Minimum-Entropy Coupling Approximation Guarantees Beyond the Majorization Barrier

Spencer Compton
Stanford University
MIT-IBM Watson AI Lab

Dmitriy Katz
MIT-IBM Watson AI Lab
IBM Research

Benjamin Qi
MIT

Kristjan Greenewald
MIT-IBM Watson AI Lab
IBM Research

Murat Kocaoglu
Purdue University

Abstract

Given a set of discrete probability distributions, the minimum entropy coupling is the minimum entropy joint distribution that has the input distributions as its marginals. This has immediate relevance to tasks such as entropic causal inference for causal graph discovery and bounding mutual information between variables that we observe separately. Since finding the minimum entropy coupling is NP-Hard, various works have studied approximation algorithms. The work of [Compton, 2022] shows that the greedy coupling algorithm of [Kocaoglu et al., 2017a] is always within $\log_2(e) \approx 1.44$ bits of the optimal coupling. Moreover, they show that it is impossible to obtain a better approximation guarantee using the majorization lower bound that all prior works have used: thus establishing a *majorization barrier*. In this work, we break the majorization barrier by designing a stronger lower bound that we call the *profile method*. Using this profile method, we are able to show that the greedy algorithm is always within $\log_2(e)/e \approx 0.53$ bits of optimal for coupling two distributions (previous best-known bound is within 1 bit), and within $\frac{1+\log_2(e)}{2} \approx 1.22$ bits for coupling any number of distributions (previous best-known bound is within 1.44 bits). We also examine a generalization of the minimum entropy coupling problem: *Concave Minimum-Cost Couplings*. We are able to obtain similar guarantees for this generalization in terms of the concave cost function. Additionally, we make progress on the open problem of [Kovačević et al., 2015] regarding NP membership of the minimum entropy coupling problem by showing that any hardness of minimum entropy coupling beyond NP comes from the difficulty of computing arithmetic in the complexity class NP. Finally, we present exponential-time algorithms for computing the *exactly optimal* solution. We experimentally observe that our new profile method lower bound is not only helpful for analyzing the greedy approximation algorithm, but also for improving the speed of our new backtracking-based exact algorithm.

1 INTRODUCTION

Entropy, particularly in the context of its maximization, has found wide application in statistics and machine learning, e.g., in the maximum entropy principle for estimation [Levine and Tribus, 1979, Nigam et al., 1999], entropic optimal transport [Cuturi, 2013], and, more generally, entropic regularization [Grandvalet and Bengio, 2006, Niu et al., 2014, Haarnoja et al., 2018].

While much more challenging to optimize due to the concavity of entropy, there recently has been increasing interest in the *minimization* of entropy, particularly in the context of the minimum entropy coupling (MEC) problem. Recent applications of the MEC include the entropic causal inference framework [Kocaoglu et al., 2017a, Javidian et al., 2021, Compton et al., 2022]; communications [Sokota et al., 2022, de Witt et al., 2022]; random number generation (discussed in [Li, 2021]); functional representation (discussed in [Cicalese et al., 2019]); and dimensionality reduction [Vidyasagar, 2012, Cicalese et al., 2016].

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

The minimum entropy coupling problem seeks to answer the question: Given a set S of m marginal (discrete-valued) distributions, each with at most n states, what is the minimum-entropy joint distribution (i.e. coupling) that explains the marginals? While simple to state and appealing from an information-theoretic viewpoint, algorithmic optimization over a concave cost (entropy) is often difficult (e.g. [Cardinal et al., 2008]). In particular, computing the minimum entropy coupling is NP-hard [Kovačević et al., 2015]. Despite this challenge, the minimum entropy coupling problem continues to attract interest and has seen an increasing number of applications.

By way of intuition, note that the entropy of the coupling is related to the mutual information between the marginal variables, through the identity $I(X;Y) = H(X) + H(Y) - H(X,Y)$. In particular, the minimum entropy coupling is the coupling that *maximizes* the mutual information, i.e. it can be viewed as the *maximum mutual information* coupling. The perspective of maximizing mutual information has been very influential in machine learning [Viola and Wells III, 1997, Torkkola, 2003, Tschannen et al., 2019, Sun et al., 2020]. In discrete-valued settings, this coupling-based upper bound on mutual information can be of interest, as it provides a tighter bound than the traditional bound by the entropy of the marginal distributions $\min(H(X), H(Y))$. Accordingly, the minimum entropy coupling enables one to upper bound the mutual information between variables that are observed separately. Additionally, couplings have immediate relevance to optimal transport (OT), where the set of all couplings is called the transport polytope. As remarked by [Cicalese et al., 2019], the minimum entropy coupling is the minimum entropy element in the transport polytope.

The entropic causal inference framework proposed in [Kocaoglu et al., 2017a], and further developed in [Kocaoglu et al., 2017b, Compton et al., 2020, Javidian et al., 2021, Compton et al., 2022] addresses the problem of orienting causal graphs when only observational data is available. In such a regime, the data is not sufficient to orient the graph and additional assumptions are required. Existing assumption frameworks (e.g. additive noise models) were not amenable to settings such as those with categorical variables, hence [Kocaoglu et al., 2017a] introduced an empirically and intuitively motivated (e.g. Occam's razor) hypothesis that the true causal direction between a pair of categorical variables often has the smallest exogenous noise entropy associated with its generative model. Finding this exogenous noise entropy from the observed joint distribution was shown to correspond to finding the minimum entropy coupling of a (possibly large) set of distributions. The later work of [Compton et al., 2020] showed an identifiability result that if the true causal direction has low entropy exogeneous noise, then the reverse (anticausal) direction will have large entropy exogenous noise with high probability. An extension to general multi-variable causal directed acyclic graphs was presented in [Compton et al., 2022]. In all of these works, solving or approximating the minimum entropy coupling problem is key to applying the entropic causal framework, and in practice the greedy algorithm of [Kocaoglu et al., 2017a] is used to approximate the minimum entropy coupling. Improved guarantees on the accuracy of the greedy algorithm for many distributions such as our Theorem 4.2, or improved algorithmic approaches, are therefore useful for improving the applicability and trustworthiness of entropic causality.

The recent work of [Sokota et al., 2022] studies Markov coding games: a generalization of source coding and referential games. At a high level, the crux of their approach is to combine reinforcement learning algorithms with minimum entropy coupling algorithms, in an effort to create an agent that can communicate well through Markov decision process trajectories. As an example, they design an agent that can efficiently communicate images just through its actions in the video game Pong and still play the game well. This agent extensively uses the coupling algorithm of [Cicalese et al., 2019] to couple pairs of distributions. The work of [de Witt et al., 2022] also uses minimum entropy coupling for communication, in the context of steganography: securely encoding secret information concealed in seemingly-regular text. They show that under an information-theoretic model of steganography, a procedure is secure if and only if it corresponds to a coupling. Moreover, the minimum entropy coupling corresponds to the maximally efficient secure procedure. They demonstrate strong performance for the minimum entropy coupling-based approach for perfectly secure steganography in experiments using GPT-2 [Radford et al., 2019] and WaveRNN [Kalchbrenner et al., 2018] as communication channels. In doing so, they extensively utilize the greedy coupling algorithm of [Kocaoglu et al., 2017a] to couple pairs of distributions. Our result of Theorem 4.1 directly improves the approximation guarantees for coupling pairs of distributions: a crucial algorithm for both works.

The work of [Vidyasagar, 2012] directly computes the minimum entropy coupling (by a different name) of two distributions in their metric for dimension-reduction of stochastic processes. Accordingly, our result of Theorem 4.1 directly improves the best-known approximation guarantee for efficiently computing this metric. The work of [Cicalese et al., 2016] likewise uses minimum entropy coupling for dimension reduction of probability distributions.

A variety of additional applications of minimum entropy couplings are discussed in [Cicalese et al., 2019, Li, 2021].

Our contributions. Many recent works have focused on proving approximation guarantees for the minimum entropy coupling problem. A unifying prop-

erty of [Cicalese et al., 2019, Rossi, 2019, Li, 2021, Compton, 2022] is that they all show their approximation guarantee in relation to the same lower bound: *the majorization lower bound*. Moreover, [Compton, 2022] shows that better guarantees for general m in relation to the majorization lower bound are impossible: thus establishing the majorization barrier.

In this work, we move away from the majorization strategy and introduce a new method for lower bounding the coupling entropy, which we call the "profile method." This approach allows us to obtain stronger bounds in both the case of coupling two distributions and for coupling an arbitrary number of distributions, as is shown in Table 1.

We note that for all applications, the task of choosing what coupling algorithm to use was previously unclear. Prior to our work, there were three distinct algorithms that all were shown to have a 1 bit additive approximation guarantee for coupling two distributions [Cicalese et al., 2019, Rossi, 2019, Li, 2021]. In this sense, there was no theoretical rationale for using one algorithm over another. For example, the recent work of [Sokota et al., 2022] used the coupling algorithm of [Cicalese et al., 2019], while the work of [de Witt et al., 2022] used the coupling algorithm of [Kocaoglu et al., 2017a]. Our work provides clarity by improving the approximation guarantee for the greedy coupling of [Kocaoglu et al., 2017a] and providing counterexamples that show the algorithms of [Cicalese et al., 2019, Li, 2021] do not match its guarantees (see Appendix H).

We also observe that our profile method is not limited to minimum entropy coupling, as we extend our approach to general concave minimum-cost couplings.

In addition to new (highly efficient to compute) lower bounds on the coupling, we also examine the problem of exact computation of the minimum entropy coupling problem. First, we make progress on the open problem of [Kovačević et al., 2015] regarding NP membership of the minimum entropy coupling problem by providing an NP-reduction from minimum entropy coupling to the problem of simply calculating the Shannon entropy of a distribution. In essence, this shows that any hardness of minimum entropy coupling beyond NP comes from the difficulty of computing arithmetic in the complexity class NP.

We then prove a structural result that enables faster (but still exponential-time) algorithms for exactly coupling two distributions. Namely, we propose a dynamic programming and backtracking algorithm that are significantly faster than the naive baseline. We observe that our profile method lower bound mentioned above is not only helpful for analysis, but also for speeding up the backtracking algorithm.

We summarize the contributions of our paper as follows:

1. A new profile method for lower bounding the entropy of the optimal coupling.
2. Approximation guarantee for coupling two distributions: $H(\mathcal{G}_S) \leq H(\text{OPT}_S) + \frac{\log_2(e)}{e} \approx H(\text{OPT}_S) + 0.53$, where $\mathcal{G}_S$ is the greedy coupling.
3. Approximation guarantee for coupling an arbitrary number of distributions: $H(\mathcal{G}_S) \leq H(\text{OPT}_S) + \frac{1+\log_2(e)}{2} \approx H(\text{OPT}_S) + 1.22$.
4. An NP-reduction from minimum entropy coupling to the problem of simply calculating the Shannon entropy of a distribution.
5. Experimentally faster (still exponential-time) algorithms for exactly coupling two distributions.
6. Experimental results showing that the profile method lower bound is not only helpful for analysis, but also for speeding up the backtracking algorithm.

Table 1: Best-Known Additive Approximation Guarantees

	Prior Work	**This Work**
$m = 2$	1 [Cicalese et al., 2019, Rossi, 2019, Li, 2021]	$\frac{1}{e}\log_2(e) \approx 0.53$
General m	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\frac{1}{2}(\log_2(e)+1) \approx 1.22$

Appendix D contains a more complete table with additional results for small m.

2 RELATED WORK AND BACKGROUND

Notation: $\ln$ and $\log$ denote $\log_e$ and $\log_2$, respectively. H denotes Shannon entropy. Throughout the paper, assume the states of any probability distribution p are sorted in non-increasing order of size; that is, $p(1) \geq p(2) \geq \cdots \geq p(|p|)$. We say that a probability distribution is *possibly partial* if the sum of its sizes may be less than one. S denotes a set of (possibly partial) probability distributions, all with the same total mass $\text{MASS}(S)$. OPT_S denotes an optimal coupling of S.

Couplings: A coupling $\mathcal{C}$ is a joint distribution over m input distributions $p_1, \ldots, p_m$ such that the marginals of $\mathcal{C}$ are the input distributions. In particular, each state of a coupling $\mathcal{C}$ maps to exactly one state for each of the input distributions, where $\mathcal{C}(i_1, i_2, \ldots, i_m)$ maps to $p_1(i_1), p_2(i_2), \ldots, p_m(i_m)$. Thus, the marginal condition means that the total mass of a coupling's states mapping to a particular state i_k of an input distribution p_k must be equal to the mass of $p_k(i_k)$, or equivalently $\mathcal{C}(\cdot, \ldots, i_k, \cdot) = p_k(i_k)$ for all k, i_k. For simplicity of presentation, we will denote every input distribution p_i as having n states $p_i(1), \ldots, p_i(n)$. However, we emphasize that all our results hold for coupling distributions with different numbers of states; we can simply "pad" the smaller distributions by appending states with probability 0.

Greedy Coupling: Our work will analyze the greedy coupling algorithm of [Kocaoglu et al., 2017a], formally described in Algorithm 1. In words, the algorithm repeatedly makes a greedy choice for the next state of the coupling, where it creates a state that maps to the maximally large state in each distribution, and has size equal to the minimum of these maximums. [Kocaoglu et al., 2017b] showed this algorithm is locally optimal, [Rossi, 2019] showed it finds a coupling within 1 bit of the optimal entropy for coupling two distributions, and [Compton, 2022] showed it is within $\log(e) \approx 1.44$ bits for coupling arbitrary numbers of distributions. We use $\mathcal{G}_S$ to denote the sequence of sizes of states the greedy algorithm produces. It is known that $|\mathcal{G}_S| \leq nm - (m-1)$ and $\mathcal{G}_S$ is non-increasing.

Algorithm 1 Greedy Coupling [Kocaoglu et al., 2017a]

Input: Marginal distributions of m variables each with n states, in matrix form $\mathbf{M} = [p_1^T; p_2^T; ...; p_m^T]$.
$\mathcal{G}_S = [\ ]$
Sort each row of $\mathbf{M}$ in non-increasing order.
Find minimum of maximum of each row: $r \leftarrow \min_i(p_i(1))$
while $r > 0$ **do**
    Append r as the next state of $\mathcal{G}_S$.
    Update maximum of each row: $p_i(1) \leftarrow p_i(1) - r, \forall i$
    Sort each row of $\mathbf{M}$ in non-increasing order.
    $r \leftarrow \min_i(p_i(1))$
end while
return $\mathcal{G}_S$.

