

FPRAS for Total Variation Distance in High Dimensions

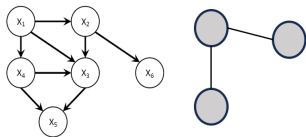
Sutanu Gayen

(Based on a Joint Work with
Arnab Bhattacharyya, Kuldeep S. Meel, Dimitrios Myrisiotis, A.
Pavan, N. V. Vinodchandran)

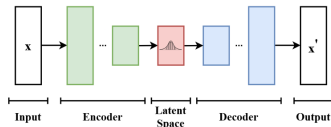
Indian Institute Technology Kanpur

July 27, 2023

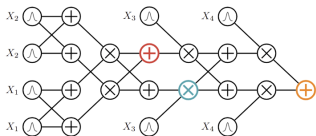
High-Dimensional Statistical Models



Probabilistic graphical models



Deep generative models (GANs, VAEs, normalizing flows, etc)



Probabilistic Circuits

Distance Estimation/Testing

Central Question

Given two models \mathcal{M}_1 and \mathcal{M}_2 decide whether their distributions are close or far with respect to a certain distance function.

- Additive testing: $\text{dist}(\mathcal{M}_1, \mathcal{M}_2) \leq \varepsilon$ or $> 2\varepsilon$.
 - ▶ Equivalent to **additive** estimation: $(\text{dist}(\mathcal{M}_1, \mathcal{M}_2) \pm \varepsilon)$
- Multiplicative testing: $\text{dist}(\mathcal{M}_1, \mathcal{M}_2) \leq \delta$ or $> (1 + \varepsilon)\delta$ for some δ such as $1/2$.
 - ▶ Equivalent to multiplicative estimation: $\text{dist}(\mathcal{M}_1, \mathcal{M}_2)(1 \pm \varepsilon)$
- Efficient algorithms: $\text{poly}(\varepsilon^{-1}, \text{size}(M_1), \text{size}(M_2))$ time, succeeds with $2/3$ probability

Distance Functions

- Several different choices: f -divergences, integral probability metrics, Wasserstein distances
- f -divergences: $\mathbb{E}_{x \sim P} \left(f \left(\frac{Q(x)}{P(x)} \right) \right)$ for some function $f(t)$ such that $f(1) = 0$.

distance	notation	$f(t)$	formula
<u>total variation</u>	$\text{TV}(P, Q)$	$\frac{1}{2} t - 1 $	$\frac{1}{2} \sum_{x \in \Omega} P(x) - Q(x) $
squared Hellinger	$H^2(P, Q)$	$\frac{1}{2}(\sqrt{t} - 1)^2$	$\frac{1}{2} \sum_{x \in \Omega} \left(\sqrt{P(x)} - \sqrt{Q(x)} \right)^2$
Kullback-Leibler	$\text{KL}(P, Q)$	$\log t$	$\sum_{x \in \Omega} P(x) \log \frac{Q(x)}{P(x)}$
chi-squared	$\chi^2(P, Q)$	$(t - 1)^2$	$\sum_{x \in \Omega} \frac{(P(x) - Q(x))^2}{P(x)}$

Total Variation Distance: Properties

$$\begin{aligned}\text{TV}(P, Q) &= \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)| \\ &= \max_{S \subseteq \Omega} [P(S) - Q(S)] \\ &= \max_{f: \Omega \rightarrow [0,1]} (\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)])\end{aligned}$$

- Only f -divergence which is also an Integral Probability Metric (satisfies triangle inequality)
- Max. difference between the probabilities of P and Q for any event
- $1 - 2 * (\text{min. error for distinguishing } P \text{ and } Q \text{ by a single sample})$
- Minimum probability that $X \neq Y$ among all couplings (X, Y) between P and Q

Prior Work

- Goldreich, Sahai, and Vadhan (1999, 2003) showed that the TV distance is hard to **additively** approximate for distributions samplable by Boolean circuits.
- Canonne and Rubinfeld (2014) showed how to **additively** approximate the TV distance for models with efficient inference and sampling.

$$\text{TV}(P, Q) = \sum_{x \in \Omega: P(x) > Q(x)} (P(x) - Q(x)) = \mathbb{E}_{x \sim P} [1_{P(x) > Q(x)} (P(x) - Q(x))]$$

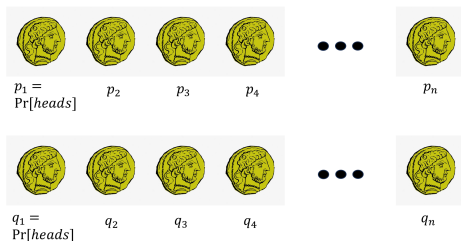
TV Distance of High-dimensional Models

$$\text{TV}(P, Q) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)|$$

- A naive computation takes $O(\Omega)$ time, intractable over $\{0, 1\}^n$
- Surprisingly, the complexity of TV distance computation between high-dimensional probabilistic models has not been studied before our work.
- Also, multiplicative approximation algorithms for TV distance has not been studied before.

