IISE TRANSACTIONS
2025, VOL. 57, NO. 4, 454–468
https://doi.org/10.1080/24725854.2024.2347918

Minimum-cost ordering for selective assembly

Thomas A. Weber

École Polytechnique Fédérale de Lausanne, Switzerland

ABSTRACT

We consider the minimization of input cost for a selective assembly system that features two random inputs and a finite number of matching classes. This setup frequently arises in high-precision manufacturing when input tolerances are not tight enough for the required output precision. We determine optimality conditions for the cost-optimal input portfolio given an expected-output target, first using a normal approximation of the multinomial binning distribution, and second employing a simple upper envelope of the output objective. We show that the relative error tends to zero as the production scale becomes sufficiently large. The envelope optimization problem also yields a closed-form solution for the cost-minimizing inputs as well as total costs, which are easy to understand for managers. A numerical study tests the practicality of the envelope approach. The latter can be used as seed for a numerical solution of the exact problem, as well as a closed-form approximation, which allows for an analysis of structural properties.

ARTICLE HISTORY
Received 20 February 2023
Accepted 18 April 2024

KEYWORDS
Cost minimization; selective assembly; stochastic production

1. Introduction

By sorting low-precision inputs into predefined tolerance classes, so as to obtain a grade-specific matching for components of different types, it is possible to produce relatively high-precision output. This process of "selective assembly" has been known for more than 100 years—as part of interchangeable manufacturing. While much of the literature thus far has been concerned with optimizing the design of the matching bins, the available work on cost-effectively procuring the inputs for a given system of matching bins with arbitrary distributions of part characteristics is rather sparse. This article provides such an analysis, characterizing the optimal order quantities for two different part types, together with a closed-form approximation of the optimal input quantities. The latter allows for structural insights and suggests a straightforward generalization to an arbitrary number of part types.

In the batch production of movements for mechanical watches, the oscillator—assembled from a hairspring and a balance wheel—must exhibit a well-defined resonance frequency within a small tolerance.[1] The accuracy of this output characteristic is key for the quality of the watch, and this is why many high-end horology companies have adopted a system of selective assembly, sorting hairsprings and balance wheels into different grades, so as to guarantee a sufficiently high clocking precision after joining any two components with matching grades. Naturally, when using selective assembly the probability that, say, of 100 hairsprings *not all* can be matched with 100 balance wheels is generally *positive*, and this mismatch likelihood *increases* in the number of matching classes. That is, a greater output precision comes at the price of a bigger loss of input components. Finding the cost-optimal input quantities needs to trade off the cost of the inputs against the benefits of the outputs, which would vary across the different assembly grades. Given two multinomial distributions for the input parts being in the various matching classes, we characterize the cost-minimizing input quantities so as to achieve a given expected output quantity. We also show that a good approximation of the output objective (with arbitrarily small relative error for sufficiently large production lots) leads to a constant ratio between inputs and outputs, taking the production process to be a weighted sum of perfectly complementary systems. The latter may then allow for the separation of input parts into "principal components" and "buffer components." In mechanical watchmaking, for example, hairsprings turn out to be buffer components for the significantly more expensive balance wheels; cf. Section 4.

1.1. Literature

The idea of sorting inputs of a given type into different bins so as to be able to interchangeably match components across corresponding bins for different part types, also termed "selective assembly," was discussed more than a century ago

CONTACT Thomas A. Weber thomas.weber@epfl.ch

Supplemental data for this article can be accessed online at https://doi.org/10.1080/24725854.2024.2347918.

[1]The hairspring features a stiffness (described by $\varkappa$ [in Nm/rad]) whereas the balance wheel is characterized by a moment of inertia (given by I [in kg m^2]). Their ratio determines the oscillation period, $T = 2\pi\sqrt{I/\varkappa}$.

by Buckingham (1921) who notes: "[t]he chief purpose of manufacturing, by selective assembly (…), is the production of large quantities of duplicate parts as economically as possible, within such limits that they may be assembled without further machining" (p. 224). Our concern here is related to the requirement of "as economically as possible," by specifically considering the tradeoff of input costs against output benefits. The latter may naturally depend on the specific "grade" of the output, that is, on the matching class from which inputs were taken to assemble a given unit of a finished product. Mansoor (1961) points out that in practice assembly processes with more than 20 matching classes can be observed, but that one of the main problems of selective assembly is evidently the possibility for mismatch. In contrast with our input–output view, which is driven by the input-cost minimization problem, as well as the dependence of input cost on the required output precision and the number of matching classes, Caputo and Di Salvo (2019) investigate different manufacturing methods, for which they disaggregate production-specific cost parameters including processing times, sorting costs, and so forth, which are essentially invariant once the production process has been fixed, leaving open the question about the optimal inputs.

There are numerous practical applications for selective assembly systems, such as hole-and-shaft combinations (Asha *et al.*, 2008; Kannan and Pandian, 2021), sheet metal assembly (Rezaei Aderiani *et al.*, 2019), flat-panel display manufacturing (Duenyas *et al.*, 1997), camshafts in the automobile industry (Mease *et al.*, 2004), as well as bimetal thermostats, Fortini's clutch, and knuckle-joint assembly (Tan and Wu, 2012). Mease *et al.* (2004) note that cost imbalances may put into question an otherwise yield-optimized selective assembly system. In the case where input prices differ substantially, as for hairsprings versus the at least 10-times more expensive balance wheels, the less costly parts should be ordered in excess, so that the expensive parts are paired more often.

One main concern of the literature thus far has been the design of these matching classes, in the form of bins in the space of part characteristics, which are often assumed to be normally distributed. For example, Pugh (1986) considers an equidistant partition of scalar part characteristics. Kwon *et al.* (1999) improve on the equidistant partition by determining matching classes, so as to minimize a square deviation of the difference of the (assumed to be) commensurate parts and some target. Although Matsuura (2011) allows for more general convex loss functions, the overall problem with the loss-function approach is that it aggregates penalizations (by taking the expectation) irrespective of whether parts fit or not. In other words, under a set of matching classes which was designed using the penalization approach it is generally possible that parts do not match, even though they were taken from corresponding bins, and therefore, should match. We therefore follow here the robust binning approach by Weber (2021) which guarantees that parts across corresponding bins always match. This allows us to concentrate on the design of the minimum-cost input portfolio. We also mention in passing that in addition to the fixed-bin selective assembly discussed thus far there is the notion of a "direct" selective assembly, which requires to keep track of every single part so as to be able to match parts one-to-one by solving a matching problem on a bipartite graph (Coullard *et al.*, 1998). Since tracking very similar individual parts would often prove impractical and very costly in production settings with significant output volumes, we restrict attention to a selective assembly where bins are given, so that for each part type it is possible to specify a discrete probability distribution of parts being associated with a given bin. As noted earlier, we also assume that parts of different types can always be matched when taken from corresponding bins.

In Economics, the perfect complementarity of inputs for production, with transformation to outputs in constant proportions, goes back to Leontief (1941) who used it as a simplifying assumption to analyze a large economy. In assembly operations, constant proportions in the conversion of inputs to outputs arise because the required amount of input for a unit of output is typically fixed. Without any significant loss of generality, the different part types are assumed to be matched one-to-one, i.e., at equal proportions. For example, if each finished ball bearing requires eight balls and one shaft, then we can think of the balls as being provided in lots of eight, so as to maintain equal proportions of inputs and outputs. Given that for any given matching class, the number of components of each type is ex-ante uncertain, the deterministic production function that maps inputs to expected outputs no longer exhibits perfect complementarity, but rather a certain input substitutability, since additional quantities of type-1 parts (e.g., "balls") provides an insurance against being long with type-2 inputs (e.g., "shafts") in a given matching class. Thus, increasing the quantity of one input type must always lead to a nonzero increase in the expected output, despite the perfect complementarity applied to any realization of the "sorting product" (i.e., the vector of measured parts in the different matching classes). This article extends the work by Weber (2022) that focuses on the special case of binary random part matching, while we here allow for a selective assembly system with an arbitrary number of matching classes and also examine the scaling behavior.

2. Model

2.1. Matching classes and sorting products

The selective assembly problem considered here involves components of two types ("1" and "2") and $M \geq 2$ matching classes. A component of type $i \in \mathcal{I} = \{1, 2\}$ features a (random) characteristic S_i, with realizations s_i in the (measurable) space $\mathcal{S}_i$ which has at least M elements; see Section 3.7. The "matching classes" $\mathcal{C}_{i,m}$, for $m \in \mathcal{M} = \{1, \ldots, M\}$, are nonempty subsets that together form a partition of $\mathcal{S}_i$, in the sense that

$$(m, \hat{m} \in \mathcal{M},\ m \neq \hat{m} \ \Rightarrow\ \mathcal{C}_{i,m} \cap \mathcal{C}_{i,\hat{m}} = \emptyset) \ \text{and} \ \bigcup_{m \in \mathcal{M}} \mathcal{C}_{i,m} = \mathcal{S}_i,$$

for all $i \in \mathcal{I}$. Let

$$p_{i,m} = \mathbb{P}(S_i \in \mathcal{C}_{i,m}) \in (0, 1) \tag{1}$$

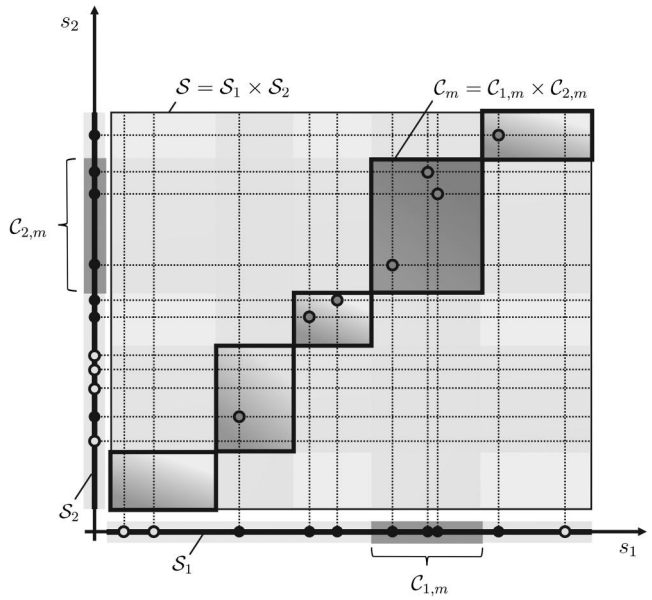


Figure 1. Sorting $x_1 = 10$ type-1 and $x_2 = 11$ type-2 parts with characteristics (s_1, s_2) in $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$; random matching within $M = 5$ matching classes $\mathcal{C}_m = \mathcal{C}_{1,m} \times \mathcal{C}_{2,m}$, for m in $\mathcal{M} = \{1, ..., M\}$; three type-1 components and four type-2 components remain unmatched.

be a given probability that a given component of type i has characteristics that lie in the matching class m. To avoid trivialities, we exclude the case where $p_{i,m} \in \{0, 1\}$. Indeed, for $p_{i,m} = 0$ no successful grade-m matching is possible and the class should be suppressed. Conversely, if $p_{i,m} = 1$, then no other matching class can have successful matches, contradicting the assumption of having at least two functional matching classes, the minimum number of classes for a meaningful selective assembly.

Two parts of characteristics s_1 and s_2 are referred to as "matched" (across types) if the vector $s = (s_1, s_2) \in \mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ is an element of the joint matching class $\mathcal{C}_m = \mathcal{C}_{1,m} \times \mathcal{C}_{2,m}$, for some $m \in \mathcal{M}$. Since by assumption the set of matching classes $\{\mathcal{C}_{i,1}, ..., \mathcal{C}_{i,M}\}$ forms a partition of the space of characteristics $\mathcal{S}_i$, all incoming parts can be sorted into matching classes; thus, taking two parts from corresponding matching classes $\mathcal{C}_{1,m}$ and $\mathcal{C}_{2,m}$ always produces a "matched pair;" see Figure 1. In practice, the characteristics of different types of parts (e.g., hairsprings and balance wheels for mechanical watches) can vary substantially and may be measured in different units (cf. footnote 1). When matching parts of corresponding grades, the number of feasible assemblies for any given matching class (referred to as "grade") corresponds to the *minimum* number of parts available from either type. To clarify this strict complementarity, Figure 1 depicts a situation where the smallest matching class (for each component type) produces zero output: indeed, the two type-1 components in $\mathcal{C}_{1,1}$ cannot be matched with any type-2 components, as $\mathcal{C}_{2,1}$ is empty.

Let $x_i \geq 0$ be the chosen input quantity for type-i components. Sorting them into matching classes yields the vector $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,M})$ (henceforth referred to as "sorting product"), where $x_{i,m} \geq 0$ denotes the number of components of type i in matching class m, so that

$$x_i = \sum_{m \in \mathcal{M}} x_{i,m}, \qquad i \in \mathcal{I}.$$

The multinomial probability that a size-x_i ordering of type-i components into M matching classes produces a particular sorting product $\boldsymbol{x}_i \in \mathbb{N}^{x_i}$ is

$$\mathbb{P}(\mathbf{X}_i = \boldsymbol{x}_i | \mathbf{p}_i, x_i) = \begin{cases} \left(\prod_{m \in \mathcal{M}} \frac{p_{i,m}^{x_{i,m}}}{x_{i,m}!} \right) (x_i!), & \text{if } x_i = \sum_{m \in \mathcal{M}} x_{i,m}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{p}_i = (p_{i,1}, ..., p_{i,M}) \in (0, 1)^M$ is a given vector of positive likelihoods $p_{i,m}$ such that

$$\sum_{m \in \mathcal{M}} p_{i,m} = 1, \qquad i \in \mathcal{I}. \tag{2}$$

The preceding normalization constraint may be relaxed to include the possibility of wasteful off-spec components (cf. Remark 5).