Majorization: A distribution p is majorized by a distribution q (denoted by $p \preceq q$) if and only if $\sum_{j=1}^{i} p(j) \leq \sum_{j=1}^{i} q(j)$ for all $i \in \{1, \ldots, |q|\}$ [Marshall et al., 1979]. It is known that any valid coupling must be majorized by all distributions in S, meaning $\text{OPT}_S \preceq p, \forall p \in S$ [Cicalese et al., 2019]. Consider the partial ordering induced by majorization. Let $\bigwedge S$ denote the "maximal" distribution in this partial ordering such that $\bigwedge S \preceq p$, $\forall p \in S$, specifically, $\bigwedge S(i) = \left(\min_{p \in S} \sum_{j=1}^{i} p(j)\right) - \bigwedge S(i-1)$. It can be shown that if $p \preceq q$ then $H(q) \leq H(p)$ (this holds for any concave function), hence by definition $H(\text{OPT}_S) \geq H(\bigwedge S)$. We thus have $H(\bigwedge S)$ as a lower bound to $H(\text{OPT}_S)$, which is itself upper bounded by the greedy algorithm. This fact can be used to provide an approximation guarantee for the greedy algorithm, by bounding the gap between $H(\bigwedge S)$ and the entropy returned by the greedy algorithm. [Compton, 2022] shows that this gap is at most $\log_2(e) \approx 1.44$ bits, and shows that this $\log_2(e)$ gap can be achieved for general m. Hence the bound is tight in the sense that no better approximation guarantee is possible for greedy coupling in terms of this $H(\bigwedge S)$ gap for general m.

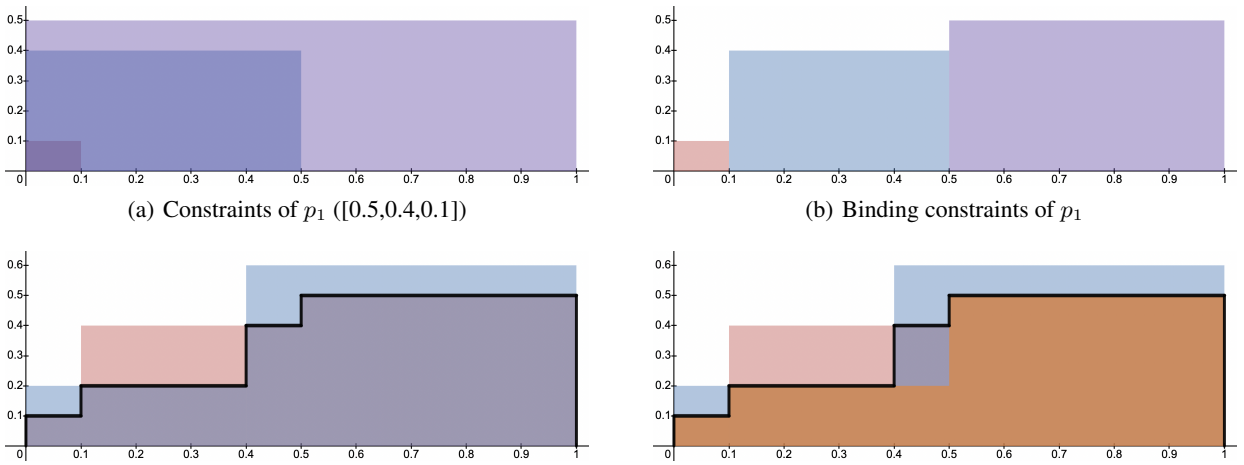
3 NOVEL LOWER BOUND: THE PROFILE METHOD

As mentioned above, prior works show guarantees for minimum entropy coupling in relation to the majorization lower bound $H(\bigwedge S)$. Our work will introduce a stronger lower bound that will enable us to show stronger approximation guarantees that break the majorization barrier.

Motivation. In designing a new lower bound, we begin from first principles and examine what properties a valid coupling must obey. For example, consider a distribution $p_1 = [0.5, 0.4, 0.1]$. We would like to analyze necessary conditions for a valid coupling including p_1. Any valid coupling must have states that map to the smallest state of p_1 of size 0.1. More concretely, these coupling states will have total size 0.1 and each such state must be of size ≤ 0.1. As such, the coupling must have at least 0.1 mass that comes from states of size ≤ 0.1. A similar statement is true from the coupling needing to have states that map to the state of p_1 of size 0.4: at least $0.1 + 0.4 = 0.5$ mass must come from states of size ≤ 0.4. There are accordingly three such necessary constraints for a valid coupling:

1. At least 0.1 mass that comes from states of size ≤ 0.1.
2. At least 0.5 mass that comes from states of size ≤ 0.4.
3. At least 1.0 mass that comes from states of size ≤ 0.5.

In Figure 1(a) we visualize these constraints with the former quantity corresponding to the x dimension, and the latter quantity as the y dimension. It is fitting that these constraints are left-aligned to maximize their overlap, as this corresponds to the constraints being minimally binding. If we clean this visualization by eliminating parts of constraints that are not binding, we obtain Figure 1(b). Notably, this visualization provides a *sketch* of the distribution p_1, where the sketch is a visualization in which the i-th state corresponds to a box of width and height $p_1(i)$. In terms of this visualization, *our necessary conditions are exactly enforcing that the sketch of any valid coupling must not exceed the sketch of* p_1. We generalize this motivation by considering the corresponding constraints and visualization in the presence of another distribution $p_2 = [0.6, 0.2, 0.2]$. Combining the constraints imposed by coupling p_1 and p_2, they enforce that the sketch of any valid coupling must not exceed the lower-envelope of the input distributions' sketches. We will call this lower-envelope the *profile* of the input distributions. In Figure 1(c), these constraints are visualized along with the profile of the sketches for p_1 and p_2. As an example, one valid coupling for p_1 and p_2 has the distribution $[0.5, 0.2, 0.2, 0.1]$. As visualized in Figure 1(d), the sketch of the coupling never exceeds the profile of p_1 and p_2, as is required by any valid coupling.



(a) Constraints of p_1 ([0.5,0.4,0.1])

(b) Binding constraints of p_1

(c) Sketches of p_1 (red), p_2 ([0.6,0.2,0.2], blue), and their profile (black)

(d) Sketches of p_1 (red), p_2 (blue), their profile (black), and a valid coupling ([0.5,0.2,0.2,0.1], orange)

Figure 1: Visualizations of the profile method.

More generally, the profile method will work to obtain a lower bound for the optimal coupling by relating this profile-related constraint on the sketch representation of any valid coupling to the required entropy for any valid coupling. In particular, we will observe how for a point on the profile of height y, this will correspond to some mass of the coupling that must come from a state of size $\leq y$, and thus must be contributing information $\geq \log(\frac{1}{y})$. This will enable us to integrate over the profile after such a logarithmic transformation to obtain a lower bound for the optimal coupling. Note that the existing lower bound of $\bigwedge S$ does not utilize the constraints we outline in this motivation, and it is by using the information provided by these constraints that the profile method is able to obtain a stronger bound.

The profile method. We now more concretely describe the profile method, starting with its sketch visualization. For a (possibly partial) distribution p, consider visualizing it in a way similar to a sketch, where the states are sorted from left to right in non-decreasing order of size, and a state is represented by a square box with width and height $p(i)$. We formally define it as follows:

Definition 3.1. $\textsc{Sketch}_p(x)$ *is defined for* $x \in (0, \textsc{Mass}(p)]$ *such that* $\textsc{Sketch}_p(x) = p(i)$ *for* $x \in (\sum_{j>i} p(j), \sum_{j\geq i} p(j)]$.

For a particular example where S contains two distributions $p_1 = [0.6, 0.2, 0.2]$ and $p_2 = [0.5, 0.4, 0.1]$, we illustrate their sketches in Figure 1(c). As previously explained, the constraints of coupling necessitate that the sketch of any valid coupling does not exceed the sketch of any input distribution. Another way in which sketches are insightful is that we can relate the entropy of a distribution to its sketch. As a point on a distribution's sketch of height y corresponds to mass contributing information $\log(\frac{1}{y})$, the entropy of a distribution is exactly obtained by integrating over its sketch after applying a logarithmic transformation:

Remark 1. $H(p) = \int_0^{\textsc{Mass}(p)} \log\left(\frac{1}{\textsc{Sketch}_p(x)}\right) dx.$

Recall how the sketch of any valid coupling must not exceed the sketch of any input distribution. Equivalently, we can say the sketch of any valid coupling must not exceed the lower-envelope of the input distributions' sketches. We refer to this lower-envelope as *the profile*:

Definition 3.2. $\textsc{Profile}_S(x) \triangleq \min_{p\in S} \textsc{Sketch}_p(x)$.

An example of the profile is illustrated in Figure 1(c). Note how (unlike $\bigwedge S$) the profile does not necessarily correspond to a partial distribution nor does it necessarily consist of squares. Intuitively, it is a much more flexible object that will allow us to obtain a better lower bound for the optimal coupling. Similar to our observation in Remark 1 that the integral of a distribution's sketch after a log-transformation is equal to its entropy, we define an analogous quantity for the profile:

Definition 3.3. *Let the profile "entropy" be* $H(\textsc{Profile}_S) \triangleq \int_0^{\textsc{Mass}(S)} \log\left(\frac{1}{\textsc{Profile}_S(x)}\right) dx.$

As the sketch of any valid coupling must not exceed the profile, every point on the profile of height y can be bijectively mapped to mass in the valid coupling from a state of size $\leq y$, and thus contributing $\geq \log(\frac{1}{y})$ entropy. This finally implies our novel lower bound for the optimal minimum entropy coupling (full proof in Appendix A.1):

Theorem 3.4. $H(\textsc{Opt}_S) \geq H(\textsc{Profile}_S)$.

4 BETTER GREEDY COUPLING APPROXIMATION GUARANTEES

In this section, we will use this new lower bound gained from the profile method to obtain better approximation guarantees for the greedy coupling algorithm.

4.1 Coupling Two Distributions

For coupling just $m = 2$ distributions, the surprising state of affairs is that there are three distinct algorithms that all attain a 1 bit additive approximation guarantee as shown in [Cicalese et al., 2019, Rossi, 2019, Li, 2021]. It is then natural to wonder if this is the best possible approximation guarantee. Further, we can construct instances where $H(\text{OPT}_S) \approx H(\bigwedge S) + 0.66$ (see Appendix H), meaning it is impossible to get a better approximation guarantee than 0.66 bits with respect to the previously used lower bound of $H(\bigwedge S)$. Yet, with a proof relating to the profile method lower bound, we break the majorization barrier and show that the greedy coupling algorithm of [Kocaoglu et al., 2017a] is always additively within ≈ 0.53 bits of the optimum:

Theorem 4.1. *For $m = 2$, $H(\mathcal{G}_S) \leq H(\text{PROFILE}_S) + \frac{\log e}{e} \approx H(\text{PROFILE}_S) + 0.53$.*

Proof intuition for Theorem 4.1. Let S^t denote the distributions of S after the t-th state of the greedy coupling. Similarly, $\mathcal{G}_S^t$ represents the vector $\mathcal{G}_S(1), \ldots, \mathcal{G}_S(t)$. We will consider the evolution of the non-decreasing monovariant $M^t \triangleq H(\text{PROFILE}_{S^t}) + H(\mathcal{G}_S^t)$. Initially, nothing is coupled and we have $M^0 = H(\text{PROFILE}_{S^0}) + H(\mathcal{G}_S^t) = H(\text{PROFILE}_S) + 0$. In the end, everything is coupled and we have $M^{|\mathcal{G}_S|} = H(\text{PROFILE}_{S^{|\mathcal{G}_S|}}) + H(\mathcal{G}_S^{|\mathcal{G}_S|}) = 0 + H(\mathcal{G}_S)$. We will show that the profile method implies

$$M^{t+1} \leq M^t + \frac{\log e}{e}\mathcal{G}_S(t+1). \tag{1}$$

Eq. (1) would suffice to prove the conclusion because then

$$\begin{aligned} H(\mathcal{G}_S) = M^{|\mathcal{G}_S|} &\leq M^0 + \frac{\log e}{e}\sum_{t=1}^{|\mathcal{G}_S|}\mathcal{G}_S(t) \\ &= H(\text{PROFILE}_S) + \frac{\log e}{e}. \end{aligned}$$

We now outline the derivation of Eq. (1). Consider the two distributions in S^t to be denoted by p^t and q^t (sorted to be in non-increasing order). Without loss of generality, suppose $p^t(1) \leq q^t(1)$. This means our greedy algorithm will select $\mathcal{G}_S(t+1) = p^t(1)$, the state $p^t(1)$ will be fully coupled, and $q^t(1)$ will afterwards have size $q^t(1) - p^t(1)$. With some calculation, we can bound the increase in our monovariant $M^{t+1} - M^t$ by $(q^t(1) - p^t(1)) \cdot \max\left(0, \log(\frac{1}{q^t(1)-p^t(1)}) - \log(\frac{1}{p^t(1)})\right)$. We can in turn bound this expression by

$$\begin{aligned} &p^t(1) \cdot \max_{p^t(1)<q^t(1)<2p^t(1)} \left[\frac{q^t(1)-p^t(1)}{p^t(1)} \log\left(\frac{p^t(1)}{q^t(1)-p^t(1)}\right)\right] \\ &= p^t(1) \cdot \max_{0<r<1}\left[r \log \frac{1}{r}\right] = p^t(1) \cdot \frac{\log e}{e}, \end{aligned} \tag{2}$$

where the maximum is attained at $r = e^{-1}$. As $\mathcal{G}_S(t+1) = p^t(1)$, Eq. (2) implies Eq. (1), thus concluding our proof outline. Hence, the profile method enables us to naturally analyze how the monovariant of the profile and greedy evolves, and attain an approximation guarantee breaking the majorization barrier. The full proof is in Appendix B.