Notations and Preliminaries

Binary Product Distributions



- Joint distribution of n independent coin flips, $\Omega = \{0, 1\}^n$.
 - ▶ $P = P_1 \otimes P_2 \otimes \dots \otimes P_n$.
- Inference and sampling is trivial.
 - ▶ $P[x_1, \dots, x_n] = P_1(x_1) \dots P_n(x_n)$ where $P_i = \text{Bern}(p_i)$
- Non-binary product distributions, $\Omega = [k]^n$.

Notations and Preliminaries

Approximation algorithms

- FPTAS: fully polynomial time multiplicative approximation algorithm for computing $\text{TV}(P, Q)$
- FPRAS: fully polynomial time *randomized* multiplicative approximation algorithm for computing $\text{TV}(P, Q)$

Recent Development

- Hardness of Computing the TV Distance between Product Distributions and FPTAS for Special Cases.
[Bhattacharyya, Gayen, Meel, Myrasiotis, Pavan, Vinodchandran; IJCAI 2023, arXiv:2206.07209]
- FPRAS for Computing the TV Distance between Product Distributions.
[Feng, Guo, Jerrum; SIAM SOSA 2023, TheoretCS 2023, arXiv:2208.00740]
- Hardness and FPRAS for Computing the TV Distance between Bayesian Networks.
[Bhattacharyya, Gayen, Meel, Myrasiotis, Pavan, Vinodchandran; 2023+]

- 1 Introduction
- 2 Hardness of Computing the TV Distance between Product Distributions and FPTAS for Special Cases
 - Hardness
 - FPTAS for Distance to Uniformity for Binary Product Distributions
- 3 FPRAS for Computing the TV Distance between Arbitrary Product Distributions
- 4 Computing the TV Distance between Bayesian Networks

Hardness of Computing the TV Distance between Product Distributions and FPTAS for Special Cases

Our Contribution

Hardness

It is $\#P$ -hard in general to exactly compute the TV distance between two binary product distributions P and Q .

FPTAS for special cases

We give an FPTAS for the TV distance between an arbitrary binary product distribution P and the uniform distribution $Q = U$.

Hardness

Hardness

It is #P-hard in general to exactly compute the TV distance between two binary product distributions P and Q .

#SUBSETPROD

Given positive integers a_1, \dots, a_n and T , find

$$\left| \left\{ S \subseteq [n] : \prod_{i \in S} a_i = T \right\} \right|.$$

(Known to be #P-hard)

#PMFEQUALS

Given a binary product distribution P with biases p_1, \dots, p_n and a $0 \leq v \leq 1$, find

$$|\{x \in \{0, 1\}^n : P(x) = v\}|.$$

$$\#SUBSETPROD \leq \#PMFEQUALS \leq \text{TV}$$

#SUBSETPROD \leq #PMFEQUALS

- Given an instance of #SUBSETPROD: a_1, \dots, a_n and T , define an instance of #PMFEQUALS as follows:

$$p_i = \frac{a_i}{a_i + 1} \quad \text{and} \quad v = T \cdot \prod_i (1 - p_i)$$