2.2. Normal approximation

We use the standard normal approximation of the multinomial distribution of the random vector $\mathbf{X}_i = (X_{i,1}, ..., X_{i,M})$ with the mean $\boldsymbol{\mu}_i = (\mu_{i,1}, ..., \mu_{i,M})$, where $\mu_{i,m} = p_{i,m} x_i$, and the symmetric positive-definite covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{M \times M}$. The latter features the diagonal elements $\sigma_{i,m}^2 = \text{var}(X_{i,m}) = x_i p_{i,m}(1 - p_{i,m})$ and the off-diagonal elements $\text{cov}(X_{i,m}, X_{i,\hat{m}}) = -x_i p_{i,m} p_{i,\hat{m}}$ for $m, \hat{m} \in \mathcal{M}$ with $m \neq \hat{m}$. In what follows, we usually suppress the dependence on $\mathbf{p}_i$ for simplicity, as it is not a decision variable. Under the proposed normal approximation,

$$\mathbb{P}(X_{i,m} \leq \xi | x_i) \approx G_{i,m}(\xi | x_i) = \Phi\left(\frac{\xi - \mu_{i,m}}{\sigma_{i,m}} \right), \qquad \xi \in [0, x_i],$$

denotes the cumulative distribution function (cdf) for the random number $X_{i,m}$ of type-i components in the mth matching class, where the cdf of the standard normal distribution is

$$\Phi(\xi) = \int_{-\infty}^{\xi} \phi(\zeta) d\zeta = \frac{1}{2} \left[1 + \text{erf}\left(\xi / \sqrt{2} \right) \right], \qquad \xi \in \mathbb{R},$$

with probability density function (pdf)

$$\phi(\xi) = \frac{\exp\left(-\xi^2/2\right)}{\sqrt{2\pi}}, \qquad \xi \in \mathbb{R}.$$

Clearly, the functions $G_{i,m}(\cdot | x_i)$, for $m \in \mathcal{M}$, constitute *marginal* distributions of the joint cdf that describes the random sorting product of type-i components under the normal approximation, but it turns out that they are all that is required for our purposes.

Remark 1 (Approximation error). By the Berry–Esseen theorem (Berry, 1941; Esseen, 1942) the absolute deviation of the marginal distribution $G_{i,m}$ from the underlying binomial distribution can be bounded as follows:

$$\sup_{\xi\in[0,x_i]}\left|\mathbb{P}(X_{i,m}\le\xi|x_i)-G_{i,m}(\xi|x_i)\right|\le\left(\frac{p_{i,m}^2+(1-p_{i,m})^2}{\sigma_{i,m}}\right)\kappa,$$

for all $(i,m)\in\mathcal{I}\times\mathcal{M}$, with $\kappa=0.4748$ (Shevtsova, 2011), which lies above the theoretical lower bound for κ of $\left(\sqrt{10}+3\right)/\left(6\sqrt{2\pi}\right)\approx 0.4097$ (Esseen, 1956). Hence, as long as

$$x_i>\left(\frac{\max\{p_{i,m},1-p_{i,m}\}}{\min\{p_{i,m},1-p_{i,m}\}}\right)\nu^2,\qquad \nu\ge 0,$$

the approximation error corresponds to the probability of "extreme events," further than ν standard deviations away from the mean.

Remark 2 (Continuous input values). The normal approximation allows for non-negative input quantities x_i, lifting the burden of the integer constraint for the purposes of optimization. A practically implementable input vector can then usually be found *ex post* by rounding to the nearest integer or the nearest feasible batch size.

2.3. Matching output

Given its type-1 and type-2 inputs, the firm produces the (sorted) output vector $\mathbf{Y}=(Y_1,\ldots,Y_M)$. The mth random *matching output* Y_m, generated by using components with joint characteristics in matching class $\mathcal{C}_m$, is

$$Y_m=\min\{X_{1,m},X_{2,m}\},\qquad m\in\mathcal{M}. \tag{3}$$

It is evident that the intrinsic value of the firm's output may not be the same for all matching classes. For example, the matching-class index m can denote different quality grades, in which case high-quality output would yield a higher price than low-quality output, without any difference in the unit cost for each random input component. To capture the differentiation of benefits across matching classes, let

$$Q=\eta\cdot\mathbf{Y}=\sum_{m\in\mathcal{M}}\eta_m Y_m \tag{4}$$

denote the firm's total (weighted) output, where $\eta=(\eta_1,\ldots,\eta_M)$ is a vector which contains the net unit value $\eta_m>0$ for the mth matching output; see Figure 2. The net unit value can be viewed as an absolute or a relative measure of benefits. When *absolute*, it may incorporate the revenues from selling the assembled grade-m product minus production costs, whereas when *relative* it serves to quantify a normalized net benefit. In the latter case, the normalizations $\eta_1+\cdots+\eta_m\in\{1,M\}$ are commonly used; for instance, by setting $\eta_m\equiv 1$ one obtains a simple physical output count, $Q=Y_1+\cdots+Y_m$, relevant for many (if not most) manufacturing applications. The assumption that η_m is strictly positive reflects the firm's concern for all units and its option of "free disposal." If a certain matching class (e.g., $m=1$) represents "junk" components, then it is

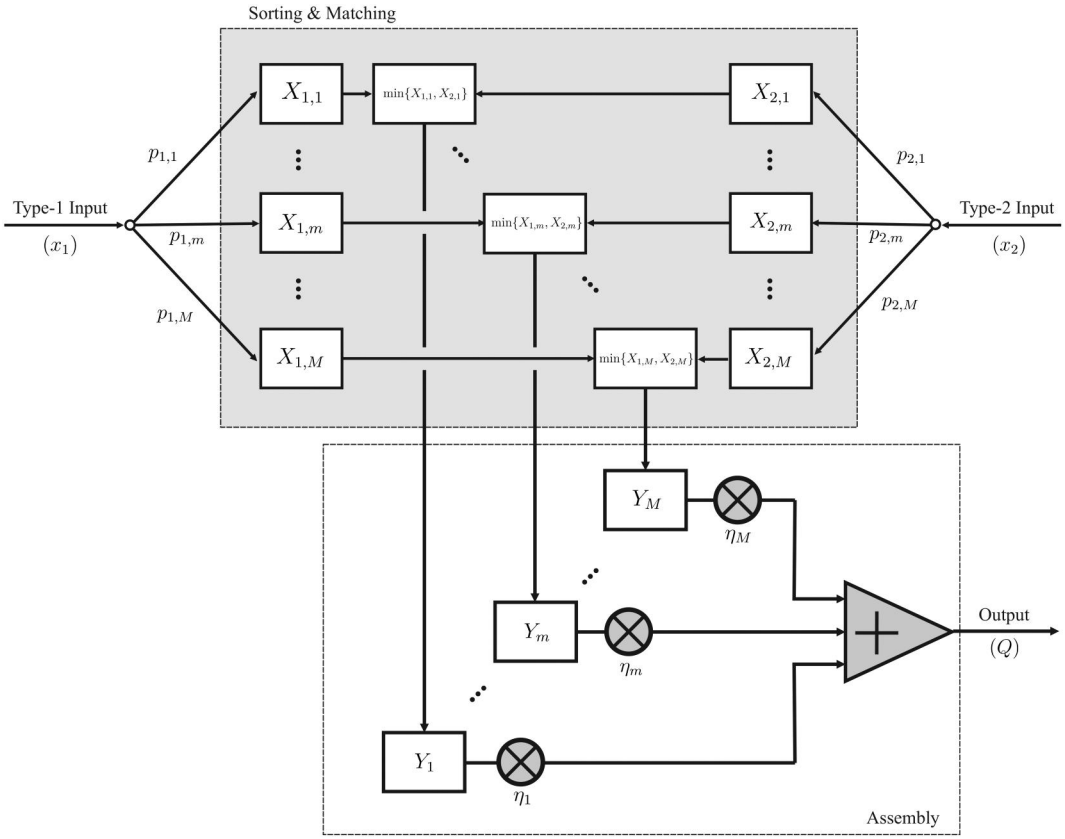


Figure 2. Selective assembly system, featuring the sorting of two types of inputs into $M\ge 2$ matching classes according to the part characteristics, random matching across corresponding classes, and subsequent assembly. The scalar output Q is a random variable.

possible to reduce that weight so as to reflect any possible scrap value.

2.4. Expected output

The need for grade-specific component matching renders the firm's production stochastic, allowing generically for the possibility of even zero production output (with positive probability), irrespective of the chosen input quantities. While this may be true in principle, of course the likelihood of such poor output realizations becomes negligible in any realistic setting. For example, the likelihood of observing a type-i quantity of two standard deviations below the expected quantity in any given matching class is below 5%; for three standard deviations it is below 0.3%.

Since the firm is risk-neutral, its focus lies on its expected total (weighted) output,

$$F(\boldsymbol{x}) = \mathbb{E}[Q|\boldsymbol{x}] = \sum_{m\in\mathcal{M}} \eta_m \mathbb{E}[Y_m|\boldsymbol{x}] = \sum_{m\in\mathcal{M}} \eta_m F_m(\boldsymbol{x}), \ \ \boldsymbol{x}\in\mathbb{R}^2_+, \tag{5}$$

which is a deterministic function of the corresponding input $\boldsymbol{x} = (x_1, x_2)$, where the total output $F(\cdot)$ is obtained as the weighted sum of the expected grade-m outputs $F_m(\cdot)$. For any given m, the expected grade-m output depends on the statistical properties of the difference, $\Delta_m = X_{2,m} - X_{1,m}$, referred to as the (grade-m) "part defect," which has the mean

$$\delta_m = \mu_{2,m} - \mu_{1,m},$$

and the standard deviation

$$\sigma_m = \sqrt{\sigma^2_{1,m} + \sigma^2_{2,m}}.$$

The expected output in any given matching class can be expressed as a convex combination of the expected type-1 and type-2 quantities, minus a positive loss term that increases superlinearly in the aforementioned standard deviation of the parts defect.

Lemma 1 (Grade-*m* output). *Given any $m\in\mathcal{M}$, the expected grade-m output is given by*

$$F_m(\boldsymbol{x}) = \Phi_m\mu_{1,m} + (1-\Phi_m)\mu_{2,m} - \phi_m\sigma_m, \tag{6}$$

for all $\boldsymbol{x}\in\mathbb{R}^2_+$, where $\Phi_m = \Phi(\delta_m/\sigma_m)$ denotes the probability that the relative grade-m part defect does not exceed δ/σ_m, and $\phi_m = \phi(\delta_m/\sigma_m)$ is the probability density at that point.

When the expected grade-m parts defect δ_m is positive (resp., negative), then the weight $\Phi_m\in[0,1]$ placed on the expected number $\mu_{1,m}$ of type-1 parts decreases (resp., increases) in σ_m, since with an increased coefficient of variation for parts defects the expected number of type-2 parts increases (resp., decreases) in relative importance. In other words, perhaps counterintuitively, the nominally expected matches (as measured by the convex combination of the expected type-specific yields $\mu_{1,m}$ and $\mu_{2,m}$) goes up in the dispersion of the grade-m parts defect. However, this effect is offset by an increased "matching loss" $\sigma_m\phi_m$: Independent of the sign of δ_m, the latter grows in σ_m superlinearly, for $\phi_m > 0$ also grows in σ_m.

Lemma 2 (Monotonicity). *Given any $m\in\mathcal{M}$, the expected grade-m output is increasing, so*

$$(\boldsymbol{x}\neq\hat{\boldsymbol{x}} \quad \text{and} \quad \boldsymbol{x}\leq\hat{\boldsymbol{x}}) \ \Rightarrow \ F_m(\boldsymbol{x}) < F_m(\hat{\boldsymbol{x}}),$$

for any two input vectors $\boldsymbol{x},\hat{\boldsymbol{x}}\in\mathbb{R}^2_+$.

By increasing the amount of any input, the expected output increases over all matching classes. By increasing the type-i input from x_i to $\hat{x}_i > x_i$ the grade-m distribution $G_{i,m}(\cdot|\hat{x}_i)$ first-order stochastically dominates the grade-m distribution $G_{i,m}(\cdot|x_i)$. Since first-order stochastic dominance is preserved by taking the minimum across the component types, the grade-m output $\hat{Y}_m$ (with x_i in $\boldsymbol{x}$) first-order stochastically dominates the output Y_m (with $\hat{x}_i$ in $\hat{\boldsymbol{x}}$), all else equal, which in turn implies the (strict) monotonicity of the firm's total output in its inputs.

