Now that we’ve explained the demand side of the market, our goal is to develop a greater understanding of the supply side. Ultimately, we want to use a theory of the firm to put foundations under the supply curve. But before we can do that, we need to talk about the physical processes of production that firms have to deal with.

I. Some Basic Notions of Production and Cost

Production = combining inputs to make outputs
Inputs = land, raw materials, labor, machinery (capital)
Outputs = goods and services

Although many inputs are used in most production processes, we usually abstract from that complexity and talk about just two: capital and labor.

We sometimes represent the relationship between inputs and output with a production function, like so: \( q = f(K,L) \). This is a mathematical representation that tells how much output (q) you’ll get from any combination of inputs (K and L). A production function can be very complex, but here’s a simple example: \( q = \sqrt[3]{K^2L} \).

Example: When I was a teenager, I worked at the Great American Chocolate Chip Cookie Company (GACC). Suppose the production function above is the production function for sheets of cookies per hour. Then 4 workers and 1 oven can produce 2 sheets of cookies per hour. The same quantity could be produced with 2 workers and 2 ovens. With 4 ovens and 9 workers, 6 sheets of cookies could be produced.

II. Short Run versus Long Run

Economists use these terms a bit differently from how people use them in everyday language. We define them functionally like so:

Short-run = a period of time during which one or more of a firm’s inputs cannot be changed.

Example: GACC signs a one-year lease for a shop-front unit in a mall. Even if the company doesn’t bake a single cookie, it will still have to pay for this unit. And if it wants to expand, it will take a long while to acquire a new unit and get it equipped and ready to use. In this case, the short-run is any period less than a year.

Long-run = a period of time during which all inputs can be changed.
Example: When planning for periods longer than a year, GACC can consider shutting down its one shop or opening more.

The concepts of long-run and short-run are closely related to the concepts of fixed inputs and variable inputs. A fixed input is one whose quantity remains constant during the time frame in question. A variable input is one whose quantity can be altered during the time frame in question.

Example: At GACC, during the short-run capital is a fixed input, but labor is a variable input. In the long-run, both capital and labor are variable inputs.

NOTE: The distinction between SR and LR, and between fixed and variable inputs, is an abstraction. In reality, a firm will generally have many inputs, and each input can be changed in its own time frame. GACC, for example, might be able to alter its number of shop fronts only once a year, its number of cash registers and ovens once a month, and its number of workers once a day. So to be more realistic, we could talk about very SR, SR, medium run, LR, very LR, etc.

(A colleague of mine puts it another way: that “short-run” and “long-run” are really just a matter of what meeting you’re at. At some meetings, certain things will be taken as given because they are not on the agenda. Maybe you’re deciding how many more workers to hire, and the possibility of opening more stores is not on the agenda; for the time being, you’re taking the number of stores as given. At a different meeting, the possibility of opening more stores might be addressed more directly.)

III. Short-Run Production: Total, Average, and Marginal Product

Until further notice, we are in a short-run world. At least one input is fixed, and we’re examining how much can be produced from various amounts of the variable inputs.

Total product: the maximum quantity of output that can be produced with a given combination of inputs (i.e., q).

Average product of labor (AP\(_L\)): output per laborer or labor hour; i.e., AP\(_L\) = q/L, where L measures labor-hours.

Marginal product of labor (MP\(_L\)): the additional output that results from adding one more laborer or labor hour; i.e., AP\(_L\) = Δq/ΔL.

Example: GACC has a fixed number of ovens, but it can vary its labor. It can produce the following number of sheets of cookies per hour:

<table>
<thead>
<tr>
<th>L</th>
<th>q</th>
<th>AP(_L)</th>
<th>MP(_L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
The last two columns are calculated from the first two columns using the definitions above.

