Mining Customer Valuations to Optimize Product Bundling Strategy

Li Ye*, Hong Xie*, Weijie Wu†, John C.S. Lui*
*The Chinese University of Hong Kong †Huawei Technologies Co., Ltd
Email: *{yli,hxie,cslui}@cse.cuhk.edu.hk †wuwjpku@gmail.com

Abstract—**Product bundling is widely adopted for information goods and online services because it can increase profit for companies. For example, cable companies often bundle Internet access and video streaming services together. However, it is challenging to obtain an optimal bundling strategy, not only because it is computationally expensive, but also that customers' private information (e.g., valuations for products) is needed for the decision, and we need to infer it from accessible datasets. As customers' purchasing data are getting richer due to the popularity of online shopping, doors are open for us to infer this information. This paper aims to address: *(1) How to infer customers' valuations from the purchasing data? (2) How to determine the optimal product bundle to maximize the profit?* We first formulate a profit maximization framework to select the optimal bundle set. We show that finding the optimal bundle set is NP-hard. We then identify key factors that impact the profitability of product bundling. These findings give us insights to develop a computationally efficient algorithm to approximate the optimal product bundle with a provable performance guarantee. To obtain the input of the bundling algorithm, we infer the distribution of customers' valuations from their purchasing data, based on which we run our bundling algorithm and conduct experiments on an Amazon co-purchasing dataset. We extensively evaluate the accuracy of our inference and the bundling algorithm. Our results reveal conditions under which bundling is highly profitable and provide insights to guide the deployment of product bundling.**

I. Introduction

How to increase revenue is an everlasting question in business. Developing new products requires significant efforts, so companies seek to improve revenue by sales strategies. One promising sales strategy is *products bundling* (a.k.a. bundling sales). By using a bundling sale, a company groups a set of products and set a single price for the whole group. It has been widely adopted, especially for information goods and online services. For example, Netflix offers all videos at a single price. Amazon provides "Amazon Prime"[1] where users pay a single price to enjoy services including free shipping, access to thousands of movies, etc. AT&T bundles phone calls, Internet access and TV services[2].

The profitability of bundling sales stems from the reduction of variance of customers' valuations of the products [1], [2], [3], [4]. The valuation refers to the highest price that a customer is willing to pay. It is also called the customer's reservation price of this product. To illustrate, consider a company selling three products (A, B and C) to two customers, and the customers' valuations are depicted in Table 1. Assume that a customer's valuation for a bundle is the sum of valuations of all the products bundled. For example, customer 1's valuation for bundle (A,B) is the sum of $5 (product A) and $15 (product B), which is $20. If products are sold separately, product A could be priced at $5 to attract both customers, resulting a maximum revenue of $10; or it could be priced at $10 to attract only customer 2, also resulting a maximum revenue of $10. Similarly, the maximum revenue of selling product B is $20. The total maximum revenue for products A and B is $30 ($10+$20). In contrast, if products A and B are bundled and priced at $20, both customers will buy the bundle and the total revenue from products A and B turns out to be $40, which is higher than the revenue of separate sales. This example shows that bundling *reduces the variance of customers' valuations of products*, and thus increases the revenue.

	Product A	Product B	Product C	Bundle (A,B)	Bundle (B,C)
Customer 1	$5	$15	$15	$20	$30
Customer 2	$10	$10	$5	$20	$15
Optimal Price	$5 (or $10)	$10	$15	$20	$15 (or $30)
Max Revenue	$10	$20	$15	$40	$30

Table 1: An example of bundle sale and separate sale.

Bundling sales may lead to a revenue drop if the products are not carefully selected. As shown in Table 1, the maximum revenue from the bundle of products B and C is only $30. Together with revenue $10 from the separate sale of product A, the total maximum revenue for "bundling (B,C), and selling A separately" is $40, which is less than selling all products separately. Table 2 depicts that different bundling strategies result in significantly different revenues, and some of them could be lower than separate sales.

Sales strategy	**Sell A, B and C separately**	**Bundle (A, B), sell C separately**	**Bundle (B, C), sell A separately**
Max Total Revenue	$45	$55	$40

Table 2: Revenue of different sales strategies

Motivated by this, we aim to address the question *how to select the optimal product bundle?* It is non-trivial and there are two key challenges. The first one is how to infer customers' valuations. In previous examples, we assume the valuations are known so that we can compare the profitability of product bundles. However, in practice, they are customers' private information and unknown to the company. The only information companies can directly obtain is customers' purchasing data, indicating whether or not a customer has bought a product. Thus, we first need to develop method to infer customers'

[1] www.amazon.com/prime

[2] www.attinternetservice.com/Bundles

valuations from such purchasing records. The second yet more difficult challenge is that selecting the optimal product bundle can be computationally expensive. This difficulty comes from the combinatorial nature in selecting products to bundle, which we will show is NP-hard with respect to the increase in the number of products. Thus, when companies face a large number products, which is often the case for information goods and online services, we need to develop efficient approximation algorithms to find bundles that are close to optimal. This paper addresses these two challenges. Our contributions are:

- We develop a probabilistic model to characterize customers' purchasing behaviors under bundling and separate sales, and formulate a profit maximization framework to select the optimal product bundle set.
- We show that finding the optimal bundle set is NP-hard. We develop a computationally efficient algorithm to approximate the optimal product bundle with a provable performance guarantee under some mild assumptions.
- We design methodology to infer the distribution of customer's valuation from purchasing data.
- We conduct experiments on an Amazon co-purchasing dataset. We show that our inference algorithm and our bundling algorithm are accurate. We also reveal conditions under which bundling is highly profitable, and provide insights to guide the deployment of bundling.

This paper organizes as follows. Section II presents the model and problem formulation. In Section III we analyze the impact of various factors and reveal insights to design efficient bundling algorithms. Section IV presents algorithms to approximate the optimal product bundle. Section V presents our method to infer model parameters. In Section VI we conduct experiments on Amazon co-purchasing data. Section VII discusses related work and Section VIII concludes.

II. Mathematical Model

We formulate a mathematical model to characterize the online market. In particular, we model how customers and company make their purchase/sales decisions. For ease of presentation, we focus on one seller who sells a set of $[N] \triangleq \{1, \ldots, N\}$ products. In fact, for online markets like Amazon, there is usually a dominating seller who sells most products. This seller has two sales strategies, which are defined as follows.

Definition 1 (Separate sale)**.** *A separate sale is a strategy to individually sell each product i at price $p_i \in \mathbb{R}_+$. Customers could choose to purchase or not each product $i \in [N]$.*

Definition 2 (Bundling sale)**.** *A bundling sale is a strategy to offer a set of products as a whole at a single price $p_b \in \mathbb{R}_+$. Customers can purchase either all products in the bundle as a whole, or none of them.*

A. Model on Separate Sales

Each customer has a valuation towards a particular product, or the maximal willingness-to-pay price. We consider a continuous spectrum of customers, where the valuation of the whole customer population to product i follows a continuous probability distribution $\mathcal{D}_i$ over $\mathbb{R}$, i.e., $V_i \sim \mathcal{D}_i$, where the random variable $V_i \in \mathbb{R}$ denotes the valuation. Once a customer buys a product i, we define her utility (a.k.a. surplus) as the difference between her valuation of the product and the price she pays to the seller:

$$U_i \triangleq V_i - p_i. \tag{1}$$

Usually, a customer has a potential to buy a product i, if her utility is non-negative. Let δ_i be the fraction of potential buyers of product i, or

$$\delta_i \triangleq \mathbb{P}(U_i \geq 0) = \mathbb{P}(V_i \geq p_i). \tag{2}$$

However, in real world, not all potential buyers will eventually purchase the product. A potential buyer decides not to buy a product for many reasons, e.g., she was not informed of the product, or she has a limited budget. We define a mapping function to capture customers' collective purchasing behaviors.

Definition 3. *Let $f(\delta_i) : [0,1] \mapsto [0,1]$ denote a mapping function, which prescribes the fraction of actual buyers given the fraction of potential buyers of product i. It increases in δ_i.*

Obviously, we have $f(\delta_i) \leq \delta_i$. Given a fraction $f(\delta_i)$ of customers who buy product i, we define the normalized profit of the seller earned from selling product i as

$$P_i(p_i) \triangleq (p_i - m_i) f(\delta_i),$$

where $m_i \in \mathbb{R}_+$ denotes the marginal cost of product i. Not limited to the revenue discussed in the example of Section I, our model also captures the profit. Considering a normalized profit does not lose any generality since the total number of customers is fixed. We impose the following assumption since the profit reduces to zero when the price is sufficiently high:

Assumption 1. *The mapping function f is well-formed such that $\lim_{p_i \to \infty} P_i(p_i) = 0$.*

Let the total profit of separate sales be $P_s(\mathbf{p})$, where $\mathbf{p} \triangleq (p_1, \ldots, p_N)^T$ denotes the price vector. It is the summation of the profit for each product, i.e., $P_s(\mathbf{p}) = \sum_i P_i(p_i)$. Given Assumption 1, there is at least one price vector $\mathbf{p}_i^*$ to attain the maximum profit $P_s^* \triangleq \sup_{\mathbf{p} \in \mathbb{R}_+^N} P_s(\mathbf{p}) = \sum_{i \in [N]} P_i^*$, where

$$P_i^* \triangleq \sup_{p_i \in \mathbb{R}+} P_i(p_i)$$

defines the maximum profit for product i.

B. Model on Bundling Sales

Now let us model bundling sales. Formally, the seller chooses a subset $\mathcal{B} \subseteq [N]$ of products to form a bundle. According to the definition of bundling sale, the bundle $\mathcal{B}$ is priced as an indivisible unit. In other words, it is regarded as a "single product" and we index it by b. A customer's valuation of the bundle $\mathcal{B}$ is defined as the summation of her valuation for each individual product within the bundle, i.e., $V_b \triangleq \sum_{i \in \mathcal{B}} V_i$. By extending Equation (1), we define the utility for a customer buying the bundle $\mathcal{B}$ as: $U_b \triangleq$

$V_b - p_b$. Also, by extending Equation (2), we can express the fraction of potential buyers of the bundle, denoted by δ_b, as $\delta_b = \mathbb{P}\left(\sum_{i \in \mathcal{B}} V_i \geq p_b\right)$. Since bundling does not affect the mapping function f, a fraction $f(\delta_b)$ of customers will eventually buy the bundle $\mathcal{B}$. The normalized profit earned from the bundle $\mathcal{B}$ is then

$$P_b(p_b) \triangleq (p_b - m_b) f(\delta_b),$$

where $m_b \triangleq \sum_{i \in \mathcal{B}} m_i$ denotes the marginal cost for the bundle $\mathcal{B}$. Given a bundle $\mathcal{B}$, the seller can set an appropriate price to maximize the profit. If Assumption 1 holds, there exists at least one optimal bundle price p_b^* to attain the maximum profit for the bundle, which is defined as

$$P_b^* \triangleq \sup_{p_b \in \mathbb{R}_+} P_b(p_b).$$

C. Instantiation on the Distribution of Valuations

In order to maximize the profit, a company needs to understand the *whole population* of customers instead of an individual one. Thus, it is important to capture the distribution of customers' valuations. In particular, let us consider customers' valuations $\mathbf{V} \triangleq (V_1, \ldots, V_N)^T$ follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with the mean vector $\boldsymbol{\mu} \triangleq \mathbb{E}[\mathbf{V}] \in \mathbb{R}^N$ and the covariance matrix $\boldsymbol{\Sigma} \triangleq \text{cov}(\mathbf{V}, \mathbf{V}) \in \mathbb{R}^{N \times N}$. It was shown in [5] that the Gaussian distribution family is a natural way to characterize a population of customers' valuations. It has been commonly used in previous works on product bundling [5], [6], [7]. Furthermore, the multivariate Gaussian distribution inherently captures the correlation among products by the covariance matrix.

As a simple consequence of the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $\mathbf{V}$, the distribution of the valuation for product i is $V_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, where $\sigma_i^2 \triangleq \text{Var}[V_i]$, and the distribution of the valuation for the bundle $\mathcal{B}$ is $V_b \sim \mathcal{N}(\mu_b, \sigma_b^2)$, where $\mu_b \triangleq \sum_{i \in \mathcal{B}} \mu_i$ and $\sigma_b^2 \triangleq \sum_{i \in \mathcal{B}} \sum_{j \in \mathcal{B}} \rho_{ij} \sigma_i \sigma_j$. Here, ρ_{ij} denotes the Pearson's correlation coefficient between random variable V_i and V_j. The fraction of potential buyers for product i and that for the bundle $\mathcal{B}$ are:

$$\delta_i = 1 - \Phi\left[(p_i - \mu_i)/\sigma_i\right], \quad \delta_b = 1 - \Phi\left[(p_b - \mu_b)/\sigma_b\right], \quad (3)$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard Gaussian distribution. Based on these closed-form expressions, we have the following property for separate sales.