3. Minimum-cost selective matching

3.1. Problem statement

Each unit of a type-i input is procured at the cost $c_i > 0$, which implies that the firm is interested in minimizing the cost of generating a desired (expected) total output $q > 0$. Thus, given the cost vector $\mathbf{c} = (c_1, c_2)$, the firm's (*minimum-cost*) *selective matching problem* is to determine the optimal input $\boldsymbol{x}^*(q) = (x_1^*(q), x_2^*(q))$ (also referred to as its "output-driven demand") which achieves its minimum cost,

$$C(q) = \min_{\boldsymbol{x}\in\mathbb{R}^2_+}\{\mathbf{c}\cdot\boldsymbol{x}\}, \quad \text{subject to}: \ \ F(\boldsymbol{x})\geq q, \tag{$*$}$$

for any output $q > 0$. Clearly, the solution to the selective matching problem does not depend on the individual input costs, but only on their ratio $\gamma = c_1/c_2$.

3.2. Optimal input

The firm's output-driven demand (or optimal input) $\boldsymbol{x}^*(q)$ minimizes the cost of its inputs to achieve a given output objective $q > 0$. It will come as no surprise that at the optimum the expected total output must exactly equal the output objective. This "output efficiency" is a consequence of the strict monotonicity of expected output in its inputs and the fact that inputs are costly at the margin. In addition, the technical rate of substitution between the two inputs must, at the optimum, be equal to the cost ratio γ. This provides optimality conditions for the otherwise nonconvex optimization problem $(*)$, summarized by the following result.

Theorem 1. *For any expected total output $q > 0$, a solution $\boldsymbol{x}^*(q)$ to the minimum-cost selective matching problem $(*)$ satisfies output efficiency,*

$$F(\boldsymbol{x}^*(q)) = q; \tag{7}$$

in addition, at $\boldsymbol{x} = \boldsymbol{x}^*(q)$ the marginal rate of technical substitution between type-1 and type-2 components is equal to the cost ratio $\gamma = c_1/c_2$, so

$$\left.\frac{\sum_{m\in\mathcal{M}}\eta_m\ p_{1,m}\left(\Phi_m-\frac{1-p_{1,m}}{2}\ \phi_m\right)}{\sum_{m\in\mathcal{M}}\eta_m\ p_{2,m}\left(1-\Phi_m-\frac{1-p_{2,m}}{2}\ \phi_m\right)}\right|_{\boldsymbol{x}=\boldsymbol{x}^*(q)}=\gamma, \tag{8}$$

where Φ_m and ϕ_m, for all $m\in\mathcal{M}$, are specified in Lemma 1.

Remark 3 (Nonconvexity). Generally, the output objective $F(\boldsymbol{x})$ can be nonconcave in the vector of inputs $\boldsymbol{x}$. Indeed, in the simple case of two grades, where the firm only cares about grade-1 output, it is $M=2$ and $\boldsymbol{\eta}=(\eta_1,\eta_2)=(1,0)$. The corresponding Hessian of F is:

$$\det\ D^2F=\det\begin{bmatrix}\partial^2_{x_1x_1}F & \partial^2_{x_1x_2}F\\ \partial^2_{x_1x_2}F & \partial^2_{x_2x_2}F\end{bmatrix}=-\frac{p_{1,1}^2p_{2,1}^2(2-p_{1,1}-p_{2,1})^2}{8\ \sigma_1^4(\boldsymbol{x})}$$
$$\phi^2\left(\frac{p_{2,1}x_2-p_{1,1}x_1}{\sigma_1(\boldsymbol{x})}\right)<0,$$

for all $\boldsymbol{x}=(x_1,x_2)\in\mathbb{R}^2_{++}$. The preceding determinant of the Hessian is equal to the product of its eigenvalues, which must have therefore different signs. As a consequence the output objective is nonconvex in this case. Thus, the solution to the optimality conditions (7)–(8) may not be unique. As shown in Section 3.4, a simplified approximate output objective may yield a reasonable approximate solution (in closed-form).

3.3. Envelope approximation

Instead of the normal approximation of the expected output F we now consider the envelope output,

$$\bar{F}(\boldsymbol{x})=\sum_{m\in\mathcal{M}}\eta_m\min\{\mu_{i,m}:i\in\mathcal{I}\}$$
$$=\sum_{m\in\mathcal{M}}\eta_m\ \min\{p_{i,m}x_i:i\in\mathcal{I}\}, \tag{9}$$

which is well-defined for all $\boldsymbol{x}\in\mathbb{R}^2_+$. The envelope output $\bar{F}$ is an upper bound for F.

Lemma 3. *$F(\boldsymbol{x})\le\bar{F}(\boldsymbol{x})$, for all $\boldsymbol{x}\in\mathbb{R}^2_+$.*

As the positive linear combination of minimum functions is concave, Jensen's inequality implies that the expectation of the minimum is weakly smaller that the minimum of the expectations. An important question that arises of course is how far, for a given input $\boldsymbol{x}$, the envelope $\bar{F}(\boldsymbol{x})$ output may deviate from the reference output $F(\boldsymbol{x})$? We thus consider the absolute deviation,

$$R(\boldsymbol{x})\triangleq|F(\boldsymbol{x})-\bar{F}(\boldsymbol{x})|=\bar{F}(\boldsymbol{x})-F(\boldsymbol{x})\ge 0,\qquad \boldsymbol{x}\in\mathbb{R}^2_+,$$

referred to as the envelope approximation error. The next result shows that this error increases only sublinearly with the inputs.

Lemma 4 (Approximation error). *For any $\boldsymbol{x}\in\mathbb{R}^2_{++}$ it is*

$$\frac{R(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\le\frac{1}{\sqrt{2\pi}},$$

where $\sigma(\boldsymbol{x})=\sum_{m\in\mathcal{M}}\eta_m\sigma_m(\boldsymbol{x})$.

In other words, the envelope approximation error is at most proportional to the weighted sum of the grade-specific standard deviations $\sigma_m(\boldsymbol{x})=\sqrt{p_{1,m}(1-p_{1,m})x_1+p_{2,m}(1-p_{2,m})x_2}$, using the given output weight η_m, for each grade $m\in\mathcal{M}$. Note also that the preceding error bound is tight when $\eta_m\equiv$ const., $p_{i,m}\equiv 1/2$, and $x_{1,m}\equiv x_{2,m}$. In a heterogeneous setting, the derived bound may be conservative as it presumes that $p_{1,m}x_1=p_{2,m}x_2$ holds for all $m\in\mathcal{M}$; cf. Weber (2022). While the absolute deviation between F and $\bar{F}$ may grow with the size of the inputs, in terms of the relative error,

$$r(\boldsymbol{x})=\frac{R(\boldsymbol{x})}{F(\boldsymbol{x})},\qquad \boldsymbol{x}\in\mathbb{R}^2_{++},$$

we obtain the following strong global approximation result.

Lemma 5 (Relative approximation). *For any $\varepsilon\in(0,1)$, there exists a finite minimal envelope output $\underline{q}(\varepsilon)$, which is such that:*[2]

$$\bar{F}(\boldsymbol{x})\ge\underline{q}(\varepsilon)\quad\Rightarrow\quad r(\boldsymbol{x})\le\varepsilon.$$

For small $\varepsilon>0$, the minimal envelope output $\underline{q}(\varepsilon)$ is $O(\varepsilon^{-2})$.

Instead of a minimal envelope output $\underline{q}(\varepsilon)$ it is possible to specify a minimal input $\underline{x}(\varepsilon)$, which is also $O(\varepsilon^{-2})$), since by the homogeneity of $\bar{F}$ it is proportional to $\underline{q}(\varepsilon)$; cf. footnote 2. Then $\min\{x_1,x_2\}\ge\underline{x}(\varepsilon)$ also implies $r(\boldsymbol{x})\le\varepsilon$. A key driver of the relative approximation result in Lemma 5 is that, as a consequence of Lemma 4, the given upper bound of the relative error decreases with the inverse square-root of the minimal input (i.e., with $(\min\{x_1,x_2\})^{-1/2}$). The given lower bounds for input or output are fairly conservative. In practice, it is preferable to perform an error correction based on the relative error at the solution to an envelope optimization problem.

3.4. Envelope optimization

Based on the upper bound $\bar{F}$ for the production function introduced in Section 3.3, we now consider the *envelope optimization problem*,

$$\bar{C}(q)=\min_{\boldsymbol{x}\in\mathbb{R}^2_+}\{\mathbf{c}\cdot\boldsymbol{x}\},\quad\text{subject to:}\quad\bar{F}(\boldsymbol{x})\ge q, \tag{**}$$

for any given target output $q>0$. By construction, this optimization problem is convex, and in order to find a solution to (**), which will be referred to as an *optimal envelope input* $\bar{\boldsymbol{x}}^*$, it is useful to combine all grades which have the same ratio of likelihoods (across component types),

$$\{\rho_\ell:\ell\in\{1,\ldots,L\},\ \rho_1<\cdots<\rho_L\}=\left\{\frac{p_{1,m}}{p_{2,m}}:m\in\mathcal{M}\right\}, \tag{10}$$

[2] As established in the proof of Lemma 5, a (very conservative) minimal envelope output is obtained by setting $\underline{q}(\varepsilon)=(\sum_{\ell=1}^{L}\hat{\eta}_\ell\min\{\rho_{\ell'},1\})\underline{x}(\varepsilon)/(1+\varepsilon)$, where $\underline{x}(\varepsilon)=(4/\pi)(1/\varepsilon^2)(1-p_{\min})/p_{\min}$ is the corresponding minimal input, with $p_{\min}=\min_{(i,m)\in\mathcal{I}\times\mathcal{M}}\{p_{i,m}\}>0$ being the smallest grade-achievement probability across all part types. For example, given $p_{\min}=10\%$ and $\varepsilon=10\%$, the minimum input becomes $\underline{x}(\varepsilon)=(60)^2/\pi\approx 1146$ units.

to obtain $L \leq M$ combined grades, where each such combined grade is characterized by a distinct likelihood ratio ρ_ℓ. We can then rewrite the envelope output in (9) as follows:

$$\bar{F}(\boldsymbol{x}) = \sum_{\ell=1}^{L} \hat{\eta}_\ell \min\{\rho_\ell x_1, x_2\}, \tag{11}$$

where

$$\hat{\eta}_\ell = \sum_{m \in \mathcal{M}} \mathbf{1}_{\{\rho_\ell = p_{1,m}/p_{2,m}\}}\ \eta_m p_{2,m}, \qquad \ell \in \{1, ..., L\},$$

are the corresponding combined weights. The alternative representation of the envelope output in (11) allows for a closed-form representation of the solution to the envelope optimization problem (∗∗).

Theorem 2. *For any $q > 0$, a solution to the envelope optimization problem (∗∗) is*

$$\bar{\boldsymbol{x}}^*(q) = \bar{\boldsymbol{\xi}}^{\ell^*} q, \tag{12}$$

where each $\bar{\boldsymbol{\xi}}^\ell = (\bar{\xi}_1^\ell, \bar{\xi}_2^\ell)$, for $\ell \in \{1, ..., L\}$, denotes a vertex of the iso-output contour for a unit output, with

$$\bar{\xi}_1^\ell = \left(\sum_{k=1}^{L} \hat{\eta}_k \min\{\rho_k, \rho_\ell\}\right)^{-1} \quad \text{and} \quad \bar{\xi}_2^\ell = \rho_\ell \bar{\xi}_1^\ell.$$

The optimal combined grade, $\ell^* = \max\{\ell \in \{1, ..., L\} : s_\ell < \gamma\}$, determines the critical vertex $\bar{\boldsymbol{\xi}}^{\ell^*}$, where the iso-output contour slopes $s_1, ..., s_L$ are such that

$$0 = s_1 < s_\ell = \left(\bar{\xi}_2^\ell - \bar{\xi}_2^{\ell-1}\right) / \left(\bar{\xi}_1^{\ell-1} - \bar{\xi}_1^\ell\right),$$

for all $\ell \in \{2, ..., L\}$.

The optimal envelope solution $\bar{\boldsymbol{x}}^*$ is such that the slope of the iso-cost curves lies in the subgradient, i.e., $-\gamma \in \partial \bar{F}(\bar{\boldsymbol{x}}^*)$, and at the same time envelope output efficiency holds, in the sense that $\bar{F}(\bar{\boldsymbol{x}}^*) = q$. The following example illustrates the result.

