is negative and so lies below ℓ_t, which is positive on $[0, a_0]$.

Let D be the maximum in Eq. (140), and set $t_0 = \log(1025/1024)$, $D_0 = At_0 + \log 1024 - \xi$, $s_0 = s_{t_0}$. The hypothesis on the weighted sum of the v_i excludes $D = 0$. If $D \ge D_0$, there is a unique $t \le t_0$ of the form above with $D = At + q(t)$. Since $v_i \le F_D(x_i) \le \ell_t(x_i)$, multiplying by the weights, summing, and using the last hypothesis gives $\sum_i w_i x_i \ge (S - A)t$. Hence

$$\sum_i w_i x_i + D \ge St + \mu(t) - \xi \ge g(S) - \xi,$$

where the last inequality holds because $g(S)$ is the minimum of $St + \mu(t)$ over $t > 0$.

If $0 < D < D_0$, then $F_D \le (D/D_0)F_{D_0}$. With $W = S - A$, the same summation gives

$$\sum_i w_i x_i + D \ge Wt_0 - W/s_0 + D + WD_0/(s_0 D).$$

Here $W \ge D_0 s_0$, since

$$A + D_0 s_0 = 1025 + \frac{A(1025 - A)\log(1025/1024)}{\log 1024 - \xi} < 1025 + \frac{12(1025/1024)}{661/100} < 1027.$$

The right side is therefore minimized over $0 < D \le D_0$ at $D = D_0$, where it equals $St_0 + \mu(t_0) - \xi \ge g(S) - \xi$. □

Lemma A.23 (Sufficient conditions on the omitted mass and its second moment)**.** *Let $L \geq 2048$ and suppose that a real number $\mathcal{K}$ satisfies*

$$\mathcal{K} \geq \log(n-1) + c + \sum_i w_i x_i + \max_i \{Ax_i + v_i[\mu(x_i) - \xi]\}, \tag{141}$$

$$W := \sum_i w_i \geq L - b - 17, \qquad \sum_i w_i v_i \geq W - E, \tag{142}$$

where $w_i \geq 0$, $0 \leq x_i \leq a_0$, $v_i \geq 0$, and $v_i = 0$ when $x_i = 0$. Suppose $b, E \geq 0$ and $A \geq 1$. Then $\mathcal{K} > B_n(L)$ in each of the following six cases:

1. *$\xi = \xi_0$, $A \leq 3$, $c \geq 1/2$, and*
$$b \leq \mathsf{B}(y), \quad E \leq \mathsf{Q}(y), \qquad y = \min\{11c/2 + 11/10,\ 4c + 4\}.$$
2. *$\xi = \xi_0$, $A \leq 2$, $c \geq 17/20$, and*
$$b \leq \mathsf{B}(y), \quad E \leq \mathsf{Q}(y) - 9/2, \qquad y = \min\{11c/2 + 7/2,\ 4c + 6\}.$$
3. *$\xi = \xi_0$, $A \leq 3$, $c - \xi \geq 11/500$, $b \leq 71/20$, and $E \leq 1681/800$.*
4. *$\xi = \xi_0$, $A \leq 3$, $c - \xi \geq 3/25$, $b \leq 55/8$, and $E \leq 349/128$.*
5. *$\xi = \xi_0$, $A \leq 2$, $473/1000 \leq c \leq 17/20$, and, with $s = (100c + 170)/31$,*
$$b \leq s\left(1 + \frac{\log s}{\log 4}\right), \qquad E \leq \frac{3}{4}s^2 - \frac{9}{2}.$$
6. *$\xi = 0$, $A \leq 3$, $c \geq 1/5$, $b \leq 8c + 6$, and $E \leq Q_0(c) := 8c^2 + 20c + 21/2$.*

Proof. If $c > g(L)$, the result follows from (141), whose last two terms are nonnegative. Otherwise put $U = 1 + \log L$, so $c < U$. The estimates at the end of this proof show that

$$W > 0, \quad \tau := 1 - E/W > 0, \quad A/\tau < 12, \quad (W + A)/\tau > 1027.$$

We apply Lemma A.22 to $(w_i/\tau, v_i/\tau, A/\tau)$. The monotonicity of g, together with $g'(x) < 1/(x-1)$, gives

$$\mathcal{K} - B_n(L) \geq c - \xi - \frac{b + 17 + Eg(L)}{L - b - 18}. \tag{143}$$

Indeed, by that lemma the sum of the last two terms in (141) is at least $\tau\{g((W + A)/\tau) - \xi\}$, and

$$\tau g((W + A)/\tau) \geq \tau g(\min\{W + A, L\}) \geq g(L) - \frac{b + 17}{L - b - 18} - \frac{Eg(L)}{L - b - 17}.$$

It remains to verify the four conditions above and to show that the right side of Eq. (143) is positive. At $L = 2048$ one has $43/5 < U < 87/10$. In cases 1 and 2, the second affine expression for y gives, respectively, the bounds

	upper bound for b	upper bound for E
1	$67203/230$	$126663/100$
2	$106967/345$	$138853/100$.

The ratios $E/(2048 - b - 18)$ are less than $3/4$ and $5/6$, respectively, and both denominators exceed 1027. Cases 3, 4, and 6 have smaller bounds than case 1, and case 5 has smaller bounds than case 2. These inequalities remain valid as L increases: each bound $B(U)$ has logarithmic derivative less than $2/U$, and each bound $Q(U)$ is a positive quadratic polynomial, so $B(U)/L$ and $Q(U)/L$ decrease. Hence $\tau > 1/4$ in cases 1, 3, 4, and 6, and $\tau > 1/6$ in cases 2 and 5, as required.

For the bounds $b \leq B(c)$ and $E \leq Q(c)$ of each case, it remains to prove

$$\mathcal{F}(L, c) := (c - \xi)[L - B(c) - 18] - B(c) - 17 - Q(c)U > 0.$$

In cases 1 and 2 we use the first affine expression for y on $c \le 2$ and the second on $c \ge 2$. On each of these intervals, and throughout case 5,

$$\mathcal{F}_{cc} = -(c+1-\xi)B'' - 2B' - Q''U < 0.$$

Hence it suffices to check the ends of these intervals. At $L = 2048$, substituting $U < 87/10$ and applying (64) with $N = 8$, after scaling the argument of each logarithm into $[1, 2]$ by a power of two, gives the following strict lower bounds:

	left endpoint	$c = 2$, first expression	$c = 2$, second expression
1	114 ($c = 1/2$)	1870	1890
2	344 ($c = 17/20$)	1370	1490

For case 5 the lower bounds are 12 at $c = 473/1000$ and 644 at $c = 17/20$. At all these ends, which do not depend on L, the values increase with L, since $\mathcal{F}_L = c - \xi - Q(c)/L > 0$ there. At the right end $c = U$, which moves with L, lower bounds at $L = 2048$ are 3000 and 1800 in cases 1 and 2. To extend them to all $L \ge 2048$, divide $\mathcal{F}(L, U)$ by L. Its leading term $U - \xi$ increases. The logarithmic derivative of $(U + 1 - \xi)B(U)$ is less than $3/U < 1$; every other subtracted term is a positive polynomial of degree at most three in U divided by $L = e^{U-1}$. Hence all these terms decrease.