Lemma 1. *Consider product i with $\mu_i > m_i$ and the maximal profit P_i^*. Suppose there is another product $\tilde{i}$ associated with parameters $\tilde{m}_i = 0$, $\tilde{\mu}_i \triangleq 1$ and $\tilde{\sigma}_i \triangleq \sigma_i/(\mu_i - m_i)$. Then the maximal profit of the product $\tilde{i}$ is $\tilde{P}_i^* = P_i^*/(\mu_i - m_i)$.*

All proofs to lemmas and theorems are in the Appendix. Lemma 1 states the scaling property of a product with zero marginal cost. This implies that the optimal profit for product i can be expressed as:

$$P_i^* = (\mu_i - m_i) \times \pi^*\left(\sigma_i/(\mu_i - m_i)\right), \quad (4)$$

where $\pi^*(\sigma) \triangleq \max_{\delta \in (0,1)} \left(\Phi^{-1}(1-\delta)\sigma + 1\right) f(\delta)$ is the maximum profit for the product with cost 0, mean of valuations 1 and variance of valuations σ. This product form of P_i^* will uncover important insights on the bundling strategy.

D. The Seller's Decision Model

If a seller decides to bundle products in set $\mathcal{B}$ and leave other ones for separate sales, the total profit consists of: (1) the profit from the bundle $\mathcal{B}$; and (2) the profit from separate sales of other products. Formally, the total profit is:

$$P(\mathcal{B}) \triangleq P_b^* + \sum_{i \in [N] \setminus \mathcal{B}} P_i^*.$$

Note that we use $P(\mathcal{B})$ to denote the total profit of all products (given that those in $\mathcal{B}$ are bundled), but not the profit from the bundle only. Let us define the optimal bundling problem to maximize the total profit.

Problem 1. *Profit maximization for bundling where $k \geq 2$:*

$$\begin{aligned} &\underset{\mathcal{B}}{\textit{maximize}} && P(\mathcal{B}), \\ &\textit{subject to} && |\,\mathcal{B}\,| = k. \end{aligned}$$

One may note that the size of the bundle is fixed under this formulation. On the one hand, this formulation can lead to an exact bundle size which a company may have a business decision a priori. This can be due to various considerations like budget, business scale, etc. For example, though updated periodically, Amazon Prime always bundles several thousand movies. On the other hand, if a company has the flexibility to vary the bundle size, one can simply add dummy services (with zero mean and variance of valuations, and zero marginal costs), so that the optimal solution is allowed to include multiple dummy products in the bundle. This is in fact a simple way to extend the condition $|\,\mathcal{B}\,| = k$ to be $|\,\mathcal{B}\,| \leq k$.

E. NP-hardness

It is worthwhile to note that it is NP-hard to compute the exact solution for Problem 1. To reach this conclusion, our approach is to show that under some mild assumption on the mapping function $f(\cdot)$, some special cases of Problem 1 is already NP-hard. In particular, the assumption is based on the elasticity of the mapping function, which is defined as follows.

Definition 4. *The elasticity of the mapping function $f(\delta_i)$ with respect to δ_i is defined as*

$$Ef(\delta_i) \triangleq \frac{df(\delta_i)}{f(\delta_i)} / \frac{d\delta_i}{\delta_i} = \frac{\delta_i f'(\delta_i)}{f(\delta_i)}.$$

The elasticity $Ef(\delta_i)$ characterizes the ratio of the relative change in the fraction of actual buyers (i.e., $df(\delta_i)/f(\delta_i)$) with respect to the relative change in the fraction of potential buyers (i.e., $d\delta_i/\delta_i$). Note that the elasticity is a standard concept in the economic literature. Using the concept of elasticity, we now present our main result on the NP-hardness of Problem 1.

Theorem 1. *Suppose the elasticity $Ef(\delta_i)$ is lower bounded, i.e., there exists a constant $c > 0$ such that $Ef(\delta_i) > c, \forall \delta_i \in [0, 1]$. Problem 1 is NP-hard when $k \geq 2$.*

Theorem 1 states that Problem 1 is NP-hard if the elasticity $Ef(\delta_i)$ is lower bounded. Thus, computing the exact solution to Problem 1 is computationally expensive especially when the bundle size is large. The elasticity $Ef(\delta_i)$ being lower bounded means that increasing or reducing the fraction of potential buyers can significantly increase or reduce the profit. In fact, a broad family of functions satisfy this assumption. For example, $f(\delta_i)=\delta_i^a$, $(a \geq 0)$ satisfies this condition with $c=a$.

III. Factors influencing Bundling Profitability

Given the NP-hardness of the profit maximization problem, it is important to develop accurate yet efficient approximation algorithms to determine the bundle set. To achieve this goal, we will first analyze some important properties of the problem, from which we can reveal fundamental principles to form bundles, and we will later use them to design our algorithms. In particular, we will investigate the impacts of (1) the intrinsic characteristics of an individual product to determine whether it is suitable to be bundled; (2) the interdependency of products to determine whether multiple products are suitable to be placed in a bundle simultaneously. We will unify the impacts of these two factors and show the tradeoff between them.

A. Impact of Individual Product

Note that some products are intrinsically suitable to bundle, while others are not. For example, we often see bundling electronic books/movies, but it is rare to bundle two cars. In other words, there are some intrinsic factors that determine whether a product itself is suitable to bundle, regardless of its correlation with others. Now let us reveal such factors.

Given a particular product, suppose we can duplicate it into many virtual products, each of them are independent and identical, i.e., for any virtual product i, we have $\mu_i = \mu, \sigma_i = \sigma, m_i = m (m > \mu)$, and the covariance matrix $\mathbf{\Sigma}$ is diagonal. If we bundle k of them, we have $\mu_b = k\mu, \sigma_b^2 = k\sigma^2, m_b = km$. From Equation (4), the optimal profit for the bundle is $P_b^* = k(\mu - m) \times \pi^* \left(\frac{\sigma}{\sqrt{k}(\mu-m)} \right)$. This implies that a bundle of sufficiently large number of independent and identical products can be viewed as a single product whose variance of valuations is 0, i.e., $\lim_{k\to\infty} \frac{\sigma}{\sqrt{k}(\mu-m)} = 0$. In other words, bundling can reduce the variance of customers' valuations. We thus define the *potential profit gain* of a product as the profit gain of bundling sufficiently large number of its duplicates averagely:

$$\gamma(\mu, \sigma, m) \triangleq \lim_{k\to+\infty} (P_b^* - \sum\nolimits_{i\in\mathcal{B}} P_i^*)/k$$
$$= (\mu - m) \left[\pi^*(0) - \pi^* \left(\sigma/(\mu - m) \right) \right]. \quad (5)$$

Note that the concept of *potential profit gain* is based on a particular product, and it is an important indicator to determine whether the product is suitable to be bundled with others, in particular, similar products. Equation (5) reveals that the scale of the potential profit gain of a product is determined by the mean of valuations, i.e. $\mu - m$. If the scaling factor $\mu - m$ is fixed, the potential profit gain is determined by the standard deviation of valuations, i.e. σ.

To illustrate, we let $f(\delta) = \delta, \delta \in [0, 1]$ and plot the normalized potential profit gain $\gamma(\mu, \sigma, m)/(\mu - m)$ w.r.t. the normalized standard deviation $\sigma/(\mu - m)$ in Figure 1. From Figure 1, we observe that the potential profit gains vary significantly when products have different normalized standard deviations. The potential profit gain reaches the highest value when the normalized standard deviation is small but non-zero. When the variance of valuations is large, potential profit gain of bundling even becomes negative. This is because when the variance is large, the seller wants to set a high price and only make a profit from customers with high valuations, but bundling reduces the amount of customers with high valuations that is away from average. We then have the following principle for the bundling decision.

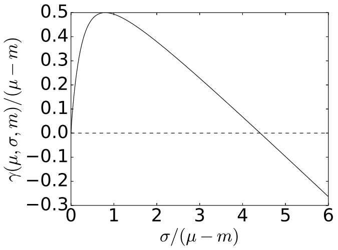

Fig. 1: Normalize potential profit gain v.s. normalized standard deviation.

Bundling Principle 1. A product is suitable to be bundled if it has a large potential profit gain, i.e., the mean of valuations is high, and the variance is relatively small but non-zero.

B. Impact of Correlation Among Products

Besides the intrinsic features of a product, the decision on whether we bundle the product also depends on what other products are in the bundle. Now we investigate the impact of correlation. We extend the setting in the last subsection (i.e., Section III-A), i.e., the covariance for any two products can be non-zero $\Sigma_{ij} \neq 0, \forall i \neq j$. We are only interested when bundling could be more profitable than separate sales; otherwise, there is no need to bundle. If this is satisfied, we have the following theorem:

Theorem 2. *Suppose the bundle $\mathcal{B}$ is more profitable than separate sales, i.e., $P(\mathcal{B}) > \sum_{i\in[N]} P_i^*$. If there exists another bundle $\widetilde{\mathcal{B}}$ with the same size and a smaller variance of valuations, i.e., $|\mathcal{B}| = |\widetilde{\mathcal{B}}|$ and $\widetilde{\sigma}_b < \sigma_b$, then $\widetilde{\mathcal{B}}$ is more profitable than $\mathcal{B}$, i.e., $P(\widetilde{\mathcal{B}}) > P(\mathcal{B})$.*

Note that we are still in the setting that $\mu_i = \mu, \forall i$. Theorem 2 shows that when the variance of valuations for the bundle is smaller, the profit of bundling is higher. This is not surprising because bundling is regarded as a way to reduce the variance of customers' valuations. Note that the variance for the bundle is $\sigma_b^2 = \sum_{i\in\mathcal{B}} \sum_{j\in\mathcal{B}} \rho_{ij}\sigma_i\sigma_j$. If the negative correlation among products in the bundle (i.e. smaller $\rho_{ij}, \forall i, j \in \mathcal{B}$) is stronger, the variance σ_b of the valuations for the bundle is lower, therefore the profit of bundling is higher. This observation motivates us to select products with stronger negative correlation to form the bundle. We summarize it in the following principle:

Bundling Principle 2. One should bundle products that are negatively correlated, so as to minimize the variance of customers' valuations of the bundle.

C. Unifying Impacts of Individual Product and Correlation

Our results thus far reveal two important factors that affect the bundling strategy. Now we aim to unify these observations for the general case. We first extend the settings in Section III-B for the general cases that products may have different costs, or different means or variances of valuations. In the following theorem we decompose the profit of a bundle.

Theorem 3. *The profit $P(\mathcal{B})$ can be decomposed as*

$$P(\mathcal{B}) = \sum\nolimits_{i \in \mathcal{B}} \gamma(\mu_i, \sigma_i, m_i) - \gamma(\mu_b, \sigma_b, m_b) + P_s^*. \quad (6)$$

Note: $LHS = \sum_{i \in [N] \setminus \mathcal{B}} P_i^* + P_b^*(\mathcal{B}) = \left[P_b^*(\mathcal{B}) - \sum_{i \in \mathcal{B}} \pi^*(0)(\mu_i - m_i) \right]$

$$+ \left[\sum_{i \in \mathcal{B}} (\pi^*(0)(\mu_i - m_i) - P_i^*) \right] + \sum_{i=1}^{N} P_i^* = RHS.$$

Theorem 3 unifies our observations in the previous two subsections. The impact of the intrinsic characteristic of products is captured by the first term $\sum_{i \in \mathcal{B}} \gamma(\mu_i, \sigma_i, m_i)$. The impact of the bundle as a unit is captured by the second term $\gamma(\mu_b, \sigma_b, m_b)$[3]. The last term P_s^* can be treated as a constant and thus ignored. In order to improve the profit gain of bundling, we should maximize the first term and minimize the second term. These two sub-objectives may not lead to the same bundling strategy: the set of products with the highest potential profit gains may not minimize the potential profit gain of the bundle. Hence one needs to balance these two sub-objectives, which is non-trivial as shown in the next section.

IV. Bundling Algorithms

We have already shown that it is computationally expensive to locate the optimal bundle set. In this section, we aim to design approximation algorithms to determine the bundle set. We begin with a homogeneous case where products have the same cost, the same mean and variance of valuations, but valuations are correlated. We design an approximation algorithm for this special case and illustrate the key idea in approximating the optimal bundle set. Later, we extend this algorithm to appropriate the optimal bundle in the general case.

A. Algorithm for Homogeneous & Correlated Products

We consider that products have the same cost m, the same mean μ (greater than the cost) of valuations, and the same variance of valuations σ^2, where the valuations are correlated. This special case simplifies the problem in that the optimal bundle set minimizes the variance of valuations of the bundle, i.e. σ_b^2 (as stated by Theorem 2). Thus, Problem 1 is equivalent to the following integer programming.

$$\underset{\mathbf{b} \in \{0,1\}^N}{\text{minimize}} \quad \mathbf{b}^T \mathbf{\Sigma} \mathbf{b}, \qquad \text{subject to} \quad \sum\nolimits_{i \in [N]} b_i = k,$$

where $b_i = 1$ if $i \in \mathcal{B}$ and $b_i = 0$ otherwise. The objective function for this integer programming is convex and quadratic, and the constraint is an affine function. This implies that this integer programming becomes the well-known quadratic programming if we relax the domain of the decision variable to a continuous set, i.e., $\mathbf{b} \in [0,1]^N$. Convex quadratic programming can be efficiently solved in polynomial time [8] by some standard convex programming algorithms (e.g., interior-point method). We utilize the randomized dependent rounding technique [9] to map the optimal solution for the relaxed problem to be integers $\{0,1\}$ added up to k with a time complexity of $O(N)$ [9]. After this rounding, we obtain an approximate bundle set. We formally describe the above idea in Algorithm 1.