Example 1. Consider a selective-assembly system with $M = 5$ matching classes featuring the matching-probability vectors $\mathbf{p}_1 = (0.4, 0.2, 0.1, 0.1, 0.2)$ for parts of type 1 and $\mathbf{p}_2 = (0.2, 0.1, 0.1, 0.2, 0.4)$ for parts of type 2. The unit-cost vector is $\mathbf{c} = (3, 1)$, so $\gamma = 3$. Since the set of likelihood ratios, $\{p_{1,m}/p_{2,m} : m \in \mathcal{M}\} = \{1/2, 1, 2\}$, features only $L = 3$ distinct elements, by (10) the relevant likelihood ratios (in increasing order) are $\rho_1 = 1/2$, $\rho_2 = 1$, and $\rho_3 = 2$. Given a uniform output-weight vector $\boldsymbol{\eta} = (1, 1, 1, 1, 1)$, the envelope output can be written as in (11), with the combined weights $\hat{\eta}_1 = 0.6$ (for $\rho_1 = p_{4,1}/p_{4,2} = p_{5,1}/p_{5,2}$), $\hat{\eta}_2 = 0.1$ (for $\rho_2 = p_{3,1}/p_{3,2}$), and $\hat{\eta}_3 = 0.3$ (for $\rho_3 = p_{1,1}/p_{1,2} = p_{2,1}/p_{2,2}$):

$$\bar{F}(x_1, x_2) = 0.6\min\{x_1/2, x_2\} + 0.1\min\{x_1, x_2\} + 0.3\min\{2x_1, x_2\}.$$

The corresponding unit-output vertices are $\bar{\boldsymbol{\xi}}^1 = (2, 1)$, $\bar{\boldsymbol{\xi}}^2 = (10/7, 10/7)$, and $\bar{\boldsymbol{\xi}}^3 = (1, 2)$, resulting in the iso-output contour slopes $s_1 = 0 < s_2 = 3/4 < s_3 = 4/3$. Comparing the latter with the cost ratio $\gamma = 3$, we find $\ell^* = \max\{\ell \in \{1, 2, 3\} : s_\ell < \gamma\} = 3$, so that $\bar{\boldsymbol{\xi}}^3$ becomes the critical vertex. By Theorem 2, the solution to the envelope optimization problem (∗∗) is therefore $\bar{\boldsymbol{x}}^*(q) = (1, 2)q = (q, 2q)$, for any $q > 0$. Figure 3 depicts the corresponding situation comparing, for $q = 100$, the iso-F contour with the iso-$\bar{F}$ contour. We also note that to $\ell^* = 3$ there correspond two critical grades, $m^* \in \{1, 2\}$. The solution to the envelope optimization problem is consistent with minimizing input cost only with respect to the critical grades.

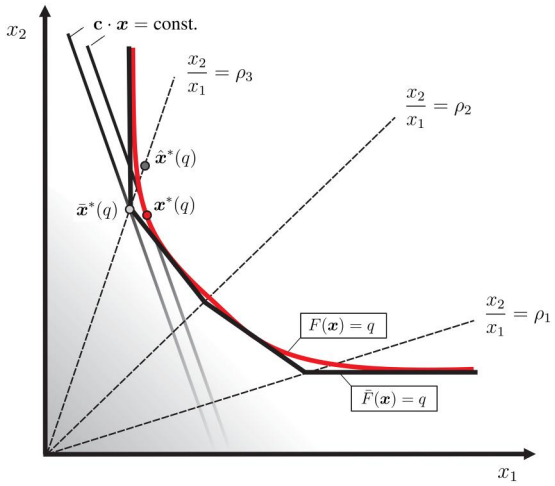


Figure 3. The solution $\boldsymbol{x}^*(q)$ to the selective matching problem (∗) can be approximated by scaling the solution $\bar{\boldsymbol{x}}^*(q)$ to the envelope optimization problem (∗∗). This yields the "approximate selective matching solution," $\hat{\boldsymbol{x}}^*(q) = [q/F(\bar{\boldsymbol{x}}^*(q))]\bar{\boldsymbol{x}}^*(q)$, feasible in (∗); cf. Examples 1–3.

Remark 4 (Alternative representation). From a computational viewpoint, instead of determining the set $\{\rho_1, ..., \rho_L\}$ from the likelihood ratios $p_{1,m}/p_{2,m}$, it is usually more convenient to compute $\bar{F}(\bar{\boldsymbol{x}}^m(q))$ for all M candidate solutions $\bar{\boldsymbol{x}}^m(q) = (\bar{x}_1^m(q), \bar{x}_2^m(q))$, with

$$\bar{x}_i^m(q) = \left(\sum_{k \in \mathcal{M}} \eta_k\ \min\left\{\frac{p_{j,k}}{p_{j,m}} : j \in \mathcal{I}\right\}\right)^{-1} \frac{q}{p_{i,m}}, \quad (i, m) \in \mathcal{I} \times \mathcal{M}, \tag{13}$$

instead of using (12), and to then select the *optimal* envelope input as the cheapest of these, so $\bar{\boldsymbol{x}}^*(q) = \bar{\boldsymbol{x}}^{m^*}(q)$, with

$$m^* \in \arg\min_{m \in \mathcal{M}} \{\mathbf{c} \cdot \bar{\boldsymbol{x}}^m(1)\}. \tag{14}$$

The critical grade m^* does thereby not depend on the magnitude of the output, since both the cost and the envelope output function are homogeneous (of degree 1).

Example 2. In the setting of Example 1, we find $\bar{\boldsymbol{x}}^1(1) = \bar{\boldsymbol{x}}^2(1) = (1, 2)$, $\bar{\boldsymbol{x}}^3(1) = (10/7, 10/7)$, and $\bar{\boldsymbol{x}}^4(1) = \bar{\boldsymbol{x}}^5(1) = (2, 1)$. Since $\mathbf{c} \cdot \bar{\boldsymbol{x}}^1(1) = 5 < \mathbf{c} \cdot \bar{\boldsymbol{x}}^3(1) = 40/7 < \mathbf{c} \cdot \bar{\boldsymbol{x}}^5(1) = 7$, this implies the critical grades $m^* \in \text{argmin}_{m \in \mathcal{M}}\{c \cdot \bar{\boldsymbol{x}}^m(1)\} = \{1, 2\}$, and thus the solution to the envelope optimization problem, $\bar{\boldsymbol{x}}^*(q) = q\ \bar{\boldsymbol{x}}^{m^*}(1) = (q, 2q)$, for all $q > 0$, as before.

3.5. Approximate solution

By Lemma 3 the envelope output $\bar{F}$ exceeds the expected output F. For any target output $q > 0$, output efficiency (i.e., the fact that $\bar{F}(\bar{\boldsymbol{x}}^*(q)) = q$) means the solution $\bar{\boldsymbol{x}}^*(q)$ to the envelope optimization problem $(**)$ is not a feasible point of the original selective matching problem $(*)$. However, since the envelope objective is homogeneous (of degree 1), it is enough to scale the optimal envelope input to achieve feasibility, leading to an *approximate solution to the selective matching problem*,[3]

$$\hat{\boldsymbol{x}}^*(q) = \left[\frac{q}{F(\bar{\boldsymbol{x}}^*(q))}\right]\bar{\boldsymbol{x}}^*(q) = \frac{\bar{\boldsymbol{x}}^*(1)q^2}{F(\bar{\boldsymbol{x}}^*(q))}, \tag{15}$$

which in itself is sublinear in q (as F is convex in q due to a risk-pooling effect). Consequently, the approximate cost function,

$$\hat{C}(q) = \left(\sum_{k\in\mathcal{M}} \eta_k \min_{i\in\mathcal{I}}\left\{\frac{p_{i,k}}{p_{i,m^*}}\right\}\right)^{-1}\left(\sum_{i\in\mathcal{I}}\frac{c_i}{p_{i,m^*}}\right)\frac{q^2}{F(\bar{\boldsymbol{x}}^*(q))}, \tag{16}$$

is an *upper bound* for $C(q)$ in $(*)$, where the critical grade m^* is determined by (14). On the other hand, the optimal envelope cost, $\bar{C}(q) = \mathbf{c}\cdot\bar{\boldsymbol{x}}^*(q)$, must constitute a *lower bound* for the optimal cost, so

$$\underbrace{\bar{C}(q)}_{\mathbf{c}\cdot\bar{\boldsymbol{x}}^*(q)} \le \underbrace{C(q)}_{\mathbf{c}\cdot\boldsymbol{x}^*(q)} \le \underbrace{\hat{C}(q)}_{\mathbf{c}\cdot\hat{\boldsymbol{x}}^*(q)}, \tag{17}$$

for all $q > 0$. At the approximate solution $\hat{\boldsymbol{x}}^*(q)$, the relative output error becomes

$$\begin{aligned}\varepsilon(q) &= \frac{\bar{F}(\hat{\boldsymbol{x}}^*(q)) - F(\hat{\boldsymbol{x}}^*(q))}{F(\hat{\boldsymbol{x}}^*(q))} = \frac{q^2}{F(\bar{\boldsymbol{x}}^*(q))\cdot F(\hat{\boldsymbol{x}}^*(q))} - 1\\ &\le \frac{q}{F(\bar{\boldsymbol{x}}^*(q))} - 1,\end{aligned} \tag{18}$$

since, by construction, $F(\hat{\boldsymbol{x}}^*(q)) \ge q$. We note that the term on the right-hand side of the preceding inequality (18) also bounds the relative cost overage,

$$\frac{\hat{C}(q) - C(q)}{C(q)} = \frac{\mathbf{c}\cdot\hat{\boldsymbol{x}}^*(q)}{\mathbf{c}\cdot\boldsymbol{x}^*(q)} - 1 \le \frac{q}{F(\bar{\boldsymbol{x}}^*(q))} - 1, \tag{19}$$

where we have used (15) and (17). Based on the preceding inequality, it is now possible to provide an *a priori* performance guarantee for the relative cost overage which depends only on the desired level of production output.

Lemma 6 (Relative cost overage). *For any target output $q > 0$, the relative cost overage is such that*

$$\frac{\hat{C}(q) - C(q)}{C(q)} \le \left(2\sqrt{\frac{1-p_{\min}}{\pi\ p_{\min}}}\right)\frac{1}{\sqrt{q}}, \tag{20}$$

where $p_{\min} = \min_{(i,m)\in\mathcal{I}\times\mathcal{M}}\{p_{i,m}\} > 0$.

In particular, the relative cost overage is $O(q^{-1/2})$, in the sense that it is inversely proportional to the square root of the target output. For example, if $p_{\min} = 10\%$, the upper bound for the relative cost overage (corresponding to the right-hand side in (20)) becomes $\left(6/\sqrt{\pi}\right)/\sqrt{q} \approx (3.3851)/\sqrt{q}$; this conservative bound drops below 10% for target outputs of 1146 units and beyond. The latter value corresponds to the input bound implied by Lemma 5; cf. footnote 2.

Example 3. In the setting of Examples 1 and 2, the solution to the envelope optimization problem, $\bar{\boldsymbol{x}}^*(q) = (q, 2q)$, produces the expected output $F(\bar{\boldsymbol{x}}^*(q))$ of about 94.6345 for $q = 100$ (resp., an expected output of about 983.2032 for $q = 1000$). Accordingly, the approximate solution of the selective matching problem is $\hat{\boldsymbol{x}}^*(100) = \lceil 100/94.6345\rceil(100, 200) \approx (105.67, 211.34)$ with relative output error $\varepsilon(100) \approx 5.499\% \le 5.67\%$ (resp., $\hat{\boldsymbol{x}}^*(1000) \approx (1017.1, 2034.2)$ with $\varepsilon(1000) \approx 1.692\% \le 1.71\%$). By (19) the upper bound of 5.67% (resp., 1.71%) also majorizes the relative cost overage incurred by the approximate solution $\hat{\boldsymbol{x}}^*(q)$ to the selective matching problem compared to the optimal solution $\boldsymbol{x}^*(q)$, for $q = 100$ (resp., for $q = 1000$).

3.6. Cost of selective matching

To gauge the cost of selective matching as a function of the grade cardinality $M = |\mathcal{M}|$, we consider a symmetric setting with maximum entropy, where all grades are attained with equal probability (i.e., $\mathbf{p}_i = (1/M, \ldots, 1/M)$ and $\gamma = 1$), and where the output weights and the unit costs are all equal (i.e., $(\eta, \gamma) = \mathbf{1}$). Assuming a symmetric input $\boldsymbol{x} = (N, N)$ for some $N \ge M$, Lemma 1 yields the expected grade-m output,

$$F_m(N, N) = \frac{N}{M} - \sqrt{\frac{N}{\pi M}\left(1 - \frac{1}{M}\right)}, \quad m \in \mathcal{M}.$$

Summing over all matching grades $m \in \mathcal{M}$, with $F(N, N) = MF_m(N, N)$, yields the (symmetric) *relative output loss* for selective-matching,

$$\mathcal{L}(N, M) = \frac{N - F(N, N)}{N} = \sqrt{\frac{M-1}{\pi N}} \le \frac{1}{\sqrt{\pi}},$$

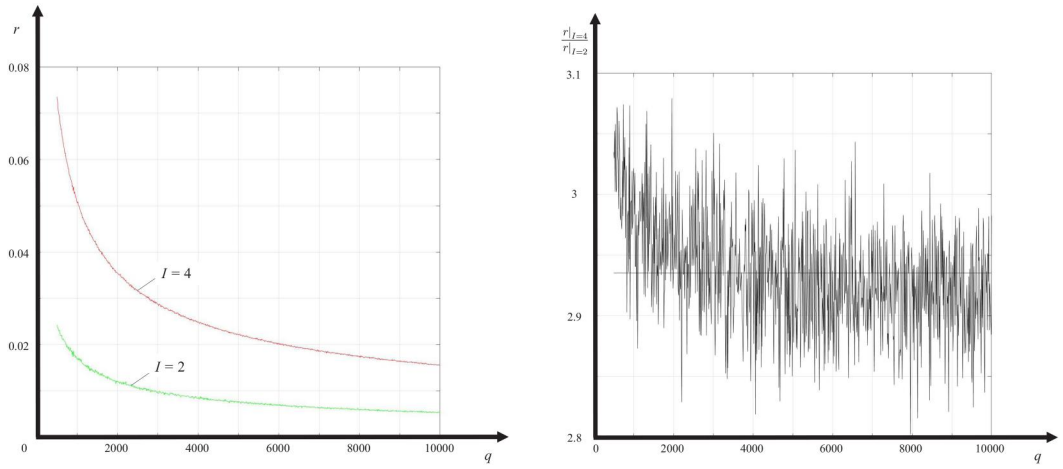
which is increasing (*sub*linearly) in $M \ge 2$ and decreasing in $N \ge M$. Since by symmetry (8) in Theorem 1 is automatically satisfied, the cost-minimizing input quantity $N^*(q)$ for both type-i inputs, given a desired output $q \ge M$, is entirely determined by the output-efficiency condition (7),

$$N^*(q) = q + \frac{M-1}{2\pi}\left(1 + \sqrt{1 + \frac{4\pi q}{M-1}}\right), \quad q \ge M.$$

[3] That $\hat{\boldsymbol{x}}^*(q) = \left[\frac{q}{F(\bar{\boldsymbol{x}}^*(q))}\right]\bar{\boldsymbol{x}}^*(q)$ is feasible in (*) can be seen as follows. By (5) the production function F is homogeneous of degree 1 (i.e., $F(\alpha\boldsymbol{x}) \equiv \alpha F(\boldsymbol{x})$, for any $\alpha > 0$). As a consequence,

$$F(\hat{\boldsymbol{x}}^*(q)) = F\left(\left[\frac{q}{F(\bar{\boldsymbol{x}}^*(q))}\right]\bar{\boldsymbol{x}}^*(q)\right) = \left[\frac{q}{F(\bar{\boldsymbol{x}}^*(q))}\right]F(\bar{\boldsymbol{x}}^*(q)) = q.$$

Thus, not only is $\hat{\boldsymbol{x}}^*(q)$ feasible in (*), but it also satisfies output efficiency (since $F(\hat{\boldsymbol{x}}^*(q)) \le q$ is in fact binding).