In cases 3 and 4 we take the least allowed value of c, since $\mathcal{F}$ increases with c (both B and Q are constant). This gives the positive lower bounds $230041/40000$ and $249829/1280$ at 2048, and their derivatives $c - \xi - Q/L$ remain positive. For case 6,

$$\mathcal{F}(L, c) = c(L - 8c - 24) - 8c - 23 - Q_0(c)U$$

is concave on $[1/5, U]$. At the left end it equals $L/5 - 743/25 - (741/50)U$, which exceeds 250 at 2048 and increases with L. At the right end it equals

$$UL - 8U^3 - 28U^2 - \frac{85}{2}U - 23,$$

which exceeds 9800 at 2048 and remains positive by the argument used for cases 1 and 2. □

Lemma A.24 (The bound for $L \ge 2048$)**.** *For $5 \le n \le 52$, $P_i > 2$, and $L \ge 2048$, every row-stochastic matrix R satisfies $K_R(P) \ge B_n(L)$.*

Proof. We may assume that all extreme-point costs are finite, since otherwise there is nothing to prove. First let $2 \le m < n$, and choose G as in Lemma A.21. Let $\beta = \beta_\kappa$, $A = A_\kappa$, $w_i = P_i - \beta$, $b = \sum_{i \notin G} P_i$, and $E = \frac{1}{2}\sum_{i \notin G}(P_i - \beta)^2$. Then

$$A \le 3, \quad m\beta < 17, \quad W = L - b - m\beta > L - b - 17, \quad \sum_{i \in G} w_i\theta_i \ge W - E.$$

Define

$$\ell = \log\frac{n-1}{\kappa}, \qquad \xi = \begin{cases} \xi_0, & r > 0, \\ 0, & r = 0, \end{cases} \qquad c = \xi M + a_0\big(k + e + k\mathbb{1}_{\{\kappa=1\}}\big) - \ell.$$

Taking the largest source average in (132) gives the two hypotheses of Lemma A.23, with $\mathcal{K} = K_R(P)$, $x_i = a_i$, and $v_i = \theta_i$. It remains to show that one of the six cases of Lemma A.23 applies.

Repeated maximizing columns, $\kappa \ge 2$. Here $r \ge 1$. Let $X = M + e$ and $Y = k + X$. Since $X \ge 2r$,

$$c \ge \xi_0 Y + (a_0 - \xi_0)k - \log(1 + (Y + k)/4) \ge \xi_0 Y - \log(1 + Y/4).$$

For the second inequality we use $Y \ge 2$ and $\log(1 + k/(Y + 4)) \le k/6 < (a_0 - \xi_0)k$ when $k > 0$. The tangent lines of the logarithmic term at $Y = 4$ and $Y = 16$ give

$$Y < \min\{11c/2 + 11/10,\ 4c + 4\}.$$

Hence, if $c \ge 1/2$, case 1 of Lemma A.23 applies.

If $c < 1/2$, the bound

$$c \ge 2\xi_0 r + a_0 k - \log(1 + (k + r)/\kappa)$$

leaves only $(k,r,m)=(0,1,m)$ with $4\le m\le 8$, or $(k,r,m)=(1,1,3)$. For fixed $k+r$ the right side is smallest at $r=1$. For $k+r\ge 3$ it increases with $k+r$, and its value at $k+r=3$ is at least $2\xi_0+2a_0-\log(5/2)>1/2$. For $k+r=2$, the choice $r=2$ is excluded by $4\xi_0-\log 2>1/2$, and the choice $k=r=1$ forces $\kappa=2$. For $k+r=1$, one has $k=0$, $r=1$, and $m\ge 4$; $m\ge 9$ is excluded by $2\xi_0-\log(9/8)>1/2$.

For $(k,r,m)=(0,1,m)$, $4\le m\le 8$, exactly one row lies outside G, $\beta>3/2$, and $c<1/2$ implies $M<51/20$. The bound for a single row in (136) gives

$$b<71/20,\qquad E<1681/800,\qquad c-\xi_0\ge \xi_0-\log(4/3)>11/500.$$

This is case 3. For $(k,r,m)=(1,1,3)$ we have $\beta=2$, $2\le X<9/4$, and $c-\xi_0>3/25$. The same single-row bound gives

$$\begin{aligned} b &\le 2+e+\max\{X-e+1,(X+3)/2\}<55/8,\\ E &\le \frac{1}{2}e^2+\frac{1}{2}\max\{X-e-1,(X-1)/2\}^2\le \frac{5}{8}X^2-\frac{1}{4}X+\frac{1}{8}<349/128. \end{aligned}$$

For the second inequality on E, the middle expression is convex in e on $0\le e\le X$, so its maximum is attained at an end of this interval, namely at $e=X$. This is case 4.

Repeated maximizing columns, $\kappa=1$. Now $m=2$, $\beta=1$, $A=2$, and $|B|=n-2\ge 3$, so $E\le \mathsf{Q}(Y)-9/2$. Since $\kappa=1$, the definition of c contains the additional term a_0k, which gives

$$c\ge \xi_0 Y+(2a_0-\xi_0)k-\log(1+(Y+k)/2)\ge \xi_0 Y-\log(1+Y/2).$$

The tangent lines of the logarithmic term at $Y=6$ and $Y=18$ give $Y<\min\{11c/2+7/2,4c+6\}$. Hence, if $c\ge 17/20$, case 2 applies. Otherwise the bound

$$c\ge 2\xi_0 r+2a_0k-\log(k+r+1),\qquad k+r\ge 3,$$

leaves only $k=0$, $r=3$, and $n=5$. For $k+r\ge 4$ we use $8\xi_0-\log 5>17/20$ and monotonicity; for $k+r=3$ and $k\ge 1$ we use $4\xi_0+2a_0-\log 4>17/20$. In the remaining case,

$$c>473/1000,\qquad S\le M+1<\frac{100c+170}{31}.$$

The bounds of Lemma A.21 for b and E_1 in the case $k=0$ now give case 5.

Distinct maximizing columns, $m<n$. Here $r=0$, $k\ge 1$, $G=H$, and $\xi=0$. Since $r=0$ and $n\ge 5$, $k=1$ forces $\kappa=n-2\ge 3$. The smallest possible lower bounds for c are therefore

$$a_0-\log(4/3),\qquad 2a_0-\log 2,\qquad 6a_0-\log 4,$$

according as $\kappa\ge 2, k=1$; $\kappa\ge 2, k\ge 2$; or $\kappa=1$. Each is greater than $1/5$, and the corresponding lower bounds increase with k. Using $\log(k+1)\le \log 4+(k-3)/4$ in the definition of c gives $k+e<4c+3$. Hence

$$b=2k+e<8c+6,\qquad E\le 2(k+e)+(k+e)^2/2<Q_0(c),$$

which is case 6.

It remains to treat $m=n$, $m=0$, and $m=1$. If $m=n$, apply Lemma A.19. If $m=0$, every entry of R is at most $3/5$, so $K_R(P)\ge a_0L>B_n(L)$.

Finally suppose $m=1$. Let j be the only maximizing column, and let $b_0=\sum_{i\notin H}P_i$. The cost of the central transport is at most $K_R(P)$, since P is the average of the row marginals at two opposite extreme points q_{uv} and q_{vu}. If $j\notin H$, the rows in H send at least $(L-2b_0)_+$ mass outside column j, so the cost of the central transport is at least

$$a_0b_0+\log(5/2)(L-2b_0)_+\ge a_0L/2.$$

If $j\in H$ but $R_{jj}\le 5/8$, this cost is at least both $a_1(P_j+b_0)$ and $\log(5/2)(L-P_j-b_0)_+$, where $a_1=\log(8/5)$. It is therefore at least $a_1L/2$. Both lower bounds exceed $B_n(L)<2(1+\log L)$ for $L\ge 2048$.

It remains to consider $z=R_{jj}>5/8$. Take an extreme point with source j, and write the mass that row j of its optimal transport sends outside column j as $1+x$, $x\ge 0$. Then row j sends P_j-x to column j, and the other rows together send x to that column. Their remaining mass is $L-P_j-1-x$,

and every entry of R that this mass uses is at most $3/5$. The cost of this extreme point is therefore at least

$$\begin{aligned} &- P_j \log z - \log(1-z) + x \log \frac{z}{1-z} + a_0(L - P_j - 1 - x) \\ &\qquad \geq g(P_j + 1) + a_0(L - P_j - 1), \end{aligned}$$

because $\log(z/(1-z)) > a_0$ and the first two terms have minimum $g(P_j + 1)$. The case $z = 1$ has infinite extreme-point cost. Writing $D = L - P_j - 1 \geq 2n - 3$, the last bound minus $B_n(L)$ increases with D for fixed L, since $g'(P_j + 1) < g'(3) < a_0$. Therefore

$$K_R(P) - B_n(L) \geq (2n-3)a_0 - \log(n-1) - \frac{2n-3}{L - 2n + 2} > 0.$$

The first two terms are at least $7a_0 - \log 4 > 2$, while the fraction is at most $101/1946$. □

A.10 The computer-assisted cases

Proposition A.25 (The remaining cases)**.** *Let $P_i > 2$ and let R be row stochastic. Then*

$$K_R(P) \geq B_n(L)$$

whenever either $n \in \{3, 4\}$ and $L \geq 2n$, or $5 \leq n \leq 52$ and $2n \leq L \leq 2048$.