Algorithm 1: Select a Size-k Bundle (Homogeneous Case)

```
function VarianceMinimization($\mathbf{\Sigma}$, $k$):
    Solve the continuous relaxation problem to obtain $\tilde{\mathbf{b}}^*$:
        $\underset{\tilde{\mathbf{b}} \in [0,1]^N}{\text{minimize}} \quad \tilde{\mathbf{b}}^T \mathbf{\Sigma} \tilde{\mathbf{b}},$
        $\text{subject to} \quad \sum\nolimits_{i \in [N]} \tilde{b}_i = k.$
    $\hat{\mathbf{b}}$ = DepRound($\tilde{\mathbf{b}}^*, k$), $\widehat{\mathcal{B}} = \{i | \hat{b}_i = 1\}$
    return $\widehat{\mathcal{B}}$
```

Algorithm 2: Randomized Dependent Rounding

```
function DepRound($\tilde{\mathbf{b}}^*$, $k$):
    $\theta$ = a random permutation of $N$ elements
    while $\exists i, \tilde{b}_i^* \in (0,1)$ do
        $i = \min\{i' | \tilde{b}^*_{\theta(i')} \in (0,1)\}$
        $j = \min\{j' | \tilde{b}^*_{\theta(j')} \in (0,1), j' \neq i\}$
        $p = \min\{1 - \tilde{b}^*_{\theta(i)}, \tilde{b}^*_{\theta(j)}\}$, $q = \min\{\tilde{b}^*_{\theta(i)}, 1 - \tilde{b}^*_{\theta(j)}\}$
        $(\tilde{b}^*_{\theta(i)}, \tilde{b}^*_{\theta(j)}) = \begin{cases} (\tilde{b}^*_{\theta(i)} + p, \tilde{b}^*_{\theta(j)} - p), \text{w.p.} \frac{q}{p+q} \\ (\tilde{b}^*_{\theta(i)} - q, \tilde{b}^*_{\theta(j)} + q), \text{w.p.} \frac{p}{p+q} \end{cases}$
    return $\tilde{b}^*$
```

To show the accuracy Algorithm 1, we provide the approximation ratio of Algorithm 1 in the following theorem.

Theorem 4. *The bundle set $\widehat{\mathcal{B}}$ computed by Algorithm 1 has the following approximation ratio on the optimal profit, provided that $\sum_{i=1}^N \min\{\tilde{b}_i^*, 1 - \tilde{b}_i^*\} > 2$ and $P(\mathcal{B}^*) > \sum_i P_i^*$:*

$$\frac{\mathbb{E}[P(\widehat{\mathcal{B}})]}{P(\mathcal{B}^*)} \geq \min_{\sigma' \geq 0, \Delta \in [0, \Delta_0]} \frac{\pi^*(\sqrt{\sigma'^2 + \Delta})}{\pi^*(\sigma')} \triangleq \xi,$$

where $\Delta_0 = \frac{\max\{\sigma, |\min \mathbf{\Sigma}|\}}{(\mu - m)^2 k} \max\{2N/(\sqrt{N} - 2) + 1, 2\sqrt{N} + N/k\}$, and $\mathcal{B}^*$ is the optimal bundle. If $k \triangleq \Omega(N^{0.5+\epsilon})$, $\epsilon > 0$, then Algorithm 1 is asymptotically accurate, i.e., $\lim_{N \to \infty} \xi = 1$.

Theorem 4 states an approximation ratio of Algorithm 1. It reveals that Algorithm 1 is asymptotically accurate when the bundle size is not too small as compared to the number of products. Table 3 presents the approximation ratio when

[3] This can be viewed as the intrinsic characteristic of this bundle, to determine whether it is suitable to be bundled with other products. For a good bundling strategy, this bundle is not supposed to be further bundled with other products, otherwise a different bundle should have been formed.

$f(\delta_i) = \delta_i$, $\max\{\sigma, |\min \mathbf{\Sigma}|\}/(\mu - m)^2 = 1$. From Table 3 we observe that the approximation ratio is high, when the number of products is large.

$\boldsymbol{N}$	$1,000$	$10,000$	$100,000$	$1,000,000$
$\boldsymbol{k}$	250	$2,500$	$25,000$	$250,000$
ξ	0.522	0.608	0.716	0.810

Table 3: Approximation ratio for different (N, k)**.**

B. Generalizations to Heterogeneous Products

Now we extend Algorithm 1 to approximate the optimal bundle set for general cases where products are correlated and may have different costs, or different means or variances of valuations. Recall that in Equation (6), we decompose the profit gain of the bundle $\mathcal{B}$ into two parts: (1) the summation of the potential profit gains of each product within the bundle, i.e., $\sum_{i \in \mathcal{B}} \gamma(\mu_i, \sigma_i, m_i)$; minus (2) the potential profit gain of the bundle, i.e., $\gamma(\mu_b, \sigma_b, m_b)$. The homogeneous case simplifies the problem in that the potential profit gain for each product is the same, i.e., $\gamma(\mu_i, \sigma_i, m_i)$ is a constant, so one only needs to minimize $\gamma(\mu_b, \sigma_b, m_b)$, which is done by minimizing the variance of bundle σ_b. However, for the general case, the potential profit gain of different products may be different. In this case, to maximize the profit of bundling, we need to maximize $\sum_{i \in \mathcal{B}} \gamma(\mu_i, \sigma_i, m_i)$ and at the same time, minimize $\gamma(\mu_b, \sigma_b, m_b)$. We need to balance these two terms as they mutually affect each other in general.

Our idea is to balance them by a two-layer selection. First, we select a subset of n ($k \leq n \leq N$) products with the highest potential profit gains. Then, from these n products selected, we apply Algorithm 1 to further select a subset of k products (denoted by $\mathcal{B}_n$) with the minimum variance of valuations. One can vary n to balance the potential profit gain $\sum_{i \in \mathcal{B}} \gamma(\mu_i, \sigma_i, m_i)$ and variance of the valuations of the bundle σ_b^2. More concretely, the case $n = k$ corresponds to maximizing the potential profit gain only (the variance σ_b^2 is not considered). In other words, we put zero weights on minimizing the variance σ_b^2. As n increases, we weigh more on minimizing the variance σ_b^2, because the variance for the bundle $\mathcal{B}_n$ decreases in n. Eventually, we hit the case $n = N$, which corresponds to minimizing the variance σ_b only (the potential profit is not considered).

To achieve a good balance, one can exhaustively try all possible values of n from k to N. However, this method is computationally expensive especially when the number of products N is large, which is often in online markets like Amzaon. To make our algorithm scalable to a large number of products, we try different configurations of n in a subset $\mathcal{C}_k \subset \{k, \ldots, N\}$ (with a step size larger than one) to obtain the optimal configuration. Intuitively, when the set $\mathcal{C}_k$ is larger, the approximation accuracy is higher, while the running time is also longer. One can vary $\mathcal{C}_k$ to attain a balance between a high approximation accuracy and a short running time. We formally describe the above idea in Algorithm 3.

Algorithm 3 is a generalization of Algorithm 1. The computational complexity of Algorithm 3 is $|\mathcal{C}_k|$ times that of Algorithm 1. We will show via experiments that it approximates the optimal product bundle with high accuracy in Section VI, although we lack a proof for the approximation ratio.

Algorithm 3: Select a Size-k Bundle (General Case)

```
function BundleHeterogeous($\mathbf{\Sigma}$, $k$):
    Sort $\{\gamma(\mu_i, \sigma_i, m_i)\}_{i=1}^N$ to get $\{\gamma(\mu_{j_i}, \sigma_{j_i}, m_{j_i})\}_{i=1}^N$
    for $n \in \mathcal{C}_k$ do
        $\mathcal{I} = \{j_1, \ldots, j_n\}$
        $\mathcal{B}_n =$ VarianceMinimization($\mathbf{\Sigma}_{\mathcal{I}}, k$)
    $\mathcal{B}^* = \arg\max_{\mathcal{B}_n, n \in \mathcal{C}_k} P(\mathcal{B}_n)$, $\widehat{\mathcal{B}} = \{j_i | i \in \mathcal{B}^*\}$
    return $\widehat{\mathcal{B}}$
```

V. Learning the Distribution of Valuations

We have designed algorithms to determine how to form a bundle. Note that we need to know the joint distribution of valuations of products, i.e., $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, as input for the algorithm. In practice, companies do not know this information, but only have customer's historical purchasing records. In this section, we present our method to infer $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ based on historical purchasing data.

A. Model on Purchasing Data

We infer the parameters of our model from publicly accessible data, which are available in online market websites like Amazon and eBay. This shows the applicability of our model and method. Formally, we use the following data model to summarize the features for parameter inference. In an online market, there are a set of $[M] \triangleq \{1, \ldots, M\}$ customers and $[N] \triangleq \{1, \ldots, N\}$ products. We have the transaction history for all customers, which is denoted by the matrix $\mathbf{A} \triangleq (A_{ij}) \in \{0, 1\}^{N \times M}$ where $A_{ij} = 1$ indicates that customer j adopts product i, or $A_{ij} = 0$ otherwise. In an online market like Amazon, prices are publicly accessible by all buyers. Thus, for each product i, its price p_i is known.

We next design algorithms to infer the joint distribution of valuation $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ from the above data. It is computationally expensive to estimate the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ using maximum likelihood methods when we only have the "0-1" choices of customers indicating whether a product is bought or not [10]. Thus, we use a two-step method to tackle this computational challenge: (1) we first infer the marginal distributions of valuations for individual products; and (2) we then infer the covariance matrix with respect to all products.

B. Inferring the Marginal Distribution of Valuations

The marginal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ of customers' valuations of product i is parameterized by the mean and variance. The mean of valuations reflects the intrinsic quality of a product, and the variance reflects the collective preferences or biases of the whole customer population. Since the customer population is fixed, one can assume that the variance of valuations is the same for all the products, i.e., $\sigma_i = \sigma$ for all $i \in [N]$. Inferring the variance σ^2 is a challenging task in general [10], as customers' valuations are subjective and serve as hidden variables that affect the product adoption decisions. We will choose typical values of variance systematically to study the profitability of bundling in Section VI. Now, let us regard σ as a fixed value and set the variance of valuations for product i as $\widehat{\sigma}_i = \sigma, \forall i \in [N]$.

The fraction of actual buyers of product i can be inferred from the adoption matrix $\mathbf{A}$ as $||\mathbf{A}_i||_1/M$, where $\mathbf{A}_i$ denotes the i-th row of the adoption matrix $\mathbf{A}$. Thus, the fraction of potential buyers of product i can be estimated as $\widehat{\delta}_i = f^{-1}\left(||\mathbf{A}_i||_1/M\right)$. By applying Equation (3), we infer the mean of valuations as $\widehat{\mu}_i = \Phi^{-1}(\widehat{\delta}_i)\widehat{\sigma}_i + p_i$. We will show how to infer $f(\cdot)$ in Section VI. Let us assume it is known.

C. Inferring the Covariance of Valuations

• **Emperial covariance matrix.** In the last subsection we have inferred the diagonal entries of the covariance matrix. Now we infer the other entries of it, namely the pairwise covariance among each pair of products $\Sigma_{ij}, \forall i \neq j$. Notice that the joint distribution of V_i and V_j is

$$(V_i, V_j) \sim \mathcal{N}\left((\widehat{\mu}_i, \widehat{\mu}_j), \begin{bmatrix} \widehat{\sigma}_i^2 & \rho_{ij}\widehat{\sigma}_i\widehat{\sigma}_j \\ \rho_{ij}\widehat{\sigma}_i\widehat{\sigma}_j & \widehat{\sigma}_j^2 \end{bmatrix}\right).$$

The fraction of potential buyers for both products i and j, which is defined as δ_{ij}, can be derived as $\delta_{ij} \triangleq 1 - F_{V_i,V_j}\left(p_i, p_j; \rho_{ij}\right)$. Then, the fraction of actual buyers that buy both products i and j is $f(\delta_{ij})$. From the adoption matrix $\mathbf{A}$, we can infer $f(\delta_{ij})$ as $[\mathbf{A}\mathbf{A}^T]_{ij}/M$, where $[\mathbf{A}\mathbf{A}^T]_{ij} = \sum_{u=1}^{M} A_{iu}A_{ju}$ is the number of customers buying both products i and j. Thus, the fraction of potential buyers of both products i and j can be estimated as $\widehat{\delta}_{ij} = f^{-1}\left([\mathbf{A}\mathbf{A}^T]_{ij}/M\right)$. The estimated correlation coefficient is $\widehat{\rho}_{ij} = F_{V_i,V_j}^{-1}\left(p_i, p_j; 1 - \widehat{\delta}_{ij}\right)$,where F_{V_i,V_j}^{-1} denotes the inverse function of F_{V_i,V_j} with respect to ρ_{ij}. It is well defined since the function F_{V_i,V_j} is monotone with respect to ρ_{ij} [11]. Finally, we estimate the empirical covariance of V_i and V_j as $\widetilde{\Sigma}_{ij} = \widehat{\sigma}_i\widehat{\sigma}_j\widehat{\rho}_{ij}$.