Returns to Labor: Looking at the table above, we can see that $MP_L$ gets larger as $L$ rises from 0 to 3. Then it’s constant as $L$ rises from 3 to 4. Then it falls as $L$ rises from 4 to 7. These three ranges correspond to three phenomena:

- Increasing marginal returns to labor. As more workers are added at first, they are able to cooperate to exploit gains from specialization and division of labor. Example: At GACC, a single worker has to scoop the dough, put the trays in the oven, take them out, detrays them, work the register, etc. When the second and third workers are added, the tasks can be divided up to save time.

- Constant marginal returns to labor. Eventually, the gains from specialization are exhausted. Adding more workers doesn’t increase the productivity of other workers.

- Decreasing marginal returns to labor. At some point, added workers don’t bring in as much new output as the old workers. Example: At GACC, the floorspace starts to get crowded, and the ovens are operating at maximum capacity. Some labor time is wasted as people bump into each other, wait in line to have their turn at the oven, etc.

NOTE: The fact that the marginal product of labor is decreasing does not mean that total production decreases. They are still adding something because their productivity is positive, and that can be seen in the $q$ column of the table. Nor does it mean that it’s inefficient to have this many workers. In order to make a judgment of efficiency, we’d have to know something about the revenues the workers bring in – but right now we’re only talking about production and cost.

Average Product of Labor. From the table, you can see that it rises for a while (as $L$ goes from 0 to 4), and then falls. This is similar to $MP_L$, but the turning point is in a different place – it occurs later. Notice that $AP_L$ gets larger as long $MP_L$ is greater than $AP_L$ was before, and it gets smaller otherwise. Why? This turns out to be a general feature of the relationship between marginal and average values of anything. It occurs because new values pull the average in their direction. In this case, the average productivity per worker is rising because each additional worker is more productive than the average of the previous workers.
Example: Consider your cumulative GPA. When you get new grades that are higher than your cumulative GPA, your cumulative GPA rises. When you get new grades that are lower than your cumulative GPA, your cumulative GPA falls. It does not matter than your most recent “marginal” grade might be smaller than your last “marginal” grade – as long as it's higher than your cumulative GPA, the effect is still to raise your GPA.

In the case of labor productivity, the average productivity per worker is rising because each additional worker is more productive than the average of the previous workers. It eventually falls when additional workers are less productive than the average of previous workers.

IV. Short-Run Production: A Graphical Explanation

The graph below will help to show the relationship between TP, AP_L, and MP_L.

The marginal product is the slope (change) in total product. In this picture, you can find the MP_L by finding the slope of the TP curve at a specific point. At L_1, the MP_L is the slope of the line m_1. At L_2, the MP_L is the slope of the line m_2. At L_3, the MP_L is the slope of the line m_3.

But these are not averages. Remember that the APL is equal to quantity (q) over the amount of labor (L). It follows that the APL is the slope of a line from the origin through a point on the TP curve.
At $L_1$, the $AP_L$ is the slope of the line $m_1$ (which is $q_1/L_1$). At $L_2$, the $AP_L$ is the slope of the line $m_2$ (which is $q_2/L_2$). At $L_3$, the $AP_L$ is the slope of the line $m_3$ (which is $q_3/L_3$). Notice that only at the last point is $AP_L$ the same as $MP_L$; this because the tangent line is also the line through the origin. Also note that $AP_L$ will begin to fall beyond this point.

In both of these graphs, we have assumed a TP curve with a specific shape. It is a typical shape, because it shows both rising marginal product (from specialization and division of labor) and falling marginal product (from crowing and related effects). But there may be goods and services for which the technology is different – perhaps showing a much longer period of rising or constant marginal product, perhaps showing none at all, perhaps showing less smooth (“lumpier”) transitions.

V. Short-Run Costs: Total and Average

Now that we know how much a firm can produce with different amounts of labor, we want to know how much it will cost to produce any given amount of output. Obviously, this depends on the prices of the inputs required to make the output.