Proof sketch. The ranges of n and L covered by each part of the verification are listed in Tables 2 and 3. In dimensions three and four the verification covers all $L \geq 2n$, including unbounded mass intervals. For dimensions five through fifty-two it covers $L \leq 2048$.

The inequality to be verified. At a minimizer with finite value, let X^{uv} be the optimal transports chosen by the rule (52). For any probability distribution λ on extreme points, Gibbs' inequality gives

$$K_R(P) \geq \sum_{u \neq v} \lambda_{uv} C_{uv}(R, P) \geq \mathcal{H}\left(\sum_{u \neq v} \lambda_{uv} X^{uv} \right). \tag{144}$$

When T_s is minimized, the same bound holds for convex combinations of source averages over at least $s-1$ sinks. For the objective $G \in \{K, T_s\}$ used in a given dimension, the stationary distribution also gives the identity and column constraints

$$G(R, P) = \sum_i v_i H(R_i), \qquad \sum_i v_i R_{ij} = P_j, \qquad v = P + \alpha - \zeta, \quad \sum_i \alpha_i = \sum_i \zeta_i = 1.$$

Each comparison below uses the marginals, the transports X^{uv} and these relations at a single minimizer.

In the normalized variables $t = 1/L$ and $p = tP$, the marginals range over a compact domain, and the normalized threshold is h_n:

$$\mathcal{D}_n = \left\{ (p, t) : \sum_i p_i = 1,\ p_i \geq 2t,\ 0 \leq t \leq \frac{1}{2n} \right\}, \tag{145}$$

$$h_n(t) = -(1-t)\log(1-t) - t \log \frac{t}{n-1}, \qquad h_n(0) = 0. \tag{146}$$

Each vertex of a transportation polytope has a forest as support. Deleting an edge of that forest expresses the entry on that edge as a signed sum of row and column marginals. Hence, for a fixed support, the normalized transport tX is affine in (p, t) on the region where it is nonnegative.

Dimensions three and four. At a positive minimum, the entries of the central transport X^0 greater than one form a complete matching. We relabel its columns and the marginal P accordingly, writing $S_j = P_{f(j)}$ and $Q_{ij} = R_{i,f(j)}$. Exchange arguments at the minimizer give

$$|P_i - S_i| < 1, \qquad Q_{ii} > \frac{1}{2}, \qquad Q_{kk} Q_{ij} \geq Q_{ik} Q_{kj}.$$

Hence $\log(Q_{jj}/Q_{ij})$ is a directed metric. Cancelling cycles and shortening paths shows that direct transports from surplus rows to deficit columns are optimal. The remaining choices are the permutation f and, in dimension four, the two ends of the interval of solutions of each residual two-by-two transportation problem.

If R has zero entries, the same reductions apply to paths through positive entries of R, with the diagonal masses along each path as capacities. In dimension three there are nine choices of a zero entry, 32 cost chambers and eleven marginal regions. These give $1{,}314$ jointly optimal families of transports and, up to symmetry, 590 entropy expressions. Dimension four has 149 classes of supports with at least one zero entry. They are organized by the number of independent cycles of the support and by the entries of X^0 greater than one; the path reductions are carried out for every matching permutation f.

In dimension four, the aligned case $S = P$ needs no path enumeration. If $S = P$, any nonzero off-diagonal flow of X^0 contains a directed cycle, which the positive diagonal would turn into a cycle of the bipartite support of X^0. That support is a forest, so $X^0 = \operatorname{diag}(P)$. Lemma A.8 and stationarity then give

$$Q_{ii} \geq \frac{P_i - 1 + \alpha_i}{P_i + \alpha_i - \zeta_i} \geq \frac{P_i - 1}{P_i} \geq \frac{1}{2},$$

and the matching estimate with equal marginals $P_i = P_{f(i)}$ applies.

For instance, with $\nu = 1/(n-1)$ and $D = L - 1 - \nu$, the source distribution of Lemma A.13 gives

$$N = \frac{1}{n-1} \sum_{u \neq v} (P_u - \nu) X^{uv}, \qquad K_R(P) \geq \frac{\mathcal{H}(N)}{D}.$$

It therefore suffices to prove $\mathcal{H}(N) \geq D B_n(L)$. Other families use fixed rational distributions on extreme points. On each marginal region the resulting matrices are affine or quadratic in the marginals. In the affine case, concavity of the entropy reduces the comparison to the vertices of the region. Quadratic dependence is handled by the identity

$$N\left(\sum_i \theta_i x_i\right) = \sum_i \theta_i^2 N(x_i) + 2 \sum_{i<j} \theta_i \theta_j C_{ij}, \qquad C_{ij} = 2N\left(\frac{x_i + x_j}{2}\right) - \frac{N(x_i) + N(x_j)}{2}.$$

The matrices C_{ij} are nonnegative on the corresponding regions. Entropy bounds for $N(x_i)$ and C_{ij}, together with convexity of $(L - 4/3)B_4(L)$ in dimension four, give the required inequality on each marginal region. The counts and unbounded intervals are recorded in Table 2.

Dimensions five through fifty-two. Fix a threshold $z_0 > 1/2$ for the entries of R; the value $z_0 = 1/2$, with strict inequality, is used in the dimension-specific calculations for $5 \leq n \leq 8$. Let $I = \{i : \max_j R_{ij} > z_0\}$ and let $f(i)$ be the unique maximizing column of row i. Write

$$a = |f(I)|, \qquad k = n - |I|, \qquad r = |I| - a.$$

For each column $j \in f(I)$, choose a row of largest marginal among the rows of I whose maximum lies in column j. The row estimates bound the source averages of these chosen rows, and the capacity of the shared columns bounds the contribution of the other rows. The cases $a = 0$ and $a = 1$ are bounded analytically. The complete-matching case, which contains the equality case, is treated separately. For unequal marginals P_i and $P_{f(i)}$, the row bounds contain, for a constant c, terms of the form

$$\frac{c}{2(L - 1 - \nu)} \big\{ (P_i - \nu)^2 - (P_{f(i)} - \nu)^2 \big\}, \qquad \sum_i \big\{ (P_i - \nu)^2 - (P_{f(i)} - \nu)^2 \big\} = 0.$$

This cancellation turns the row bounds into the sharp bound for the sum. Below $L = 128$ the resulting scalar inequalities are verified on rational domains, with a different source distribution for the exceptional five-dimensional cases. For $128 \leq L \leq 2048$, the full-matching bound is analytic.

For the remaining cases we take $z_0 = 2/3$ and work on the domains defined by the counts (a, k, r). Write the marginal mass of the k rows outside I as $2k + \xi$. If $r > 0$, let M be the minimum, over the chosen rows as sources, of the average mass outside the cells $(i, f(i))$, $i \in I$; the average is over the

other chosen rows as sinks. Set $X = M + \xi$; for $r = 0$ it suffices to set $X = \xi$. The row and column estimates imply

$$K_R(P) \geq \frac{2}{5}(k + X), \qquad X \geq 2r. \tag{147}$$

Hence a minimizer with $K_R(P) < B_n(L)$ on a mass interval where $B_n(L) < 2C/5$ belongs to one of the closed domains

$$\begin{gathered} 2 \leq a < n, \qquad k \geq 0, \qquad r = n - a - k \geq 0, \qquad k + 2r < C, \\ L_0 \leq L \leq L_1, \qquad 2r \leq X \leq C - k. \end{gathered} \tag{148}$$

For fixed n, C, the number of these initial domains is

$$\sum_{b=1}^{\min(n-2,C-1)} \big(\min\{b, C - 1 - b\} + 1\big).$$

Where the counts (a, k, r) are too coarse, the dimension-specific calculations use more information about f. Since the domains are closed, equality at a threshold and shared boundaries are included.