Remark. This proof method can also show guarantees for small m that are better than the guarantee for general m in Theorem 4.2. More details are included in Appendix D.

4.2 Coupling an Arbitrary Number of Distributions

For coupling many distributions, the result of [Compton, 2022] showed that the greedy coupling algorithm of [Kocaoglu et al., 2017a] is within ≈ 1.44 bits of the optimal and that it is impossible to get a better guarantee with respect to the $H(\bigwedge S)$ lower bound. Here we use the profile lower bound to break the majorization barrier and show the greedy coupling is always within ≈ 1.22 bits of the optimal:

Theorem 4.2. $H(\mathcal{G}_S) \leq H(\text{PROFILE}_S) + \frac{1+\log e}{2} \approx H(\text{PROFILE}_S) + 1.22$.

Proof intuition for Theorem 4.2. For this proof, we will find it helpful to upper bound the total remaining mass to be coupled as a function of the size of the greedy's next state. More concretely, we will define a function $\text{REM-MASS}_S(y)$ that corresponds to an upper bound on the total remaining mass to be coupled if the greedy coupling is about to create a state of size $\leq y$. For any non-decreasing function REM-MASS_S satisfying this condition, the following quantity upper bounds the cost of the greedy coupling solution:

$$H(\mathcal{G}_S) \leq \int_0^1 \text{REM-MASS}'_S(y) \cdot \log\left(\frac{1}{y}\right) dy. \tag{3}$$

This holds because, for any y, the increase in our remaining mass upper bound denoted by $\text{REM-MASS}'_S(y)$ corresponds to mass that must come from greedy states of size $\geq y$. Adversarially, we say this mass contributes information $\log(1/y)$. It remains to be answered how one can construct a REM-MASS_S that upper bounds the remaining mass given the next greedy state is of size $\leq y$. Our most intuitive bound is:

$$\text{REM-MASS}_S^{\text{SIMPLE}}(y) \triangleq \tag{4}$$

$$\max_{p \in S} \text{REM-MASS}_p^{\text{SIMPLE}}(y) \triangleq \max_{p \in S} \sum_{j=1}^{n} \min(p(j), y).$$

We design the bound $\text{REM-MASS}_{p'}^{\text{SIMPLE}}(y)$ for a particular distribution p' such that if the largest remaining state in p' is $\leq y$, then $\text{REM-MASS}_{p'}^{\text{SIMPLE}}(y)$ will be a valid upper bound on the total remaining mass in p'. Recall that there must be such a p', given how the greedy is defined in Algorithm 1. So, consider a p' where its largest remaining state has size $\leq y$. Then the amount of mass remaining in any particular state j of p' is $\leq \min(p'(j), y)$, hence the bound given in Eq. (4). As an example, if the largest remaining state of p_2 (specified in Figure 1(c)) is at most 0.3, then the remaining mass in p_2 is at most $\text{REM-MASS}_{p_2}^{\text{SIMPLE}}(0.3) = \min(0.5, 0.3) + \min(0.4, 0.3) + \min(0.1, 0.3) = 0.7$. Through further non-trivial analysis, one can relate the bound of Eq. (3) when using $\text{REM-MASS}^{\text{SIMPLE}}$ defined in Eq. (4) to show that $H(\text{OPT}_S) \leq H(\text{PROFILE}_S) + \log(e) \approx H(\text{PROFILE}_S) + 1.44$, matching the guarantee of [Compton, 2022].

However, we can improve upon $\text{REM-MASS}^{\text{SIMPLE}}$. In our example with p_2, we stated that if the largest remaining state of p_2 is at most 0.3, then the remaining mass in its state $p_2(1)$ was $\leq \min(0.5, 0.3) = 0.3$. While this is true, it is not tight. If the largest remaining state in p_2 is r_{max} where $r_{\text{max}} < 0.5$, then given the monotonicity in the size of greedy states, there must have been a greedy state that removed at least r_{max} mass from 0.5. More formally, we can bound that the remaining mass in 0.5 is at most $0.5 - r_{\text{max}}$. Combining this with the prior upper bound that its remaining mass is $\leq r_{\text{max}}$, we can show how the remaining mass in $p_2(1)$ is at most $\max_{0 \leq r_{\text{max}} \leq 0.3} \min(r_{\text{max}}, 0.5 - r_{\text{max}}) = 0.25$. We can generalize this reasoning as follows:

$$\begin{aligned}
&\text{REM-MASS}_S^{\text{ADVANCED}}(y) \\
&\triangleq \max_{p \in S} \text{REM-MASS}_p^{\text{ADVANCED}}(y) \\
&\triangleq \max_{p \in S} \sum_{j=1}^{n} \begin{cases} y & y \leq \frac{p(j)}{2} \\ p(j)/2 & \frac{p(j)}{2} < y < p(j) \\ p(j) & p(j) \leq y \end{cases} \qquad (5)
\end{aligned}$$

In our example, this attains $\text{REM-MASS}_{p_2}^{\text{ADVANCED}}(0.3) = 0.25 + 0.2 + 0.1 = 0.55$, improving upon $\text{REM-MASS}_{p_2}^{\text{SIMPLE}}(0.3) = 0.7$. Through further non-trivial analysis, one can relate the bound of Eq. (3) when using $\text{REM-MASS}^{\text{ADVANCED}}$ defined in Eq. (5) to show that $H(\text{OPT}_S) \leq H(\text{PROFILE}_S) + \frac{1+\log(e)}{2} \approx H(\text{PROFILE}_S) + 1.22$, attaining a new best guarantee for coupling many distributions and breaking the majorization barrier. We defer the full proof to Appendix C.

5 COMPUTING EXACTLY OPTIMAL SOLUTIONS

While earlier sections discuss approximating the minimum entropy coupling, this section focuses on approaches to compute an *exact* solution. This is known to be generally intractable as it is NP-Hard. It was posed as an open problem in [Kovačević et al., 2015] whether the problem is in NP. In viewing the minimum entropy coupling as an optimization problem with linear constraints, there are exponentially many (n^m) variables. Perhaps surprisingly, we show there is always an optimal coupling that uses relatively few states:

Lemma 5.1. *There always exists an optimal solution to* MEC *with support size of at most* $nm - (m-1)$.

Proof. We can formulate minimum entropy coupling as an optimization problem over the $d \triangleq n^m$-dimensional polytope $\mathbf{P}(i_1, \ldots, i_m)$:

$$\begin{aligned}
\underset{\mathbf{P}}{\text{minimize}} \quad & \sum_{i=(i_1,\ldots,i_m)} \mathbf{P}(i) \log(1/\mathbf{P}(i)) \\
\text{subject to} \quad & \mathbf{P}(i) \geq 0, \forall i; \sum_i \mathbf{P}(i) = 1 \\
& \sum_{i:i_j=k} \mathbf{P}(i) = \mathbf{p}_j(k), \forall 1 \leq j \leq m, 1 \leq k < n.
\end{aligned}$$

Observe that there are a total of $q \triangleq m(n-1)+1 = nm-(m-1)$ equality constraints. We have omitted the constraints for $\mathbf{p}_j(k=n)$ because they are linear combinations of the constraint $\sum_i \mathbf{P}(i) = 1$ and the constraints for $\mathbf{p}_j(k<n)$.

Since the objective is concave, there exists a vertex of $\mathbf{P}$ at which the objective is minimized. To finish, it suffices to show every vertex of this polytope has at most q nonzero entries. This is true because every vertex is the unique point in the intersection of some d constraint hyperplanes. At least $d-q$ of these hyperplanes must be of the form $\mathbf{P}(i) = 0$, so every vertex must have at least this number of zeros. □

This bound of $nm-(m-1)$ is particularly significant, as we know the number of states the greedy coupling algorithm uses is upper bounded by the same value (shown in [Kocaoglu et al., 2017a]). More generally, it is notable that there is always an optimal solution with support size of at most $nm-(m-1)$, because for almost all instances (in the measure 1 sense) *there is no possible coupling that uses fewer than* $nm-(m-1)$ *states*. Furthermore, this enables an exponential-time algorithm (simply enumerating over all polytope vertices). This also makes progress on the open problem of [Kovačević et al., 2015] regarding the NP membership of minimum entropy coupling (MEC), as one could use the constraints identifying the polytope vertex as

a certificate. Given the polytope vertex, all that remains is the seemingly trivial task of computing the Shannon entropy of a discrete distribution:

Corollary 5.2. MEC $\leq_{NP}$ COMPUTE-ENTROPY.

For reasons involving complexity of finite precision arithmetic (e.g. [Kayal and Saha, 2012]), it is unknown whether COMPUTE-ENTROPY is in NP, despite the fact in practice this appears to be a simple task. We note that Corollary 5.2 assumes all inputs are represented as rational numbers.

5.1 Speeding Up Backtracking

Recent works such as [Sokota et al., 2022, de Witt et al., 2022] have applied minimum entropy couplings for $m = 2$ distributions in settings where finding better couplings has concrete benefits. It is thus natural to wonder how quickly we can exactly couple two distributions. As enumerating over the exponentially many polytope vertices is computationally expensive, we propose other algorithms that can more efficiently couple two distributions (although they are still exponential-time algorithms). In Appendix F.1, we introduce a new dynamic programming algorithm that runs in $O(9^n \cdot \text{poly}(n))$ time. This approach utilizes a new characterization of couplings as spanning trees over the distributions, similar to the perspective of Lemma 5.3. In this section, we present a backtracking algorithm that recursively chooses the next state of the coupling and eliminates the smallest size marginal the state maps to. This is a generalization of how the greedy coupling algorithm always eliminates the state corresponding to the minimum over all remaining distributions' maximums. Dynamic programming and backtracking can be seen as two ways of leveraging our new spanning tree perspective. While the dynamic programming approach is relatively final, the backtracking approach is crucially improved by new lower bounds for coupling as they help prune its search space, meaning *better lower bounds directly improve its speed*.

Our motivation is to enable the computation of higher-quality couplings than the polynomial-time greedy coupling in settings where enumerating over all polytope vertices is computationally infeasible. Backtracking is exactly optimal for $m = 2$. Further details are provided in Appendix F.2.

We use the following lemma to prove the validity of our backtracking algorithm for $m = 2$.

Lemma 5.3. *Consider the weighted undirected bipartite graph G with vertex set $\{\ell_{1\ldots n}, r_{1\ldots n}\}$ and edge set $\{(\ell_i, r_j, \mathbf{p}(i,j))) \mid \mathbf{p}(i,j) > 0\}$ induced by some minimum-entropy coupling $\mathbf{p}$. This graph is a forest, and any edge with the maximum weight must have a leaf of the forest as one of its endpoints.*

Our backtracking algorithm is guaranteed to find any coupling satisfying the lemma above, because it can repeatedly peel away the maximum weight edge of the forest. The proof of Lemma 5.3 is deferred to Appendix G.

Combining lower bounds. We now discuss a better lower bound we use to improve the speed of backtracking. Recall that the intuitions behind the lower bounds $H(\bigwedge S)$ and $H(\text{PROFILE}_S)$ are somewhat different. Interestingly, we are able to combine them to get an even better lower bound than simply taking $\max(H(\bigwedge S), H(\text{PROFILE}_S))$:

Definition 5.4. $\text{MAJOR-PROFILE}_S \triangleq \arg\min_d H(d)$ $s.t.\ \text{SKETCH}_d(x) \leq \text{PROFILE}_S(x)\ \forall x \in (0, \text{MASS}(S)]$

Theorem 5.5. $H(\text{OPT}_S) \geq H(\text{MAJOR-PROFILE}_S) \geq H(\text{PROFILE}_S), H(\bigwedge S)$

In words, MAJOR-PROFILE_S is the minimum entropy distribution whose sketch never exceeds the profile (as is shown to be a necessary condition for a valid coupling in the proof of Theorem 3.4). On the surface, this seems only related to the $H(\text{PROFILE}_S)$ lower bound and not $H(\bigwedge S)$. Yet, one can show that this constraint actually enforces $\text{MAJOR-PROFILE}_S \preceq \bigwedge S$. This means we have a lower bound that is always at least as good as both $H(\bigwedge S)$ and $H(\text{PROFILE}_S)$. The full proof of Theorem 5.5 is deferred to Appendix G.3.

Note that $H(\text{PROFILE}_S)$ and $H(\text{MAJOR-PROFILE}_S)$ can both be efficiently computed in near-linear time, as will be leveraged by our backtracking algorithm. Details about their computation are deferred to Appendix F.3.

Experimental results. In Table 2 we show experiments comparing the average running time of coupling two distributions with different (provably optimal) algorithms. All algorithms we present have stronger performance than the naive baseline of enumerating over all polytope vertices. We also observe clear improvements in the speed of backtracking with better lower bounds than the previously best-known lower bound of $H(\bigwedge S)$. It is particularly noteworthy that MAJOR-PROFILE_S provides better performance than PROFILE_S, as our theoretical approximation guarantees in Sections 4.1 and 4.2 do not immediately yield any benefit for using MAJOR-PROFILE_S in place of PROFILE_S. This seems to indicate that MAJOR-PROFILE_S is a meaningfully stronger lower bound for many instances, and may be a helpful reference for future work on minimum entropy coupling approximation guarantees. We observe the best performance from our dynamic programming algorithm and backtracking algorithm with MAJOR-PROFILE_S. Due to the nature of its pruning, the backtracking has higher variance. Finally, we emphasize that any future work providing new lower bounds for minimum entropy coupling could immediately improve our backtracking method. Additional experiment details are provided in Appendix E.