(a) (b)

Figure 4. (a) Relative error $r(\bar{x}^*(q))$, for $I = 2$ and $I = 4$, as a function of $q \in [500, 10,000]$; (b) Ratio of relative errors, for $I = 4$ to $I = 2$, as a function of $q \in [500, 10,000]$.

This output-driven demand leads to a *relative input overage*,

$$\mathcal{O}(q) = \frac{N^*(q) - q}{q} = \frac{M-1}{2\pi q}\left(1 + \sqrt{1 + \frac{4\pi q}{M-1}}\right),$$

which is decreasing in the target output quantity q, due to risk pooling across matching classes. The value of $\mathcal{O}(q)$ is also equal to the *relative cost overage* (i.e., $\mathcal{O}(q) = (C(q) - 2cq)/(2cq)$, for any $\mathbf{c} = (c, c)$ with $c > 0$) compared with perfect matching.[4] However, this relative overage increases almost linearly in the number of excess matching classes $M - 1$.

Example 4. When transitioning from $M = 20$ to $M = 40$ matching classes, the relative overage in input or cost increases by a factor of at least 1.9279 (for $q \geq 1$), all else equal. For $q \to \infty$, the factor approximates $(40 - 1)/(20 - 1) \approx 2.0526$. That is, for small target outputs q we observe a slight *sub*linear increase of the relative overage in the number of matching classes, which becomes somewhat *super*linear when target outputs are relatively large.

3.7. Multi-type matching

When there are more than two part types (i.e., for $\mathcal{I} = \{1, ..., I\}$ with $I > 2$), an explicit formula for the expected output under the normal approximation is no longer available. This in turn makes it more difficult to derive solid approximation results. The corresponding I-type envelope optimization problem can still be solved explicitly. Equations (13) and (14) remain valid, and they determine an optimal solution $\bar{\boldsymbol{x}}^*(q)$ of the I-type envelope optimization problem (**). The re-scaling in (15) still applies and so does (16) for the approximate cost. To illustrate the behavior of the approximation error $r(\bar{x}^*(q))$ for $I = 4$ compared to $I = 2$, we can "double" Example 1 by setting $\mathbf{p}_3 = \mathbf{p}_1$ and $\mathbf{p}_4 = \mathbf{p}_2$ for the additional component types $i = 3$ and $i = 4$. This ensures a fair comparison when the number of component types increases by a factor of two. Accordingly, the optimal solution to the envelope optimization problem for $I = 4$ becomes $\bar{x}^*(q) = (q, 2q, q, 2q)$ (compared to $\bar{x}^*(q) = (q, 2q)$ for $I = 2$; cf. Example 2). Figure 4(a) shows the behavior of the relative approximation error, $r(\bar{x}^*(q))$ (defined after Lemma 4), as a function of $q \in [500, 10000]$. The (true) expected output is obtained (approximately) as the sample-average using the corresponding normal distribution (for $N = 10,000$ joint samples) for each matching class, according to (9). Indeed, the law of large numbers guarantees a consistent approximation of the population mean by the sample mean.

The relative error, while increasing on average by a factor of about 2.94 on the given range (decreasing in q; see Figure 4(b)) drops below 5% shortly after passing $q = 1000$, so that for industrial-size problems the provided approximation continues to be reasonable. A more precise treatment of multi-type matching remains an interesting topic for further research.

4. Application

Consider the production of the oscillator for a mechanical watch movement where hairsprings (i.e., type-1 components) and balance wheels (i.e., type-2 components) are mated across $M \geq 2$ matching classes, where M is an even number. Modern hairsprings are often made of silicon for its antimagnetic properties and are delivered in wafers, with substantial variations in stiffness across any sample. We assume that the spring coefficient S_1 [in Nm/rad] is random with realizations in the interval $[a_1, b_1]$. Balance wheels are usually made from metal alloys (e.g., containing beryllium or gold among other constituents) and are characterized by

[4] Perfect matching happens when there is only a single matching class (i.e., for $M = 1$); then all parts are matched. Naturally, this indiscriminate approach propagates lax input tolerances into the assembly, which drives up other costs (e.g., in terms of rework or rejected output parts for a failure to satisfy tight output tolerances).

the moment of inertia S_2 [in kg m^2]. The latter can be considered a random variable with realizations in the interval $[a_2, b_2]$. The prespecified constants a_i and b_i are such that $0 < a_i < b_i < \infty$. As already pointed out in footnote 1, the key characteristic of the assembled oscillator is the oscillation period,

$$T = 2\pi\sqrt{S_2/S_1},$$

which, in itself a random variable, should have realizations close to the target period t_0 for the particular movement under consideration. For example, $1/t_0 \in \{2.5, 3, 3.5, 4, 5\}$Hz corresponds to today's most common beat rates, BR $= 2\ (3600\text{s})/t_0$ in $\{18,000, 21,600, 25,200, 28,800, 36,000\}$, which are measured in half-periods ("ticks") per hour.

4.1. Model identification

The four sets of primitives for the selective-assembly model, (i)–(iv), are now identified in turn, first the matching classes (i), then the parameters of the multinomial distribution for the sorting products (ii), the output values (iii), and finally the unit input costs (iv).

(i) Matching classes. The M type-i matching classes $\mathcal{C}_{i,m} = [s_{i,m-1}, s_{i,m})$, for $m \in \mathcal{M} \setminus \{M\}$, and $\mathcal{C}_{i,M} = [s_{i,M-1}, s_{i,M}]$ for $m = M$ feature $M+1$ breakpoints $s_{i,0}, ..., s_{i,M}$ such that

$$a_i = s_{i,0} < s_{i,1} < \cdots < s_{i,M-1} < s_{i,M} = b_i,$$

for $i \in \mathcal{I}$. According to Weber (2021), the optimal set of breakpoints, which minimize the maximum absolute deviation of T from t_0, in this setup is such that the maximum matching error e is equal over all matching classes, so

$$e = \max_{(s_1, s_2) \in \mathcal{C}_m} \left| 2\pi\sqrt{s_2/s_1} - t_0 \right|, \qquad m \in \mathcal{M},$$

where $\mathcal{C}_m = \mathcal{C}_{1,m} \times \mathcal{C}_{2,m}$ denotes the mth joint matching class. This approach requires a consistency condition, namely that the ratio of the endpoints $\beta = b_i/a_i$ is constant across the component types $i \in \mathcal{I}$, so

$$\frac{e}{t_0} = \frac{\beta^{1/M} - 1}{\beta^{1/M} + 1}.$$

The latter can always be achieved by increasing (resp., decreasing) one of the four endpoints by extending the range of one characteristic (or by adding a zero-value matching class). The optimal set of breakpoints is then given by

$$s_{i,2k} = \left(\tfrac{t_0+e}{t_0-e}\right)^{2k} a_i, \qquad s_{i,2k+1} = \begin{cases} s_{2,2k}[(2\pi)/(t_0-e)]^2, & \text{if } i = 1, \\ s_{1,2k}[(t_0+e)/(2\pi)]^2, & \text{if } i = 2, \end{cases}$$

for $k \in \{0, ..., M/2\}$ and $i \in \mathcal{I}$, where we limit attention to the case where M is even.

(ii) Distributional parameters. Given a distribution of the random part characteristics, the multinomial probabilities $p_{i,m}$, for $(i,m) \in \mathcal{I} \times \mathcal{M}$, follow directly from (1) in Section 2.1. Alternatively, it is possible to avoid distributional assumptions about the underlying distribution of characteristics and proceed in a fully data-driven way, based on repeated observation of realizations $\boldsymbol{x}_i^{(k)} = (x_{i,1}^{(k)}, ..., x_{i,M}^{(k)})$ for the random sorting product $\mathbf{X}_i$. This yields the maximum-likelihood estimates $\hat{\mathbf{p}}_i = (\hat{p}_{i,1}, ..., \hat{p}_{i,M})$ of the unknown parameters $\mathbf{p}_i = (p_{i,1}, ..., p_{i,M})$ of the multinomial distribution,

$$\hat{p}_{i,m} = \frac{\sum_{k=1}^{K} x_{i,m}^{(k)}}{\sum_{m \in \mathcal{M}} \sum_{k=1}^{K} x_{i,m}^{(k)}}, \tag{21}$$

where $k \in \{1, ..., K\}$ is the batch number, $x_{i,m}^{(k)}$ denotes the number of grade-m components in batch k, and K the number of observed batches. That is, the maximum-likelihood estimates for the component likelihoods in their respective type–grade combinations correspond simply to their respective frequencies over all observed samples.

Remark 5 (Off-spec components). Due to the empirical regularity that a received batch of input parts may well contain items that do not fit into any of the matching classes, the following question arises: What happens when relaxing the normalization in (2) to account for "off-spec" components? Specifically, if $x_i^{(k)}$ denotes the size of batch k, then the expression for $\hat{p}_{i,m}$ in (21) implicitly allows for off-spec components (which do not fit in any $\mathcal{C}_{i,m}$), since generally $\sum_{m \in \mathcal{M}} \sum_{k=1}^{K} x_{i,m}^{(k)} \le x_i^{(k)}$, while at the same time $\hat{p}_{i,1} + \cdots + \hat{p}_{i,M} = 1$ satisfying (2). However, to explicitly account for the possibility of "off-spec" components let

$$\hat{p}_{i,0} = 1 - \frac{\sum_{m \in \mathcal{M}} \sum_{k=1}^{K} x_{i,m}^{(k)}}{\sum_{k=1}^{K} x_i^{(k)}}, \qquad i \in \mathcal{I},$$

denote the observed frequency of type-i reject components, as the maximum-likelihood estimate of the true underlying reject probability $p_{0,i}$. Then by adding a matching class 0 with output weight η_0, by (13) one can simply rescale the approximate solution to the selective matching problem in (15) to account for the reject probability:

$$\hat{x}_i^*(q) = \frac{1}{1 - p_{i,0}} \left[\frac{q}{F(\bar{\boldsymbol{x}}^*(q))} \right] \bar{x}_i^*(q), \qquad i \in \mathcal{I}, \tag{22}$$

where $q > 0$ is any given output target and $\bar{x}_i^*(q)$ is given in Theorem 2 (for $p_{i,m} \equiv \hat{p}_{i,m}$). Hence, we can decouple the consideration of off-spec components from the main analysis (which is purely "on-spec") and then rescale the approximate solution *ex post*.

(iii) Net unit values. The weighted total output in (4) requires a vector of positive weights $\boldsymbol{\eta} = (\eta_1, ..., \eta_M)$. Since the quality of the assembled mechanical oscillators does not depend on the index m of the matching grade, it is most natural for the firm to set all $\eta_m = 1$ and consider standard output, i.e., $Q = Y_1 + \cdots + Y_M$. In standard selective-manufacturing applications, no grade is privileged over others (except possibly for a designated "reject" grade of negligible net unit value); in that case, the grade sorting satisfies the sole purpose of sharpening the precision of the assembly

output given the quality imprecisions of the component inputs.

(iv) Unit input costs. The cost function in the selective matching problem $(*)$ depends on the vector $\mathbf{c} = (c_1, c_2)$ of the unit costs for the two component types. However, as evident from Theorem 1, for any given output quantity $q > 0$ the optimal input vector $\boldsymbol{x}^*(q)$ depends only on the ratio $\gamma = c_1/c_2$ of the unit costs, so that there is no need to know their absolute values to determine the composition of the firm's optimal input portfolio. We assume that the unit cost of a balance wheel is about 10 to 20 times the unit cost of a silicon hairspring, so $\gamma \in [0.05, 0.1]$. In the numerical example, we set $\mathbf{c} = (1, 10)$, so $\gamma = 0.1$; the results for $\gamma = 0.05$ are qualitatively very similar.