On each final domain the inequality follows from capacity bounds for the extreme points and the columns, affine row and shared-column estimates, bounds on the mass outside the row maxima and on moments of the marginals of the omitted rows, or the entropy of a convex combination of the averages that define G. Estimates that depend on the support of a transport are checked for every forest support allowed on the domain. A domain is discarded when necessary conditions at the same minimizer are inconsistent on it. Table 3 lists the initial and final domain counts. The five-dimensional case is completed on all mass intervals, including the analytic unbounded interval, before its diagonal-row reduction is used in dimension six.

Exact comparisons and coverage. Logarithms are enclosed by the rational bounds in Eq. (64). If a mixture $M(t)$ of normalized transports is affine on $[a, b] \subset (0, 1)$ and $\mathcal{H}(M(t)) - h_n(t)$ exceeds

$$\frac{(b - a)^2}{8a(1 - b)}$$

at both ends of the interval, then $\mathcal{H}(M(t)) \geq h_n(t)$ throughout the interval, by concavity of the entropy and $h_n''(t) = -1/[t(1 - t)]$. Quadratic matrices are handled by the identity above. The unbounded three- and four-dimensional intervals use expansions with explicit remainders, followed by a reciprocal change of variable and nonnegative Bernstein coefficients. Each domain is the union of its subdomains, so the inequalities proved on the final domains cover every initial domain, including degenerate faces. Hence no class in the two tables contains a minimizer with $K_R(P) < B_n(L)$. □

Code availability. The code for the computer-assisted verification is available at `https://github.com/realgourmet/e-watermark`.

Dimension and family	Finite classification	Verification and size
$n = 3$, positive support	Marginal-preserving permutations, transpositions, and the two orientations of a three-cycle	Entropy of the direct mixture; concavity in the gap of the middle marginal. Four rational polynomial inequalities settle the cycle cases.
$n = 3$, proper support	9 forbidden positions, 32 cost chambers, 11 marginal regions; 1,314 families of optimal transports in 590 entropy classes	576 entropy classes compared on marginal regions; 7 classes with diagonal central transport (9 families); 7 classes with mass rerouted along a path (10 families).
$n = 4$, all supports	36,317 labeled connected spanning supports; 150 classes, including full support	The 149 proper classes have 7–15 edges. Every permutation found numerically is included.
$n = 4$, full support	572 direct families of optimal transports, 4,996 labeled quadratic controls, 657 entropy classes	4 classes containing equality cases; the other 653 classes use 11,754 comparisons on finite intervals and 5,224 coefficients for the bounds on unbounded intervals.
$n = 4$, no perfect matching, or an exceptional central transport without a perfect matching	Shared-neighbor supports and the exceptional seven-edge tree supporting the central transport, in both orientations	Analytic entropy bounds exclude these cases on their complete marginal domains.
$n = 4$, one support cycle	28 classes; 49,482 quadratic controls	21,505 finite intervals and 14,374 unbounded intervals. Seven-edge trees follow by support inclusion.
$n = 4$, two edge-disjoint support cycles	3 classes; 18,697 quadratic controls	10,774 finite intervals and 6,620 unbounded intervals.
$n = 4$, three paths joining two vertices	29 classes; 1,329,008 quadratic controls	The six orderings of the three path costs give all extreme mixtures; 499,333 finite and 237,736 unbounded intervals.
$n = 4$, five entries of X^0 greater than one	108 placements in ten-edge supports; 41 in eleven-edge supports, with all permutations found numerically	Contracting the trees of these entries gives three components. The ten-edge case has 4,892 cost chambers, 58,742 vertex curves and 222,206 controls.
$n = 4$, aligned matching	A forest central transport with $S = P$ is diagonal	Stationarity gives $Q_{ii} = R_{i,f(i)} \geq (P_i - 1)/P_i \geq 1/2$; the marginal-preserving matching estimate is analytic.
$n = 4$, remaining supported paths	3,141 path systems with $\lvert P_i - S_i \rvert \leq 1$; 660 for each sign of one larger displacement; 56 with two larger displacements	All supports with 10–15 edges. The ten-edge case has 263,060 entropy curves, 754,808 finite intervals and 263,060 unbounded intervals.

Table 2: The computer-assisted cases in dimensions three and four, for every $P_i \geq 2$ and every $L \geq 2n$. The classifications include all families of transports at extreme points that are optimal for the same cost matrix. A quadratic control is one of the matrices in the convex representation of N in Appendix A.10. The unbounded intervals are part of the computer-assisted entropy verification. The support counts by edge number $7, 8, \ldots, 16$ are $14, 28, 32, 32, 19, 14, 6, 3, 1, 1$.

Dimension	Mass interval	Classification or count cap	Cases, or initial / final domains
5	$[10, 112]$	Threshold $1/2$; at most three selected columns, injective maps, and the remaining $(\lvert I\rvert, a) = (5, 4)$ maps	14 remaining graph types
6	$[12, 112]$	Threshold $1/2$; at most four selected columns, injective maps, and $(\lvert I\rvert, a) = (6, 5)$; diagonal-row reduction uses $n = 5$	26 remaining graph types
7	$[14, 112]$	Threshold $1/2$; at most five selected columns, injective maps, and $(\lvert I\rvert, a) = (7, 6)$	One repeated-column pair
8	$[16, 112]$	Threshold $1/2$; residual counts $(8, 7), (8, 6), (7, 6)$	37 maximizing-column cases
9–14	$[2n, 112]$	Threshold $2/3$; $n - 6 \le a < n$, $0 \le r \le 2$, $k = n - a - r \ge 0$, with all lower-threshold refinements	17 count classes
$15, 16$	$[2n, 112]$	Minimize T_9, T_{10}, respectively; $n - 6 \le a < n$, $0 \le r \le 3$, $k = n - a - r \ge 0$	21 count classes
7–20	$[112, 128]$	$C = 25$ and $a + k > c_n$, with the c_n listed below; complementary classes use source and capacity bounds	751 / 15,787
17	$[34, 128]$	$C = 25$; the additional $(a, k, r) = (14, 2, 1)$ class is split into ten group-count types	122 plus one domain
$18, 19$	$[2n, 128]$	$C = 22$	237 / 7,660
$20, 21$	$[2n, 128]$	$C = 25$	300 / 11,747
22	$[44, 128]$	$C = 25$	158 / 3,990
23–25	$[2n, 128]$	$C = 25$	494 / 5,963
26	$[52, 128]$	$C = 25$	168 / 2,623
27	$[54, 128]$	$C = 25$	168 / 2,059
28–52	$[2n, 128]$	$C = 25$	4,200 / 12,838
5–52	$[128, 144]$	$C = 25$	6,533 / 36,275
5–52	$[144, 192]$	$C = 26$	6,949 / 40,437
5–52	$[192, 448]$	$C = 28$	7,817 / 51,528
5–63	$[448, 2048]$	$C = 32$	12,579 / 38,374

Table 3: The computer-assisted cases in dimensions five through fifty-two. The counts a, k, r are defined in Appendix A.10, and the domains with cap C in Eq. (148). For the $[112, 128]$ row, $(c_7, \dots, c_{20}) = (5, 5, 5, 5, 6, 7, 8, 9, 10, 11, 11, 12, 13, 14)$. The source estimates in dimensions five and six and the domains for $n \ge 18$ also cover this interval. The cases $a = 0$, $a = 1$, and $a = n$ are handled separately in the proof of Proposition A.25; they are not part of the displayed count subdivisions. On every final domain, either the inequality holds throughout or its necessary constraints are inconsistent. In the last row, only $n \le 52$ is needed.