Table 2: Average Runtime (in Seconds) for Exactly Coupling Two Distributions

Algorithm	n							
	4	5	6	7	8	9	10	11
Naive Polytope Vertex Enumeration	0.002	0.189	33.79	> 120	> 120	> 120	> 120	> 120
Backtracking [0]	0.0003	0.004	0.106	3.576	> 120	> 120	> 120	> 120
Backtracking $[H(\bigwedge S)]$	0.0001	0.001	0.013	0.256	4.907	> 120	> 120	> 120
Backtracking $[H(\textsc{Profile}_S)]$	0.0001	0.0009	0.009	0.153	2.206	> 120	> 120	> 120
Backtracking $[H(\textsc{Major-Profile}_S)]$	0.00006	0.0004	0.003	0.035	0.344	5.398	> 120	> 120
Dynamic Programming	0.00007	0.0004	0.002	0.012	0.093	0.802	9.769	> 120

6 CONCLUSION

In this work, we showed advances in algorithms, lower bounds, and approximation guarantees for the minimum entropy coupling problem. In Appendix I, we discuss how our methods generalize to any concave minimum-cost coupling. We do not foresee any adverse societal impacts of our results. There are many avenues for future work, such as examining how previously discussed applications of minimum entropy coupling may concretely benefit from our results. For example, the algorithms for entropic causal inference used in [Kocaoglu et al., 2017a, Compton et al., 2020, Compton et al., 2022] do not use the construction of the coupling, just the value of its entropy. As our new lower bounds on the value are more efficient to compute than the greedy coupling and have the same theoretical guarantees, using these instead would directly improve the speed of these causal discovery algorithms. There are many remaining theoretical questions. Hardness of approximation remains an open problem. It is still unknown the best approximation guarantee that greedy coupling achieves, or whether there exists a polynomial-time algorithm with strictly better worst-case approximation guarantees than greedy coupling. Deriving an analysis that more smoothly interpolates best-known guarantees from $m = 2$ to $m = \infty$ is also an open question. Another open question is whether one can obtain a better approximation guarantee with respect to $\textsc{Major-Profile}_S$ than with $\textsc{Profile}_S$. Finally, in Appendix H we provide some discussion on gaps between various relevant quantities, with counter-examples that were computationally discovered by local search. For example, we provide instances where the algorithms of [Cicalese et al., 2019] and [Li, 2021] have optimality gaps greater than $\log(e)/e \approx 0.53$ for coupling two distributions, whereas Theorem 4.1 shows this is impossible for greedy coupling. It is our hope that this information will help guide algorithm selection and be insightful for future conjectures.

Acknowledgements

This work was supported in part by the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program, NSF Grant CAREER 2239375, and the MIT-IBM Watson AI Lab.

References

[Cardinal et al., 2008] Cardinal, J., Fiorini, S., and Joret, G. (2008). Tight results on minimum entropy set cover. *Algorithmica*, 51(1):49–60.

[Cicalese et al., 2016] Cicalese, F., Gargano, L., and Vaccaro, U. (2016). Approximating probability distributions with short vectors, via information theoretic distance measures. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1138–1142. IEEE.

[Cicalese et al., 2019] Cicalese, F., Gargano, L., and Vaccaro, U. (2019). Minimum-entropy couplings and their applications. *IEEE Transactions on Information Theory*, 65(6):3436–3451.

[Compton, 2022] Compton, S. (2022). A tighter approximation guarantee for greedy minimum entropy coupling. In *IEEE International Symposium on Information Theory, ISIT 2022, Espoo, Finland, June 26 - July 1, 2022*, pages 168–173. IEEE.

[Compton et al., 2022] Compton, S., Greenewald, K., Katz, D. A., and Kocaoglu, M. (2022). Entropic causal inference: Graph identifiability. In *International Conference on Machine Learning*, pages 4311–4343. PMLR.

[Compton et al., 2020] Compton, S., Kocaoglu, M., Greenewald, K., and Katz, D. (2020). Entropic causal inference: Identifiability and finite sample results. *Advances in Neural Information Processing Systems*, 33:14772–14782.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

[de Witt et al., 2022] de Witt, C. S., Sokota, S., Kolter, J. Z., Foerster, J., and Strohmeier, M. (2022). Perfectly secure steganography using minimum entropy coupling. *arXiv preprint arXiv:2210.14889*.

[Grandvalet and Bengio, 2006] Grandvalet, Y. and Bengio, Y. (2006). Entropy regularization.

[Haarnoja et al., 2018] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

[Javidian et al., 2021] Javidian, M. A., Aggarwal, V., Bao, F., and Jacob, Z. (2021). Quantum entropic causal inference. In *Quantum Information and Measurement*, pages F2C–3. Optical Society of America.

[Kalchbrenner et al., 2018] Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A., Dieleman, S., and Kavukcuoglu, K. (2018). Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.

[Kayal and Saha, 2012] Kayal, N. and Saha, C. (2012). On the sum of square roots of polynomials and related problems. *ACM Transactions on Computation Theory (TOCT)*, 4(4):1–15.

[Kocaoglu et al., 2017a] Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. (2017a). Entropic causal inference. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[Kocaoglu et al., 2017b] Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. (2017b). Entropic causality and greedy minimum entropy coupling. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1465–1469. IEEE.

[Kovačević et al., 2015] Kovačević, M., Stanojević, I., and Šenk, V. (2015). On the entropy of couplings. *Information and Computation*, 242:369–382.

[Levine and Tribus, 1979] Levine, R. D. and Tribus, M. (1979). Maximum entropy formalism. In *Maximum Entropy Formalism Conference (1978: Massachusetts Institute of Technology)*. Mit Press.

[Li, 2021] Li, C. T. (2021). Efficient approximate minimum entropy coupling of multiple probability distributions. *IEEE Transactions on Information Theory*, 67(8):5259–5268.

[Marshall et al., 1979] Marshall, A. W., Olkin, I., and Arnold, B. C. (1979). *Inequalities: theory of majorization and its applications*, volume 143. Springer.

[Nigam et al., 1999] Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Stockholom, Sweden.

[Niu et al., 2014] Niu, G., Dai, B., Yamada, M., and Sugiyama, M. (2014). Information-theoretic semi-supervised metric learning via entropy regularization. *Neural computation*, 26(8):1717–1762.

[Radford et al., 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Rossi, 2019] Rossi, M. (2019). Greedy additive approximation algorithms for minimum-entropy coupling problem. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1127–1131. IEEE.

[Sokota et al., 2022] Sokota, S., De Witt, C. A. S., Igl, M., Zintgraf, L. M., Torr, P., Strohmeier, M., Kolter, Z., Whiteson, S., and Foerster, J. (2022). Communicating via markov decision processes. In *International Conference on Machine Learning*, pages 20314–20328. PMLR.

[Sun et al., 2020] Sun, F.-Y., Hoffman, J., Verma, V., and Tang, J. (2020). Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *International Conference on Learning Representations*.

[Torkkola, 2003] Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *Journal of machine learning research*, 3(Mar):1415–1438.

[Tschannen et al., 2019] Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. (2019). On mutual information maximization for representation learning. In *International Conference on Learning Representations*.

[Vidyasagar, 2012] Vidyasagar, M. (2012). A metric between probability distributions on finite sets of different cardinalities and applications to order reduction. *IEEE Transactions on Automatic Control*, 57(10):2464–2477.

[Viola and Wells III, 1997] Viola, P. and Wells III, W. M. (1997). Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154.

Supplementary Materials for: Minimum-Entropy Coupling Approximation Guarantees Beyond the Majorization Barrier

A PROOFS FOR SECTION 3: THE PROFILE METHOD LOWER BOUND

A.1 Proof of Theorem 3.4

Theorem 3.4. $H(\text{OPT}_S) \geq H(\textsc{Profile}_S)$.

While in our proof outline in the main text we discussed **sketches** and **profiles**, our formal proof in this section is simplified by analyzing their "inverses," which we define next.

Definition A.1. *For a (possibly partial) distribution p_i, define* $\textsc{Inv-Sketch}_{p_i}(y)$ *to be the non-decreasing function* $\textsc{Inv-Sketch}_{p_i}\colon [0,1] \to [0,1]$ *such that* $\textsc{Inv-Sketch}_{p_i}(y)$ *equals the sum of all probability masses of p_i less than or equal to y. That is,*

$$\textsc{Inv-Sketch}_{p_i}(y) \triangleq \sum_{j=1}^{|p|} p_i(j) \cdot [p_i(j) \leq y].$$

For a set of probability distributions S, we define $\textsc{Inv-Prof}_S(y)$ *as*

$$\textsc{Inv-Prof}_S(y) \triangleq \max_{p_i \in S} \textsc{Inv-Sketch}_{p_i}(x).$$

We refer to $\textsc{Inv-Sketch}_p$ and $\textsc{Inv-Prof}_S$ as "inverse sketches" and "inverse profiles," respectively. Note that for any S, the graph of $\textsc{Inv-Prof}_S$ may be obtained by transposing the graph of $\textsc{Profile}_S$, though they are not strictly inverses because neither function is strictly increasing.

We also extend the definition of entropy to inverse profiles.

Definition A.2. *For an inverse profile (or inverse sketch) h, we define the entropy of h to be:*

$$H(h) \triangleq \int_0^1 \frac{h(y)}{y \ln 2}\, dy. \tag{6}$$

Lemma A.3. *The definition of entropy is consistent for profiles and inverse profiles (as well as sketches and inverse sketches); that is, $H(\textsc{Profile}_S) = H(\textsc{Inv-Prof}_S)$.*

Proof. Start with Definition 3.3. First, we substitute x in place of y:

$$\begin{aligned} \textsc{Profile}_S(x) &= y, \\ x &= \textsc{Inv-Prof}_S(y), \\ dx &= \textsc{Inv-Prof}'_S(y)\, dy, \end{aligned}$$ [1]

which implies that

$$\begin{aligned} \int_0^{\textsc{Mass}(S)} \log\left(\frac{1}{\textsc{Profile}_S(x)}\right) dx &= \frac{1}{\ln 2} \int_0^{\textsc{Mass}(S)} \ln\left(\frac{1}{\textsc{Profile}_S(x)}\right) dx, \\ &= \frac{1}{\ln 2} \int_0^1 \textsc{Inv-Prof}'_S(y) \ln\left(\frac{1}{y}\right) dy. \end{aligned}$$

Next we integrate by parts:

$$= \frac{1}{\ln 2}\left(\text{INV-PROF}_S(y)\ln\left(\frac{1}{y}\right)\Big|_0^1 + \int_0^1 \frac{\text{INV-PROF}_S(y)}{y}\right),$$
$$= \frac{1}{\ln 2}\int_0^1 \frac{\text{INV-PROF}_S(y)}{y},$$

as desired. □

Lemma A.4. $\text{INV-SKETCH}_{OPT_S}(y) \geq \text{INV-PROF}_S(y)$ *for all* y.

Proof. By definition of INV-PROF_S it suffices to show that $\text{INV-SKETCH}_{\text{OPT}_S}(y) \geq \text{INV-SKETCH}_p(y)$ for all $p \in S$. Note that OPT_S may be derived from p by repeatedly performing the following operation: take some state in p and split it into two smaller states, transforming p into p'.

Suppose that an operation splits a state of mass $m_1 + m_2$ into two states of masses $m_1 \leq m_2$. Then it is easy to see that

- $\text{INV-SKETCH}_{p'}(y) = \text{INV-SKETCH}_p(y)$ for $y < m_1$
- $\text{INV-SKETCH}_{p'}(y) > \text{INV-SKETCH}_p(y)$ for $m_1 \leq y < m_1 + m_2$
- $\text{INV-SKETCH}_{p'}(y) = \text{INV-SKETCH}_p(y)$ for $m_1 + m_2 \leq y$.

It follows that if p' can be derived from p via some sequence of operations, $\text{INV-SKETCH}_{p'}(y) \geq \text{INV-SKETCH}_p(y)$ for all y. Substituting OPT_S in place of p' yields the desired result.

□

Proof of Theorem 3.4 We can now prove Theorem 3.4. By consistency of our definitions of entropy, $H(\text{OPT}_S) = H(\text{INV-SKETCH}_{\text{OPT}_S})$ and $H(\text{PROFILE}_S) = H(\text{INV-PROF}_S)$. To finish, $H(\text{INV-SKETCH}_{\text{OPT}_S}) \geq H(\text{INV-PROF}_S)$ follows by combining Lemma A.4 with Definition A.2.

B PROOF OF THEOREM 4.1: BETTER GREEDY COUPLING APPROXIMATION GUARANTEE FOR TWO DISTRIBUTIONS

Theorem 4.1. *For* $m = 2$, $H(\mathcal{G}_S) \leq H(\text{PROFILE}_S) + \frac{\log e}{e} \approx H(\text{PROFILE}_S) + 0.53$.

It suffices to derive Eq. (1) in the main text. As in Appendix A.1, we find it more convenient to prove our results in terms of $H(\text{INV-PROF}_S)$ instead of $H(\text{PROFILE}_S)$ (which is allowed by Lemma A.3 in Appendix A.1).