Fixed inputs imply fixed costs. $TFC = \text{total cost of fixed inputs}$. Example: GACC has signed a lease committing it to $\$2400/\text{month}$ in rent for all its facilities. This is the fixed cost of capital. You can divide this up more finely to find the fixed cost per week, day, or hour.
Variable inputs imply variable costs. TVC = total cost of all variable inputs. Example: GACC has to hire workers, whom it can hire and fire at will. The wages paid to labor are part of variable cost.

Total cost is the sum of fixed and variable costs: TC = TFC + TVC. Example: At GACC, the fixed cost per hour is $10 ($2400/30 days/8 hours), and the firm decides to hire 5 workers at $6/hour. Then TC = 10 + 30 = 40.

Average total cost (ATC) is the cost of per unit; i.e., ATC = TC/q. Using the definition of TC above, we can also write ATC = TFC/q + TVC/q. Example: With 5 workers, GACC produces 16 sheets of cookies. ATC = 40/16 = 2.5. Alternately, ATC = 10/16 + 30/16 = .625 + 1.875 = 2.5.

Average fixed cost (AFC) is always falling. Why? The numerator (TFC) is, by definition, fixed. As the denominator (q) gets larger, the AFC must fall. This is known as fixed cost spreading or overhead spreading.

Average variable cost (AVC) is shaped like a bowl – it falls for a while, and then eventually rises. Why? Its closely related to the average product of labor. AVC results from payments to variable inputs -- in our simplified world, labor. The total payments to labor are wL, where w is the wage, and so AVC = wL/q. We already know that AP_L = q/L, which we saw earlier rises and then falls. When q/L rises, that means L/q falls, and that means it takes fewer workers per unit of output. And that means the labor cost per unit of output, wL/q is also falling.

Example: With 1 worker, AP_L = 2 sheets/worker, or ½ worker per sheet. At 3 workers, AP_L = 3 sheets per worker, or 1/3 worker per sheet.

So long as AP_L is rising, the number of workers per unit of output is falling. And that means labor cost per unit is falling as well. The reverse is also true: when AP_L is falling, the labor cost per unit of output is rising. Thus:

rising AP_L → falling AVC
falling AP_L → rising AVC

Average total cost (ATC) is the sum of AVC and AFC, whose shapes we’ve just discussed. So what does ATC look like? A lot like AVC, a bowl shape. But the existence of overhead spreading (through AFC) causes it to be decreasing for longer.

Note that the fall and rise of ATC do not indicate economies and diseconomies of scale. Those are long-run concepts we will get to later. The textbook incorrectly indicates that fixed-cost spreading is one cause of economies of scale. This is a misuse of terminology, because fixed-cost spreading happens only when there are fixed costs, and that is only in the short run, and scale is a long-run concept.
VI. Short-Run Marginal Cost

Marginal cost is the additional cost of producing one more unit of output; i.e., \( MC = \frac{\Delta TC}{\Delta q} \), where the change in \( q \) is usually one.

<table>
<thead>
<tr>
<th>L</th>
<th>q</th>
<th>TFC</th>
<th>TVC</th>
<th>TC</th>
<th>ATC</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>--</td>
<td>6/2 = 3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>10</td>
<td>6</td>
<td>16</td>
<td>8</td>
<td>6/3 = 2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>10</td>
<td>12</td>
<td>22</td>
<td>4.4</td>
<td>6/4 = 1.5</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>10</td>
<td>18</td>
<td>28</td>
<td>3.11</td>
<td>6/4 = 1.5</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>10</td>
<td>24</td>
<td>34</td>
<td>2.62</td>
<td>6/3 = 2</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>10</td>
<td>30</td>
<td>40</td>
<td>2.5</td>
<td>6/2 = 3</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>10</td>
<td>36</td>
<td>46</td>
<td>2.56</td>
<td>6/1 = 6</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>10</td>
<td>42</td>
<td>52</td>
<td>2.74</td>
<td></td>
</tr>
</tbody>
</table>