A.11 Completion of the proof

Lemma A.26 (The bound in all dimensions)**.** *For $n \geq 2$, $P_i \geq 2$, and every row-stochastic R,*

$$K_R(P) \geq B_n(L), \qquad L = \sum_i P_i. \tag{149}$$

Proof. First let $P_i > 2$. The binary case is Lemma A.10, and the case $n \geq 53$ is Lemma A.18. For $n = 3, 4$, the bound is Proposition A.25. For $5 \leq n \leq 52$, the same proposition applies when $L \leq 2048$, and Lemma A.24 applies when $L \geq 2048$. Now let R be positive and $P_i \geq 2$. Adding εI_n to any optimal transport at an extreme point gives

$$K_R(P + \varepsilon \mathbf{1}) \leq K_R(P) + \varepsilon \sum_i (-\log R_{ii}). \tag{150}$$

Applying the case $P_i > 2$ at total mass $L + n\varepsilon$ and letting $\varepsilon \downarrow 0$ gives the bound for positive R. For arbitrary R, we apply it to $(1-\eta)R + \eta \mathbf{1}\mathbf{1}^\top/n$ and let $\eta \downarrow 0$ in (63). This proves, also for R with zero entries,

$$K_R(P) \geq g(L) + \log(n-1) \qquad (P_i \geq 2). \tag{151}$$

□

Proof of Proposition A.1 and Theorem 4.1. Let $t = \delta/2$, $P = p/t$ and $L = 1/t$. Then $P_i \geq 2$ and $q_i \geq p_i - t \geq t > 0$ throughout the uncertainty set. The normalization of Appendix A.2, Lemma A.7, and Lemma A.26 give

$$F_{p,t}(R) = -tK_R(P) \leq (1-t)\log(1-t) + t \log \frac{t}{n-1}. \tag{152}$$

For attainment, take

$$R^*_{ii} = 1 - t, \qquad R^*_{ij} = t/(n-1) \quad (i \neq j). \tag{153}$$

Since $t \leq 1/(2n)$, the payoff of a coupling W under R^* increases with its diagonal mass:

$$\langle W, \log R^* \rangle = D(W)\log(1-t) + (1 - D(W)) \log \frac{t}{n-1}. \tag{154}$$

Writing $\tau = \mathtt{TV}(q, p)$, every coupling has

$$D(W) \leq \sum_i \min(q_i, p_i) = 1 - \tau. \tag{155}$$

The coupling Eq. (11) attains this bound: the surplus and the deficit of q relative to p have disjoint supports and the same total τ, so its off-diagonal row and column sums are $(q-p)_+$ and $(p-q)_+$. At $\tau = 0$ it is $\mathrm{diag}(p)$. Hence

$$T_{R^*}(q, p) = (1-\tau)\log(1-t) + \tau \log \frac{t}{n-1}. \tag{156}$$

This is strictly decreasing in τ, and the extreme points have $\tau = t$. The infimum is therefore attained exactly at the q with $\mathtt{TV}(q, p) = t$, and

$$F_{p,t}(R^*) = (1-t)\log(1-t) + t \log \frac{t}{n-1}. \tag{157}$$

Together with Eq. (152) and (41), this proves the claimed value and the optimal detector $e^*_{ij} = R^*_{ij}/p_j$. The optimal generator is $W^{\mathrm{M}}_{ij}(q)/p_j$; at $q = p + t(\mathbf{e}_a - \mathbf{e}_b)$, its coupling is $\mathrm{diag}(p) + t(\mathbf{e}_a\mathbf{e}_b^\top - \mathbf{e}_b\mathbf{e}_b^\top)$, as in Eq. (4). □

Nonuniqueness. Any permutation σ preserving the anchor distribution, $p_{\sigma(j)} = p_j$, maps an optimal detector to another one by $e^\sigma(i, j) = e(i, \sigma(j))$: permuting the columns of a coupling by σ preserves both of its marginals and leaves its payoff unchanged when e is replaced by e^σ. For the uniform binary anchor distribution, e^* and the detector obtained by swapping its columns are distinct optimal detectors.

B Proof of Theorem 4.3

Proof. Write

$$J_t^* = H(p_0^t) - H(\nu_\delta), \qquad K_T = \sum_{t=1}^{T} J_t^*.$$

By Eq. (6), $K_T/T \to J_\infty^*$ almost surely under every admissible choice of strategies. Let

$$u = (1-(n-1)\delta, \delta, \dots, \delta), \qquad j_0 = H(u) - H(\nu_\delta).$$

The vector u majorizes every distribution with coordinates at least δ, and ν_δ strictly majorizes u. Hence $J_t^* \geq j_0 > 0$ for all t, and $J_\infty^* \geq j_0 > 0$. For a detector $e = (e_t)_{t\geq 1}$, set

$$Y_t = \log e_t(v^t, s^t), \qquad S_T = \sum_{t=1}^{T} Y_t, \qquad \tau_B = \inf\{T \geq 1 : S_T \geq B\},$$

with $S_0 = 0$ and $\log 0 = -\infty$.

We first prove the converse. Fix a predictable e-value strategy e and $\lambda > 0$, and define

$$\widetilde{e}_t(v,s) = \frac{e_t(v,s) + \lambda}{1+\lambda}.$$

Each $\widetilde{e}_t$ is a valid conditional e-value. Since validity implies $e_t(v,s) \leq 1/p_0^t(s) \leq 1/\delta$, we have

$$\frac{\lambda}{1+\lambda} \leq \widetilde{e}_t(v,s) \leq \frac{1}{\delta}.$$

Thus $\widetilde{Y}_t = \log \widetilde{e}_t(v^t, s^t)$ is uniformly bounded. At each history, define

$$F_t(q) = \sup_{w \in \mathcal{P}(p_0^t, q)} \sum_{v,s} w(v,s) \log \widetilde{e}_t(v,s).$$

Mixing feasible couplings shows that F_t is concave in q. Since $\mathcal{Q}(p_0^t, \delta)$ is the convex hull of the finitely many points

$$q_{a,b}^t = p_0^t + \frac{\delta}{2}(\mathbf{e}_a - \mathbf{e}_b), \qquad a \neq b,$$

its minimum is attained at one of these points. By Theorem 4.1,

$$\min_{a \neq b} F_t(q_{a,b}^t) \leq J_t^*.$$

Let $\mathcal{A}_\lambda$ choose a minimizing point at each step, breaking ties by a fixed ordering of the pairs (a,b). This defines a predictable adversary independently of B, and for every generator,

$$\mathbb{E}[\widetilde{Y}_t \mid \mathcal{F}_{t-1}] \leq J_t^* \qquad \text{a.s. for all } t.$$

The martingale

$$\widetilde{D}_T = \sum_{t=1}^{T} \Big(\widetilde{Y}_t - \mathbb{E}[\widetilde{Y}_t \mid \mathcal{F}_{t-1}] \Big)$$

has uniformly bounded increments. The martingale strong law gives $\widetilde{D}_T/T \to 0$ almost surely, so

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \widetilde{Y}_t \leq J_\infty^* \qquad \text{a.s.}$$

Since $\log e_t(v,s) \leq \log \widetilde{e}_t(v,s) + \log(1+\lambda)$, it follows that

$$\limsup_{T\to\infty} \frac{S_T}{T} \leq J_\infty^* + \log(1+\lambda) \qquad \text{a.s.}$$

Also $Y_t \leq \log(1/\delta)$, so $\tau_B \geq B/\log(1/\delta)$. Applying the preceding growth bound at the crossing times yields

$$\liminf_{B\to\infty} \frac{\tau_B}{B} \geq \frac{1}{J_\infty^* + \log(1+\lambda)} \qquad \text{a.s.}$$

Fatou's lemma therefore gives

$$\inf_{\mathcal{G}} \liminf_{\alpha \downarrow 0} \frac{\mathbb{E}_{\mu(\mathcal{A}_\lambda, \mathcal{G})}[\tau_\alpha(e)]}{\log(1/\alpha)} \geq \frac{1}{J_\infty^* + \log(1+\lambda)}.$$

Taking the supremum over adversaries and then letting $\lambda \downarrow 0$ proves the converse.