• **Refined estimation via symmetric matrix factorization.** The empirical covariance matrix has two drawbacks: (1) it may not be positive semidefinite; and (2) the pairwise covariance estimation is based on limited and local data, so it may not be accurate enough. To address these two issues, we apply a symmetric matrix factorization approach [12]. The basic idea is that we believe the covariance matrix $\mathbf{\Sigma}$ is of low rank. In particular, it has the following low rank decomposition: $\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X}$, where $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_N]$ with $\mathbf{X}_i \in \mathbb{R}^r$, $r \ll N$. We denote $\mathbf{X}_i \triangleq \widehat{\sigma}_i\mathbf{x}_i$ and $||\mathbf{x}_i||_2 = 1$. Our objective is to infer $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ is as close to the empirical covariance $\widetilde{\mathbf{\Sigma}}$ as possible:

$$\min_{||\mathbf{x}_i||_2=1, \forall i \in [N]} \psi(\mathbf{X}) = \sum_{i \in [N]}\sum_{j \in [N]} w_{ij}(\widehat{\sigma}_i\widehat{\sigma}_j\mathbf{x}_i^T\mathbf{x}_i - \widetilde{\Sigma}_{ij})^2, \quad (7)$$

where $w_{ij} \in \mathbb{R}^+$ is a weight representing our confidence of the empirical correlation coefficient $\widehat{\rho}_{ij}$. For those product pairs with a larger number of co-purchasing occurrences (i.e., $[\mathbf{A}\mathbf{A}^T]_{ij}$ is large), we are more confident on the empirical covariance. Hence, if the value of $[\mathbf{A}\mathbf{A}^T]_{ij}$ is larger, the weight w_{ij} is larger. We will choose specific weights based on this rule in Section VI. We apply the projected gradient method [13] to solve the optimization problem stated in Equation (7). We first derive the gradient as

$$\nabla_{\mathbf{x}_k}\psi(\mathbf{X}) = 2\sum\nolimits_{i \neq k} w_{ik}(\widehat{\sigma}_k\widehat{\sigma}_i\mathbf{x}_k^T\mathbf{x}_i - \widetilde{\Sigma}_{ik})\mathbf{x}_i.$$

Then we outline the projected gradient method in Algorithm 4. Through Algorithm 4, we obtain $\widehat{\mathbf{X}}$. The final estimation of the covariance matrix is $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{X}}^T\widehat{\mathbf{X}}$.

Algorithm 4: Projected Gradient Descent

```
Initialize $\mathbf{X}^0$.
for $\tau = 0$ to $t - 1$ do
    for $k = 1$ to $N$ do
        $\mathbf{y}_k^\tau = \mathbf{x}_k^\tau - \eta_\tau \nabla_{\mathbf{x}_k}\psi(\mathbf{X}^\tau)$   // $\eta_\tau$ is the step size
        $\mathbf{x}_k^{\tau+1} = \mathbf{y}_k^\tau/||\mathbf{y}_k^\tau||_2$ // project onto the unit sphere
    $\mathbf{X}^{\tau+1} = [\widehat{\sigma}_1\mathbf{x}_1^{\tau+1}, \ldots, \widehat{\sigma}_N\mathbf{x}_N^{\tau+1}]$
$\widehat{\mathbf{X}} = \mathbf{X}^t$
return $\widehat{\mathbf{X}}$
```

VI. Experiments

In the previous section, we have inferred parameters as inputs to our bundling algorithm. In this section, we perform experiments using our inference method and bundling algorithms on the Amazon co-purchasing data set to investigate the profitability of bundling sales. We show that under general settings, our inference results are accurate in estimating the model parameters, and our bundling algorithm approximates the optimal bundle with high accuracy. Experimental results show that bundling sale is highly profitable when the bundle size is relatively large. As we increase (or reduce) the bundle size, the optimal bundle set expands (or shrinks) incrementally. For reproducibility, we release the code and data in [14].

A. Amazon Co-purchasing Dataset

We conduct experiments on the Amazon product co-purchasing data[4]. This dataset contains product metadata and reviews for $548,552$ products, consisting of music CDs, DVDs and VHS video tapes, which are all information goods. Each product review corresponds to one purchase, and it contains the buyer's ID and the product's ID. If customer j wrote a review to product i, then $A_{ij} = 1$. The price for each product is missing for the co-purchasing dataset, and we crawled the price from *The tracktor*[5] website, which tracks the prices of products on Amazon. We select all the products in the "music" category with no less than 5 reviews and available price information. In total we select $N = 7,783$ products. These products have a total number of $515,129$ reviews assigned by $228,195$ customers, i.e., $M = 228,195$. The marginal cost m_i is not accessible in general. The marginal cost of reproducing an information good (e.g., music) is usually minimal, and thus

[4] https://snap.stanford.edu/data/amazon-meta.html

[5] https://thetracktor.com/

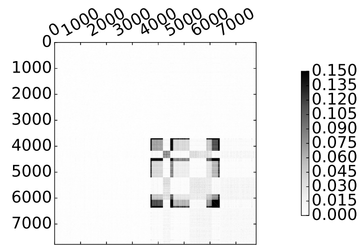

Fig. 2: Similarity matrix.

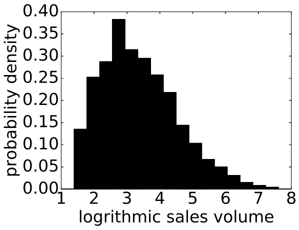

Fig. 3: Dist. of logrithmic sales volume of products.

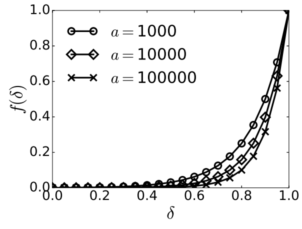

Fig. 4: Impact of a on $f(\delta_i)$.

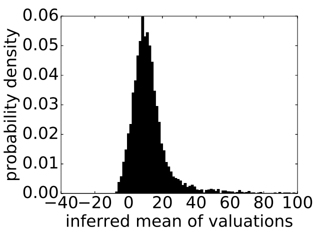

Fig. 5: Dist. of inferred mean of valuations (σ=15).

is usually considered to be zero [2]. We also set the marginal cost as zero, i.e., $m_i = 0, \forall i \in [N]$.

Let us show some statistics on customers' co-purchasing behaviors in order to gain some insights on the covariance of valuations. Intuitively, two products i and j will have similar adoption vectors (i.e., the i-th row $\mathbf{A}_i$ and j-th row $\mathbf{A}_j$ of adoption matrix $\mathbf{A}$), if customers' valuations of these two products are positively correlated. Thus, the similarity between $\mathbf{A}_i$ and $\mathbf{A}_j$ reflects the correlation of valuations. Formally, we use the cosine similarity to quantify it:

$$\cos(i,j) \triangleq \frac{\langle \mathbf{A}_i, \mathbf{A}_j \rangle}{||\mathbf{A}_i||_2 ||\mathbf{A}_j||_2}.$$

When the cosine similarity is higher, the valuations of these two products are more likely to be positively correlated. To visualize the similarity matrix, we apply the spectral clustering [15] on it and obtain 15 clusters. We plot the similarity matrix in Figure 2, where we rearrange the indices of the products so that those in a cluster have consecutive indices. From Figure 2, we observe that a small number of products have high similarity. It implies that customers' valuations are positively correlated only for a small number of products, while most of the others are negatively correlated. As we have revealed in Section III, this implies a significant potential to improve the profit of Amazon using the bundling sale strategy.

B. Inferring Model Parameters

• **Inferring the mapping function.** We infer the mapping function $f(\delta_i)$ from the sales volume distribution across products. We count the number of reviews of a product as its sales volume, which refers to the total number of times this product is sold. We plot the distribution of sales volume in Figure 3. From Figure 3 we observe that the distribution of the number of actual buyers across products follows a log-normal distribution. Naturally, the distribution of the number of potential buyers across products is a normal distribution. Thus, the mapping function $f(\delta_i)$ has an exponential form. Since $f(0) = 0$, the mapping function is in the form of $f(\delta_i) = c\frac{a^{\delta_i}-1}{a-1}$, where a and c are constants. The remaining task is to determine the parameters a and c.

In real life, some products are attractive to all customers, i.e., $\max_{i\in[N]} \delta_i = 1$. In the dataset, the maximum number of actual buyers for a product is 3,815. Since the mapping function $f(\delta)$ is increasing in δ, the fraction of potential buyers for the product with the maximum sales volume should be 1, and we have $c = f(1) = 3815/228195 = 0.0167$. Due to the variance of valuations, some products may attract a large fraction of potential buyers, while others attract a small fraction. Thus, we assume on average, the fraction of potential buyers for all products is 0.5, i.e., $\bar{\delta} \triangleq (\sum_{i=1}^{N} \delta_i)/N = 0.5$. This holds when the price equals the mean of valuations of the product. Then it follows that $\bar{\delta} \triangleq (\sum_{i=1}^{N} \delta_i)/N = \sum_{i=1}^{N} f^{-1}(||\mathbf{A}_i||_1/M) = 0.5$, which yields $a = 16674.27$. One may note that our setting on the value of $\bar{\delta}$ is quite artificial, which impacts the value of a. However, fortunately, the value of the mapping function $f(\cdot)$ is not very sensitive to the value of a, as we show in Figure 4. This implies our selection of parameters is reasonable.

• **Inferring the distribution of valuations.** Now we apply the algorithms in Section V to infer the distribution of valuations $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In Algorithm 4, we set the weight $w_{ij} = 0.1 + [\mathbf{A}\mathbf{A}^T]_{ij}$, the number of iterations $t = 100$, and the updating step size $\eta_\tau = 0.00005/\sqrt{\tau+1}$. Figure 5 shows the histogram of the inferred mean value $\widehat{\mu}_i$. It is observed that the histogram is well represented by a truncated normal distribution.

• **Measuring the accuracy of inference.** Now we show the accuracy of the inferred parameters $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$. To measure the accuracy, we adopt the recall-based approach [16]. We compare the co-purchasing of products predicted by the inferred distribution of valuation $\mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, with the co-purchasing of products observed from the dataset (i.e., the adoption matrix $\mathbf{A}$). If they coincide, then $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are accurate. Formally, the prediction accuracy is measured by [16]:

$$\overline{rank} \triangleq \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{A}\mathbf{A}^T]_{ij} rank_{j|i}}{\sum_{i=1}^{N}\sum_{j=1}^{N}[\mathbf{A}\mathbf{A}^T]_{ij}},$$

where $rank_{j|i}$ denotes the percentile rank of the co-purchasing probability $\mathbb{P}(j|i)$ among the sequence $\mathbb{P}(1|i), \ldots, \mathbb{P}(N|i)$,

$$\begin{aligned}\mathbb{P}(j|i) &\triangleq \mathbb{P}(\text{A customer buys } j \mid \text{A customer buys } i)\\ &= f\left(1 - F_{V_i,V_j}(p_i, p_j; \widehat{\Sigma}_{ij}/(\widehat{\sigma}_i\widehat{\sigma}_j))\right)/f(\widehat{\delta}_i),\end{aligned}$$

where $\widehat{\delta}_i$ and $F_{V_i,V_j}(p_i,p_j)$ is achieved based on the inferred distribution $\mathcal{N}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$. If the average rank is lower, the prediction accuracy is higher. Figure 6 plots the average rank $\overline{rank}$, where we vary the number of factors r from 5 to 80 in the matrix factorization. From Figure 6 we observe that when r increases, the average rank $\overline{rank}$ decreases, or the prediction accuracy increases. When r is larger than 20, increasing r will only reduce the average rank slightly. We want to emphasize

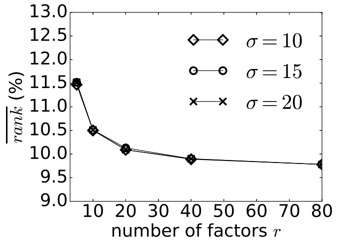

Fig. 6: Accuracy of the predictions.

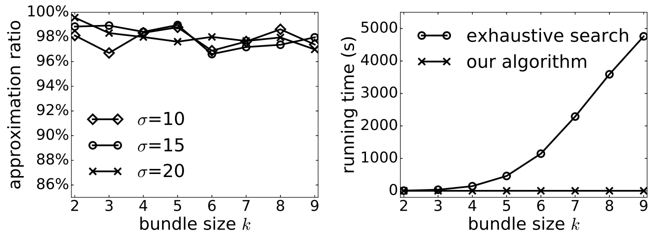

Fig. 7: Approximation ratio and running time for our bundling algorithm ($N = 20$, k varied).

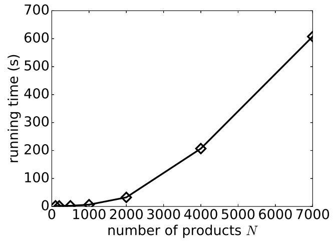

Fig. 8: Running time of our algorithm (k=$\lceil \frac{N}{4} \rceil$, N varied).

that when $r = 20$, the average rank is $\overline{rank} = 10.1\%$, indicating our prediction is accurate, or the inferred parameters $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are accurate. Thus, we set $r = 20$.