After the greedy algorithm selects $\mathcal{G}_S(t+1) = p^t(1)$, we have

- $H(\mathcal{G}_S^{t+1}) = H(\mathcal{G}_S^t) + p^t(1)\log(1/p^t(1))$.
- $\text{INV-SKETCH}_{p^{t+1}}(x) = \begin{cases} \text{INV-SKETCH}_{p^t}(y) & y < p^t(1) \\ \text{INV-SKETCH}_{p^t}(y) - p^t(1) & p^t(1) \leq y \end{cases}$
- $\text{INV-SKETCH}_{q^{t+1}}(y) = \begin{cases} \text{INV-SKETCH}_{q^t}(y) & y < q^t(1) - p^t(1) \\ \text{INV-SKETCH}_{q^t}(y) + q^t(1) - p^t(1) & q^t(1) - p^t(1) \leq y < q^t(1) \\ \text{INV-SKETCH}_{q^t}(y) - p^t(1) & q^t(1) \leq y \end{cases}$

From the above equations, we conclude that:

[1]Technically, $\text{INV-PROF}'_S$ is not defined at the points where INV-PROF_S is discontinuous. We can circumvent this issue by letting $\text{INV-PROF}'_S$ take on the value of the appropriate multiple of the Dirac delta function at such points.

$$\textsc{Inv-Prof}_{S^{t+1}}(y) \begin{cases} = \textsc{Inv-Prof}_{S^t}(y) & y < q^t(1) - p^t(1) \\ \leq \textsc{Inv-Prof}_{S^t}(y) + q^t(1) - p^t(1) & q^t(1) - p^t(1) \leq y < p^t(1) \\ = \textsc{Inv-Prof}_{S^t}(y) - p^t(1) = \textsc{Mass}(S^{t+1}) & p^t(1) \leq y \end{cases}, \tag{7}$$

which in turn implies

$$\begin{aligned} H(\textsc{Inv-Prof}_{S^{t+1}}) &\leq H(\textsc{Inv-Prof}_{S^t}) - p^t(1)\log(1/p^t(1)) + \begin{cases} (q^t(1) - p^t(1)) \int_{q^t(1)-p^t(1)}^{p^t(1)} \frac{1}{y \ln 2}\, dy & q^t(1) < 2p^t(1) \\ 0 & \text{otherwise} \end{cases} \\ &\leq H(\textsc{Inv-Prof}_{S^t}) - p^t(1)\log(1/p^t(1)) + \begin{cases} \frac{q^t(1)-p^t(1)}{\ln 2} \cdot \ln \frac{p^t(1)}{q^t(1)-p^t(1)} & q^t(1) < 2p^t(1) \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

Combining this inequality with our expression for $H(\mathcal{G}_S^{t+1})$ gives

$$\begin{aligned} M^{t+1} &\leq M^t + p^t(1) \cdot \max_{p<q<2p} \frac{1}{\ln 2} \cdot \frac{q - p^t(1)}{p^t(1)} \ln \frac{p^t(1)}{q - p^t(1)} \\ &= M^t + \mathcal{G}_S(t+1) \cdot \frac{1}{\ln 2} \cdot \max_{0<r<1} \left[r \ln(1/r)\right]. \end{aligned} \tag{8}$$

As the function in brackets is concave, we can attempt to maximize it by setting its derivative to 0:

$$0 = \frac{d}{dr}(r \ln(1/r)) = \ln(1/r) - 1 \implies r = e^{-1}.$$

Since $e^{-1} \in (0, 1)$, this maximizes the function in brackets in the specified range. Plugging this value of r into Eq. (8) gives

$$M^{t+1} \leq M^t + \mathcal{G}_S(t+1) \cdot \frac{1}{e \ln 2} = M^t + \mathcal{G}_S(t+1) \cdot \frac{\log e}{e},$$

as desired.

C PROOF OF THEOREM 4.2: BETTER GREEDY COUPLING APPROXIMATION GUARANTEE FOR ARBITRARILY MANY DISTRIBUTIONS

Theorem 4.2. $H(\mathcal{G}_S) \leq H(\textsc{Profile}_S) + \frac{1+\log e}{2} \approx H(\textsc{Profile}_S) + 1.22$.

We prove this theorem using the following three lemmas. The first step is to upper bound $H(\mathcal{G}_S)$ by an integral.

Lemma C.1. *Let $C_S(x)\colon [0,1] \to [0,1]$ be any non-decreasing function such that if there is x mass left while running the greedy algorithm on S, the next state created by the greedy algorithm has mass at least $C_S(x)$. Then $H(\mathcal{G}_S) \leq \int_0^1 \log \frac{1}{C_S(x)}\, dx$.*

Proof. First rewrite $H(\mathcal{G}_S) = H(\textsc{Sketch}_{\mathcal{G}_S})$. Next, we show that C_S lower bounds $\textsc{Sketch}_{\mathcal{G}_S}$. By definition of C_S, $C_S(x) \leq \textsc{Sketch}_{\mathcal{G}_S}(x)$ whenever $x = 1 - \sum_{j=1}^{i} \mathcal{G}_S(j)$. Since C_S is non-decreasing and $\textsc{Sketch}_{\mathcal{G}_S}(x)$ is constant on the interval $(1 - \sum_{j=1}^{i+1} \mathcal{G}_S(j), 1 - \sum_{j=1}^{i} \mathcal{G}_S(j)]$, it follows that $C_S(x) \leq \textsc{Sketch}_{\mathcal{G}_S}(x)$ for all $x \in (0, 1]$. The conclusion now follows from recalling the definition of $H(\textsc{Sketch}_{G_S})$.

□

We next aim to design a C_S. First, recall our definition of $\textsc{Rem-Mass}_S^{\text{ADVANCED}}(y)$.

$$\triangleq \max_{p \in S} \textsc{Rem-Mass}_p^{\textsc{advanced}}(y)$$
$$\triangleq \max_{p \in S} \sum_{j=1}^{n} \begin{cases} y & y \le \frac{p(j)}{2} \\ p(j)/2 & \frac{p(j)}{2} < y < p(j) \\ p(j) & p(j) \le y \end{cases}$$

Informally, recall that $\textsc{Rem-Mass}_S(y)$ is designed to correspond to an upper bound on the total remaining mass to be coupled if the greedy coupling is about to create a state of size $\le y$. Now, we describe our choice of C_S.

Lemma C.2. *Define* $C_S(x) \triangleq \arg\min_y \left[\textsc{Rem-Mass}_S^{\textsc{advanced}}(y) \ge x\right]$. *Then* C_S *satisfies the preconditions of Lemma C.1.*

Proof. Suppose that there is x mass remaining while running the greedy algorithm, and then the greedy algorithm allocates a state of mass y. Then there exists $p_{i'} \in S$ such that the maximum mass remaining over all states of $p_{i'}$ is equal to y. Observe that showing $\textsc{Rem-Mass}_{p_{i'}}^{\textsc{advanced}}(y) \ge x$ suffices to prove the lemma.

We can naively bound that the amount of mass remaining in the j-th state of $p_{i'}$ is at least $\min(y, p_{i'}(j))$, giving the bound $\sum_{j=1}^{n} \min(y, p_{i'}(j)) \ge x$. However, this alone is not sufficient to obtain the above inequality. We can improve this bound by noting that the masses of the states allocated by the greedy algorithm are non-increasing. Hence for all j with $p_{i'}(j) > y$, we must have previously subtracted at least y from it. Therefore, the j-th state of $p_{i'}$ will have at least $p_{i'}(j)$ mass remaining if $p_{i'}(j) \le y$, and otherwise will have at least $\min(y, p_{i'}(j) - y)$ mass remaining. In total, this gives the bound

$$\sum_{j=1}^{n} \begin{cases} p_{i'}(j) & p_{i'}(j) \le y \\ \min(y, p_{i'}(j) - y) & p_{i'}(j) > y \end{cases} \ge x.$$

We finish by observing that $\textsc{Rem-Mass}_{p_{i'}}^{\textsc{advanced}}(x)$ is at least the LHS of this inequality. □

Remark 2. *While* C_S *corresponds to* $\textsc{Sketch}_{\mathcal{G}_S}$, $\textsc{Rem-Mass}_S^{\textsc{advanced}}$ *corresponds to* $\textsc{Inv-Sketch}_{\mathcal{G}_S}$.

Finally, we prove Theorem 4.2 by upper bounding the integral from Lemma C.1 by $H(\textsc{Inv-Prof}_S)$ plus a small constant.

Lemma C.3. *Let* C_S *be as defined in Lemma C.2. Then*

$$\int_0^1 \log \frac{1}{C_S(x)}\, dx \le H(\textsc{Inv-Prof}_S) + \frac{1 + \log e}{2}. \tag{9}$$

Proof. As C_S is defined in terms of $\textsc{Rem-Mass}_S^{\textsc{advanced}}$, we need to upper bound $\textsc{Rem-Mass}_S^{\textsc{advanced}}(y)$ in terms of $\textsc{Inv-Prof}_S$. Defining $\textsc{Inv-Sketch}_p(y) \triangleq 1$ for $y > 1$, we can check that

$$\begin{aligned} \textsc{Rem-Mass}_p^{\textsc{advanced}}(y) &= \textsc{Inv-Sketch}_p(y) + \frac{1}{2}(\textsc{Inv-Sketch}_p(2y) - \textsc{Inv-Sketch}_p(y)) \\ &\quad + y \int_{2y}^{1} \textsc{Inv-Sketch}_p'(t)/t\, dt \\ &= \frac{\textsc{Inv-Sketch}_p(y) + \textsc{Inv-Sketch}_p(2y)}{2} + y \int_{2y}^{1} \textsc{Inv-Sketch}_p'(t)/t\, dt \end{aligned} \tag{10}$$

As Equation (10) holds for all $p \in S$, it follows that $\textsc{Rem-Mass}_S^{\textsc{advanced}}(x)$ is bounded above by the analogous expression with $\textsc{Inv-Prof}_S$ in place of $\textsc{Inv-Sketch}_p$. Now we finish by computing the integral:

$$\begin{aligned} \int_0^1 \log\left(\frac{1}{C_S(x)}\right) dx &= \int_0^1 \textsc{Rem-Mass}_S^{\textsc{advanced}\prime}(y) \log(1/y)\, dy \qquad (11) \\ &= \textsc{Rem-Mass}_S^{\textsc{advanced}}(y) \log(1/y)\Big|_{y=0}^{1} + \int_0^1 \frac{\textsc{Rem-Mass}_S^{\textsc{advanced}}(y)}{y \ln 2}\, dy \\ &= \int_0^1 \frac{\textsc{Rem-Mass}_S^{\textsc{advanced}}(y)}{y \ln 2}\, dy \end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\ln 2}\int_0^1 \frac{\textsc{Inv-Prof}_S(y)+\textsc{Inv-Prof}_S(2y)}{2y} + \left(\int_{2y}^1 \textsc{Inv-Prof}'_S(t)/t\,dt\right) dy \\
&= \frac{1}{2}H(\textsc{Inv-Prof}_S) + \frac{1}{2}H(\textsc{Inv-Prof}_S) + \frac{1}{\ln 2}\int_{1/2}^1 \frac{1}{2y}\,dy + \frac{1}{2\ln 2}\int_0^1 \textsc{Inv-Prof}'(y)\,dy \\
&= H(\textsc{Inv-Prof}_S) + \frac{1}{2} + \frac{1}{2\ln 2} \\
&= H(\textsc{Inv-Prof}_S) + \frac{1+\log e}{2}.
\end{aligned}
$$

□

D COUPLING GUARANTEES FOR SMALL m

We need to generalize Eq. (7) to the setting of general m. Without loss of generality assume $p_1^t(1) \leq p_2^t(1) \leq \cdots \leq p_m^t(1)$. Then

$$
\textsc{Inv-Sketch}_{p_i^{t+1}}(y) = \begin{cases} \textsc{Inv-Sketch}_{p_i^t}(y) & y < p_i^t(1) - p_1^t(1) \\ \textsc{Inv-Sketch}_{p_i^t}(y) + p_i^t(1) - p_1^t(1) & p_i^t(1) - p_1^t(1) \leq y < p_i^t(1) \\ \textsc{Inv-Sketch}_{p_i^t}(y) - p_1^t(1) & p_i^t(1) \leq y \end{cases}
$$

and Eq. (7) becomes

$$
\textsc{Inv-Prof}_{S^{t+1}} \begin{cases} \leq \textsc{Inv-Prof}_{S^t}(y) + p_i^t(1) - p_1^t(1) & p_i^t(1) - p_1^t(1) \leq y < \min(p_{i+1}^t(1) - p_1^t(1), p_i^t(1)) \\ = \textsc{Inv-Prof}_{S^t}(y) - p^t(1) = \textsc{Mass}(S^{t+1}) & p_i^t(1) \leq y \end{cases}
$$

Letting $d_i \triangleq \frac{p_i^t(1)-p_1^t(1)}{p_1^t(1)}$ for $i \in [2,m]$ and $d_{m+1} = 1$, we then find that the generalization of Eq. (8) is as follows:

$$
\begin{aligned}
M^{t+1} &\leq M^t + \mathcal{G}_S(t+1) \cdot \max_{0<d_2<d_3<\cdots<d_{m+1}=1} \frac{1}{\ln 2}\sum_{i=2}^m d_i \ln(d_{i+1}/d_i) \\
&= M^t + \mathcal{G}_S(t+1) \cdot \max_{0<d_2<d_3<\cdots<d_{m+1}=1} \frac{1}{\ln 2}\sum_{i=2}^m d_i \left(\ln(1/d_i) - \ln(1/d_{i+1})\right).
\end{aligned}
$$

The quantity

$$
\sum_{i=2}^m d_i \left(\ln(1/d_i) - \ln(1/d_{i+1})\right) \tag{12}
$$

can be more intuitively interpreted as follows: "Given $m-1$ rectangles indexed from $2 \ldots m$ all with lower-left corner at the origin, and the i-th has width d_i and height $\ln(1/d_i)$, what is the maximum possible area of their union?"

Remark 3. *As $m \to \infty$, Eq.* (12) *approaches* $\frac{1}{\ln 2}\int_0^1 \ln(1/x)\,dx = \frac{1}{\ln 2}\left(x - x\ln x\right)\Big|_0^1 = \log e \approx 1.44$.

Table 3: Best-Known Additive Approximation Guarantees

	Prior Work	**This Work**
$m = 2$	1 [Cicalese et al., 2019, Rossi, 2019, Li, 2021]	$\frac{1}{e}\log_2(e) \approx 0.53$
$m = 3$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<d_3<d_4=1} \sum_{i=2}^{3} d_i \log_2(d_{i+1}/d_i) \approx 0.77$
$m = 4$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_5=1} \sum_{i=2}^{4} d_i \log_2(d_{i+1}/d_i) \approx 0.90$
$m = 5$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_6=1} \sum_{i=2}^{5} d_i \log_2(d_{i+1}/d_i) \approx 0.99$
$m = 6$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_7=1} \sum_{i=2}^{6} d_i \log_2(d_{i+1}/d_i) \approx 1.06$
$m = 7$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_8=1} \sum_{i=2}^{7} d_i \log_2(d_{i+1}/d_i) \approx 1.10$
$m = 8$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_9=1} \sum_{i=2}^{8} d_i \log_2(d_{i+1}/d_i) \approx 1.14$
$m = 9$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_{10}=1} \sum_{i=2}^{9} d_i \log_2(d_{i+1}/d_i) \approx 1.17$
$m = 10$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_{11}=1} \sum_{i=2}^{10} d_i \log_2(d_{i+1}/d_i) \approx 1.19$
$m = 11$	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\max_{0<d_2<\cdots<d_{12}=1} \sum_{i=2}^{11} d_i \log_2(d_{i+1}/d_i) \approx 1.21$
General m	$\log_2(e) \approx 1.44$ [Compton, 2022]	$\frac{1}{2}(\log_2(e)+1) \approx 1.22$

E EXPERIMENT DETAILS

Experiments were run on a laptop with a 3.49GHz 8-Core Apple M2 processor. Each entry of Table 2 corresponds to the average of 100 runs over the specified configuration. If any of the runs exceeded 120 seconds, the entry is marked as "> 120". In Table 5, we provide a corresponding experiment when coupling two distributions with different numbers of states n_1, n_2. In Table 4, we provide the same data as Table 2, but accompanied with the standard deviation. Each input distribution was sampled from the Dirichlet distribution with parameter 1 (meaning, uniform over the simplex). The algorithm "Naive Polytope Vertex Enumeration" corresponds to naively enumerating over all vertices in the corresponding optimization polytope. We were particularly generous to this baseline, and implemented it using insights regarding induced spanning trees (that were not discussed before our work) to avoid requiring it to solve systems of linear equations. Essentially, this would run even slower if done extremely naively. Rows of the form "Backtracking [X]" denotes using the backtracking algorithm of Appendix F.2 with the lower bound X for its subroutine that helps prune the search space. The algorithm "Dynamic Programming" corresponds to Appendix F.1. All code is implemented in C++.