Notice again the relationship between average and marginal: average is falling whenever marginal is below the previous average, and rising whenever marginal is above the previous average. As a result, MC must cross ATC at its lowest point, as shown below. MC also crosses AVC at its lowest point, which is to the left of ATC's lowest point.
Recall that we found a relationship between \( APL \) and \( AVC \): when \( APL \) was rising, \( AVC \) fell, and vice versa. Now we find a similar relationship between \( MPL \) and \( MC \): When \( MPL \) is rising, \( MC \) is falling, and vice versa. Why? The change in payments to labor is \( \Delta L \), where \( w \) is the wage, and so \( MC = w \Delta L / \Delta q \). We already know that \( MPL = q / L \), which we saw earlier rises and then falls. When \( \Delta q / \Delta L \) rises, that means \( \Delta L / \Delta q \) falls, and that means \( w \Delta L / \Delta q \) must also fall, and vice versa when \( \Delta q / \Delta L \) falls.

VII. Short-Run Costs: A Graphical Explanation

We can show the relationship between total, average, and marginal costs in much the same way we show the relationship between total, average, and marginal product.
Here, the slopes of the lines $m_1$, $m_2$, and $m_3$ correspond to the MC at $L_1$, $L_2$, and $L_3$ (respectively).
Here, the slopes of the lines m1, m2, and m3 correspond to the ATC at L1, L2, and L3, respectively. Note that only at the last point are MC and ATC the same, and that after this point ATC will begin to rise.

VI. Long-Run Production

So far, we’ve taken capital (K) as fixed. But in the LR, firms can choose their level of capital. More broadly, in the LR all inputs are variable. As a result, all costs are variable costs. There are no fixed costs in the long run.

When the firm is able to vary all its inputs, it must choose a combination of inputs to produce its chosen level of output. Typically, there will be many combinations of inputs that can produce the same amount of output. So the cost-minimizing firm must pick the input combination that produces its chosen level of output at the lowest cost.

Example: Suppose the production function is \( q = \sqrt{KL} \), and the firm wishes to produce \( q = 4 \). Assume the wage for labor is \( w = 6 \), and the rental price of capital is \( r = 10 \). Then we can construct the following table of input combinations:

<table>
<thead>
<tr>
<th>K</th>
<th>L</th>
<th>TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16</td>
<td>106</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>64</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>92</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>166</td>
</tr>
</tbody>
</table>

Notice that each (K,L) combination produces \( q = 4 \). TC for each line is constructed using \( TC = rK + wL \). If the firm wishes to produce \( q = 4 \) at the lowest cost, \((4, 4)\) turns out to be the best input combination of those listed. (It actually turns out that the very lowest cost combination is \( K = 3, L = 16/3 \).)

From this point forward, when we are talking about firms operating in the LR, we will assume they have minimized their input costs in this fashion.

So, what does this do to the firm’s cost curves? Let’s take them in turn.

Total cost is still fixed cost plus variable cost. But in the long run, FC = 0. How will the LRTC compare to the SRTC? It turns out that for any quantity, LRTC must be less than or equal to SRTC. Why? Because in the long run, the firm can choose a cost-minimizing input combination, which it cannot do in the short run. It certainly has the option, in the long run, of choosing the exact same level of capital and labor as it did in the short run, so it couldn’t possibly do worse. But it can usually do better by altering the combination of capital and labor.
**Average total cost** is still defined as total cost divided by quantity. \( LRATC = LRTC/q \). Like SRATC, it will typically be bowl-shaped. Just as LRTC was always less than or equal to SRTC because of cost-minimization, LRATC is always less than or equal to SRATC for the same reason.

Note that SRATC is always greater than or equal to LRATC. What's happening at \( q' \), the one point where \( SRATC = LRATC \)? That's where the amount of capital (fixed inputs) the firm happened to have in the short run is exactly the same amount of capital (fixed inputs) the firm would choose to have in the long run, if it wished to produce that quantity.