- Then,

$$\prod_{i \in S} a_i = T \iff P(1_S) = v$$

#PMFEQUALS \leq TV

- Given p_1, \dots, p_n and v , find $|\{x \in \{0, 1\}^n : P(x) = v\}|$
 - ▶ assume: $v < 2^{-n}$ (other case: $v \geq 2^{-n}$)
- Define distributions \hat{P} and \hat{Q} on $(n + 1)$ bits as follows:
 - ▶ $\hat{p}_i = p_i$ for $i \in [n]$, $\hat{p}_{n+1} = 1$
 - ▶ $\hat{q}_i = \frac{1}{2}$ for $i \in [n]$, $\hat{q}_{n+1} = v \cdot 2^n$ (other case: $1/(v \cdot 2^n)$)
- Define distributions P' and Q' on $(n + 2)$ bits as follows:
 - ▶ $p'_i = p_i$ for $i \in [n]$, $p'_{n+1} = 1$, $p'_{n+2} = \frac{1}{2} + \beta$
 - ▶ $q'_i = \frac{1}{2}$ for $i \in [n]$, $q'_{n+1} = v \cdot 2^n$ (other case: $1/(v \cdot 2^n)$), $q'_{n+2} = \frac{1}{2} - \beta$
 - ▶ β is small depending on the granularity of precision

Claim

$$\text{TV}(P', Q') = \text{TV}(\hat{P}, \hat{Q}) + |\{x \in \{0, 1\}^n : P(x) = v\}| \cdot 2\beta v$$

Approximation Algorithms for $\text{TV}(P, Q)$

- Zero vs non-zero testing is easy!
- Factor n -approximation is easy!
 - ▶ $\text{TV}(P_i, Q_i) \leq \text{TV}(P, Q) \leq \sum_i \text{TV}(P_i, Q_i)$

FPTAS for Distance to Uniformity for Binary Product Distributions

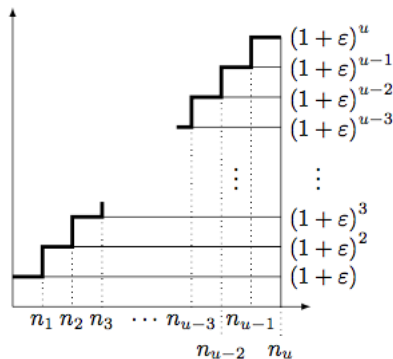
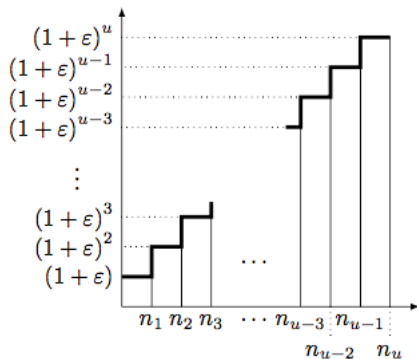
We give an FPTAS that returns $(1 \pm \varepsilon)\text{TV}(P, U)$ where U is the uniform distribution over $\{0, 1\}^n$. w.l.o.g. $\frac{1}{2} < p_i < 1$ for every i .

$$\begin{aligned}\text{TV}(P, U) &= \sum_{x \in \{0, 1\}^n} \max(0, P(x) - 1/2^n) \\ &= \sum_{S \subseteq [n]} \max\left(0, \prod_{i \in S} p_i \prod_{i \notin S} (1 - p_i) - 1/2^n\right) \\ &= \prod_{i \in [n]} (1 - p_i) \underbrace{\sum_{S \subseteq [n]} \max\left(0, \prod_{i \in S} \frac{p_i}{1 - p_i} - \prod_{i \in [n]} \frac{1}{2(1 - p_i)}\right)}_{Y_S}\end{aligned}$$

Approximating $\sum_{S \subseteq [n]} Y_S$

- $Y_S > 0$ lies in some range $[m, M]$ for each S (depending on precision)
- We create $u = \log_{1+\varepsilon} \frac{M}{m}$ levels: $[m(1+\varepsilon)^j, m(1+\varepsilon)^{j+1}]$ depending on the contribution of Y_S .
- Let n_j be the *count of sets* $S \subseteq [n]$ which contributes in the range $[m, m(1+\varepsilon)^j]$
 - ▶ $(n_{j+1} - n_j)$ sets contribute in the range $[m(1+\varepsilon)^j, m(1+\varepsilon)^{j+1}]$
 - ▶ $\sum_j (n_{j+1} - n_j) m(1+\varepsilon)^j$ is a $(1+\varepsilon)$ -factor approximation of $\sum_{S \subseteq [n]} Y_S$

Reorganization trick



$$\sum_j (n_{j+1} - n_j) m (1 + \epsilon)^j = \sum_j (n_u - n_j) ((1 + \epsilon)^{j+1} - (1 + \epsilon)^j)$$