Table 4: Average Runtime and Standard Deviation (in Seconds) for Exactly Coupling Two Distributions

Algorithm	n							
	4	5	6	7	8	9	10	11
Naive Polytope Vertex Enumeration	0.002 ± 0.0006	0.189 ± 0.002	33.79 ± 0.240	> 120	> 120	> 120	> 120	> 120
Backtracking [0]	0.0003 ± 0.0001	0.004 ± 0.002	0.106 ± 0.041	3.576 ± 1.839	> 120	> 120	> 120	> 120
Backtracking [$H(\bigwedge S)$]	0.0001 ± 0.00006	0.001 ± 0.0005	0.013 ± 0.010	0.256 ± 0.196	4.907 ± 3.896	> 120	> 120	> 120
Backtracking [$H(\text{PROFILE}_S)$]	0.0001 ± 0.00006	0.0009 ± 0.0005	0.009 ± 0.008	0.153 ± 0.133	2.206 ± 1.995	> 120	> 120	> 120
Backtracking [$H(\text{MAJOR-PROFILE}_S)$]	0.00006 ± 0.00003	0.0004 ± 0.0002	0.003 ± 0.003	0.035 ± 0.035	0.344 ± 0.352	5.398 ± 7.761	> 120	> 120
Dynamic Programming	0.00007 ± 0.00002	0.0004 ± 0.00002	0.002 ± 0.00004	0.012 ± 0.0003	0.093 ± 0.001	0.802 ± 0.016	9.769 ± 0.278	> 120

Table 5: Average Runtime (in Seconds) for Exactly Coupling Two Distributions with Different Cardinalities

Algorithm	n_1, n_2										
	6,3	7,3	8,3	9,3	10,4	11,4	12,4	13,5	14,5	15,5	16,6
Naive Polytope Vertex Enumeration	0.0003	0.038	0.219	1.172	> 120	> 120	> 120	> 120	> 120	> 120	> 120
Backtracking [0]	0.0001	0.002	0.008	0.030	1.147	4.538	19.841	> 120	> 120	> 120	> 120
Backtracking [$H(\bigwedge S)$]	0.00008	0.0008	0.003	0.010	0.157	0.498	2.139	> 120	> 120	> 120	> 120
Backtracking [$H(\text{PROFILE}_S)$]	0.00009	0.0008	0.003	0.008	0.104	0.295	1.066	> 120	> 120	> 120	> 120
Backtracking [$H(\text{MAJOR-PROFILE}_S)$]	0.00006	0.0005	0.001	0.005	0.046	0.139	0.566	5.855	19.927	> 120	> 120
Dynamic Programming	0.00005	0.0003	0.0008	0.002	0.011	0.031	0.084	0.724	2.509	8.984	> 120

F ALGORITHMS FOR SECTION 5: EXACTLY COUPLING TWO DISTRIBUTIONS, AND EFFICIENT LOWER BOUNDS

F.1 Dynamic Programming

First, we note that by Lemma 5.3, any minimum-entropy coupling of $\mathbf{p}$ can be associated with some spanning forest of the unweighted graph with vertex set $V = \{\ell_{1\ldots n}, r_{1\ldots n}\}$; then we can add 0-weight edges to change this forest into a spanning tree.

Our dynamic programming solution searches over all possible spanning trees. It keeps track of $O(2^{2n} \cdot n)$ states: for every $S \subseteq V$ and $v \in S$, we store the minimum entropy over all partial couplings of S such that v is the only vertex in S with any mass remaining in $dp[S][v]$. Note that the remaining mass of v must equal $\sum_{w \in S} (-1)^{\text{side}(w) \neq \text{side}(v)} \text{orig\_mass}(w)$ across all such partial couplings. Effectively, the state (S, v) corresponds to all partial couplings that induce a spanning tree on S rooted at v.

There are two types of transitions between states: the first attaches a new root to a spanning tree (taking $O(2^{2n} \cdot n^2)$ time), while the second merges two spanning trees with the same root (taking $O(3^{2n} \cdot n)$ time). The minimum entropy over all couplings is $dp[V][v]$ for any $v \in V$.

F.2 Backtracking

Algorithm 2 Backtracking Algorithm

1: **Input:** Marginal distributions of $m = 2$ variables each with n states $\{\mathbf{p}, \mathbf{q}\}$.
2: Initialize $best\_entropy = \infty$.
3: **Backtrack**$(\{\mathbf{p}, \mathbf{q}\}, entropy\_so\_far = 0, last\_mass = \infty)$.
4: **return** $best\_entropy$.
5: **procedure BACKTRACK**$(\{\mathbf{p}, \mathbf{q}\}, entropy\_so\_far, last\_mass)$
6: **if** $entropy\_so\_far + X(\{\mathbf{p}, \mathbf{q}\}) \geq best\_entropy$ **then**
7: **return**
8: **end if**
9: **if** Mass$(\{\mathbf{p}, \mathbf{q}\}) = 0$ **then**
10: $best\_entropy = \min(best\_entropy, entropy\_so\_far)$.
11: **return**
12: **end if**
13: **for** (i, j) s.t. $0 < \min(\mathbf{p}(i), \mathbf{q}(j)) \leq last\_mass$ **do**
14: $m = \min(\mathbf{p}(i), \mathbf{q}(j))$. $\mathbf{p}' = \mathbf{p}$. $\mathbf{q}' = \mathbf{q}$.
15: $\mathbf{p}'(i) -= m$. $\mathbf{q}'(j) -= m$.
16: **Backtrack**$(\{\mathbf{p}', \mathbf{q}'\}, entropy\_so\_far + m \log(1/m), m)$.
17: **end for**
18: **return**
19: **end procedure**

Now we argue that this algorithm is correct. First, note that line 6 does not affect the correctness of the algorithm as long as $X(\{p, q\}) \leq \text{OPT}(\{p, q\})$, so we can ignore it. Next, we show that as long as $\mathbf{p}(i, j)$ is some minimum entropy coupling, our algorithm will construct it. Let (i, j) be some pair such that $\mathbf{p}(i, j)$ is maximized. Then by Lemma 5.3, our algorithm will compute $m = \min(p(i), q(j))$ on the (i, j)-th iteration of step 4. Then by induction, our algorithm will construct the remainder of $\mathbf{p}$ after recursing into $\{p', q'\}$ on that loop iteration.

F.3 Lower Bounds in Almost-Linear Time

Computing PROFILE$_S$. Recall from Definition 3.1 that for all $p \in S$ and $i \leq |p|$, we must have $\text{PROFILE}_S\left(\sum_{j \geq i} p(j)\right) \leq p(i)$. We can compute PROFILE_S using the following procedure:

1. Define $T = \{(0, 0)\} \cup \{(\sum_{j \geq i} p(j), p(i)) \mid p \in S, 1 \leq i \leq |p|\}$.
2. For each pair $(x, y) \in T$ in decreasing order of x, if y is greater or equal to the smallest y seen so far, remove the pair from T. In other words, there exists another pair $(x', y') \in T$ such that $x' \geq x$ and $y' \leq y$, so the pair (x, y) is redundant and removing it does not affect the profile.

3. Reverse T. Now the entries of T are increasing in both x and y.

Now $H(\text{PROFILE}_S)$ is the sum of $(x_2 - x_1)\log(1/y_2)$ over all adjacent pairs (x_1, y_1) and (x_2, y_2) in T. The time complexity is $\tilde{O}(nm)$.

Computing MAJOR-PROFILE$_S$. Assume we have already computed $T = [(x_1, y_1), \ldots, (x_{|T|}, y_{|T|})]$ as in the previous part. Then it suffices to describe how to implement the algorithm described in Appendix G.3 in $O(|T|)$ additional time. We recall that the algorithm essentially asks us to answer the following query for at most $|T|$ values of t in decreasing order: how large can the side length of a square with lower-right corner at $(t, 0)$ be without the square exceeding the profile?

We first describe how to answer each query in $O(|T|)$ time, resulting in an $O(|T|^2)$ algorithm. If we are growing a square with lower-right corner at $(t, 0)$, then its side length must be

$$\min_{(x,y)\in T} \max(y, t - x) = \min_{(x,y)\in T} \begin{cases} t - x & x + y \le t \\ y & x + y > t \end{cases}, \tag{13}$$

which can be evaluated in $O(|T|)$ time.

We can speed up the above algorithm by noting that T is sorted in increasing order of $x + y$. Then define j be the minimum index such that $x_j + y_j > t$. Now Equation (13) simplifies to:

$$\min_{(x,y)\in T} \begin{cases} t - x & x + y \le t \\ y & x + y > t \end{cases} = \min\left(\min_{1\le i<j} t - x_i, \min_{j\le i\le |T|} y_j\right) = \min(t - x_{j-1}, y_j).$$

As the queries come in decreasing order of t, j must stay the same or decrease between queries. Thus the time required to answer all queries is proportional to the number of queries plus the number of times j moves between queries, both of which are $O(|T|)$.

G PROOFS FOR SECTION 5: EXACTLY OPTIMAL SOLUTIONS

G.1 Proof of Corollary 5.2

Corollary 5.2. MEC $\le_{NP}$ COMPUTE-ENTROPY.

By Lemma 5.1 it suffices to provide the coordinates of some vertex $v \in \mathbf{P}$ at which the objective is minimized. We note that Lemma 5.1 shows that every vertex has at most $e \le nm - m + 1$ nonzero coordinates. Furthermore, if the input probabilities are rational numbers with bounded bit complexity, then v' (the vector consisting only of the nonzero entries of v) must be the unique solution to a linear system of the form $Av' = b$, where A consists solely of zeros and ones, and A has shape $e \times e$. The solution to this system can easily be seen to have bit complexity bounded above by the bit complexity of $\mathbf{p}$ times poly(nm).

G.2 Proof of Lemma 5.3

Lemma 5.3. *Consider the weighted undirected bipartite graph G with vertex set $\{\ell_{1\ldots n}, r_{1\ldots n}\}$ and edge set $\{(\ell_i, r_j, \mathbf{p}(i, j))) \mid \mathbf{p}(i, j) > 0\}$ induced by some minimum-entropy coupling $\mathbf{p}$. This graph is a forest, and any edge with the maximum weight must have a leaf of the forest as one of its endpoints.*

For the first part of the lemma, it suffices to check that G cannot contain any cycles. Suppose for the sake of contradiction that G contains a simple cycle $(i_1, j_1, i_2, j_2, \ldots, i_k, j_k)$, and define $\delta > 0$ to be the minimum weight along the cycle. Then because entropy is concave, we can obtain a smaller entropy by alternately adding and subtracting δ along the edges of the cycle.

For the second part, suppose for the sake of contradiction that there exists an edge (i, j) of the maximum weight such that i is connected to some other vertex j' and j is connected to some other vertex i' in G. Let the weights $w(i, j)$, $w(i, j')$, and $w(i', j)$ be a, b, and c, respectively; WLOG $a \ge b \ge c$. Then we claim that the entropy would decrease if we perform the following four modifications:

1. increase $w(i, j)$ by c

2. decrease $w(i, j')$ by c
3. increase $w(i', j')$ by c
4. set $w(i', j) = 0$

Indeed, the change in entropy is upper bounded by

$$\begin{aligned}&[(a+c)\log(1/(a+c)) + (b-c)\log(1/(b-c)) + c\log(1/c)] - [a\log(1/a) + b\log(1/b) + c\log(1/c)]\\&= [(a+c)\log(1/(a+c)) + (b-c)\log(1/(b-c))] - [a\log(1/a) + b\log(1/b)]\,,\end{aligned}$$

which is negative because because $(a + c, b - c)$ majorizes (a, b).

G.3 Proof of Theorem 5.5

Theorem 5.5. $H(\text{OPT}_S) \geq H(\textsc{Major-Profile}_S) \geq H(\textsc{Profile}_S), H(\bigwedge S)$

We first describe how to construct $\textsc{Major-Profile}_S$. In general, we define $\textsc{Major-Profile}_S(i)$ in terms of $\textsc{Major-Profile}_S(1)$, $\textsc{Major-Profile}_S(2)$, $\ldots$, $\textsc{Major-Profile}_S(i-1)$. Specifically, if $\sum_{j=1}^{i-1} \textsc{Major-Profile}_S(j) < \textsc{Mass}(S)$, then $\textsc{Major-Profile}_S(i)$ is defined to be the maximum positive real r_i such that $\textsc{Profile}_S(x) \geq r_i$ for all $x > \textsc{Mass}(S) - \sum_{j=1}^{i-1} \textsc{Major-Profile}_S(j) - r_i$. Intuitively, this corresponds to growing a square with side length $r_i \times r_i$ and lower-right corner at $(1 - \sum_{j=1}^{i-1} \textsc{Major-Profile}_S(j), 0)$ until it touches $\textsc{Profile}_S$. Note that $r_{i+1} > r_i$ would lead to a contradiction, because setting $r_i = r_{i+1}$ would lead to the square associated with r_i lying below the profile, meaning that r_i could have been larger. Thus $r_i \geq r_{i+1}$, meaning that our construction produces the elements of a probability distribution in non-increasing order.