In the above, we have been acting like there is just one SRATC, but we could draw more. The one we drew was based on some fixed level of capital, which implied a particular level of fixed costs. But what if the firm had a different fixed level of capital in the short run? What if GACC happened to have two ovens instead of just one? Then we could construct a different set of short-run curves based on that assumption. We could do this for any level of fixed inputs that a firm might have, so we could really have an infinite number of SRATC’s.
If we kept on drawing in more SRATCs, we would eventually find that the LRATC forms an envelope of them all (encompassing them all from below).

We can also talk about long-run MC (or LRMC). It is calculated just like short-run MC, except we find the change in LRTC instead of short-run TC. It tells you how much it costs to increase output by one more unit, assuming that you’re able to adjust your combination of inputs optimally.

VII. Economies and Diseconomies of Scale

In the pictures we’ve drawn so far, we’ve implicitly assumed a certain type of technology – specifically, a technology that results in a bowl-shaped LRATC curve. That means average costs decline over some range of production, but eventually start rising again. Why? What have we assumed here?

Notice that the shape of the LRATC and SRATC are similar. But the reasons are different. The SRATC’s bowl shape resulted from changing average product of labor, combined with fixed cost spreading. The SRATC falls because of increasing average product of labor and fixed cost spreading, and it eventually rises because of decreasing average product of labor. But the LRATC’s shape is related to returns to scale.

Economies of scale occur when output increases proportionally more than cost increases. E.g., you can double output by less-than-doubling cost. Economies of scale occur mainly because of specialization – not just specialization of workers among tasks, but also specialization in capital usage. A greater amount of capital may allow better specialization, because the capital can be catered to their specific tasks. Economies of scale usually accompany mass production techniques (which is really just another way of saying “specialization combined with appropriate capital”).

Other reasons that economies of scale occur: (a) Lower inventory costs. Inventory is maintained in order to reduce the likelihood of running out if quantity demanded is higher than average. Small-scale firms are likely to experience greater variance in their sales, which means they need to keep relatively high inventories when measured as a fraction of sales. A large-scale firm will experience less variance in its sales, and thus can get by with a smaller ratio of inventories to sales. (b) Physical properties of production. Sometimes the physical nature of the production process implies a falling ratio of cost to output at output increases. For example, as the radius of a pipeline increases, the surface area (which determines the amount of metal required) increases as a linear function of radius (circumference = \(2\pi r\)) while volume increases as a function of the square of the radius (area = \(\pi r^2\)). (c) Volume discounts, resulting from contract-writing savings or the greater buying power of a larger purchaser. (d) Marketing advantages. A larger scale firm with more outlets is more likely to benefit from a given amount of advertising, because each viewer will be more able to act on the desire stoked by the ad.
**Diseconomies of scale** occur when output increases proportionally less than cost increases. This happens when the benefits of specialization and mass production begin to be overwhelmed by the costs of large-scale activity: layers of management, difficulty of communication, problems with monitoring, etc.

Other reasons diseconomies of scale occur: (a) Large-scale firms often pay higher wages, in part because their workforces are more likely to be unionized, in part because they may have to pay compensating differentials to overcome a taste workers have for working in smaller firms. (b) Spreading specialized resources too thin. Some resources are unique, and thus cannot really be scaled up in the scaling-up process. For instance, if a firm’s biggest asset is its very clever owner, that owner cannot be duplicated, and his attention may get spread too thin as the firm grows in size.

Economies and diseconomies of scale correspond to the downward-sloping and upward-sloping portions, respectively, of the LRATC curve. This is apparent from their definitions. Economies of scale means q increases at a higher rate than LRTC, so LRTC/q must decline, and vice versa for diseconomies of scale. In addition, a LRATC may have a flat section in the middle; this section corresponds to constant returns to scale.

Could we have made different assumptions about technology? Certainly. For example, we could have assumed never-ending economies of scale, which implies an always downward-sloping LRATC curve. Karl Marx believed that all major industries were like this, and predicted for that reason that all industries would end up being monopolies. The USSR agreed that all industries must have downward-sloping LRATC’s, and that’s why they consolidated all their industries into massive plants. But history has shown this to have been a poor choice. However, there are some industries, especially certain utilities, that do seem to have this feature. We call such industries natural monopolies.