It suffices to approximate $(n_u - n_j)!$

Approximating $(n_u - n_j)$

- $n_u = 2^n$, $n_j = \#$ sets with $Y_S \leq m(1 + \varepsilon)^j$.
- Therefore, $(n_u - n_j) = \#$ sets with $Y_S > m(1 + \varepsilon)^j > 0$

$$\left| \left\{ S \subseteq [n] : \prod_{i \in S} \frac{p_i}{1 - p_i} > \underbrace{m(1 + \varepsilon)^j + \prod_{i \in [n]} \frac{1}{2(1 - p_i)}}_A \right\} \right|$$

Reduction to #KNAPSACK

- Define weights $w_i = \log \frac{p_i}{1-p_i} > 0$ for every $i \in [n]$.
- $\left\{ S : \prod_{i \in S} \frac{p_i}{1-p_i} > A \right\} = \left\{ S : \sum_{i \in S} w_i > \log A \right\}$
- $|\{S \subseteq [n] : \sum_{i \in S} w_i > \log A\}| = \underbrace{\left| \left\{ T \subseteq [n] : \sum_{j \in T} w_j \leq B \right\} \right|}_{\text{\#KNAPSACK}}$
 - ▶ $T = [n] \setminus S$, $B = \sum_{i \in [n]} w_i - \log A$

[Gopalan, Klivans, Meka] [Stefanovic, Vempala, Vigoda]

Use existing FPTAS for #KNAPSACK.

Extensions for Binary Product Distributions

- FPTAS when Q has constantly many different biases
- FPRAS when $p_i \geq \frac{1}{2}, q_i \leq p_i$ for every i

FPRAS for Computing the TV Distance between Arbitrary Product Distributions

Gives an FPRAS for computing the TV distance between any two product distributions P and Q .
(could be non-binary but we focus on binary for simplicity).

Coupling interpretation of $\text{TV}(P, Q)$

A coupling between P and Q is a joint distribution (X, Y) such that $X \sim P$ and $Y \sim Q$.

$$\text{TV}(P, Q) = \min_{\text{couplings } (X, Y) \text{ between } P, Q} \Pr[X \neq Y]$$

Optimal Coupling O Given $TV(P, Q)$

- $TV(P, Q) = 1 - \sum_{w \in \Omega} \min(P(w), Q(w)) = \Pr_O(X \neq Y)$
- Make sure, $\Pr_O[X = Y] = \sum_{w \in \Omega} \min(P(w), Q(w))$
 - ▶ define $\Pr_O[X = Y = w] = \min(P(w), Q(w))$

Optimal Coupling O

$$\begin{aligned} \Pr[X = x, Y = y] &= \min(P(w), Q(w)) && \text{(if } X = Y = w) \\ &= 0 && \text{(if } P(x) < Q(x) \text{ or } Q(y) < P(y)) \\ &= \frac{(P(x) - Q(x))(Q(y) - P(y))}{TV(P, Q)} && \text{(otherwise)} \end{aligned}$$

Optimal Coupling for Product Distributions

- Computing $\sum_{w \in \Omega} \min(P(w), Q(w))$ is hard for product distributions
- Is the coordinate wise optimal coupling C also optimal overall?
 - ▶ Let (X_i, Y_i) be the optimal coupling for (P_i, Q_i) . Is the coupling $C = ((X_1, \dots, X_n), (Y_1, \dots, Y_n))$ optimal?
 - ▶ If so, computing $\Pr_C(X = Y)$ is easy! Since $(X_i, Y_i) \perp (X_j, Y_j)$.
- No! Let us see an example.

Globally Optimal (O) \neq Coordinate-wise Optimal (C)

$$(X_1, X_2) \sim P = \text{Bern}\left(\frac{1}{2} + \delta\right) \otimes \text{Bern}\left(\frac{1}{2} - \delta\right)$$

$$(Y_1, Y_2) \sim Q = \text{Bern}\left(\frac{1}{2}\right) \otimes \text{Bern}\left(\frac{1}{2}\right)$$

$$\begin{array}{ll} C_1 : \Pr[X_1 = 0, Y_1 = 0] = \frac{1}{2} - \delta & C_2 : \Pr[X_2 = 0, Y_2 = 0] = \frac{1}{2} \\ \Pr[X_1 = 1, Y_1 = 1] = \frac{1}{2} & \Pr[X_2 = 1, Y_2 = 1] = \frac{1}{2} - \delta \\ \Pr[X_1 = 0, Y_1 = 1] = 0 & \Pr[X_2 = 0, Y_2 = 1] = \delta \\ \Pr[X_1 = 1, Y_1 = 0] = \delta & \Pr[X_2 = 1, Y_2 = 0] = 0 \end{array}$$