For achievability, fix an adversary and use e_t^* from Eq. (3) with anchor p_0^t. Let the generator choose a maximal coupling of q^t and p_0^t at each step. By Theorem 4.1,

$$m_t := \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] \geq J_t^* \geq j_0 \qquad \text{a.s. for all } t.$$

The log-increments satisfy

$$|Y_t| \leq M := \log \frac{2(n-1)}{\delta}.$$

Consequently, $D_T = \sum_{t=1}^T (Y_t - m_t)$ has increments bounded in absolute value by $2M$, and $D_T/T \to 0$ almost surely. Since $S_T \geq K_T + D_T$,

$$\liminf_{T\to\infty} \frac{S_T}{T} \geq J_\infty^*, \qquad \limsup_{B\to\infty} \frac{\tau_B}{B} \leq \frac{1}{J_\infty^*} \qquad \text{a.s.}$$

For every integer $m \geq 2B/j_0$, Azuma's inequality gives

$$\begin{aligned}
\mathbb{P}(\tau_B > m) &\leq \mathbb{P}(S_m < B) \\
&\leq \mathbb{P}(D_m \leq B - j_0 m) \\
&\leq \exp\left(-\frac{j_0^2 m}{32M^2}\right).
\end{aligned}$$

This bound implies that τ_B is integrable and

$$\sup_{B \geq 1} \mathbb{E}\left[\left(\frac{\tau_B}{B}\right)^2\right] < \infty.$$

Thus $\{\tau_B/B : B \geq 1\}$ is uniformly integrable, and the almost-sure upper bound yields

$$\limsup_{\alpha \downarrow 0} \frac{\mathbb{E}_{\mu(\mathcal{A}, \mathcal{G})}[\tau_\alpha(e^*)]}{\log(1/\alpha)} \leq \frac{1}{J_\infty^*}.$$

The maximal-coupling generator is available for every adversary. Taking the infimum over generators and the supremum over adversaries proves achievability. Together with the converse, this establishes the stated minimax equality. □

C Useful Claims

Lemma C.1. *Let $\mathcal{V}$ be the set $\{1, \ldots, n\}$ and $p_0 \in \Delta(\mathcal{V})$ be a distribution over $\mathcal{V}$ such that $\min_{v \in \mathcal{V}} p_0(v) > \delta$. Define*

$$\mathcal{Q}(p_0, \delta) = \{q \in \Delta(\mathcal{V}) : \|q - p_0\|_1 \leq \delta\}$$

Then $\mathcal{Q}(p_0, \delta)$ is a convex polytope whose vertex set is given by:

$$\mathcal{Q}_{\text{ext}}(p_0, \delta) := \{p_0 + (\mathbf{e}_i - \mathbf{e}_j) \cdot \delta/2 : (i, j) \in \mathcal{V} \times \mathcal{V}, i \neq j\}.$$

Proof. Recall $\mathcal{Q}(p_0, \delta) := \{q \in \Delta^n : \|q - p_0\|_{\ell_1} \leq \delta\}$. First, we note that $\mathcal{Q}(p_0, \delta)$ is the intersection of two convex polytopes, hence it must also be a convex polytope. We claim that $\mathcal{Q}(p_0, \delta) = \text{Conv}(\mathcal{Q}_{\text{ext}}(p_0, \delta))$. Suppose $q \in \mathcal{Q}(p_0, \delta)$, then we can write that $q_i - p_i = s_i$ for $s_i \in [-\delta/2, \delta/2]$, $\sum s_i = 0$, and $\sum |s_i| \leq \delta$. Next, suppose for contradiction that $s_j > \delta/2$, then $\sum_{i \neq j} s_i = -s_j$ and hence,

$$\sum_k |s_k| \geq |s_j| + |\sum_{i \neq j} s_i| > \delta,$$

which violates the TV constraint.

Now we show that every $q \in \mathcal{Q}(p_0, \delta)$ lies in the convex hull of $\mathcal{Q}_{\text{ext}}(p_0, \delta)$. Write $s = q - p_0$ and set

$$\eta := \frac{\|q - p_0\|_1}{\delta} \in [0, 1].$$

Let $P = \{i : s_i > 0\}$ and $N = \{j : s_j < 0\}$. Choose nonnegative α_{ij} for $i \in P$, $j \in N$ such that

$$\sum_{j \in N} \alpha_{ij} = \frac{s_i}{\delta/2}, \qquad \sum_{i \in P} \alpha_{ij} = \frac{-s_j}{\delta/2}.$$

Then $\sum_{i,j} \alpha_{ij} = \eta$ and

$$p_0 + \frac{\delta}{2} \sum_{i \in P, j \in N} \alpha_{ij}(\mathbf{e}_i - \mathbf{e}_j) = q.$$

To make a convex combination, choose any symmetric probability mass ρ_{ij} on ordered pairs $i \neq j$, i.e., $\rho_{ij} = \rho_{ji}$ and $\sum_{i \neq j} \rho_{ij} = 1$. Then

$$\sum_{i \neq j} \rho_{ij} \left(p_0 + \frac{\delta}{2}(\mathbf{e}_i - \mathbf{e}_j) \right) = p_0.$$

Define

$$\lambda_{ij} = \begin{cases} \alpha_{ij} + (1 - \eta)\rho_{ij}, & i \in P,\ j \in N, \\ (1 - \eta)\rho_{ij}, & \text{otherwise.} \end{cases}$$

The coefficients λ_{ij} are nonnegative, sum to one, and satisfy

$$q = \sum_{i \neq j} \lambda_{ij} \left(p_0 + \frac{\delta}{2}(\mathbf{e}_i - \mathbf{e}_j) \right).$$

It is now enough to show that any $v \in \mathcal{Q}_{\text{ext}}(p_0, \delta)$ cannot be generated by a convex combination of two distinct points in $\mathcal{Q}(p_0, \delta)$ which will prove that $v \in \mathcal{Q}_{\text{ext}}(p_0, \delta)$ is a vertex of $\mathcal{Q}(p_0, \delta)$ and hence, $\mathcal{Q}(p_0, \delta)$ is generated by the convex hull of $\mathcal{Q}_{\text{ext}}(p_0, \delta)$. Suppose for contradiction that this is the case. Then there exists $q, q' \in \mathcal{Q}(p_0, \delta)$ and $\lambda \in (0, 1)$ such that

$$p_0 + (\mathbf{e}_i - \mathbf{e}_j)\frac{\delta}{2} = \lambda q + (1 - \lambda)q'.$$

Let $q = p_0 + \eta \frac{\delta}{2}$ and $q' = p_0 + \eta' \frac{\delta}{2}$ where $\eta \neq \eta'$ then we have that

$$\mathbf{e}_i - \mathbf{e}_j = \lambda\eta + (1 - \lambda)\eta'.$$

Then we must have that

$$\lambda\eta_i + (1 - \lambda)\eta'_i = 1, \quad \lambda\eta_j + (1 - \lambda)\eta'_j = -1,$$

and

$$\lambda\eta_k + (1 - \lambda)\eta'_k = 0.$$

But for either of the first two conditions to be true, we must have that $\eta_i = \eta'_i = 1$ and $\eta_j = \eta'_j = -1$ Since all elements η_k and η'_k are bounded in magnitude by 1. Furthermore, because $|\eta_i| + |\eta_j| = 2$, all other $\eta_k = 0$ (same for η'). Thus, $\eta = \eta'$ which is the desired contradiction. □

D Additional Experiments

D.1 Full results of Section 5

This section summarizes the complete experiment results on the MARKMYWORDS benchmark [35] under three temperature settings $(0.3, 0.7, 1.0)$. For the generation scheme, we follow the hyper-parameter configurations in Huang et al. [23] with $K = 20, |\Omega_h| = 2, \delta = 0.3$. Table 4 reports generation quality across watermarking schemes. Overall, quality is relatively stable across schemes, and our method achieves quality that is comparable to (and in some cases close to the best among) the baselines at each temperature. Table 5 reports detection size, measuring the median number of tokens needed to detect the watermark where lower is better. Here, our scheme consistently attains the smallest size at every temperature, substantially outperforming prior methods. This indicates that our detector can reliably identify watermarked text using fewer tokens. Together, these results confirm that our e-value-based scheme improves detection efficiency without sacrificing generation quality.