Next, we show that $\textsc{Major-Profile}_S$ majorizes any distribution p such that $\textsc{Sketch}_p(x) \leq \textsc{Profile}_S(x)$ for all x using induction. Suppose that we have already proven that $\sum_{j=1}^{i} p(j) \leq \sum_{j=1}^{i} \textsc{Major-Profile}_S(j)$; it remains to show this inequality for $i + 1$. Indeed, the inequality for $i + 1$ rearranges to

$$r_i \geq \sum_{j=1}^{i+1} p(j) - \sum_{j=1}^{i} \textsc{Major-Profile}_S(j).$$

This inequality is true because the entirety of the square with lower-right corner at $(1-\sum_{j=1}^{i} \textsc{Major-Profile}_S(j), 0)$ and side length $\max\left(\sum_{j=1}^{i+1} p(j) - \sum_{j=1}^{i} \textsc{Major-Profile}_S(j), 0\right)$ is contained within the square with lower-right corner at $(\sum_{j=1}^{i} p(j), 0)$ and side length $p(i + 1)$. The entirety of this second square lies below $\textsc{Profile}_S$ by the assumption on p.

Now we are ready to show the inequalities in Theorem 5.5.

- Because $\textsc{Sketch}_{\text{OPT}_S}$ lies below $\textsc{Profile}_S$, the work above shows that OPT_S is majorized by $\textsc{Major-Profile}_S$. Thus, $H(\text{OPT}_S) \geq H(\textsc{Major-Profile}_S)$.
- Next,

$$\begin{aligned}H(\textsc{Major-Profile}_S) &= \int_0^1 \log\left(1/\textsc{Sketch}_{\textsc{Major-Profile}_S}(x)\right)\, dx\\&\geq \int_0^1 \log\left(1/\textsc{Profile}_S(x)\right)\, dx\\&= H(\textsc{Profile}_S),\end{aligned}$$

  where the inequality holds because $\textsc{Sketch}_{\textsc{Major-Profile}_S}(x) \leq \textsc{Profile}_S(x)$ for all x.
- Finally, we claim that $H(\textsc{Major-Profile}_S) \geq H(\bigwedge S)$ because $\bigwedge S$ majorizes $\textsc{Major-Profile}_S$. Suppose for the sake of contradiction that $\bigwedge S$ does not majorize $\textsc{Major-Profile}_S$; then there exists $p \in S$ such that p does not majorize $\textsc{Major-Profile}_S$. This in turn implies that there exists some i such that $\sum_{j=1}^{i} p(j) \geq \sum_{j=1}^{i} \textsc{Major-Profile}_S(j)$ and $\sum_{j=1}^{i+1} p(j) < \sum_{j=1}^{i+1} \textsc{Major-Profile}_S(j)$. But then there exists some x such that $\textsc{Sketch}_{\textsc{Major-Profile}_S}(x) > \textsc{Sketch}_p(x) \geq \textsc{Profile}_S(x)$, contradicting the assumption that $\textsc{Sketch}_{\textsc{Major-Profile}_S}(x) \leq \textsc{Profile}_S(x)$ for all x. Thus, $\bigwedge S$ must majorize $\textsc{Major-Profile}_S$.

Remark 4. *It is natural to ask whether we can in fact show that $H(\text{PROFILE}_S) \geq H(\bigwedge_S)$, particularly given that this inequality often holds for randomly generated instances (e.g. all the runs included in Table 4). It turns out that this is not true in general; $S = \{[0.5, 0.5], [0.75, 0.05, 0.05, 0.05, 0.05, 0.05]\}$ is a counterexample.*

H OPTIMALITY GAPS: BOUNDS ON POSSIBLE IMPROVEMENT

In this section, we examine the open landscape of upper bounds for greedy coupling in reference to particular lower bounds. In Table 6, we recall the best-known proven upper bounds on the entropy of the greedy coupling with respect to particular lower bounds. Whether or not these bounds can be improved (and if so, by how much) are important open questions. In the first column ("Lower Bound on $H(\mathcal{G}_S)$") of Table 7, we detail counter-examples that show limits for how much the upper bounds of the greedy algorithm in Table 6 can be improved. In the second column ("Lower Bound on $H(\text{OPT}_S)$") of Table 7, we detail counter-examples that show limits for how much the upper bounds of *any algorithm* can be improved. The counter-examples are a mixture of results from prior work, new hand-constructed instances, and instances we computationally discovered with a local search. Note that the entries in the first and fifth row of the second column of Table 7 correspond to the majorization barrier. Finally, in Appendix H.4 we discuss counter-examples that show the algorithms of [Cicalese et al., 2019, Li, 2021] for coupling $m = 2$ distributions do not match the approximation guarantee for the greedy coupling we show in Theorem 4.1.

We emphasize that for all upper bounds in Table 7 it is open whether they can be improved (other than $H(\bigwedge_S)$ for general m, which is tight). We note that this section shows how the gap for improving the upper bound of $H(\mathcal{G}_S) - H(\text{OPT}_S)$ for $m = 2$ is comparatively small, as we are able to provide an input where $H(\mathcal{G}_S) - H(\text{OPT}_S) = 0.40$, which is relatively close to the upper bound provided by Theorem 4.1 (≈ 0.53). On the other hand, for general m, we were unable to generate instances with $H(\mathcal{G}_S) - H(\text{OPT}_S) > 0.71$, which is relatively far away from the upper bound provided by Theorem 4.2 (≈ 1.22). We did not bound $H(\text{OPT}_S) - H(\text{PROFILE}_S)$ or $H(\text{OPT}_S) - H(\text{MAJOR-PROFILE}_S)$ for general m due to how it is computationally infeasible to compute optimal solutions for general m (making local search similarly infeasible).

Please note that all lower bounds in Table 7 are simply the best counter-examples that are currently known (some by a fairly limited computational search), and do not correspond to any conjecture of the best lower bound.

Reference Lower Bound	m	Best Upper Bound on $H(\mathcal{G}_S)$
$H(\bigwedge_S)$	$m = 2$	$H(\mathcal{G}_S) \leq H(\bigwedge_S) + 1$ [Rossi, 2019]
$H(\text{PROFILE}_S)$	$m = 2$	$H(\mathcal{G}_S) \leq H(\text{PROFILE}_S) + 0.53$ (Theorem 4.1)
$H(\text{MAJOR-PROFILE}_S)$	$m = 2$	same as above
$H(\text{OPT}_S)$	$m = 2$	same as above
$H(\bigwedge_S)$	general m	$H(\mathcal{G}_S) \leq H(\bigwedge_S) + 1.44$ [Compton, 2022]
$H(\text{PROFILE}_S)$	general m	$H(\mathcal{G}_S) \leq H(\text{PROFILE}_S) + 1.22$ (Theorem 4.2)
$H(\text{MAJOR-PROFILE}_S)$	general m	same as above
$H(\text{OPT}_S)$	general m	same as above

Table 6: Upper Bounds on Greedy

Reference Lower Bound	m	Lower Bound on $H(\mathcal{G}_S)$	Lower Bound on $H(\text{OPT}_S)$
$H(\bigwedge_S)$	$m = 2$	$H(\mathcal{G}_S) \approx H(\bigwedge_S) + 0.66$ (Appendix H.2)	$H(\text{OPT}_S) \approx H(\bigwedge_S) + 0.66$ (Appendix H.2)
$H(\text{PROFILE}_S)$	$m = 2$	$H(\mathcal{G}_S) \approx H(\text{PROFILE}_S) + 0.46$ (Appendix H.1)	$H(\text{OPT}_S) \approx H(\text{PROFILE}_S) + 0.39$ (Appendix H.2)
$H(\text{MAJOR-PROFILE}_S)$	$m = 2$	same as above	$H(\text{OPT}_S) \approx H(\text{MAJOR-PROFILE}_S) + 0.35$ (Appendix H.2)
$H(\text{OPT}_S)$	$m = 2$	$H(\mathcal{G}_S) = H(\text{OPT}_S) + 0.40$ (Appendix H.2)	$=$
$H(\bigwedge_S)$	general m	$H(\mathcal{G}_S) \approx H(\bigwedge_S) + 1.44$ [Compton, 2022]	$H(\text{OPT}_S) \approx H(\bigwedge_S) + 1.44$ [Compton, 2022]
$H(\text{PROFILE}_S)$	general m	$H(\mathcal{G}_S) \approx H(\text{PROFILE}_S) + 0.89$ (Appendix H.1)	same as $m = 2$
$H(\text{MAJOR-PROFILE}_S)$	general m	same as above	same as $m = 2$
$H(\text{OPT}_S)$	general m	$H(\mathcal{G}_S) \approx H(\text{OPT}_S) + 0.71$ (Appendix H.3)	$=$

Table 7: Lower Bounds on Greedy and OPT

H.1 Gaps Between $H(\mathcal{G}_S)$ and $H(\text{PROFILE}_S)$

Our lower bounds are achieved by letting $\text{PROFILE}_S = \text{SKETCH}_{p^*}$, where $p^* \in S$ is a uniform distribution. We note that $H(\text{PROFILE}_S) = H(\text{MAJOR-PROFILE}_S) = H(p^*)$.

Let U_s denote the uniform distribution over s states.

Case $m = 2$. We used $S = \{U_{F_t}, U_{L_{t-1}}\}$. Here, $F_n = \text{round}\left(\frac{\varphi^n}{\sqrt{5}}\right)$ denotes the n-th Fibonacci number, while $L_n = \text{round}\left(\varphi^n\right)$ denotes the n-th Lucas number. For $t = 12$, we have $F_t = 144$, $L_{t-1} = 199$, and $H(\mathcal{G}_S) - H(p^*) > 0.457$.

Case general m. We started with $S = \{U_1, U_2, \ldots, U_n\}$, where $n = 2000$. This construction already gives $H(\mathcal{G}_S) - H(p^*) > 0.805$. To construct instances where $H(\mathcal{G}_S) - H(p^*)$ is even greater, we first state an inequality characterizing the behavior of the greedy algorithm on S:

$$\mathcal{G}_S(t+1) \geq \min\left(\frac{1 - \sum_{i=1}^{t} \mathcal{G}_S(i)}{\min(t, n)}, \frac{1}{n}\right). \tag{14}$$

The first expression in the right hand side of Equation (14) comes from all $p \in S$ with $|p| \leq t$, while the second expression comes from all $p \in S$ with $|p| > t$. Thus, $\mathcal{G}_S$ majorizes the distribution $\mathcal{G}'_S$ satisfying $\mathcal{G}'_S(t+1) = \min\left(\frac{1-\sum_{i=1}^{t} \mathcal{G}'_S(i)}{\min(t,n)}, \frac{1}{n}\right)$ for all $t \geq 0$. If we add to S a distribution p_t with $|p_t| = t$ to make Equation (14) tight for each $t \in (1000, 1414]$, then we obtain $H(\mathcal{G}_S) - H(p^*) > 0.887$. It is likely possible to construct instances with even greater gaps between $\mathcal{G}_S$ and p^*, as

$$H(\mathcal{G}_S) - H(\textsc{Profile}_S) \leq H(\mathcal{G}'_S) - H(\textsc{Profile}_S) \approx 1.082. \tag{15}$$

However, Equation (14) cannot be made tight for $t = 1415$, so we cannot have equality in Equation (15). It is unclear what the value of $H(\mathcal{G}_S) - H(\textsc{Profile})$ is if we let S consist of *all* distributions with sketches lying above $\textsc{Sketch}_{p^*}$.

H.2 Local Search for $m = 2$

In this section, we detail "counter-examples" found with a computational local search, i.e. example distributions achieving large gaps between various coupling entropy bounds. These example empirical gaps provide limits on any future theoretical bounds on the size of these gaps. At a high-level, this search works by repeatedly making random perturbations to the input distributions and accepting changes that increase the gap we are finding a lower bound for.

$\underline{H(\mathcal{G}_S) - H(\bigwedge_S) \approx 0.662463}$
$p_1 = [0.3199940773, 0.3199844734, 0.1200540976, 0.1200022716, 0.1199650801]$
$p_2 = [0.2000218248, 0.2000211548, 0.2000202369, 0.1999737730, 0.1999630105]$

$\underline{H(\text{OPT}_S) - H(\bigwedge_S) \approx 0.662405.}$
$p_1 = [0.3199940773, 0.3199844734, 0.1200540976, 0.1200022716, 0.1199650801]$
$p_2 = [0.2000218248, 0.2000211548, 0.2000202369, 0.1999737730, 0.1999630105]$

$\underline{H(\text{OPT}_S) - H(\textsc{Profile}_S) \approx 0.389941.}$
$p_1 = [0.2128275903, 0.2122898591, 0.2119627146, 0.2119384365, 0.1509813995]$
$p_2 = [0.2747693214, 0.2739951951, 0.2739769942, 0.1161585898, 0.0610998995]$

$\underline{H(\text{OPT}_S) - H(\textsc{Major-Profile}_S) \approx 0.354485.}$
$p_1 = [0.2128275903, 0.2122898591, 0.2119627146, 0.2119384365, 0.1509813995]$
$p_2 = [0.2747693214, 0.2739951951, 0.2739769942, 0.1161585898, 0.0610998995]$

$\underline{H(\mathcal{G}_S) - H(\text{OPT}_S) \approx 0.395053.}$
$p_1 = [0.4081266587, 0.3060949942, 0.1530474970, 0.0765237476, 0.0382618746, 0.0179452279]$
$p_2 = [0.3060949942, 0.2040633294, 0.2040633294, 0.1530474970, 0.0765237476, 0.0382618746, 0.0179452278]$

We can modify this last example to obtain $H(\mathcal{G}_S) - H(\text{OPT}_S) = 0.4$ exactly. Specifically, we can let
$p_1 = [0.4, 0.3, 0.15, 0.075, 0.0375, \ldots]$
$p_2 = [0.3, 0.2, 0.2, 0.15, 0.075, 0.0375, \ldots]$

where the dots represent a geometric sequence with ratio 0.5. Then $\text{OPT}_S = p_2$, but $\mathcal{G}_S$ replaces one occurrence of 0.2 in OPT_S with $[0.1, 0.05, 0.025, \ldots]$.