Coordinate-wise Optimal Coupling ($C = C_1 \otimes C_2$)

X_1	X_2	Y_1	Y_2	Pr
0	0	0	0	$\frac{1}{2}(\frac{1}{2} - \delta)$
0	0	0	1	$\delta(\frac{1}{2} - \delta)$
0	0	1	0	0
0	0	1	1	$(\frac{1}{2} - \delta)^2$
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0

X_1	X_2	Y_1	Y_2	Pr
1	0	0	0	$\frac{\delta}{2}$
1	0	0	1	δ^2
1	0	1	0	0
1	0	1	1	$\delta(\frac{1}{2} - \delta)$
1	1	0	0	$\frac{1}{4}$
1	1	0	1	$\frac{\delta}{2}$
1	1	1	0	0
1	1	1	1	$\frac{1}{2}(\frac{1}{2} - \delta)$

Overall Optimal Copling (O)

X_1	X_2	Y_1	Y_2	Pr
0	0	0	0	$\frac{1}{4} - \delta^2$
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	$(\frac{1}{2} - \delta)^2$
0	1	1	0	0
0	1	1	1	0

X_1	X_2	Y_1	Y_2	Pr
1	0	0	0	δ^2
1	0	0	1	$\delta(1 - \delta)$
1	0	1	0	$\frac{1}{4}$
1	0	1	1	δ^2
1	1	0	0	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	$\frac{1}{4} - \delta^2$

- We will multiplicatively approximate the ratio: $\frac{\Pr_O[X \neq Y]}{\Pr_C[X \neq Y]}$.
- $\Pr_C[X \neq Y]$ can be exactly computed.

$$\begin{aligned}
 \Pr_C[X \neq Y] &= 1 - \Pr_C[X = Y] \\
 &= 1 - \prod_{i \in [n]} \Pr_{C_i}[X_i = Y_i] \\
 &= 1 - \prod_{i \in [n]} (1 - \text{TV}(P_i, Q_i))
 \end{aligned}$$

Estimator

- Let Π be the distribution of $X \sim C \mid X \neq Y$. i.e.
 $\Pi(w) = \Pr_C [X = w \mid X \neq Y]$
- Let

$$f(w) := \frac{\Pr_O [X \neq Y \wedge X = w]}{\Pr_C [X \neq Y \wedge X = w]}$$

Claim

$$\mathbb{E}_{w \sim \Pi} [f(w)] = \frac{\Pr_O [X \neq Y]}{\Pr_C [X \neq Y]}$$

Proof of Claim

$$\begin{aligned}\mathbb{E}_{w \sim \Pi} [f(w)] &= \sum_{w \in \Omega} \Pr_C [X = w \mid X \neq Y] \cdot \frac{\Pr_O [X \neq Y \wedge X = w]}{\Pr_C [X \neq Y \wedge X = w]} \\ &= \sum_{w \in \Omega} \frac{\Pr_C [X = w \wedge X \neq Y]}{\Pr_C [X \neq Y]} \cdot \frac{\Pr_O [X \neq Y \wedge X = w]}{\Pr_C [X \neq Y \wedge X = w]} \\ &= \sum_{w \in \Omega} \frac{\Pr_O [X \neq Y \wedge X = w]}{\Pr_C [X \neq Y]} \\ &= \frac{\Pr_O [X \neq Y]}{\Pr_C [X \neq Y]}\end{aligned}$$

Properties of the Estimator

- Sampling $w \sim \Pi$ is efficient
- Computing $f(w)$ is efficient

$$0 \leq f(w) \leq 1$$

$$\frac{1}{n} \leq \mathbb{E}_{w \sim \Pi} [f(w)] \leq 1$$

$$\Pi(w) = \Pr_C [X = w \mid X \neq Y]$$

$$f(w) := \frac{\Pr_O [X \neq Y \wedge X = w]}{\Pr_C [X \neq Y \wedge X = w]}$$

$$\mathbb{E}_{w \sim \Pi} [f(w)] = \frac{\Pr_O [X \neq Y]}{\Pr_C [X \neq Y]}$$

FPRAS using monte-carlo sampling.