C. Evaluating the Bundling Algorithms

Now we evaluate our bundling algorithm, i.e., Algorithm 3, in terms of the running time and approximation ratio where we set $\mathcal{C}_k = \{\lfloor\{k(1+i/10)\rfloor\}_{i=0}^{20}$. The baseline algorithm for comparison is exhaustive search, which locates the optimal bundle set via enumerating all possible bundles. We run both algorithms based on our inferred parameters. We randomly select 20 products from the data, i.e., N=20, because exhaustive search is computationally inhibitive when N is large. For each bundle size from 2 to 9 we compare the profit and the running time of these two algorithms, where take the average of 24 runs. From Figure 7 we observe that the average approximation ratio is above 95%, or, our algorithm achieves at least 95% of the maximal profit and is thus highly accurate. Also, our approximation reduces the running time dramatically to 0.017% of that of exhaustive search when $k = 9$. We also plot the running time of our algorithm for different N in Figure 8, where we select 1/4 of the products to bundle. We could see the running time increases almost linearly with N. This enables us to apply our algorithm to large-scale datasets.

D. Case Studies on the Amazon Co-purchasing Dataset

We first study the profit improvement of bundling sale as compared to separate sales. Formally, we define the relative profit improvement ratio as

$$\text{PftImp} \triangleq (P(\widehat{\mathcal{B}}) - P_s^*)/P_s^*.$$

Figure 9 plots the profit improvement as the bundle size varies. From Figure 9 we observe that when the bundle size is small, the profit improvement is marginal. When it becomes large, the profit improvement is significant, i.e. 145% for $k = 2000$ and $\sigma = 15$. Figure 10 shows the intersection of bundle sets when the value of σ are different. We can see that a large number of products are always in the bundle when σ takes different values. The profit improvement curve has a single peak, or the profit improvement first increases and then decreases as we increase the bundle size. In summary, the seller should consider a bundle size that is relatively (but not extremely) large to improve the profit. If we bundle products randomly, the profit improvement is minimal or even negative, as shown in Figure 11. Thus, a carefully-designed bundling strategy is essential to the success of bundling.

Second, we investigate the evolution of the bundle set when the bundle size increases. Let us denote $\widehat{\mathcal{B}}_k$ as the bundle set selected by Algorithm 3 given a bundle size of k. Figure 12 depicts $\widehat{\mathcal{B}}_k$ as k varies. As illustrated in Figure 12, the optimal bundle set expands *incrementally* as we increase the bundle size, i.e., $\widehat{\mathcal{B}}_{1000} \subset \ldots \subset \widehat{\mathcal{B}}_{6000}$ holds roughly. This implies that when a seller wants to reduce (or increase) the bundle size, he only needs to detach (or include) a small subset of products from (into) the current bundle.

Last but not least, let us investigate which products should be selected to the bundle, in particular, what are the sales volumes of such products in the bundle when they are separately sold, which reflect their popularity to customers. Figure 13 plots the distribution of the separate-sales volume for products in the bundle. From Figure 13 we observe that most products in the bundle have a separate-sales volume close to the average , i.e. 66.2, while a small fraction of them have a particularly large sales volume. In other words, most of the products in the bundle are with average popularity, while a few others are really popular. An important indication is that a good bundle could consist of a large number of products with average popularity, and a few very popular products. Such a composition is optimal because on one hand, the bundle is attractive due to the inclusion of some star products, and on the other, having many mediocre products increases the bundle size so that the total sales volume of all products increases.

VII. Related Work

Bundling strategies have been extensively studied from economic perspectives. Adams and Yellen [17] showed that a bundling sale can be more profitable than separate sales, but it also can lead to oversupply or undersupply of some products. Schmalensee [5] extended Adams-Yellen framework [17] to consider Gaussian distributed reservation prices. They showed that pure bundling can be more profitable if the average reservation price is high enough. Hanson and Martin [18] designed algorithms to find the optimal bundle prices. Different from these works on bundling physical goods, several works [1], [2], [3] studied bundling information goods. The main difference is that the marginal cost of information goods is negligible while not for physical goods. Besides studying bundling from a theoretical perspective, Kamel *et al.* [10] and Chu *et al.* [6] collected survey data to validate theoretical assumptions of bundling strategies and further explored bundling strategies based on parameters inferred from data.

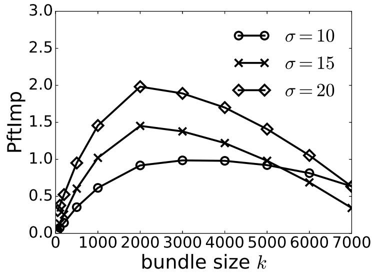

Fig. 9: Profit of bundles selected by our algorithm.

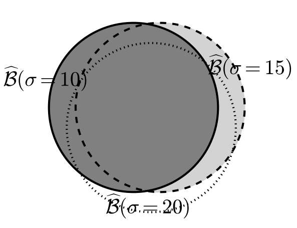
$\widehat{\mathcal{B}}(\sigma = 10)$
$\widehat{\mathcal{B}}(\sigma = 15)$
$\widehat{\mathcal{B}}(\sigma = 20)$

Fig. 10: Bundle sets of different σ ($k = 2000$).

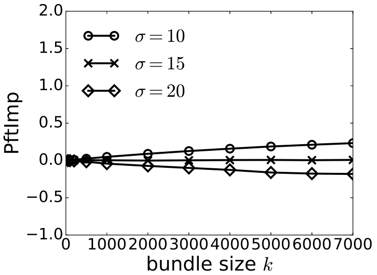

Fig. 11: Profit of bundles selected randomly (average of 50 runs).

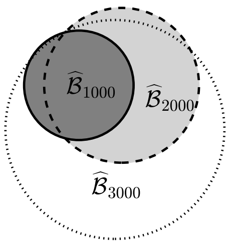
$\widehat{\mathcal{B}}_{1000}$
$\widehat{\mathcal{B}}_{2000}$
$\widehat{\mathcal{B}}_{3000}$

Fig. 12: Bundle sets of different size ($\sigma = 15$).

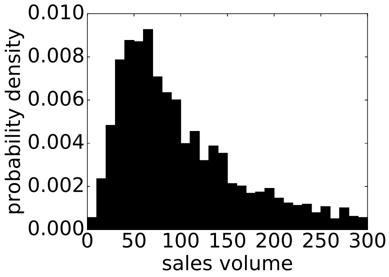

Fig. 13: Distribution of the sales volume of bundled products ($\sigma = 15$, $k = 2000$).

Although bundling has been extensively studied, most of these works considered a small number of (usually two) products. Among the few works on bundling a large number of products, Bakos and Brynjolfsson [2] studied strategies for bundling infinite number of products which have limited correlation of reservation prices. Benisch and Sandholm [7] proposed a framework for automated bundling by mining shopping cart data. They developed a novel methodology to utilize customers' purchasing data to set the optimal bundle price via heuristic algorithms, but their maximum likelihood algorithm to estimate parameters does not scale for a large number of products. Our paper differs from these works in that (1) we design an approximation algorithm to find the optimal bundling strategy for a large number of products, and (2) we propose a computationally efficient algorithm to learn customers' purchasing behaviors from accessible dataset.

VIII. Conclusion

This paper develops a statistical framework to infer customers' valuations and find the optimal product bundle from customers' purchasing data. We first formulate a profit maximization framework to select the optimal bundle set, which we show is NP-hard. We identify key factors that influence the profitability of bundling sale. They give us insights to develop a computationally efficient algorithm to approximate the optimal bundle set with provable performance guarantee. We design algorithms to infer model parameters from customers' purchasing data, and carry out experiments on an Amazon co-purchasing dataset. We show that our inference algorithm is accurate in estimating the model parameters, and that our bundling algorithms approximate the optimal bundle set with high accuracy. Our analysis and experimental results show that the bundling sale is highly profitable when the bundle size is relatively large and the valuation of customers to products are negatively correlated. As we increase (or decrease) the bundle size, the optimal bundle set expands (or shrinks) incrementally.

Acknowledgement: The research of John C.S. Lui is supported in part by the GRF 14208816.

References

[1] Y. Bakos and E. Brynjolfsson, "Bundling and competition on the internet," *Marketing Science*, vol. 19, no. 1, pp. 63–82, 2000.

[2] ——, "Bundling information goods: Pricing, profits, and efficiency," *Management Science*, vol. 45, no. 12, pp. 1613–1630, 1999.

[3] A. Prasad, R. Venkatesh, and V. Mahajan, "Optimal bundling of technological products with network externality," *Management Science*, vol. 56, no. 12, pp. 2224–2236, 2010.

[4] R. Venkatesh and W. Kamakura, "Optimal bundling and pricing under a monopoly: Contrasting complements and substitutes from independently valued products," *The Journal of Business*, vol. 76, no. 2, pp. 211–231, 2003.

[5] R. Schmalensee, "Gaussian demand and commodity bundling," *Journal of business*, pp. 211–230, 1984.

[6] C. S. Chu, P. Leslie, and A. Sorensen, "Bundle-size pricing as an approximation to mixed bundling," *American Economic Review*, vol. 101, no. 1, pp. 263–303, 2011.

[7] M. Benisch and T. Sandholm, "A framework for automated bundling and pricing using purchase data," in *International Conference on Auctions, Market Mechanisms and Their Applications*. Springer, 2011.

[8] Y. Ye and E. Tse, "An extension of karmarkar projective algorithm for convex quadratic programming," *Mathmatical Programming*, vol. 44, no. 2, pp. 157–179, 1989.

[9] J. Byrka, T. Pensyl, B. Rybicki, A. Srinivasan, and K. Trinh, "An improved approximation for k-median, and positive correlation in budgeted optimization," in *Proc. ACM-SIAM SODA*, 2015.

[10] K. Jedidi, S. Jagpal, and P. Manchanda, "Measuring heterogeneous reservation prices for product bundles," *Marketing Science*, vol. 22, no. 1, pp. 107–130, 2003.

[11] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous multivariate distributions, models and applications*. John Wiley & Sons, 2004.

[12] D. Kuang, C. Ding, and H. Park, "Symmetric nonnegative matrix factorization for graph clustering," in *Proc. of SIAM SDM*, 2012.

[13] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[14] *The link for the codes: https://github.com/lonyle/bundling.*

[15] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[16] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. of IEEE ICDM*, 2008.

[17] W. Adams and J. Yellen, "Commodity bundling and the burden of monopoly," *The Quarterly Journal of Economics*, pp. 475–498, 1976.

[18] W. Hanson and R. K. Martin, "Optimal bundle pricing," *Management Science*, vol. 36, no. 2, pp. 155–174, 1990.

[19] J. X. Panos M. Pardalos, "The maximum clique problem," 1992.

Appendix

A. Proof of Lemma 1

Proof. For a product with marCustomers valuations to product ginal cost m_i, mean of valuations μ_i and standard deviation of valuations σ_i, use the potential fraction of buyers δ_i as the decision variable rather than the price, then $p_i = \Phi^{-1}(1-\delta_i)\sigma_i + \mu_i$, and the profit maximization problem becomes

$$\max_{\delta_i \in (0,1)} \left(\Phi^{-1}(1-\delta_i)\sigma_i + \mu_i - m_i\right) f(\delta_i). \tag{8}$$

It is the same as

$$(\mu_i - m_i) \max_{\delta_i \in (0,1)} \left(\Phi^{-1}(1-\delta_i)\frac{\sigma_i}{\mu_i - m_i} + 1\right) f(\delta_i). \tag{9}$$

It is just $\mu_i - m_i$ times the optimal profit of selling the product with marginal cost 0, mean of valuations 1 and standard deviation of valuations $(\frac{\sigma_i}{\mu_i - m_i})$, i.e. $P_i^* = (\mu_i - m_i)\tilde{P}_i^*$. Thus, $\tilde{P}_i^* = P_i^*/(\mu_i - m_i)$. ∎

B. Proof of Theorem 1

Proof. In fact, we are going to see the special case of Problem 1 is already NP-hard. Let us start with the following lemma:

Lemma 2. *Suppose the elasticity $Ef(x)$ is lower bounded, i.e. there exists a constant c such that $Ef(x) \geq c$ for all $x \in (0,1)$. Suppose $\mu_i = 1$, $m_i = 0$ and $\sigma_i = \sigma$ for all $i \in [N]$, where $\sigma \in [0, c\sqrt{2/\pi})$. Then Problem 1 is equivalent to*

$$\begin{aligned} &\textit{minimize}_{\mathcal{B}} && \sigma_b^2, \\ &\textit{subject to} && |\mathcal{B}| = k. \end{aligned} \tag{10}$$

Proof. Let us first show that the maximium profit P_i^* of product i has a monotone property with respect to the standard deviation of customers' valuations σ_i.

Lemma 3. *Consider a product that has mean valuation $= 1$, variance of valuation σ, and marginal cost 0.*
(1) If $\frac{f(x)}{xf'(x)}$ is upper bounded, i.e., there exists a $c > 0$ such that $\frac{f(x)}{xf'(x)} \leq \frac{1}{c}$, profit for this product is a strictly decreasing function of σ in the region $\sigma \in [0, c\sqrt{2/\pi})$.