Temp.	Exponential	Inverse Transform	Binary	Distribution Shift	SEAL	Theorem 4.1
0.3	0.906	**0.910**	0.905	0.900	0.905	0.909
0.7	0.907	0.917	**0.919**	0.912	0.901	**0.919**
1.0	0.898	**0.917**	0.905	0.907	0.871	0.902

Table 4: Quality (↑) across different temperature settings. Best (per temperature) is highlighted in bold.

Temp.	Exponential	Inverse Transform	Binary	Distribution Shift	SEAL	Theorem 4.1
0.3	∞	∞	∞	112.5	106.0	**97.0**
0.7	∞	734.0	∞	145.0	84.5	**72.0**
1.0	240.5	163.5	∞	317.0	133.0	**97.5**

Table 5: Size (↓) across different temperature settings. Best (per temperature) is highlighted in bold.

D.2 Synthetic experiments

In this section, we verify our theoretical results with synthetic experiments on a family of simple target distributions. Here, we consider the two-token case where $|\mathcal{V}| = 2$ and for simplicity of notation, we let $\mathcal{V} = \{0, 1\}$. Any anchor distribution we consider belongs to a one-parameter Bernoulli family $p_0 = [p \quad 1-p]$ for $p \in (0, 1)$.

Log-growth optimality. We solve the optimization problem in Eq. (2) numerically and compare the objective curve with the theoretical prediction. With $n = 2$, the problem can be simplified to the following maxmin problem:

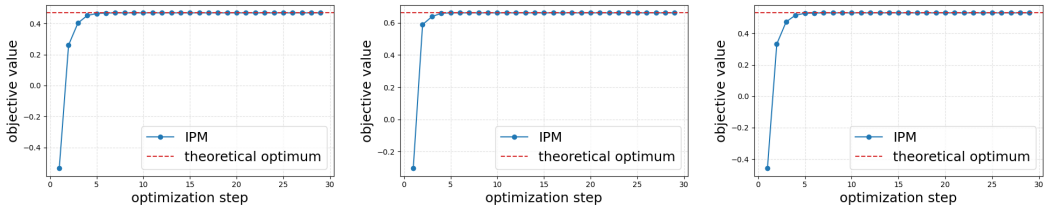
$$\sup_{e} \sum_{v} p_0(v) \log e(v, v) + \frac{\delta}{2} \min \left\{ \log \frac{e(0,1)}{e(1,1)}, \log \frac{e(1,0)}{e(0,0)} \right\}$$

subject to the constraint in Eq. (1) with $\mathcal{Q}(p_0, \delta)$ replaced with the set of vertices $\mathcal{Q}_{\text{ext}}(p_0, \delta) := \{p_0 \pm \frac{\delta}{2}(\mathbf{e}_0 - \mathbf{e}_1)\}$. Using the CLARABEL interior point method solver [17], we obtain the results in Fig. 1. We choose $\delta = 0.01$ and $p = 0.2, 0.5, 0.75$ corresponding to three separate anchor distributions p_0. For each p_0, we cold start the IPM solver and run it for 30 steps with maximum allowed step size 0.99. We observe that our numerical solution to Eq. (2) converges to the proposed theoretical value in Eq. (5), hence verifying our theoretical results in the two-token case.

Stopping time. In this setting, we simulate sequential testing with an optimal e-value e^*, as in Eq. (3). Following Theorem 4.1, we let $\mathcal{A}^*$ be the adversary that chooses $q_t = q^*$ for all t and $\mathcal{G}^*$ be the generator that selects $w_t = w^*$ in response to $\mathcal{A}^*$ for all t. Under this setting, we simulate the stopping time over several α values to obtain the results in Fig. 2. Here, we choose the parameters $\delta = 0.1$ and $p = 0.2, 0.5, 0.75$. For each anchor p_0, we compute the corresponding q^*, w^*, and e^* using the closed-form equations in Theorem 4.1. We select 30 values of α evenly log-spaced from 10^{-2} to 10^{-120}. For each α, we compute 10000 stopping-times by generating sequences $(V^t, S^t)_{t \geq 1} \sim w^*$ and computing the resulting $\log e^*(V^t, S^t)$. We then average the stopping times to form an estimate of $\mathbb{E}_{\mu(\mathcal{A}^*, \mathcal{G}^*)}[\tau_\alpha(e^*)]$. In Fig. 2, we plot our estimates of $\mathbb{E}_{\mu(\mathcal{A}^*, \mathcal{G}^*)[\tau_\alpha(e^*)]}/(\log(1/\alpha))$ which show that as $\alpha \downarrow 0$, the estimates converge to the theoretical rate of $1/J^*$, thereby validating the result in Theorem 4.3.

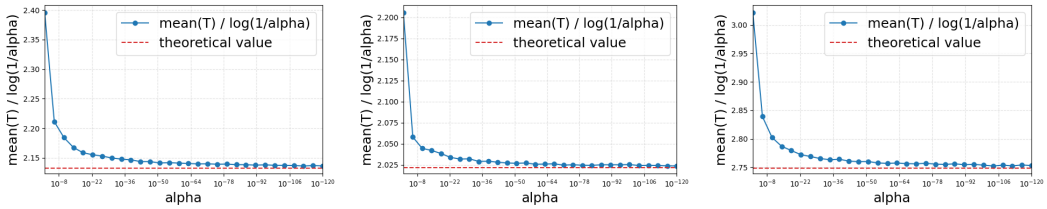
D.3 Empirical Validity of the Anchor Distribution Assumption

The growth guarantees require $\|q - p_0\|_1 \leq \delta$ and $0 < \delta \leq \min_{v \in \mathcal{V}} p_0(v)$. We assess these conditions on the finite alphabet used by the watermark. Following the green–red list construction of Kirchenbauer et al. [28], our implementation uses a partition into fewer than five groups; p_0 and q below denote their partition-level distributions.



(a) $p = 0.2$ (b) $p = 0.5$ (c) $p = 0.75$

Figure 1: Simulation of the two-token case for the log growth problem in Eq. (2). Three separate anchor distributions are used each with parameter $\delta = 0.01$. We solve the simplified maxmin problem using the CLARABEL interior point method solver which is run for 30 steps. The theoretical optimum is computed as in Eq. (5).



(a) $p = 0.2$ (b) $p = 0.5$ (c) $p = 0.75$

Figure 2: Simulation of the two-token case for the stopping time problem $\mathrm{SC}(e^*)$. We simulate for three different anchor distributions $p_0 = [p \quad 1-p]$ with $\delta = 0.1$ and estimate average stopping times $\mathbb{E}[\tau_\alpha]$ for α values ranging from 10^{-2} to 10^{-120}. Each $\mathbb{E}[\tau_\alpha]$ is estimated by simulating 10000 stopping times τ_α. The plots above display graphs of $\frac{\mathbb{E}[\tau_\alpha]}{\log(1/\alpha)}$ with red dashed lines equal to $1/J^*$ where J^* is as in Eq. (5). We observe that convergence to the theoretical optimum is obtained for sufficiently small α.

Table 6 reports the median ℓ_1 distance between the partition-level target (Llama2-7B-chat) and anchor (Phi-3-mini-128k-instruct) distributions, together with the median of $\min_v p_0(v)$, evaluated on the MARKMYWORDS prompts.

Temp.	**Median ℓ_1 distance**	**Median $\min_v p_0(v)$**
0.3	0.32	0.33
0.7	0.24	0.30
1.0	0.21	0.33

Table 6: Partition-level statistics of the target and anchor distributions on MARKMYWORDS.