The textbook uses the term “economies of scale” to refer to any decrease in ATC (thus, to any downward-sloping portion of an ATC curve), regardless of the time period. As I indicated earlier, this is an error of terminology. However, there is a good argument from a real-world perspective for the textbook’s approach. For many industries, there will be some inputs that will remain fixed for a very long period of time – for instance, the number of power generators operated by an electric utility. So by the usual economist’s definition of terms, the firm is not really in the long-run in the meantime, and thus scale economies don’t apply (the term “scale” being reserved for situations in which all inputs are variable). But there are many variable inputs besides labor, and the firm can use the process of cost minimization to choose the optimal combination of the variable inputs. As a result, many of the lessons of scale economies apply even in situations where some inputs are fixed for the time being.

**VIII. Economies of Scope**
Economies of scope exist when a firm can produce two or more products at lower cost than the two products could be produced separately. One way to represent this mathematically is like so:

\[ \text{TC}(x, y) < \text{TC}(x, 0) + \text{TC}(0, y). \]

The joint production function (shown on the left) is lower than the sum of the separate production functions (shown on the right) for some positive levels of \(x\) and \(y\).

There are two basic reasons this may occur. First, a positive level of production of one good \((x)\) may lower the marginal cost of producing another product \((y)\). Second, some fixed inputs can be shared between the two production processes, thereby allowing greater fixed-cost spreading. (Notice that economies of scope is not, like economies of scale, a purely long-run concept.)

Some more specific reasons why economies of scope occur: (a) Volume discounts. As discussed in the section on economies of scale, purchasing in bulk can lower transaction costs. A firm might not need to produce enough of a single good to justify a large purchase, but if two or more goods share some of the same inputs, then purchasing in bulk can make sense. (b) Marketing advantages. If a single producer makes multiple products, the same amount of advertising can benefit all the product lines. For example, if McDonald’s runs an advertisement, it can benefit all the food products associated with McDonald’s. This is called “umbrella branding.” (c) Research and development spillovers. The research done on one product can generate discoveries relevant to other products. For example, research on new paint thinners might yield results useful for nail polish removers, carpet cleaners, and so on.

**IX. The Learning Curve**

The learning curve represents the fact that a firm’s costs of production tend to fall as the firm gains more experience in the production of its product. The usual picture shows a downward-sloping curve as a function of the firm’s cumulative production.
People often confuse the learning curve, shown above, with an average (total or variable) cost curve. But they are quite different. Both curves show a firm’s average cost of production for some product. But the ATC curve shows the firm’s average cost of production at a fixed point in time, i.e., for a fixed level of experience, at many different quantities of output. The learning curve, on the other hand, shows the firm’s average cost of production at a fixed level of output for many different levels of experience. Notice that the horizontal axis is “Cq,” meaning cumulative quantity – the total amount of production the firm has accomplished over its lifetime. Cumulative quantity is usually a good proxy for the firm’s experience.

On p. 100 of the book, you’ll find a graph showing the relationship between the learning curve and the average cost curve. In order to show the relationship cleanly, the authors have assumed constant returns to scale, which imply horizontal long-run ATC curves. The important point is that, as time passes and experience accumulates, the firm’s long-run ATC changes, usually for the better – and that’s true whether or not there are economies (or diseconomies) of scale.

Notice that I said a learning curve assumes a fixed level of output. In the graph on p. 100, the given output assumption is unnecessary, because long-run ATC is the same for all levels of output. But if the long-run ATC is not horizontal, it is necessary to talk about a fixed level of output. In general, we assume (for simplicity) that the learning effect lowers average cost at every level of output. But it’s possible for the learning effect to impact only average cost at certain levels of output, specifically, the levels of output the firm has had experience with.