Computing the TV Distance between Bayesian Networks

Bayesian Networks

- A joint distribution over X_1, \dots, X_n that is a product of conditional probabilities (as opposed to marginal probabilities as in a product distribution)
- Defined with respect to a DAG G over $[n]$
- Notations:
 - ▶ $\Pi(i)$ = parents of node i
 - ▶ $nde(i)$ = non-descendants of node i
 - ▶ $X_S = \{X_i\}_{i \in S}$
 - ▶ $\max_i |\Pi(i)|$ is called the in-degree of the Bayes net

Bayesian Networks Factorization

$$\Pr[X_1, \dots, X_n] = \prod_{i \in [n]} \Pr[X_i \mid X_{\Pi(i)}]$$

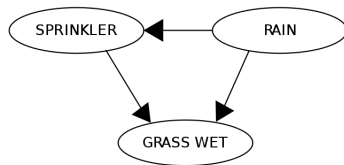
- Any node X_i is independent of its non-descendants conditioned on its parents

$$X_i \perp X_{nde(i)} \mid X_{\Pi(i)}$$

$$\implies \Pr[X_i, X_{nde(i)} \mid X_{\Pi(i)}] = \Pr[X_i \mid X_{\Pi(i)}] \Pr[X_{nde(i)} \mid X_{\Pi(i)}]$$

Example

RAIN	SPRINKLER	
	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	T	F
	0.2	0.8

SPRINKLER	RAIN	GRASS WET	
		T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

$$\Pr[R, S, G] = \Pr[R] \Pr[S | R] \Pr[G | S, R]$$

TV Approximation for Bayes Nets

Let $P[X_1, \dots, X_n]$ and $Q[Y_1, \dots, Y_n]$ be two Bayes nets over the same DAG G over $[n]$. Return:

$$d \in (1 \pm \varepsilon)\text{TV}(P, Q)$$

Our Results

- Deciding $\text{TV}(P, Q) = 0$ or not is NP-hard for Bayes nets of indegree 2
- We give an FPRAS for $\text{TV}(P, Q)$ for Bayes nets of indegree 1 (tree distributions)
- More generally, FPRAS whenever inference is feasible (computing $\Pr[X_i = 1]$ is feasible e.g. fixed treewidth)

Proof Sketch: Hardness

- Given a sat formula, create two Bayes nets P and $Q = U$ such that $\text{TV}(P, Q)$ counts the number of satisfying assignments
- The Bayes net mimicks the formula computation. Inputs are n random bits.

A Coarse Multiplicative Approximation for Trees

$$\begin{aligned} & \text{TV}(P, Q) \\ & \leq \sum_{i \in [n]} \sum_{a \in \{0, 1\}^{|\Pi(i)|}} \Pr_P[X_{\Pi(i)} = a] \text{TV}(P[X_i | X_{\Pi(i)} = a], Q[Y_i | Y_{\Pi(i)} = a]) \\ & \leq 2 \sum_{i \in [n]} \text{TV}(P[X_i, X_{\Pi(i)}], Q[Y_i, Y_{\Pi(i)}]) \end{aligned}$$

$$\begin{aligned} & \text{TV}(P, Q) \\ & \geq \text{TV}(P[X_i, X_{\Pi(i)}], Q[Y_i, Y_{\Pi(i)}]) \end{aligned}$$

Therefore, $\max_i \text{TV}(P[X_i, X_{\Pi(i)}], Q[Y_i, Y_{\Pi(i)}])$ is a $2n$ -factor approximation.

Proof Sketch: FPRAS

- We give an estimator for $\frac{\Pr_{\mathcal{O}}[X \neq Y]}{\Pr_C[X \neq Y]}$. Infer: $\Pr_C[X \neq Y]$
- Except now, C is not the product coupling
 - ▶ If we couple each factor individually, it need not be a valid coupling overall!
- C is a *partial coupling*, a joint distribution over (X, Y) :
 - ▶ corresponding factors are still coupled:

$$\begin{aligned} & \Pr[X_i = Y_i = w | X_{\Pi(i)} = a, Y_{\Pi(i)} = b] \\ & = \min\{\Pr[X_i = w | X_{\Pi(i)} = a], \Pr[Y_i = w | Y_{\Pi(i)} = b]\} \end{aligned}$$

- ▶ Only $X \sim P$.

The 4 Required properties of [Feng, Guo, Jerrum] still goes through!



Thank you!

Questions?