4.2. Data generation

To generate synthetic samples representing the observed stochastic supply characteristics, assume that the true distribution of the normalized type-i component characteristic,

$$Z_i = \frac{S_i - \mu_{S_i}}{\sigma_{S_i}},$$

follows a standard normal distribution with cdf $\Phi(\cdot)$, where $\mu_{S_i} \in (a_i, b_i)$ is the mean and $\sigma_{S_i} > 0$ the standard deviation of the (stationary) component population. Given the distributional assumption, the multinomial probabilities (conditional on parts being on-spec) can be obtained by adjusting (1) as follows:

$$p_{i,m} = \frac{1}{1 - p_{i,0}} \left(\Phi\left(\frac{s_{i,m} - \mu_{S_i}}{\sigma_{S_i}} \right) - \Phi\left(\frac{s_{i,m-1} - \mu_{S_i}}{\sigma_{S_i}} \right) \right), \quad (i, m) \in \mathcal{I} \times \mathcal{M}.$$

Note that with positive probability some realizations s_i of the random component characteristic S_i may be "off-spec," in which case we use input scaling as noted in Section 4.1 (ii). The off-spec probability in the component population is $p_{i,0} = 1 - (\Phi((b_i - \mu_{S_i})/\sigma_{S_i}) - \Phi((a_i - \mu_{S_i})/\sigma_{S_i}))$. As explained in Remark 5 on off-spec components, the renormalization of the multinomial probabilities has no effect on the likelihood ratios in (13), so that it is enough to rescale the approximate solution as in (22), an approach which is implemented here. Finally, we assume that both distributions are centered, so $\mu_{S_i} = (a_i + b_i)/2$ for $i \in \mathcal{I}$, but that the coefficient of variation is twice as large for hairsprings than for balance wheels, with $\sigma_{S_1}/\mu_{S_1} = 2\%$ and $\sigma_{S_2}/\mu_{S_2} = 1\%$.

4.3. Performance analysis

Let $(a_1, b_1) = (2.94, 3.06) \times 10^{-7}$ [in Nm/rad] be the boundaries of the hairspring stiffness and $(a_2, b_2) = (4.655, 4.845) \times 10^{-10}$ [in kg m^2] the boundaries of the balance-wheel inertia (in either case corresponding to a spread of 2% around a central value), so $\beta = a_2/a_1 = b_2/b_1$, as required in Section 4.1 (i) where we also provide the optimal set of breakpoints for the movement frequency of $1/t_0 =$ 4 Hz, and thus the optimal matching classes $\mathcal{C}_{i,m}$. Clearly, the maximum matching error e decreases in the number of matching classes $M \geq 2$, which is a choice variable. For practical purposes it is useful to interpret e/t_0 in terms of "deviation per day" (dpd) [in seconds/day, i.e., s/d] by multiplying numerator and denominator by 345,600 $(= 24 \times 3600 \times 4)$. Thus, since $4 \times t_0 = 1$s, we obtain that $e/t_0 = \text{dpd} = (345,600 \times e)/(1\text{d})$, so

$$M(\text{dpd}) = \left\lceil \left(\log\left(\frac{1 + \text{dpd}}{1 - \text{dpd}} \right) \right)^{-1} \log(\beta) \right\rceil, \quad (23)$$

where $\beta \approx 1.040816327$. That is, (23) provides the number of matching classes necessary to guarantee a given dpd, as shown in Figure 5. For example, it is $M(60) = 29$, $M(30) = 58$, and $M(15) = 116$. Finally, for a rather fine partition of part characteristics into $M(5) = 346$ matching classes it would be possible (at least theoretically) to achieve the Contrôle Officiel Suisse des Chronomètres (COSC) standard of ± 5 s/d without any watchmaker intervention after assembly. Indeed, the COSC requires mechanical watch movements to run within −4 s and +6 s per day to issue a "chronometer" certification, which is equivalent to a dpd of five. For practical purposes, and to stay exactly in the framework of Weber (2021) used in Section 4.1 (i), we round up to the smallest even number of matching classes, thus considering $2\lceil M(\text{dpd})/2 \rceil$ instead of $M(\text{dpd})$ in (23), as indicated in Figure 5.

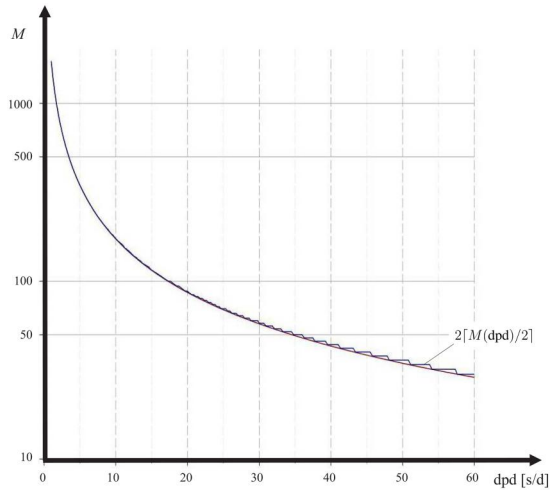


Figure 5. Smallest (even) number of matching classes to guarantee a maximum daily deviation, dpd $\in \{1, \ldots, 60\}$ [s/d]. COSC chronometer standard is achieved for dpd ≤ 5 [s/d].

To guarantee practical implementability and feasibility, the approximate solution $\hat{\boldsymbol{x}}^*(q)$ in (15) is rounded to an input vector $\lceil \hat{\boldsymbol{x}}^*(q) \rceil$, with the next-highest integers as components (i.e., $\lceil \hat{x}_i^*(q) \rceil$, for $i \in \mathcal{I}$). An integer discretization (determined by optimality) is also applied to the solution $\boldsymbol{x}^*(q)$ of the original selective matching problem $(*)$. This solution discretization accounts for some of the fluctuations in Figure 6, which depicts the optimal costs (in terms of C and $\hat{C}$; cf. (17)) and optimal inputs. To ensure comparability, the costs are normalized to the input cost $C_0(q) = (c_1 + c_2)q$ without matching classes (where $M = 1$), and the input solutions are

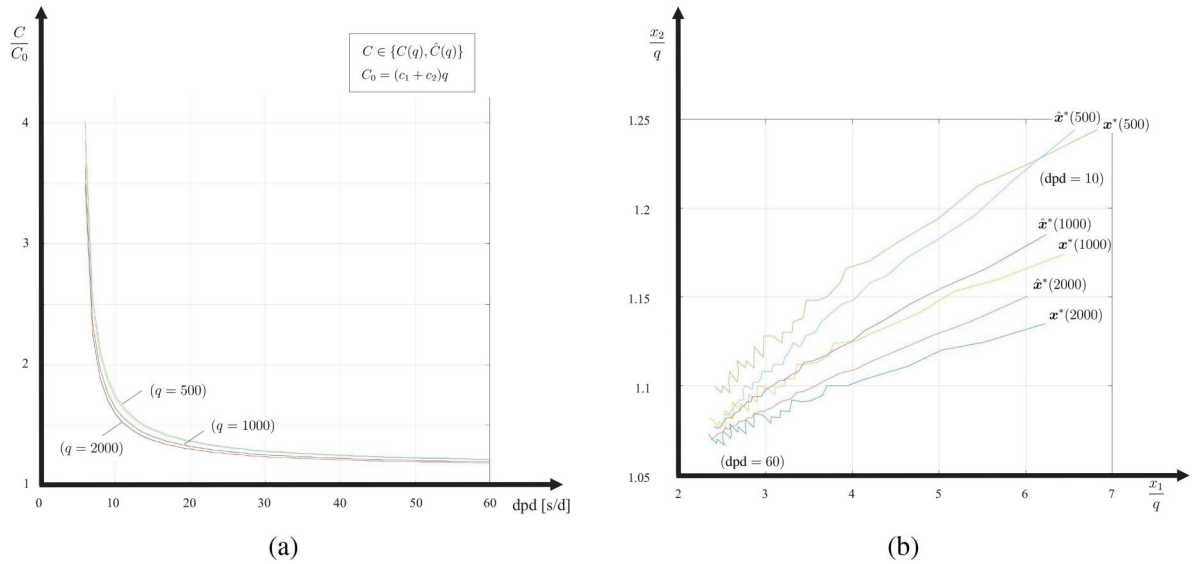


Figure 6. Optimal and approximately optimal cost (a) and input (b) as a solution to $(*)$ and $(**)$, for a target quantity $q \in \{500, 1000, 2000\}$ units and for an output tolerance dpd ≤ 60 [s/d].

normalized to the "overhead-free" input q for all types $i \in \mathcal{I}$ in that case. As shown in Figure 6(a), the normalized cost decreases in the target quantity q, which reflects a risk-pooling effect implied by the sublinear increase of the standard deviation of the part characteristics in the amount ordered. It is also apparent that the deviation between the optimal cost C (for optimal integer inputs determined via grid search) and the cost approximation $\hat{C}$ is quite negligible for realistic output targets. It is also apparent that as the dpd-precision requirement becomes more demanding (i.e., smaller), the normalized cost increases very fast—well exceeding four for COSC chronometer specifications. Figure 6(b) illustrates (for $\gamma = 0.1$) that these cost increases are driven by a massive over-ordering of the 10-times cheaper hairsprings as "buffer components." The over-ordering increases both in the production target and the precision requirement.

5. Conclusion

The envelope approximation of the firm's output objective proved effective in determining a simple closed-form solution for the cost-optimal input portfolio in (13) and (22), which includes the possibility of scrap parts. In that solution, somewhat intentionally, the notation of the index set $\mathcal{I}$ for the different part types (e.g., hairsprings and balance wheels in the watchmaking application) has been used for displaying the results. Indeed, an extension to $\mathcal{I} = \{1, ..., I\}$ with $I > 2$ part types (e.g., balls, inner/outer rings, and cages, in the case of a ball-bearing assembly) is quite straightforward, and requires no adjustment of these formulas. Although this naturally applies to the envelope output in (9), the same cannot be said for the expected output in (6), for which a simple closed-form expression may not be within easy reach for more than two input part types.

From a practical viewpoint, a doubling of the number of matching classes tends to roughly double the precision of output parts.[5] As noted in Section 3.6, it also leads to almost a doubling of the relative input overage, which is the number of extra units required compared with perfect matching, where the latter corresponds to random matching with equal inputs without quality concerns—equivalent to using a single matching class. That is, all else equal, improving the output tolerance by a certain factor basically multiplies the overage cost by that factor. In addition, one should not overlook the fact that unused parts appear in matching classes for which the counterpart of the other type is stocked-out and on back-order. However, the residual inventory vector serves as a buffer for the next order and thus increases expected output relative to an empty stock. A formulation of the corresponding dynamic reordering policy, taking into account parts residuals, is an interesting topic for future research.

Acknowledgments

The author would like to thank several anonymous referees, as well as participants of the 2022 CORS/INFORMS International Conference in Vancouver, Canada, and the 2022 EURO Conference in Espoo, Finland, for helpful suggestions.

Notes on contributor

Thomas A. Weber is full professor of Operations, Economics and Strategy at the Swiss Federal Institute of Technology in Lausanne (EPFL). Earlier he was a faculty member in the Economics and Finance Group of the Department of Management Science and Engineering at Stanford University. Prof. Weber is an Ingénieur des Arts et Manufactures (École Centrale Paris) and a Diplom-Ingenieur in Electrical Engineering (Technical University Aachen). He holds master's degrees in Technology and Policy and Electrical Engineering and Computer Science from MIT, and a PhD in Applied Economics and Managerial Science from the Wharton School of the University of Pennsylvania. He was a visiting faculty member in Economics at Cambridge University and in Mathematics at Moscow State University.

[5]See, e.g., the deviation-per-day (dpd) precision of a mechanical watch movement in Figure 5, which increases by a factor of two (from 37 s to 17.5 s) when doubling the matching classes (from $M = 50$ to $M = 100$).

Between 1998 and 2002, he was a senior consultant with the Boston Consulting Group. His current research interests include the economics of information and uncertainty, robust optimization, and strategy. Prof. Weber's more than 100 scholarly publications have appeared, for example, in *American Economic Journal: Microeconomics*, *Decision Support Systems*, *Economics Letters*, *Economic Theory*, *Health Care Management Science*, *Information Systems Research*, *Journal of Economic Dynamics and Control*, *Journal of Environmental Economics and Management*, *Journal of Management Information Systems*, *Journal of Mathematical Economics*, *Journal of Optimization Theory and Applications*, *Journal of Regulatory Economics*, *Management Science*, *Mathematical Reviews*, *Medical Decision Making*, *Operations Research*, *Operations Research Letters*, *Optimal Control: Applications and Methods*, *Physical Review E*, *Sustainable Production and Consumption*, as well as *Theory and Decision*. He has been associate editor for *Management Science* and is the author of *Optimal Control Theory with Applications in Economics* (MIT Press, 2011).