Proof. Let us denote $\pi(\sigma, \delta) \triangleq \left(\Phi^{-1}(1-\delta)\sigma + 1\right) f(\delta)$ as the profit of the seller when the fraction of potential buyers is δ. In fact, $\pi^*(\sigma) = \max_{\delta \in (0,1)} \pi(\sigma, \delta)$. Under Assumption 1, the optimal is attainable. Hence, for the optimal δ, we have:

$$\delta^* \in \arg\max_{\delta \in (0,1)} \pi(\sigma, \delta).$$

The first-order necessary condition for optimality is

$$\begin{aligned} \frac{\partial \pi(\sigma, \delta)}{\partial \delta} =& f'(\delta)\left(1 + \Phi^{-1}(1-\delta)\sigma\right) \\ &+ f(\delta)\left(\frac{-\sigma}{\phi(\Phi^{-1}(1-\delta))}\right) = 0, \end{aligned} \tag{11}$$

where $\phi(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$ is the p.d.f. of the standard normal distribution. Then we can derive:

$$\sigma = \frac{1}{\frac{f(\delta^*)}{f'(\delta^*)\phi(\Phi^{-1}(1-\delta^*))} - \Phi^{-1}(1-\delta^*)}. \tag{12}$$

The above equation indicates that σ is a function of the optimal fraction δ^*. Now, we bound the denominator in Equation (12) with the following lemmas.

Lemma 4. *For $x \in (0, \frac{1}{2}]$, if $\frac{f(x)}{xf'(x)} \leq \frac{1}{c}$, then*

$$\frac{f(x)}{f'(x)\phi(\Phi^{-1}(1-x))} - \Phi^{-1}(1-x) \leq \frac{1}{c}\sqrt{\pi/2}.$$

Proof. Because $\frac{f(x)}{xf'(x)} \leq \frac{1}{c}$ and $\Phi^{-1}(1-x) \geq 0$, we have

$$\begin{aligned} &\frac{f(x)}{f'(x)\phi(\Phi^{-1}(1-x))} - \Phi^{-1}(1-x) \\ \leq & \frac{1}{c}\frac{x}{\phi(\Phi^{-1}(1-x))} - 0. \end{aligned} \tag{13}$$

Now, we are going to show $\frac{x}{\phi(\Phi^{-1}(1-x))} \leq \sqrt{\pi/2}$. In fact, consider $g(t) = \frac{1-\Phi(t)}{\phi(t)}$ for $t \geq 0$. Then,

$$\begin{aligned} g'(t) =& \frac{-\phi(t)\phi(t) - (1-\Phi(t))(-t\phi(t))}{\phi^2(t)} \\ =& \frac{t\int_t^{+\infty}\phi(x)}{\phi(t)} - 1 \leq \frac{\int_t^{+\infty} x\phi(x)dx}{\phi(t)} - 1 \\ =& \frac{\int_t^{+\infty} -d\phi(x)}{\phi(t)} - 1 = \frac{-\phi(+\infty) + \phi(t)}{\phi(t)} - 1 = 0. \end{aligned} \tag{14}$$

We can see that $g(t)$ is a monotonically decreasing function for $t \in [0, +\infty)$. As a result, $g(t) \leq g(0) = \sqrt{\pi/2}$. Now, plugging in $t = \Phi^{-1}(1-x)$, we know $\frac{x}{\phi(\Phi^{-1}(1-x))} \leq \sqrt{\pi/2}$. Combining it with Equation (13), we conclude our lemma. ∎

Based on Lemma 4, we know when $\sigma < c\sqrt{2/\pi}$, $\delta^* > 1/2$. Otherwise, suppose $\delta^* \leq 1/2$, we have a contradiction that

$$\sigma = \frac{1}{\frac{f(\delta^*)}{f'(\delta^*)\phi(\Phi^{-1}(1-\delta^*))} - \Phi^{-1}(1-\delta^*)} \geq \frac{1}{\frac{1}{c}\sqrt{\pi/2}} = c\sqrt{2/\pi}.$$

Now, we want to show the monotonicity of $\pi^*(\sigma)$ in σ.

First, consider $\sigma_1 < \sigma_2 < c\sqrt{2/\pi}$. If we denote

$$\delta_2^* \triangleq \arg\max_{\delta \in (0,1)} \pi(\sigma_2, \delta),$$

then $\delta_2^* > 1/2$ and therefore $\Phi^{-1}(1-\delta_2^*) < 0$. Hence,

$$\begin{aligned} &\pi^*(\sigma_2) = \pi(\sigma_2, \delta_2^*) = f(\delta_2^*)(\Phi^{-1}(1-\delta_2^*)\sigma_2 + 1) \\ <& f(\delta_2^*)(\Phi^{-1}(1-\delta_2^*)\sigma_1 + 1) = \pi(\sigma_1, \delta_2^*) \leq \pi^*(\sigma_1) \end{aligned} \tag{15}$$

According to the definition, the $\pi^*(\sigma)$ is strictly decreasing for $\sigma \in [0, c\sqrt{2/\pi})$.

∎

Now, we come back to our Lemma 2. Recall that all the products have the same optimal profit $P_0 \triangleq \pi^*(\sigma)$ if sold separately.

For the bundle, notice that $\mu_b = k$. Then V_b follows a Gaussian distribution $\mathcal{N}(k, \sigma_b^2)$. By the scaling property of Lemma 1, the total profit

$$P(\mathcal{B}) = (N-k)P_0 + k\pi^*(\frac{\sigma_b}{k}). \tag{16}$$

We claim that $\frac{\sigma_b}{k} \leq \sigma$. In fact, since $\mathbf{\Sigma}$ is symmetric positive semidefinite, $\sigma_i\sigma_j - \Sigma_{ij}\Sigma_{ij} \geq 0$, which means $\Sigma_{ij} \leq \sigma^2, \forall i, j$. Therefore, $\sigma_b^2 = \sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{B}}\Sigma_{ij} \leq k^2\sigma^2$, which concludes our claim.

Now, if $Ef(x) \geq c$, then $\frac{f(x)}{xf'(x)} \leq \frac{1}{c}$. Therefore by Lemma 3, $\pi^*(\frac{\sigma_b}{k})$ is strictly increasing as σ_b^2 decreases. Hence, $P(\mathcal{B})$ is strictly increasing as σ_b^2 decreases. In other words, maximizing the profit is equivalent to minimizing σ_b^2 with the constraint $|\,\mathcal{B}\,| = k$, which concludes our Lemma 2. ∎

Now we show the NP-hardness of the following problem, which is also mentioned in Section IV-A.

Lemma 5. *The following problem with strict cardinality constraint is NP-hard even if we only consider the special case where the matrix $\mathbf{\Sigma} \in \mathbb{R}^{N\times N}$ is positive semidefinite:*

$$\begin{aligned} \textit{minimize} \quad & \mathbf{b}^T\mathbf{\Sigma}\mathbf{b}, \\ \textit{subject to} \quad & \mathbf{b} \in \{0,1\}^N, \\ & \sum_{i=1}^{N} b_i = k, \end{aligned} \tag{17}$$

where $k \geq 2$ is a positive integer.

Proof. From Theorem 1.9 of [19], we know the maximum weight independent set problem is equivalent to the following quadratic 0-1 problem

$$\begin{aligned} \text{minimize} \quad & \mathbf{x}^T G\mathbf{x}, \\ \text{subject to} \quad & \mathbf{x} \in \{0,1\}^N, \end{aligned} \tag{18}$$

where G is a matrix corresponding to the weights of the graph. Also, the maximum weight independent set problem is a well-known NP-hard problem. Then the following problem is NP-hard for an arbitrary matrix $C \in \mathbb{R}^{n\times n}$:

$$\begin{aligned} \text{minimize} \quad & \mathbf{x}^T C\mathbf{x}, \\ \text{subject to} \quad & \mathbf{x} \in \{0,1\}^N. \end{aligned} \tag{19}$$

Now, we want to show the problem for arbitrary matrix (19) could be reduced to problem (17) by computation within polynomial time complexity. In fact, for arbitrary matrix C, there exists a constant $\lambda > 0$, such that $C + \lambda I$ is a positive semidefinite matrix by adding to the diagonal elements. Note that when $\mathbf{x} \in \{0,1\}^N$ and $\sum_{i=1}^N x_i = k$,

$$\begin{aligned} \mathbf{x}^T(C+\lambda I)\mathbf{x} = \mathbf{x}^T C\mathbf{x} + \lambda(\mathbf{x}^T\mathbf{x}) &= \mathbf{x}^T C\mathbf{x} + \lambda(\sum_{i=1}^N x_i^2) \\ = \mathbf{x}^T C\mathbf{x} + \lambda(\sum_{i=1}^N x_i) &= \mathbf{x}^T C\mathbf{x} + \lambda k \end{aligned}$$

As a result, the following problems (20) and (21) are equivalent:

$$\begin{aligned} f(\mathbf{x}) = & \lambda k + \min \mathbf{x}^T C\mathbf{x}, \\ \text{subject to} \quad & \mathbf{x} \in \{0,1\}^N, \\ & \sum_{i=1}^N x_i = k, \end{aligned} \tag{20}$$

and

$$\begin{aligned} f'(\mathbf{x}) = & \min \mathbf{x}^T(C+\lambda I)\mathbf{x}, \\ \text{subject to} \quad & \mathbf{x} \in \{0,1\}^N, \\ & \sum_{i=1}^N \mathbf{x}_i = k. \end{aligned} \tag{21}$$

The reduction from (19) to (17) is constructed as follows. If we are able to find a polynomial-time algorithm to solve problem (17), then we can apply this algorithm to solve an instance of problem (21) where $\sum_{i=1}^N x_i^{(k)} = k$ and denote the optimal solution as $\mathbf{x}^{(k)}$. Then $\mathbf{x}^{(k)}$ is also the optimal solution for problem (20). Repeat the process for N times, we get $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$. Denote $\mathbf{x}^{(0)} = \{0, \ldots, 0\}$ as an all-zero vector, and the optimal solution of problem (19) would be $\arg\max_{\mathbf{x}\in\{\mathbf{x}^{(0)},\ldots,\mathbf{x}^{(N)}\}} f(\mathbf{x})$ because all possible values of $\sum_{i=1}^N x_i$ are in $\{0, 1, \ldots, N\}$. Hence, we are able to find an polynomial-time algorithm to solve problem (19).

Because problem (19) is NP-hard, the problem (17) should also be NP-hard, even if C is positive semidefinite. ∎

Back to Theorem 1. Notice that (10) and (17) are the same problem with different notations. In other words, the sepcial case described in Lemma 2 is equivalent to the problem in Lemma 5 which is NP-hard. Obviously, the more general Problem 1 with arbitrary multivariate Gaussian valuation distribution is NP-hard. ∎

C. Proof of Theorem 2

Before further analysis, let us reveal of convexity of the $\pi^*(\cdot)$ function.

Lemma 6. *$\pi^*(\sigma)$ is a convex function of σ.*

Proof. Note that $\pi(\sigma, \delta) = \left(\Phi^{-1}(1-\delta)\sigma + \mu\right) f(\delta)$ is an affine function with respect to σ. Then, $\forall t \in [0,1]$ we have

$$\pi\left(t\sigma_1 + (1-t)\sigma_2, \delta\right) = t\pi(\sigma_1, \delta) + (1-t)\pi(\sigma_2, \delta). \tag{22}$$

By definition, we know $\pi(\sigma_1, \delta) \leq \pi^*(\sigma_1)$ and $\pi(\sigma_1, \delta) \leq \pi^*(\sigma_1)$ for all $\delta \in (0,1)$. Hence, for all $\delta \in (0,1)$,

$$\pi\left(t\sigma_1 + (1-t)\sigma_2, \delta\right) \leq t\pi^*(\sigma_1) + (1-t)\pi^*(\sigma_2). \tag{23}$$

By our Assumption 1, the maximum profit is attainable for bounded p^* hence $\pi^*(\sigma) = \pi(\sigma, \delta^*)$ is attainable for some $\delta^* \in (0,1)$. As a result,

$$\pi^*\left(t\sigma_1 + (1-t)\sigma_2\right) = \max_{\delta\in(0,1)} \pi\left(t\sigma_1 + (1-t)\sigma_2, \delta\right) \tag{24}$$

$$\leq t\pi^*(\sigma_1) + (1-t)\pi^*(\sigma_2), \forall t \in [0,1]. \tag{25}$$

According to definition, $\pi^*(\sigma)$ is a convex function of σ. ∎

Next, before the proof of Theorem 2, we introduce the following necessary condition for a bundle to be more profitable.