The values are consistent with the intuition that higher-temperature decoding yields more diffuse distributions, simultaneously shrinking the ℓ_1 gap and lifting $\min_v p_0(v)$ away from zero. The radius δ is a robustness/power parameter: increasing it enlarges $\mathcal{Q}(p_0, \delta)$ but decreases the worst-case log-growth rate J^* in Theorem 4.1. When the configured radius underestimates the empirical gap, Ville's inequality still guarantees Type-I error control, but the power and stopping-time guarantees from Theorem 4.1 and Theorem 4.3 may degrade.

D.4 Empirical Type-I Error

Type-I error control is theoretically guaranteed by the validity of the e-value and Ville's inequality, but it is informative to verify the guarantee empirically. We evaluate Theorem 4.1 on MARKMYWORDS at the nominal level $\alpha = 0.02$ used in the main experiments, running the detector on unwatermarked text generated from the same prompts.

Temp.	Type-I error
0.3	0.00
0.7	0.00
1.0	0.00

Table 7: Empirical Type-I error of Theorem 4.1 on MARKMYWORDS at $\alpha = 0.02$. The observed false-positive rate is zero at every temperature.

The observed false-positive rate is strictly below $\alpha = 0.02$ in every setting. This slack does not come from looseness of the one-step null expectation: for an optimal e-value,

$$\sum_{s \in \mathcal{V}} p_0(s) e^*(v, s) = 1, \qquad \forall v \in \mathcal{V},$$

and therefore

$$\mathbb{E}_{v \sim q,\, s \sim p_0}[e^*(v, s)] = 1, \qquad \forall q \in \Delta(\mathcal{V}).$$

The gap between the observed false-positive rate and α instead reflects the conservativeness of Ville's inequality for finite-time threshold crossings and the finite inspection budget used in the experiments.

D.5 Robustness evaluation

We provide the robustness evaluation in each attack setting in Table 8. The attacks are applied independently.

temp	ContractionAttack	SynonymAttack_0.25	MisspellingAttack_(0.25,)	SwapAttack_0.0_0.05_0	SwapAttack_0.0_0.1_0	SynonymAttack_0.5	SwapAttack_0.05_0.05_0	LowercaseAttack	MisspellingAttack_(0.5,)	SynonymAttack_0.75	SynonymAttack_1	TranslationAttack_French	TranslationAttack_Russian	TypoAttack_(0.05,)	TypoAttack_(0.1,)
0.3	0.9167	0.5833	0.6667	0.5000	0.3333	0.1667	0.2500	0.0833	0.1667	0.0833	0.0833	0.0000	0.0000	0.0000	0.0000
0.7	1.0000	1.0000	0.8333	0.6667	0.5000	0.4167	0.2500	0.2500	0.0833	0.1667	0.1667	0.1667	0.0833	0.0000	0.0000
1.0	0.9167	0.7500	0.5000	0.4167	0.3333	0.0000	0.0000	0.1667	0.0000	0.0000	0.0000	0.0833	0.0000	0.0000	0.0000

Table 8: Watermark detection rate at each temperature and attack type.

D.6 Additional Target Model: Qwen3-4B

The main experiments in Section 5 use Llama2-7B-chat as the target model. We additionally instantiate the target with Qwen3-4B [46], while keeping the decoding and detection parameters fixed, to assess whether the sample-efficiency gains transfer across model families. Table 9 reports the size of Theorem 4.1 against the same five baselines as in the main text. As in Table 1, an entry of ∞ indicates that more than half of the watermarked generations fail to be detected within the inspection budget.

Temp.	Exponential	Inverse Transform	Binary	Distribution Shift	SEAL	Theorem 4.1
0.3	∞	∞	∞	**111.0**	128.5	127.5
0.7	400.0	232.0	∞	151.0	110.0	**80.0**
1.0	207.5	169.0	∞	272.5	149.0	**129.5**

Table 9: Size (↓) on MARKMYWORDS with Qwen3-4B as the target model and Qwen3-1.7B as the anchor. Best (per temperature) is highlighted in bold.

The pattern mirrors the Llama2-7B-chat results: Theorem 4.1 attains the smallest size at temperatures 0.7 and 1.0, and remains competitive with the best baseline at temperature 0.3. The efficiency benefits of the optimal anchored e-value therefore do not depend on a specific target model.

D.7 Comparison Against Chunk-Based Sequential Baselines

We compare with another competitive sequential baseline which restricts the detector to a finite, pre-declared set of inspection times and applies a tighter Bonferroni correction over that set. Concretely, fix a chunk size L and a chunk count M, defining an inspection budget of ML tokens. For each baseline, the detector evaluates the cumulative p-value p_k only at the chunk endpoints $k \in \{L, 2L, \ldots, ML\}$ and applies the fixed correction $p_k \mapsto M \cdot p_k$. Detection is declared at the first endpoint $k = mL$ for which $M \cdot p_{mL} < \alpha$, equivalently $p_{mL} < \alpha/M$, and the reported size is therefore an integer multiple of L. By union bound applied to the M inspection times, this procedure controls Type-I error at level α. In our experiments we set $M = 20$ and $L = 20$, giving an inspection budget of 400 tokens.

Temp.	Exponential	Inverse Transform	Binary	Distribution Shift	SEAL	Theorem 4.1
0.3	∞	∞	∞	100.0	100.0	**97.0**
0.7	260.0	380.0	∞	100.0	80.0	**72.0**
1.0	130.0	200.0	∞	160.0	100.0	**97.5**

Table 10: Size (↓) under the chunk-based sequential protocol with $M = 20$ chunks of size $L = 20$ tokens (correction $p \mapsto M \cdot p$) on MARKMYWORDS. Best (per temperature) is highlighted in bold.

The chunk-based correction tightens several baselines relative to the per-token Bonferroni results of Table 5. Nevertheless, Theorem 4.1 remains the most efficient method at every temperature.

D.8 Hyperparameter Setup and Evaluation Details

For full reproducibility we report the hyperparameters and evaluation details in the MARKMYWORDS experiments of Section 5 and Appendix D.3–D.7.

The anchor distribution p_0 in Section 5 is Phi-3-mini-128k-instruct [33]. For Qwen3-4B experiments in Appendix D.6, we use Qwen3-1.7B as the anchor distribution. Both draft models are restricted to attend only to the previous $K = 20$ tokens to ensure prompt agnosticity. This choice is made to align the anchor with the target model's token space: Qwen3 models use a substantially larger ($4\times$) tokenizer vocabulary than the Llama-2/Phi-3 models, and statistical watermarking is defined over the target LLM's token alphabet. Since watermark generation and detection already require knowing the target token space, this alignment does not impose an additional assumption. Limiting the anchor context window makes p_0 at detection time depend only on a short suffix of the observed text, so the detector remains prompt-agnostic.

Following the configuration of Huang et al. [23], we use:

- `bias` (watermark strength applied to the partition-level logits, inherited from the green–red list construction of Kirchenbauer et al. [28]): 4, decreased to 2.5 at temperature 1.0;
- `l1_dist` (the theoretical robustness radius δ defining $\mathcal{Q}(p_0, \delta)$ in Theorem 4.1): 0.3, decreased to 0.25 at temperature 0.3;
- `gamma` (partition-level mass parameter inherited from the green–red list construction of Kirchenbauer et al. [28]): 0.5;
- Anchor context window: $K = 20$ tokens;
- `proposal_temp` (sampling temperature used by the draft model when generating the anchor): 1.0 at target temperature 0.3, 1.2 at 0.7, and 1.25 at 1.0.

The same hyperparameters are used at both generation and detection. All experiments run on $8\times$A100 80GB GPUs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope on anytime-valid statistical watermarking.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitations in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides the full set of assumptions and complete proofs in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper disclose all the information needed to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the data and codes.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all test details, and does not train models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports other appropriate information about the statistical significance of the experiments, including Type I error and aggregations over multiple task settings and target models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the computer resources used to run the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive and negative societal impacts in Section 6.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the original assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the usage of LLMs in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.