References

Asha, A., Kannan, S.M. and Jayabalan, V. (2008) Optimization of clearance variation in selective assembly for components with multiple characteristics. *International Journal of Advanced Manufacturing Technology*, **38**(9–10), 1026–1044.

Berge, C. (1963) *Topological Spaces*. Oliver and Boyd, Edinburgh, UK.

Berry, A.C. (1941) The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, **49**(1), 122–136.

Buckingham, E. (1921) *Principles of Interchangeable Manufacturing*. Industrial Press, New York, NY.

Caputo, A.C. and Di Salvo, G. (2019) An economic decision model for selective assembly. *International Journal of Production Economics*, **207**, 56–69.

Coullard, C.R., Gamble, A.B. and Jones, P.C. (1998) Matching problems in selective assembly operations. *Annals of Operations Research*, **76**, 95–107.

Duenyas, I., Keblis, M.F. and Pollock, S.M. (1997) Dynamic type mating. *Management Science*, **43**(6), 751–763.

Esseen, C.-G. (1942) On the Liapunoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik*, **A28**(9), 1–19.

Esseen, C.-G. (1956) A moment inequality with an application to the central limit theorem. *Skandinavisk Aktuarietidskrift*, **39**, 160–170.

Kannan, S.M. and Raja Pandian, G. (2021) A new selective assembly model for achieving specified clearance in radial assembly. *Materials Today: Proceedings*, **46**(17), 7411–7417.

Kwon, H.-M., Kim, K.-J. and Jeya Chandra, M. (1999) An economic selective assembly procedure for two mating components with equal variance. *Naval Research Logistics*, **46**(7), 809–821.

Leontief, W.W. (1941) *The Structure of American Economy 1919–1929: An Empirical Application of Equilibrium Analysis*. Harvard University Press, Cambridge, MA.

Mansoor, E.M. (1961) Selective assembly – Its analysis and applications. *International Journal of Production Research*, **1**(1), 13–24.

Matsuura, S. (2011) Optimal partitioning of probability distributions under general convex loss functions in selective assembly. *IIE Transactions*, **40**(9), 1545–1560.

Mease, D., Nair, V.N. and Sudjianto, A. (2004) Selective assembly in manufacturing: Statistical issues and optimal binning strategies. *Technometrics*, **46**(2), 165–175.

Pugh, G.A. (1986) Partitioning for selective assembly. *Computers and Industrial Engineering*, **11**(1–4), 175–179.

Rezaei Aderiani, A., Wärmefjord, K., Söderberg, R. and Lindkvist, L. (2019) Developing a selective assembly technique for sheet metal assemblies. *International Journal of Production Research*, **57**(22), 7174–7188.

Shevtsova, I. (2011) On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. Working Paper, Lomonosov Moscow State University, Moscow, Russia. *arXiv: 1111.6554v1*.

Tan, M.H.Y. and Wu, C.F.J. (2012) Generalized selective assembly. *IIE Transactions*, **44**(1), 27–42.

Weber, T.A. (2021) Minimum-error classes for matching parts. *Operations Research Letters*, **49**(1), 106–112.

Weber, T.A. (2022) Optimal matching of random parts. *Journal of Mathematical Economics*, **101**, 102688:1–14.

Zorich, V.A. (2004) *Mathematical Analysis I*. Springer, New York, NY.

Appendix: Proofs

Proof of Lemma 1. Let $m \in \mathcal{M}$ be a given matching grade and let $\boldsymbol{x} \in \mathbb{R}^2_+$ be an arbitrary input vector. By (3) the firm's random grade-m output (conditional on $\boldsymbol{x}$),

$$\begin{aligned} Y_m &= \min\{X_{1,m}, X_{2,m}\} = X_{1,m} + \min\{0, X_{2,m} - X_{1,m}\} \\ &= X_{1,m} + \min\{0, \Delta_m\}, \end{aligned}$$

can be viewed as a sum of the grade-m type-1 output and the negative part of the grade-m parts defect, $\Delta_m = X_{2,m} - X_{1,m}$. As the difference of two correlated and normally distributed random variables, the latter follows a normal distribution with mean $\delta_m = \mu_{2,m} - \mu_{1,m}$, and variance $\sigma_m^2 = \sigma_{1,m}^2 + \sigma_{2,m}^2$. Thus, the expected grade-m output is $F_m(\boldsymbol{x}) = \mathbb{E}[Y_m|\boldsymbol{x}] = \mu_{1,m} + \mathbb{E}[\min\{0, \Delta_m\}|\boldsymbol{x}]$, where

$$\mathbb{E}\left[\min\{0, \Delta_m\}\middle|\boldsymbol{x}\right] = \delta_m + \sigma_m \ \mathbb{E}\left[\min\left\{-\frac{\delta_m}{\sigma_m}, \frac{\Delta_m - \delta_m}{\sigma_m}\right\}\middle|\boldsymbol{x}\right].$$

On the other hand, with $(\Delta_m - \delta_m)/\sigma_m$ following a standard normal distribution,

$$\mathbb{E}\left[\min\left\{-\frac{\delta_m}{\sigma_m}, \frac{\Delta_m - \delta_m}{\sigma_m}\right\}\middle|\boldsymbol{x}\right] = -\frac{\delta_m}{\sigma_m}\left(1 - \Phi\left(-\frac{\delta_m}{\sigma_m}\right)\right) + \int_{-\infty}^{-\delta_m/\sigma_m} \xi \ d\Phi(\xi).$$

Taking into account the symmetry of $\phi(\cdot)$ the last term can be computed explicitly,

$$\begin{aligned} \int_{-\infty}^{-\delta/\sigma_m} \xi \ d\Phi(\xi) &= \left[-\frac{\exp(-\xi^2/2)}{\sqrt{2\pi}}\right]_{\xi\to-\infty}^{\xi=-\delta_m/\sigma_m} = -\phi\left(-\frac{\delta_m}{\sigma_m}\right) \\ &= -\phi\left(\frac{\delta_m}{\sigma_m}\right). \end{aligned}$$

Hence, the expected grade-m output becomes

$$F_m(\boldsymbol{x}) = \mu_{1,m} + \delta_m - \delta_m\left(1 - \Phi\left(-\frac{\delta_m}{\sigma_m}\right)\right) - \sigma_m\phi\left(\frac{\delta_m}{\sigma_m}\right).$$

By virtue of the fact that $\Phi(-\xi) = 1 - \Phi(\xi)$, for all $\xi \in \mathbb{R}$, the preceding expression simplifies to

$$F_m(\boldsymbol{x}) = \Phi_m \ \mu_{1,m} + (1 - \Phi_m)\mu_{2,m} - \phi_m\sigma_m,$$

where we have set $\Phi_m = \Phi(\delta_m/\sigma_m)$ and $\phi_m = \phi(\delta_m/\sigma_m)$. This establishes the claim. □

Proof of Lemma 2. Let $m \in \mathcal{M}$ be a given matching grade and let $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_+$ be an arbitrary input vector. For any $i \in \mathcal{I}$, the number of type-i components in grade-m follows a normal distribution with cdf $G_{i,m}(\cdot|x_i)$. The cdf of the random grade-m output $Y_m = \min\{X_{1,m}, X_{2,m}\}$ becomes

$$\begin{aligned} \mathbb{P}(Y_m \le \xi|\theta) &= G_{1,m}(\xi|x_1) + G_{2,m}(\xi|x_2) - G_{1,m}(\xi|x_1)G_{2,m}(\xi|x_2), \\ &\xi \ge 0. \end{aligned}$$

Consider now an increase of type-i input from x_i to $\hat{x}_i > x_i$, and note that

$$\frac{\partial G_{i,m}(\xi|x_i)}{\partial x_i} = -\frac{p_{i,m} + (\xi/x_i)}{2\ \sigma_{i,m}}\phi\left(\frac{\xi - \mu_{i,m}}{\sigma_{i,m}}\right) < 0, \quad x_i \ge 0.$$

This implies First-Order Stochastic Dominance (FOSD) of the type-i quantity distribution after the quantity increase, $G_{i,m}(\xi|\hat{x}_i) < G_{i,m}(\xi|x_i)$, $\xi \ge 0$. Thus, if we denote by $(\hat{X}_{1,m}, \hat{X}_{2,m})$ the grade-m

yields under $\hat{\boldsymbol{x}}$, then $(X_{1,m}, X_{2,m}) \prec_{\text{FOSD}} (\hat{X}_{1,m}, \hat{X}_{2,m})$, where $\prec_{\text{FOSD}}$ marks the (strict) FOSD-order. But this implies that $Y_m = \min\{X_{1,m}, X_{2,m}\} \prec_{\text{FOSD}} \min\{\hat{X}_{1,m}, \hat{X}_{2,m}\} = \hat{Y}_m$, so that we can now compare the expected values, $F_m(\boldsymbol{x}) = \mathbb{E}[Y_m|\boldsymbol{x}] < \mathbb{E}[\hat{Y}_m|\hat{\boldsymbol{x}}] = F_m(\hat{\boldsymbol{x}})$, establishing the claimed monotonicity. □

Proof of Theorem 1. Let $(q, c) \in \mathbb{R}_{++} \times \mathbb{R}^2_{++}$. The Lagrangian associated with the cost-minimization problem $(*)$ is $L(\boldsymbol{x}; \lambda) = c_1 x_1 + c_2 x_2 - \lambda(F(\boldsymbol{x}) - q)$, where the adjoint variable $\lambda \geq 0$ measures the "shadow cost" of relaxing the output-attainment constraint. The first-order necessary optimality conditions are therefore

$$\frac{\partial L(\boldsymbol{x}; \lambda)}{\partial x_i} = c_i - \lambda \frac{\partial F(\boldsymbol{x})}{\partial x_i} = 0, \tag{24}$$

for all $i \in \mathcal{I}$ (Zorich, 2004, p. 527ff). Since by Lemma 2 it is $\partial F(\boldsymbol{x})/\partial x_i > 0$, which in particular applies to input vectors $\boldsymbol{x}$ such that $F(\boldsymbol{x}) = q$, the output-attainment constraint $F(\boldsymbol{x}) \geq q$ must be binding, so that (by complementary slackness) necessarily $\lambda > 0$ and (7) must be satisfied if $\boldsymbol{x} = \boldsymbol{x}^*(q)$ is to be a solution candidate for the selective matching problem $(*)$. Substitution of (5) and (6) (in Lemma 1) into Eq. (24) for $i \in \{1, 2\}$ yields the marginal rate of technical substitution,

$$\frac{\partial F(\boldsymbol{x})/\partial x_1}{\partial F(\boldsymbol{x})/\partial x_2} = \frac{\sum_{m \in \mathcal{M}} \eta_m \partial F_m(\boldsymbol{x})/\partial x_1}{\sum_{m \in \mathcal{M}} \eta_m \partial F_m(\boldsymbol{x})/\partial x_2} = \frac{c_1}{c_2} = \gamma,$$

where

$$\begin{aligned}
\frac{\partial F_m(\boldsymbol{x})}{\partial x_1} &= p_{1,m}\Phi_m - \delta_m \frac{\partial \Phi_m}{\partial x_1} - \sigma_m \frac{\partial \phi_m}{\partial x_1} - \frac{p_{1,m}(1 - p_{1,m})}{2\ \sigma_m}\ \phi_m, \\
\frac{\partial F_m(\boldsymbol{x})}{\partial x_2} &= p_{2,m}(1 - \Phi_m) - \delta_m \frac{\partial \Phi_m}{\partial x_2} - \sigma_m \frac{\partial \phi_m}{\partial x_2} - \frac{p_{2,m}(1 - p_{2,m})}{2\ \sigma_m}\ \phi_m,
\end{aligned}$$

and

$$-\delta_m \frac{\partial \Phi_m}{\partial x_i} = \sigma_m \frac{\partial \phi_m}{\partial x_i} = \frac{\delta_m}{\sigma_m}\left((-1)^{i+1}\ p_{i,m} + \frac{\delta_m}{\sigma_m}\frac{p_{i,m}(1 - p_{i,m}) - 2\ C_{i,m}}{2\ \sigma_m}\right)\phi_m,\ \ i \in \mathcal{I}.$$

As a result,

$$\begin{aligned}
\frac{\partial F_m(\boldsymbol{x})}{\partial x_1} &= p_{1,m}\left(\Phi_m - \frac{1 - p_{1,m}}{2\ \sigma_m}\ \phi_m\right) \text{ and} \\
\frac{\partial F_m(\boldsymbol{x})}{\partial x_2} &= p_{2,m}\left(1 - \Phi_m - \frac{1 - p_{2,m}}{2\ \sigma_m}\ \phi_m\right),
\end{aligned}$$

which yields Eq. (8) for $\boldsymbol{x} = \boldsymbol{x}^*(q)$. This concludes our proof. □

Proof of Lemma 3. Fix any input bundle $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_+$. By Jensen's inequality, it is

$$\mathbb{E}[Y_m|\boldsymbol{x}] = \mathbb{E}\big[\min\{X_{1,m}, X_{2,m}\}|\boldsymbol{x}\big] \leq \min\big\{\mathbb{E}[X_{1,m}|x_1], \mathbb{E}[X_{2,m}|x_2]\big\},\ \ m \in \mathcal{M}.$$