Lemma 7. *If the bundle $\mathcal{B}$ is more profitable than the separate sale, i.e., $P(\mathcal{B}) > \sum_{i\in[N]} P_i^*$, and $\mu_i - m_i \geq 0$ for all $i \in \mathcal{B}$, then the fraction of potential customers for the bundle under the optimal price (denoted by δ_b^*) satisfies $\delta_b^* > \frac{1}{2}$.*

Proof. Suppose the bundle set is $\mathcal{B}$. Denote $\tilde{\sigma}_i = \frac{\sigma_i}{\mu_i - m_i}$, then the optimal profit of separate sale of all products is

$$\sum_{i\in[N]} P_i^* = \sum_{i\in\mathcal{B}} (\mu_i - m_i)\pi^*(\tilde{\sigma}_i) + \sum_{j\in[N]\setminus\mathcal{B}} (\mu_j - m_j)\pi^*(\tilde{\sigma}_j). \tag{26}$$

Also the optimal profit after bundling is

$$P(\mathcal{B}) = (\sum_{i\in\mathcal{B}} (\mu_i - m_i))\pi^*(\tilde{\sigma}_b) + \sum_{j\in[N]\setminus\mathcal{B}} (\mu_j - m_j)\pi^*(\sigma_j). \tag{27}$$

where $\tilde{\sigma}_b \triangleq \frac{\sqrt{\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{B}}\Sigma_{ij}}}{\sum_{i\in\mathcal{B}}(\mu_i - m_i)}$ is the normalized standard deviation for the bundle.
Define $\beta_i = \frac{\mu_i - m_i}{\sum_{i\in\mathcal{B}}(\mu_i - m_i)}$ for $i \in \mathcal{B}$, so that $\sum_{i\in\mathcal{B}} \beta_i = 1$. Then, by the convexity of $\pi^*(\cdot)$,

$$\sum_{i\in\mathcal{B}} \beta_i \pi^*(\tilde{\sigma}_i) \geq \pi^*(\sum_{i\in\mathcal{B}} \beta_i \tilde{\sigma}_i). \tag{28}$$

Now, we are going to show $\tilde{\sigma}_b \leq \sum_{i\in\mathcal{B}} \beta_i \tilde{\sigma}_i$. In fact,

$$\begin{aligned}&\left((\sum_{i\in\mathcal{B}} (\mu_i - m_i))\tilde{\sigma}_b \right)^2 = \sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{B}} \Sigma_{ij} \leq \sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{B}} \sigma_i\sigma_j \\ =&\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{B}} (\mu_i - m_i)\tilde{\sigma}_i(\mu_j - m_j)\tilde{\sigma}_j = \left(\sum_{i\in\mathcal{B}} (\mu_i - m_i)\tilde{\sigma}_i \right)^2.\end{aligned} \tag{29}$$

Hence, $(\sum_{i\in\mathcal{B}}(\mu_i - m_i))\tilde{\sigma}_b \leq \sum_{i\in\mathcal{B}}(\mu_i - m_i)\tilde{\sigma}_i$. Thus,

$$\tilde{\sigma}_b \leq \sum_{i\in\mathcal{B}} \beta_i \tilde{\sigma}_i.$$

We prove our lemma by contradiction, suppose otherwise $\delta_b^* \leq \frac{1}{2}$, then $\Phi^{-1}(1 - \delta_b^*) \geq 0$. As a result,

$$\begin{aligned}\pi^*(\tilde{\sigma}_b) &= f(\delta_b^*)(\Phi^{-1}(1 - \delta_b^*)\tilde{\sigma}_b + 1) \\ &\leq f(\delta_b^*)(\Phi^{-1}(1 - \delta_b^*)(\sum_{i\in\mathcal{B}} \beta_i\tilde{\sigma}_i) + 1) \\ &=\pi(\sum_{i\in\mathcal{B}} \beta_i\tilde{\sigma}_i, \delta_b^*) =\leq \pi^*(\sum_{i\in\mathcal{B}} \beta_i\tilde{\sigma}_i).\end{aligned} \tag{30}$$

Comparing $\sum_{i\in[N]} P_i^*$ and $P(\mathcal{B})$ by substracting,

$$\begin{aligned}&\sum_{i\in[N]} P_i^* - P(\mathcal{B}) = \sum_{i\in\mathcal{B}} (\mu_i - m_i)\pi^*(\tilde{\sigma}_i) - (\sum_{i\in\mathcal{B}} (\mu_i - m_i))\pi^*(\tilde{\sigma}_b) \\ =&\left(\sum_{i\in\mathcal{B}} (\mu_i - m_i) \right) (\beta_i\pi^*(\tilde{\sigma}_i) - \pi^*(\tilde{\sigma}_b)) \\ \geq&\left(\sum_{i\in\mathcal{B}} (\mu_i - m_i) \right) \left(\pi^*(\sum_{i\in\mathcal{B}} \beta_i\tilde{\sigma}_i) - \pi^*(\tilde{\sigma}_b) \right) \geq 0,\end{aligned} \tag{31}$$

where the first inequality is from (28) and the second inequality is from (30). Note that (31) is a contradiction to $P(\mathcal{B}) > \sum_{i\in[N]} P_i^*$, concluding our lemma. ∎

Proof of Theorem 2. With Lemma 6 and Lemma 7, now we come back to Theorem 2. Note that

$$\sum_{i\in\widetilde{\mathcal{B}}} (\mu_i - m_i) = \sum_{i\in\mathcal{B}} (\mu_i - m_i) = k(\mu - m) \triangleq \mu_b - m_b.$$

Then, denote $\delta_b^* \triangleq \arg\max_{\delta\in(0,1)} \pi(\frac{\sigma_b}{\mu_b - m_b}, \delta)$ as the fraction of potential customers for the bundle $\mathcal{B}$. Then, according to Lemma 7, $\delta_b^* > \frac{1}{2}$, therefore $\Phi^{-1}(1 - \delta_b^*) < 0$. As a result,

$$\begin{aligned}&\pi^*(\frac{\sigma_b}{\mu_b - m_b}) = f(\delta_b^*)(\Phi^{-1}(1 - \delta_b^*)\frac{\sigma_b}{\mu_b - m_b} + 1) \\ <&f(\delta_b^*)(\Phi^{-1}(1 - \delta_b^*)\frac{\widetilde{\sigma}_b}{\mu_b - m_b} + 1) = \pi(\frac{\widetilde{\sigma}_b}{\mu_b - m_b}, \delta^*) \\ \leq&\pi^*(\frac{\widetilde{\sigma}_b}{\mu_b - m_b}).\end{aligned} \tag{32}$$

We conclude our theorem that

$$\begin{aligned}&P^*(\widetilde{\mathcal{B}}) - P^*(\mathcal{B}) \\ =&(\mu_b - m_b)\pi^*(\frac{\widetilde{\sigma}_b}{\mu_b - m_b}) - (\mu_b - m_b)\pi^*(\frac{\sigma_b}{\mu_b - m_b}) > 0.\end{aligned} \tag{33}$$

∎

D. *Proof of Theorem 4*

The idea of the proof of Theorem 4 is that the random variables rounded by `DepRound` are nearly independent, so that the rounded integer solution is close to the minimum.

Lemma 8. *Let $\mathbf{X} = \texttt{DepRound}(\mathbf{x}^*, k)$, for any matrix $\mathbf{\Sigma}$, we have:*

$$\begin{aligned}&\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] \leq \mathbf{x}^{*T}\mathbf{\Sigma}\mathbf{x}^* \\ &\quad+ \max\{\sigma, |\min\mathbf{\Sigma}|\} \left(\frac{2Nk}{\sum_{i=1}^N \min\{x_i^*, 1 - x_i^*\} - 2} + k \right),\end{aligned}$$

where $\mathbf{X} = [X_1, \ldots, X_N]$, and $\mathbf{x}^* = [x_1^*, \ldots, x_N^*]$.

Proof. Frist, let $\alpha_i \triangleq \min\{x_i^*, 1 - x_i^*\}$, for all $i \in [N]$. According to Theorem 3.2 of [9], we set $I = I^+ = \{i, j\}$, $J = [N]\backslash\{i, j\}$, and $t = 2$. In here, $(x_1^*, x_2^*, \ldots, x_n^*)$ corresponds to the vector $(p_1^*, p_2^*, \ldots, p_n^*)$ in [9]. Then we have the following inequality to bound the dependency of X_i and X_j for any i, j:

$$(1 - \frac{2}{N\hat{q}\hat{\alpha}})x_i^*x_j^* \leq \mathbb{E}[X_iX_j] \leq x_i^*x_j^*. \tag{34}$$

where $\frac{1}{\hat{q}} = \frac{1}{2}(\frac{1}{x_i^*} + \frac{1}{x_j^*})$, and $\hat{\alpha} = \frac{1}{N-2}(\sum_{k=1}^n \alpha_k - \alpha_i - \alpha_j)$. Provided that $\sum_{k=1}^N \alpha_k > 2$, we have

$$\begin{aligned}\mathbb{E}[X_iX_j] &\geq (1 - \frac{2}{N\hat{q}\hat{\alpha}})x_i^*x_j^* \\ &= x_i^*x_j^* - \frac{N-2}{N(\sum_{k=1}^N \alpha_k - \alpha_i - \alpha_j)}(x_i^* + x_j^*) \\ &\geq x_i^*x_j^* - \frac{N-2}{N(\sum_{k=1}^N \alpha_k - 2)}(x_i^* + x_j^*) \\ &\geq x_i^*x_j^* - \frac{1}{(\sum_{k=1}^N \alpha_k - 2)}(x_i^* + x_j^*). \end{aligned} \tag{35}$$

Denote $Z \triangleq \sum_{i=1}^N \alpha_i - 2$ as the normalizer, then

$$x_i^*x_j^* - \mathbb{E}[X_iX_j] \leq \frac{1}{Z}(x_i^* + x_j^*). \tag{36}$$

Now, we give upper bound for $\mathbb{E}[X^T\mathbf{\Sigma}X] - (x^*)^T\mathbf{\Sigma}(x^*)$. Use $\mathbb{I}[\cdot] \in \{0, 1\}$ as the indicator function for an event, then

$$\begin{aligned}&\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] - \mathbf{x}^{*T}\mathbf{\Sigma}\mathbf{x}^* = \sum_{i=1}^N\sum_{j=1}^N \Sigma_{ij}(\mathbb{E}[X_iX_j] - x_i^*x_j^*) \\ =&\sum_{i=1}^N \sum_{j=1, j\neq i}^N \mathbb{I}[\Sigma_{ij} \geq 0]\Sigma_{ij}(\mathbb{E}[X_iX_j] - x_i^*x_j^*) \\ &+ \sum_{i=1}^N (\mathbb{E}[X_i^2] - (x_i^*)^2)\Sigma_{ii} \\ &+ \sum_{i=1}^N \sum_{j=1, j\neq i}^N \mathbb{I}[\Sigma_{ij} < 0]\Sigma_{ij}(\mathbb{E}[X_iX_j] - x_i^*x_j^*) \\ \leq& 0 + \sum_{i=1}^N (\mathbb{E}[X_i])\Sigma_{ii} + \sum_{i=1}^N \sum_{j=1, j\neq i}^N \mathbb{I}[\Sigma_{ij} < 0]|\Sigma_{ij}|(x_i^*x_j^* - \mathbb{E}[X_iX_j]) \\ =& k\sigma + \sum_{i=1}^N \sum_{j=1, j\neq i}^N \mathbb{I}[\Sigma_{ij} < 0]|\Sigma_{ij}|(x_i^*x_j^* - \mathbb{E}[X_iX_j]) \\ \leq& k\sigma + \sum_{i=1}^N \sum_{j=1, j\neq i}^N \mathbb{I}[\Sigma_{ij} < 0]|\Sigma_{ij}|\frac{1}{Z}(x_i^* + x_j^*). \end{aligned}$$

In the above, we divide the summation according to three cases: (1) diagonal; (2) non-diagonal, and $\Sigma_{ij} < 0$ (3) non-diagonal, and $\Sigma_{ij} \geq 0$. Now, note that $\min \mathbf{\Sigma}$ is the upper bound for $|\Sigma_{ij}|$ when $\Sigma_{ij} < 0$, then

$$\begin{aligned}&\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] - \mathbf{x}^{*T}\mathbf{\Sigma}\mathbf{x}^* \leq \left(k\sigma + \sum_{i=1}^N\sum_{j=1}^N \frac{|\min \mathbf{\Sigma}|}{Z}(x_i^* + x_j^*)\right) \\ &= \left(k\sigma + 2N\frac{|\min \mathbf{\Sigma}|}{Z}\sum_{i=1}^N (x_i^*)\right) \\ &= \max\{\sigma, |\min \mathbf{\Sigma}|\}\left(k + \frac{2Nk}{\sum_{i=1}^N \min\{x_i^*, 1 - x_i^*\} - 2}\right). \end{aligned}$$

■

Lemma 9. *Let* $\mathbf{X} = \texttt{DepRound}(\mathbf{x}^*, k)$*, for any matrix* $\mathbf{\Sigma}$*,*

$$\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] \leq \mathbf{x}^{*T}\mathbf{\Sigma}\mathbf{x}^* + \max\{\sigma, |\min \mathbf{\Sigma}|\}\max\left\{\frac{2Nk}{\sqrt{N}-2} + k, 2\sqrt{N}k + N\right\}.$$

This is derived from Lemma 8, but it gives the bound that holds uniformly independent of $\mathbf{x}^*$.

Proof. Again, we let $\alpha_i \triangleq \min\{x_i^*, 1 - x_i^*\}$, for all $i \in [N]$. For notational convenience, we denote $\mathcal{M} \triangleq \max\{\sigma, |\min \mathbf{\Sigma}|\}$. Now, we study the following two cases.

Case 1: $\sum_{i=1}^N \alpha_i \geq \sqrt{N}$.

Then, directly from Lemma 8:

$$\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] - (\mathbf{x}^*)^T\mathbf{\Sigma}(\mathbf{x}^*) \leq \max\{\sigma, |\min \mathbf{\Sigma}|\}(\frac{2Nk}{\sqrt{N}-2} + k). \tag{37}$$

Case 2: $\sum_{i=1}^N \alpha_i < \sqrt{N}$.