H.3 Gap Between $H(\mathcal{G}_S)$ and $H(\text{OPT}_S)$ for General m

We can generate instances where $H(\mathcal{G}_S)$ and $H(\text{OPT}_S)$ differ by starting with $S = \{\text{OPT}_S\}$ and repeatedly inserting into S any probability distribution generated by combining states of OPT_S. We generated 2000 instances with $|\text{OPT}_S| = 10$

from the simplex and $m = 4001$. The instance that resulted in the maximum difference between $H(\mathcal{G}_S)$ and $H(\text{OPT}_S)$ gave $H(\mathcal{G}_S) \approx H(\text{OPT}_S) + 0.707147$.

H.4 Counter-Examples for Other Algorithms

The works of [Cicalese et al., 2016, Li, 2021] introduce algorithms for coupling $m = 2$ distributions. Prior to our work, the algorithms of [Kocaoglu et al., 2017a, Cicalese et al., 2019, Li, 2021] all had the best additive guarantee of 1 bit. By Theorem 4.1, we show the greedy coupling algorithm of [Kocaoglu et al., 2017a] is always within $\log(e)/e \approx 0.53$ bits of the optimum. We provide a counter-example where the algorithms of [Cicalese et al., 2016, Li, 2021] do not meet this approximation guarantee (found with local search). We do not conjecture whether these algorithms have counter-examples where their optimality gap is larger. Interestingly, for this counter-example, the algorithms of [Cicalese et al., 2016, Li, 2021] both produce a coupling with the same output, where both algorithms have an optimality gap of ≈ 0.6265549682 for:
$p_1 = [0.5540050843, 0.1984459548, 0.1288780486, 0.0396890356, 0.0789189118, 0.0000629649]$
$p_2 = [0.2770967899, 0.2769100227, 0.1984729975, 0.1288783386, 0.0789194408, 0.0397224105]$

I GENERALIZING TO CONCAVE MINIMUM-COST COUPLINGS

We emphasize that our methods in all previous sections are not unique to the entropy function but broadly apply to concave functions. These methods also have the flexibility to yield multiplicative guarantees for functions where an additive approximation guarantee is unexpected. We will define the cost function for coupling with distribution p as $F_{\text{cost}}(p) \triangleq \sum_i f_{\text{cost}}(p(i))$. The necessary assumptions are that f_{cost} is non-negative and concave and $f_{\text{cost}}(0) = 0$. Note that entropy is a specific example of this where $f_{\text{cost}}(x) = x\log(1/x)$. The corresponding lower bound given by the profile is $F_{\text{cost}}(\textsc{Profile}_S) = \int_0^1 \frac{f_{\text{cost}}(\textsc{Profile}_S(x))}{\textsc{Profile}_S(x)}\,dx$.

For cost functions of the form $f_{\text{cost}}(x) = x^c$ where $c \in (0,1)$, we can use the same methods as Section 4.1 and Section 4.2 to show that the cost achieved by the greedy algorithm is within a multiplicative factor of the optimum.

Theorem I.1. *For $m = 2$, let $r \triangleq \max_{0<q<p\le 1} \frac{f_{cost}(q)}{f_{cost}(p)} - \frac{q}{p}$. If $r < 1$, then the greedy algorithm achieves a $\frac{1}{1-r}$-multiplicative approximation.*

Theorem I.2. *For general m, when $f_{cost}(x) = x^c$, the greedy algorithm obtains a $\left(\frac{1}{2} + \frac{1}{c2^c}\right)$-multiplicative approximation.*

Corollary I.3. *For $f_{cost}(x) = \sqrt{x}$, the greedy algorithm obtains a $4/3$-multiplicative approximation for $m = 2$ and a ≈ 1.91-multiplicative approximation for general m.*

Proof. For $m = 2$, substituting f_{cost} into Theorem I.1, we obtain $r = (q/p)^{0.5} - q/p$. This is maximized when $q/p = 1/4$, which gives us $r = 1/4$. Thus, the greedy algorithm gives us a $\frac{1}{1-1/4} = \frac{4}{3}$-multiplicative approximation. For general m, we simply substitute $c = 0.5$ into Theorem I.2, giving a $(\frac{1}{2} + \sqrt{2})$-multiplicative approximation.

□

We note that the part of Corollary I.3 corresponding to $m = 2$ can easily be generalized to other cost functions of the form $f_{\text{cost}}(x) = x^c$ for $c \in (0,1)$. For general m, one could obtain guarantees for general concave costs by integrating over $\int_0^1 \textsc{Rem-Mass}_S^{\text{ADVANCED}}(y) \cdot \frac{f_{\text{cost}}(y)}{y}\,dy$, analogous to Eq. (3).

I.1 Proof of Theorem I.1

We first extend the definition of F_{cost} to inverse sketches (and profiles):

$$F_{\text{cost}}(\textsc{Inv-Sketch}_p) \triangleq \textsc{Inv-Sketch}_p(1) f_{\text{unit}}(1) + \int_0^1 \textsc{Inv-Sketch}_p(y)(-f'_{\text{unit}}(y))\,dy \tag{16}$$

$$= \int_0^1 \textsc{Inv-Sketch}'_p(y) f_{\text{unit}}(y)\,dy. \tag{17}$$

We can verify that Eq. (16) is consistent with $F_{\text{cost}}(p)$:

$$F_{\text{cost}}(\textsc{Inv-Sketch}_p) = \sum_{j=1}^{n} p_j \left(f_{\text{unit}}(1) + \int_{p_j}^1 (-f'_{\text{unit}}(y))\,dy \right)$$

$$= \sum_{j=1}^{n} p_j f_{\text{unit}}(p_j) = \sum_{j=1}^{n} f_{\text{cost}}(p_j) = F_{\text{cost}}(p).$$

The proof is very similar to that of Theorem 4.1 after replacing all occurrences of H, $x\ln(1/x)$, and $\ln(1/x)$ with F_{cost}, $f_{\text{cost}}(x)$, and $f_{\text{unit}}(x)$, but this time we declare the monovariant to be $M = F_{\text{cost}}(\text{INV-PROF}_S) + (1-r)F_{\text{cost}}(\mathcal{G}_S)$, where $r \in (0,1)$ will be chosen such that the monovariant is nonincreasing. After the greedy step, $F_{\text{cost}}(\mathcal{G}_S)$ increases by $f_{\text{cost}}(p(1))$ and $F_{\text{cost}}(\text{INV-PROF}_S)$ increases by at most

$$-f_{\text{cost}}(p(1)) + \begin{cases} (q(1)-p(1))\left(f_{\text{unit}}(q(1)-p(1)) - f_{\text{unit}}(p(1))\right) & q(1) < 2p(1) \\ 0 & \text{otherwise} \end{cases}.$$

So M increases by at most

$$\begin{aligned} &f_{\text{cost}}(p(1)) + \max_{p(1)<q(1)<2p(1)} (q(1)-p(1))(f_{\text{unit}}(q(1)-p(1)) - f_{\text{unit}}(p)) + (1-r)f_{\text{cost}}(p(1)) \\ &= -rf_{\text{cost}}(p(1)) + \max_{0<a<p(1)} a(f_{\text{unit}}(a) - f_{\text{unit}}(p(1))). \end{aligned}$$

Now it is clear that we must choose

$$r \geq \max_{0<a<p(1)} \frac{a(f_{\text{unit}}(a) - f_{\text{unit}}(p(1)))}{f_{\text{cost}}(p(1))} = \max_{0<a<p(1)} \frac{af_{\text{unit}}(a) - af_{\text{unit}}(p(1))}{pf_{\text{unit}}(p)} = \max_{0<a<p(1)} \left[\frac{f_{\text{cost}}(a)}{f_{\text{cost}}(p)} - \frac{a}{p}\right],$$

as desired.

I.2 Generalization of Corollary I.3

Corollary I.4. *When $m = 2$ and $f_{cost}(x) = x^c$, the greedy algorithm obtains a $\left(1 - c^{1/(1-c)}(1/c-1)\right)^{-1}$-multiplicative approximation.*

Note that for $c = \frac{1}{2}$ this is consistent with Corollary I.3.

Proof. To apply Theorem I.1 we need to compute $r = \max_{0<q<p\leq 1}(q/p)^c - q/p$. Letting $t = q/p \in (0,1)$, we find

$$\begin{aligned} r &= t^c - t \\ \frac{dr}{dt} &= ct^{c-1} - 1. \end{aligned}$$

As $\frac{dr}{dt}$ is a decreasing function of t, we find that r is maximized when $t = c^{1/(1-c)}$. Then $r = t(t^{c-1}-1) = c^{1/(1-c)}(1/c-1)$. By Theorem I.1, the greedy algorithm obtains a $(1-r)^{-1}$-multiplicative approximation, done.

□

I.3 Proof of Theorem I.2

We use the same C_S and $\text{REM-MASS}_S^{\text{ADVANCED}}$ as in the proof of Theorem 4.2. Mirroring the steps of the proof starting from Eq. (11), the cost of the solution found by the greedy algorithm is bounded above by

$$\begin{aligned} &\int_0^1 f_{\text{unit}}(C_S(x))\,dx = \int_0^1 \text{REM-MASS}_S^{\text{ADVANCED}\prime}(y) f_{\text{unit}}(y)\,dy \\ &= \text{REM-MASS}_S^{\text{ADVANCED}}(y) f_{\text{unit}}(y)\Big|_{y=0}^{1} + \int_0^1 \text{REM-MASS}_S^{\text{ADVANCED}}(y)(-f'_{\text{unit}}(y))\,dy \\ &= f_{\text{cost}}(1) + \int_0^1 \text{REM-MASS}_S^{\text{ADVANCED}}(y)(-f'_{\text{unit}}(y))\,dy \qquad (18) \\ &= f_{\text{cost}}(1) + \Big(\int_0^1 \frac{\text{INV-PROF}_S(y) + \text{INV-PROF}_S(2y)}{2}(-f'_{\text{unit}}(y)) + (-f'_{\text{unit}}(y))y\int_{2y}^1 \text{INV-PROF}'_S(t)/t\,dt\,dy\Big) \end{aligned}$$

$$= \frac{1}{2}\left(f_{\text{cost}}(1) + \int_0^1 \text{INV-PROF}_S(y)(-f'_{\text{unit}}(y))\,dy\right) + \frac{1}{2}\left(f_{\text{cost}}(1) + \int_0^1 \text{INV-PROF}_S(2y)(-f'_{\text{unit}}(y))\,dy\right)$$
$$+ \int_0^1 \frac{\text{INV-PROF}'_S(t)}{t} \cdot \int_0^{t/2} -f'_{\text{unit}}(y)y\,dy\,dt. \tag{19}$$

It remains to show that each of the three summands is bounded above by a multiple of $F_{\text{cost}}(\text{INV-PROF}_S)$.

1. Using Eq. (16), twice the first summand equals $F_{\text{cost}}(\text{INV-PROF}_S)$.
2. Twice the second summand is

$$\begin{aligned} f_{\text{cost}}(1) + \int_0^1 \text{INV-PROF}_S(2y)(-f'_{\text{unit}}(y)) &= \int_0^1 \text{INV-PROF}'_S(y) f_{\text{unit}}(y/2)\,dy \\ &= 2^{1-c}\int_0^1 \text{INV-PROF}'_S(y) f_{\text{unit}}(y)\,dy \\ &= 2^{1-c} F_{\text{cost}}(\text{INV-PROF}_S), \end{aligned}$$

   where the first equality holds using integration by parts, the second by assuming $f_{\text{unit}}(y) = y^{c-1}$, and the third by Eq. (17).
3. First we can write

$$\int_0^{t/2} -f'_{\text{unit}}(y)y\,dy = \int_0^{t/2} (1-c)y^{c-1}\,dy = \frac{1-c}{c}y^c\Big|_0^{t/2} = \frac{1-c}{c2^c}f_{\text{cost}}(t).$$

   So using Eq. (17), the third summand simplifies to

$$\frac{1-c}{c2^c}\int_0^1 \text{INV-PROF}'(t) f_{\text{unit}}(t)\,dt = \frac{1-c}{c2^c}F_{\text{cost}}(\text{INV-PROF}_S).$$

Putting these three summands together gives

$$(19) = \left(\frac{1}{2} + \frac{1}{2^c} + \frac{1-c}{c2^c}\right)F_{\text{cost}}(\text{INV-PROF}_S) = \left(\frac{1}{2} + \frac{1}{c2^c}\right)F_{\text{cost}}(\text{INV-PROF}_S).$$

Remark 5. *For $f_{cost}(x) = x\log(1/x)$, we were able to show an additive rather than a multiplicative guarantee because twice the second summand was bounded above by $F_{cost}(\text{INV-PROF}_S)$ plus a constant, and the third summand was bounded above by a constant.*

Corollary I.5. *For general f_{cost}, we can use the method in the above proof to show that the greedy algorithm achieves a $(1 + a/2 + b)$ multiplicative guarantee if a and b are constants such that $\frac{f_{unit}(y/2)}{f_{unit}(y)} \le a$ and $\frac{\int_0^{t/2} -f'_{unit}(y)y\,dy}{f_{cost}(t)} \le b$ for all values of $y \in (0,1]$ and $t \in (0,1]$, respectively.*

Corollary I.6. *For any $a < 2$ such that $\frac{f_{unit}(y/2)}{f_{unit}(y)} \le a$ for all $y \in (0,1]$, the greedy algorithm achieves an $O\left(\frac{1}{1-a/2}\right)$-multiplicative guarantee.*

Proof. First, using integration by parts, $\int_0^{t/2} -f'_{\text{unit}}(y)y\,dy = -f_{\text{cost}}(t/2) + \int_0^{t/2} f_{\text{unit}}(y)\,dy$. The latter summand can be rewritten as $\sum_{j=1}^{\infty}\int_{t/2^{j+1}}^{t/2^j} f_{\text{unit}}(y)\,dy$, which in turn is bounded above by a geometric series with ratio $a/2$ and leading term $\int_{t/4}^{t/2} f_{\text{unit}}(y)\,dy \le a^2 t/4 \cdot f_{\text{unit}}(t) \le O\left(f_{\text{cost}}(t)\right)$. Thus we can take $b \le O\left(\frac{1}{1-a/2}\right)$ in Corollary I.5. □