Hence, using (5) and the linearity of the expectation operator, one obtains

$$\begin{aligned}
F(\boldsymbol{x}) = \mathbb{E}[Q|\boldsymbol{x}] &= \sum_{m \in \mathcal{M}} \eta_m \mathbb{E}[Y_m|\boldsymbol{x}] \leq \sum_{m \in \mathcal{M}} \eta_m \min\{p_{1,m}x_1, p_{2,m}x_2\} \\
&= \bar{F}(\boldsymbol{x}),
\end{aligned}$$

which concludes our proof. □

Proof of Lemma 4. Let $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_{++}$. Consider first the case where $\delta_m(\boldsymbol{x}) = p_{2,m}x_2 - p_{1,m}x_1 \geq 0$, for some $m \in \mathcal{M}$. Then by Lemma 1 it is

$$\begin{aligned}
0 &\leq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\} - F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \\
&\leq \phi\left(\frac{\delta_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})}\right) - \frac{\delta_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})}\left(1 - \Phi\left(\frac{\delta_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})}\right)\right).
\end{aligned}$$

On the other hand, since $(d/d\xi)(\phi(\xi) - \xi(1 - \Phi(\xi))) = -(1 - \Phi(\xi)) < 0$, for all $\xi \in \mathbb{R}$, the expression on the right-hand side of the preceding equality is downward-sloping in $\delta_m/\sigma_m > 0$, so

$$\delta_m(\boldsymbol{x}) \geq 0 \ \ \Rightarrow \ \ 0 \leq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\} - F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \leq \frac{1}{\sqrt{2\pi}} = \phi(0). \tag{25}$$

A similar reasoning applies for $\delta_m(\boldsymbol{x}) \leq 0$, in which case Lemma 1 yields

$$0 \leq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\} - F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \leq \phi\left(\frac{\delta_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})}\right) + \frac{\delta_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})}\ \Phi\left(\frac{\delta_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})}\right).$$

Thus, taking account of the fact that $(d/d\xi)(\phi(\xi) + \xi\Phi(\xi)) = \Phi(\xi) > 0$, for all $\xi \in \mathbb{R}$, the expression on the right-hand side of the preceding inequality must be upward-sloping in $\delta_m/\sigma_m < 0$, whence

$$\delta_m(\boldsymbol{x}) \leq 0 \ \ \Rightarrow \ \ 0 \leq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\} - F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \leq \frac{1}{\sqrt{2\pi}} = \phi(0). \tag{26}$$

Combining (25) and (26), and repeating this exercise for all grades, yields that in fact

$$0 \leq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\} - F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \leq \frac{1}{\sqrt{2\pi}}, \quad m \in \mathcal{M}. \tag{27}$$

Using the definitions of the weighted output F and the (equally weighted) envelope output $\bar{F}$, one obtains by mere summation that $0 \leq \bar{F}(\boldsymbol{x}) - F(\boldsymbol{x}) \leq \sigma(\boldsymbol{x})/\sqrt{2\pi}$, where $\sigma(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} \sigma_m(\boldsymbol{x})$. The (maximum) approximation error is therefore at most linear in the aggregate standard deviation $\sigma(\boldsymbol{x})$, so $\frac{R(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \leq 1/\sqrt{2\pi}$, which concludes our proof. □

Proof of Lemma 5. To exclude trivialities, we restrict attention to positive inputs, so $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_{++}$. By (27) we have

$$0 \leq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\} - F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \leq \frac{1}{\sqrt{2\pi}}, \quad m \in \mathcal{M},$$

so that

$$\begin{aligned}
\frac{F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} &\geq \frac{\min\{p_{1,m}x_1, p_{2,m}x_2\}}{\sigma_m(\boldsymbol{x})} - \frac{1}{\sqrt{2\pi}} \\
&\geq \sqrt{\frac{\min\{p_{1,m}x_1, p_{2,m}x_2\}}{2 - p_{1,m} - p_{2,m}}} - \frac{1}{\sqrt{2\pi}}, \quad m \in \mathcal{M}.
\end{aligned}$$

On the other hand, if we set $p_{\min} = \min_{(i,m) \in \mathcal{I} \times \mathcal{M}}\{p_{i,m}\} > 0$, then

$$\sqrt{\frac{\min\{p_{1,m}x_1, p_{2,m}x_2\}}{2 - p_{1,m} - p_{2,m}}} \geq \sqrt{\frac{p_{\min}/2}{1 - p_{\min}}}\sqrt{\min\{x_1, x_2\}}, \quad m \in \mathcal{M},$$

where the right-hand side of the preceding inequality does not depend on the grade m. Hence, it is easy to verify that

$$\begin{aligned}
\min\{x_1, x_2\} \geq \frac{4}{\pi}\left(\frac{1 - p_0}{p_0}\right)^2 \\
\Rightarrow \quad \frac{F_m(\boldsymbol{x})}{\sigma_m(\boldsymbol{x})} \geq \sqrt{\frac{p_0/8}{1 - p_0}}\sqrt{\min\{x_1, x_2\}}, \quad m \in \mathcal{M}.
\end{aligned}$$

The latter implies:

$$\min\{x_1, x_2\} \geq \frac{4}{\pi}\left(\frac{1 - p_{\min}}{p_{\min}}\right) \quad \Rightarrow \quad \frac{F(\boldsymbol{x})}{\sigma(\boldsymbol{x})} \geq \sqrt{\frac{p_{\min}/8}{1 - p_{\min}}}\sqrt{\min\{x_1, x_2\}}.$$

Combining this with Lemma 4, the relative error decreases in the smallest input,

$$r(\boldsymbol{x}) = \frac{R(\boldsymbol{x})/\sigma(\boldsymbol{x})}{F(\boldsymbol{x})/\sigma(\boldsymbol{x})} \leq \sqrt{\frac{1 - p_{\min}}{p_{\min}}}\frac{2/\sqrt{\pi}}{\sqrt{\min\{x_1, x_2\}}},$$

as long as the scale of the application is sufficiently large, so

$$\min\{x_1, x_2\} \geq \frac{4}{\pi}\left(\frac{1-p_{\min}}{p_{\min}}\right).$$

Thus, for any $\varepsilon \in (0,1)$:

$$\min\{x_1, x_2\} \geq \underline{x}(\varepsilon) = \frac{4}{\pi}\left(\frac{1-p_{\min}}{p_{\min}}\right)\max\left\{1, \frac{1}{\varepsilon^2}\right\} = \frac{4}{\pi\varepsilon^2}\left(\frac{1-p_{\min}}{p_{\min}}\right)$$
$$\Rightarrow \quad r(\boldsymbol{x}) \leq \varepsilon,$$

as claimed. But this means that

$$\min\{x_1, x_2\} \geq \underline{x}(\varepsilon) \quad \Rightarrow \quad \frac{\bar{F}(x_1, x_2)}{1+\varepsilon} \leq F(x_1, x_2),$$

which in turn implies the minimal output objective,

$$\underline{q}(\varepsilon) = \frac{\bar{F}(\underline{x}(\varepsilon), \underline{x}(\varepsilon))}{1+\varepsilon} = \left(\sum_{\ell=1}^{L} \hat{\eta}_\ell \min\{\rho_\ell, 1\}\right)\frac{\underline{x}(\varepsilon)}{1+\varepsilon}.$$

Finally, we note that $\underline{x}(\varepsilon)$ is $O(\varepsilon^{-2})$, so that for small $\varepsilon > 0$ the minimum envelope input $\underline{q}(\varepsilon)$ is also $O(\varepsilon^{-2})$, as claimed. This concludes the proof. □

Proof of Theorem 2. As a positive linear combination of concave functions, the envelope output,

$$\bar{F}(\boldsymbol{x}) = \sum_{\ell=1}^{L} \hat{\eta}_\ell \min\{\rho_\ell x_1, x_2\}, \quad \boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_+,$$

is also a concave function. Thus, for any $q > 0$ the upper contour set, $\mathcal{U}(q) = \left\{\boldsymbol{x} \in \mathbb{R}^2_+ : \bar{F}(\boldsymbol{x}) \geq q\right\}$, is convex. In addition, it is also possible to achieve at least as much envelope output with more inputs, so

$$\boldsymbol{x} \in \mathcal{U}(q) \quad \Rightarrow \quad \boldsymbol{x} + \mathbb{R}^2_+ \subset \mathcal{U}(q).$$

As a result, the upper contour set $\mathcal{U}(q)$ is a polygonal chain (i.e., a connected series of line segments),

$$\mathcal{P}(q) = \left\{(x_1, x_2) \in \mathbb{R}^2_+ : x_1 \in \left[q\bar{x}_1^L, q\bar{x}_1^1\right],\ x_2 = q \max_{\ell \in \{1,\ldots,L\}} \left\{\bar{x}_2^\ell - s_\ell\left(x_1 - \bar{x}_1^\ell\right)\right\}\right\},$$

in the sense that $\mathcal{U}(q) = \mathcal{P}(q) + \mathbb{R}^2_+$, where the vertices $\bar{\boldsymbol{x}}^\ell = (\bar{x}_1^\ell, \bar{x}_2^\ell)$ of the polygonal chain $\mathcal{P}(1)$ are implied by the conditions

$$\rho_\ell \bar{x}_1^\ell = \bar{x}_2^\ell \quad \text{and} \quad \bar{F}(\bar{\boldsymbol{x}}^\ell) = 1, \tag{28}$$

for $\ell \in \{1, \ldots, L\}$; the edges of $\mathcal{P}(1)$ have the slopes of absolute values $s_1 = 0$, and $s_\ell = (\bar{x}_2^\ell - \bar{x}_1^{\ell-1})/(\bar{x}_1^{\ell-1} - \bar{x}_1^\ell)$, for all $\ell \in \{2, \ldots, L\}$. The vertex conditions in (28) imply

$$\sum_{k=1}^{\ell} \hat{\eta}_k \rho_k \bar{x}_1^\ell + \sum_{k=\ell+1}^{L} \hat{\eta}_k \bar{x}_2^\ell = \bar{x}_1^\ell\left(\sum_{k=1}^{L} \hat{\eta}_k \min\{\rho_k, \rho_\ell\}\right) = 1,$$
$$\ell \in \{1, \ldots, L\},$$

taking account of the fact that the likelihood ratios ρ_ℓ are size-ordered (i.e., $\rho_1 < \cdots < \rho_L$) by construction. Hence, the coordinates of the vertices of $\mathcal{P}(1)$ are fully described by

$$\bar{x}_1^\ell = \left(\sum_{k=1}^{L} \hat{\eta}_k \min\{\rho_k, \rho_\ell\}\right)^{-1} \quad \text{and} \quad \bar{x}_2^\ell = \rho_\ell \bar{x}_1^\ell,$$

for all $\ell \in \{1, \ldots, L\}$. By the Hahn–Banach theorem (Berge, 1963, p. 157) there exists a hyperplane separating the convex set $\{\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_+ : c_1 x_1 + c_2 x_2 < C\}$, for some positive constant C, from the convex set $\mathcal{U}(q) = q\ \mathcal{U}(1)$ as long as they are disjoint, i.e., as long as

$$c_1 x_1 + c_2 x_2 < C \quad \Rightarrow \quad \bar{F}(x_1, x_2) < q.$$

In particular, if $c_1 \bar{x}_1^\ell + c_2 \bar{x}_2^\ell = \bar{C}(1)$, then necessarily $\bar{F}(\bar{\boldsymbol{x}}^\ell) = 1$. The aforementioned separation for $C = \bar{C}(1)$ occurs at the vertex $\bar{\boldsymbol{x}}^{\ell^*}$, with

$$\ell^* = \max\left\{\ell \in \{1, \ldots, L\} : \gamma = s_\ell < \frac{c_2}{c_1}\right\},$$

since that avoids an intersection of the separating hyperplane with the budget set $\{\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2_+ : c_1 x_1 + c_2 x_2 < \bar{C}(1)\}$. Using the fact that the cost objective and the envelope output are each homogeneous of degree 1, we obtain that the input vector $\bar{\boldsymbol{x}}^* = q\bar{\boldsymbol{x}}^{\ell^*}$ solves the envelope optimization problem (**). □

Proof of Lemma 6. Consider any fixed target output $q > 0$ and let $p_{\min} = \min_{(i,m) \in \mathcal{I} \times \mathcal{M}}\{p_{i,m}\}$ be the minimal matching probability (which is positive). By Lemma 5, the right-hand side of (19) is such that

$$q \geq \underline{q}(\varepsilon) = \frac{4}{\pi\varepsilon^2}\left(\frac{1-p_{\min}}{p_{\min}}\right) \quad \Rightarrow \quad \frac{q}{F(\bar{\boldsymbol{x}}^*(q))} - 1 \leq \frac{q(1+\varepsilon)}{\bar{F}(\bar{\boldsymbol{x}}^*(q))} - 1 = \varepsilon,$$

where we have used the fact that $\bar{F}(\bar{\boldsymbol{x}}^*(q)) = q$, as pointed out in footnote 3. Thus, by eliminating ε, the relative cost overage is such that

$$\frac{\hat{C}(q) - C(q)}{C(q)} \leq \left(\sqrt{\frac{1-p_{\min}}{(\pi/4)p_{\min}}}\right)\frac{1}{\sqrt{q}},$$

which corresponds to (20). This concludes our proof. □