It is obvious that when $x_i^* \leq \frac{1}{2}$ and $x_j^* \leq \frac{1}{2}$,

$$x_i^*x_j^* - \mathbb{E}[X_iX_j] \leq x_i^*x_j^* = \alpha_i\alpha_j.$$

When $x_i^* > \frac{1}{2}$ and $x_j^* > \frac{1}{2}$, we claim that

$$\mathbb{P}[X_i = 1, X_j = 1] \geq \mathbb{P}[X_i = 1] + \mathbb{P}[X_j = 1] - 1.$$

In fact,

$$\begin{aligned}&\mathbb{P}[X_i = 1] + \mathbb{P}[X_j = 1] \\ =&2\mathbb{P}[X_i = 1, X_j = 1] + \mathbb{P}[X_i = 1, X_j = 0] + \mathbb{P}[X_i = 0, X_j = 1] \\ \leq&2\mathbb{P}[X_i = 1, X_j = 1] + \mathbb{P}[X_i = 1, X_j = 0] \\ &+ \mathbb{P}[X_i = 0, X_j = 1] + \mathbb{P}[X_i = 0, X_j = 0] \\ =&\mathbb{P}[X_i = 1, X_j = 1] + 1. \end{aligned}$$

Hence, when $x_i^* > \frac{1}{2}$ and $x_j^* > \frac{1}{2}$,

$$\begin{aligned}&x_i^*x_j^* - \mathbb{E}[X_iX_j] = x_i^*x_j^* - \mathbb{P}[X_i = 1, X_j = 1] \\ \leq& x_i^*x_j^* - (\mathbb{P}[X_i = 1] + \mathbb{P}[X_j = 1] - 1) \\ =& x_i^*x_j^* - (x_i^* + x_j^* - 1) = (1 - x_i^*)(1 - x_j^*) = \alpha_i\alpha_j. \end{aligned} \tag{38}$$

As a result,

$$\begin{aligned}&\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] - (\mathbf{x}^*)^T\mathbf{\Sigma}(\mathbf{x}^*) = \sum_{i=1}^N\sum_{j=1}^N \Sigma_{ij}(\mathbb{E}[X_iX_j] - x_i^*x_j^*) \\ =&\sum_{i=1}^N\sum_{j=1}^N \mathbb{I}[\Sigma_{ij} \geq 0, \text{and } i \neq j]\Sigma_{ij}(\mathbb{E}[X_iX_j] - x_i^*x_j^*) \\ &+ \sum_{i=1}^N\sum_{j=1}^N \mathbb{I}[\Sigma_{ij} \geq 0, \text{or } i = j]\Sigma_{ij}(\mathbb{E}[X_iX_j] - x_i^*x_j^*) \\ \leq& 0 + \mathcal{M}\sum_{i=1}^N\sum_{j=1}^N (x_i^*x_j^* - \mathbb{E}[X_iX_j]) \\ =&\mathcal{M}\sum_{i=1}^N\sum_{j=1}^N (\mathbb{I}[x_i^* > \frac{1}{2}, x_j^* > \frac{1}{2}] + \mathbb{I}[x_i^* \leq \frac{1}{2}, x_j^* \leq \frac{1}{2}] \\ &+ 2\mathbb{I}[x_i^* > \frac{1}{2}, x_j^* \leq \frac{1}{2}])(x_i^*x_j^* - \mathbb{E}[X_iX_j]) \\ \leq&\mathcal{M}\sum_{i=1}^N\sum_{j=1}^N (\mathbb{I}[x_i^* > \frac{1}{2}, x_j^* > \frac{1}{2}]\alpha_i\alpha_j \\ &+ \mathbb{I}[x_i^* \leq \frac{1}{2}, x_j^* \leq \frac{1}{2}]\alpha_i\alpha_j + 2\mathbb{I}[x_i^* > \frac{1}{2}, x_j^* \leq \frac{1}{2}]x_i^*x_j^*). \end{aligned} \tag{39}$$

Now, we handle the term $\mathbb{I}[x_i^* > \frac{1}{2}, x_j^* \le \frac{1}{2}]x_i^* x_j^*$ in (39). Continue from (39), we have

$$\begin{aligned}&\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] - (\mathbf{x}^*)^T\mathbf{\Sigma}(\mathbf{x}^*)\\ \le&\mathcal{M}\sum_{i=1}^N\sum_{j=1}^N(\alpha_i\alpha_j + 2x_i^*\mathbb{I}[x_i^*>\frac{1}{2}]x_j^*\mathbb{I}[x_j^*\le\frac{1}{2}])\\ =&\mathcal{M}\left((\sum_{i=1}^N\alpha_i)^2 + 2(\sum_{i\in[N],x_i^*>\frac{1}{2}}x_i^*)(\sum_{j\in[N],x_j^*\le\frac{1}{2}}x_j^*)\right)\\ <&\mathcal{M}\left((\sqrt{N})^2+2(\sum_{i=1}^N x_i^*)(\sum_{j\in[N],x_j^*\le\frac{1}{2}}x_j^*)\right)\\ =&\mathcal{M}\left(N+2k(\sum_{i\in[N],x_i^*\le\frac{1}{2}}x_i^*)\right). \qquad (40)\end{aligned}$$

We claim that $\sum_{i\in[N],x_i^*\le\frac{1}{2}} x_i^* < \sqrt{N}$. In fact,

$$\sum_{i=1}^N x_i^* = k, \text{ and } \sum_{i=1}^N \min\{x_i^*, 1-x_i^*\} = \sum_{i=1}^N\alpha_i < \sqrt{N}.$$

Hence,

$$\begin{aligned}k-\sqrt{N} &< \sum_{i=1}^N(x_i^*-\alpha_i) = \sum_{i=1}^N \mathbb{I}[x_i^*>\frac{1}{2}](2x_i^*-1)\\ &\le \sum_{i\in[N],x_i^*>\frac{1}{2}}(x_i^*+1-1) = \sum_{i\in[N],x_i^*>\frac{1}{2}} x_i^*. \qquad (41)\end{aligned}$$

It implies that

$$\sum_{i\in[N],x_i^*\le\frac{1}{2}} x_i^* = k - \sum_{i\in[N],x_i^*>\frac{1}{2}} x_i^* < \sqrt{N}. \qquad (42)$$

Combine (40) and (42), and we know

$$\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}] - (\mathbf{x}^*)^T\mathbf{\Sigma}(\mathbf{x}^*) \le \mathcal{M}\left(2\sqrt{N}k+N\right).$$

Considering all possible cases, i.e. case 1 and case 2,

$$\begin{aligned}&\mathbb{E}[\mathbf{X}^T\mathbf{\Sigma}\mathbf{X}]\le\mathbf{x}^{*T}\mathbf{\Sigma}\mathbf{x}^*+\\ &\qquad\max\{\sigma,|\min\mathbf{\Sigma}|\}\max\left\{\frac{2Nk}{\sqrt{N}-2}+k, 2\sqrt{N}k+N\right\}.\end{aligned}$$

Here, Lemma 9 is proven. ■

Lemma 10. *Let $P_b(\mathcal{B})$ be the optimal profit from bundle $\mathcal{B}$, i.e. P_b^* when the bundle set is $\mathcal{B}$. Recall that*

$$\Delta_0 \triangleq \frac{\max\{\sigma,|\min\mathbf{\Sigma}|\}}{(\mu-m)^2}\max\{\frac{2N}{(\sqrt{N}-2)k}+\frac{1}{k}, \frac{2\sqrt{N}}{k}+\frac{N}{k^2}\}. \qquad (43)$$

If $P(\mathcal{B}^*) > \sum_i P_i^*$, then for the output $\widehat{\mathcal{B}}$ of Algorithm 1,

$$\mathbb{E}[P_b(\widehat{\mathcal{B}})] \ge k(\mu-m)\min_{\Delta\in[0,\Delta_0]}\pi^*(\sqrt{\frac{\sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}}{(k\mu-km)^2}+\Delta}).$$

Proof. The proof consists of two parts. Denote $\bar{\sigma}_b = \mathbb{E}[\sqrt{\hat{b}^T\mathbf{\Sigma}\hat{b}}]$ as the expected standard deviation of V_b. Without loss of generality, we set $m=0$ so that we use μ as $\mu-m$.

Firstly, because of convexity of $\pi^*(\cdot)$,

$$\begin{aligned}\mathbb{E}[P_b(\hat{\mathcal{B}})] &= k\mu\mathbb{E}[\pi^*(\frac{1}{k\mu})\sqrt{\hat{\mathbf{b}}^T\mathbf{\Sigma}\hat{\mathbf{b}}})]\\ &\ge k\mu\pi^*(\frac{1}{k\mu}\mathbb{E}[\sqrt{\hat{\mathbf{b}}^T\mathbf{\Sigma}\hat{\mathbf{b}}}]) = k\mu\pi^*(\frac{1}{k\mu}\bar{\sigma}_b). \qquad (44)\end{aligned}$$

Secondly, we are going to show

$$\left(\frac{1}{k\mu}\bar{\sigma}_b\right)^2 \in [\frac{\sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}}{(k\mu)^2}, \frac{\sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}}{(k\mu)^2}+\Delta_0].$$

For the second part, we claim that

$$\hat{\mathbf{b}}^T\mathbf{\Sigma}\hat{\mathbf{b}} \ge \sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}, \qquad (45)$$

or otherwise, $\mathcal{B}^*$ is not the optimal bundle by Theorem 2. Also, according to Lemma 9,

$$\begin{aligned}\mathbb{E}[\sqrt{\hat{\mathbf{b}}^T\mathbf{\Sigma}\hat{\mathbf{b}}}] \le&\sqrt{\mathbb{E}[\hat{\mathbf{b}}^T\mathbf{\Sigma}\hat{\mathbf{b}}]} \le \sqrt{\mathbb{E}[(\tilde{\mathbf{b}}^*)\mathbf{\Sigma}(\tilde{\mathbf{b}}^*)]+(k\mu)^2\Delta_0}\\ \le&\sqrt{\sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}+(k\mu)^2\Delta_0}, \qquad (46)\end{aligned}$$

where the last inequality is because the minimized value for the relaxed problem is lower than the minimized value for the original problem. The second inequality is obtained by substituting the notation $\mathbf{x}^*$ with $\tilde{\mathbf{b}}^*$ and substituting $\mathbf{X}$ with $\hat{\mathbf{b}}$ in Lemma 9. Hence, the second part holds:

$$\left(\frac{1}{k\mu}\bar{\sigma}_b\right)^2 \in [\frac{\sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}}{(k\mu)^2}, \frac{\sum_{i\in\mathcal{B}^*}\sum_{j\in\mathcal{B}^*}\Sigma_{ij}}{(k\mu)^2}+\Delta_0]. \qquad (47)$$

Combined with (44), $\pi^*(\frac{1}{k\mu}\bar{\sigma}_b)$ should be greater than the minimum of possible values, then we proved our Lemma 10. ■

Proof of Theorem 4. Let $\tilde{\sigma}_b^* = \frac{1}{k\mu}\sigma_b$ be the normalized standard deviation of the bundle. Then,

$$P_b(\mathcal{B}^*) = k(\mu-m)\times\pi^*(\tilde{\sigma}_b^*).$$

Further, according to Lemma 10,

$$\mathbb{E}[P_b(\widehat{B})] \ge k(\mu-m)\min_{\Delta\in[0,\Delta_0]}\pi^*(\sqrt{(\tilde{\sigma}_b^*)^2+\Delta}).$$

Recall that $P_i^* \triangleq P_0$, then:

$$\begin{aligned}\frac{\mathbb{E}[P(\widehat{\mathcal{B}})]}{P(\mathcal{B}^*)} &= \frac{\mathbb{E}[P_b(\widehat{\mathcal{B}})+(N-k)P_0]}{P_b(\mathcal{B}^*)+(N-k)P_0} \ge \frac{\mathbb{E}[P_b(\widehat{\mathcal{B}})]}{P_b(\mathcal{B}^*)}\\ &\ge \min_{\Delta\in[0,\Delta_0]}\frac{\pi^*(\sqrt{(\tilde{\sigma}_b^*)^2+\Delta})}{\pi^*(\tilde{\sigma}_b^*)}. \qquad (48)\end{aligned}$$

Note that $\tilde{\sigma}_b^* \ge 0$, considering all possible values for $\tilde{\sigma}_b^*$,

$$\frac{\mathbb{E}[P(\widehat{\mathcal{B}})]}{P(\mathcal{B}^*)} \ge \min_{\sigma'\ge0,\Delta\in[0,\Delta_0]}\frac{\pi^*(\sqrt{\sigma'^2+\Delta})}{\pi^*(\sigma')}.$$

Moreover, if $k=\Omega(N^{0.5+\epsilon})$, where $\epsilon>0$, then one can easily verify $\lim_{N\to+\infty}\Delta_0 = 0$. Also, from the definition of $\pi^*(\cdot)$, $\pi^*(\cdot)$ is a continuous function. As a result,

$$\begin{aligned}\lim_{N\to+\infty}\xi &= \lim_{N\to+\infty}\min_{\sigma'\ge0,\Delta\in[0,\Delta_0]}\frac{\pi^*(\sqrt{\sigma'^2+\Delta})}{\pi^*(\sigma')}\\ &= \min_{\sigma'\ge0}\frac{\pi^*(\sigma'^2+0)}{\pi^*(\sigma')} = 1.\end{aligned}$$

■