The Viner–Wong Envelope Theorem

Eugene Silberberg

The envelope theorem, now the fundamental tool in modern duality analysis, had its beginnings in Jacob Viner's classic 1931 article on short- and long-run cost curves. It seemed wrong to Viner that at any given point along the long-run cost curve, the long-run average cost curve should have the same slope as the short-run curve, where capital was being held constant. This is still a puzzle to many people, along with the various other envelope theorem results (see, for example, the discussion of a related issue by Sexton, Graves, and Lee 1993). Viner instructed his draftsman, Wong, to draw the long-run curve through the minimum points of the short-run average cost curves. Curiously, Samuelson's resolution of the puzzle in *Foundations* (1947) dealt only with a simple unconstrained maximum problem, *maximize* $y = f(x_1, x_2, \ldots, x_n, \alpha)$, where the x_i's are the decision variables and α is a vector of parameters. Its relation to the original Viner–Wong diagram seems a bit remote at first, and, curiously, a discussion of the envelope theorem in the explicit Viner–Wong context seems missing from the literature.[1] This is unfortunate because it is possible to communicate the essence of this result (and more general results) with a simple cost diagram that goes back to the roots of cost theory.

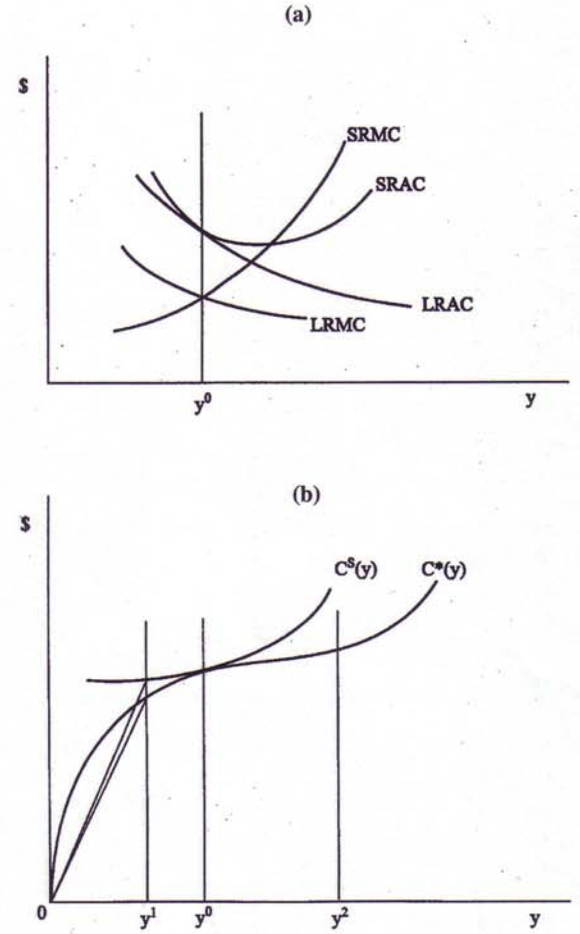
THE MODEL: TWO VIEWS

In Figure 1, I show in panel (a) a section of the classic Viner–Wong diagram, in the neighborhood of some output y^0. In panel (b) I show the corresponding total cost curve, labeled $C^*(y)$ with the usual shape, that is, it first rises at a decreasing rate and then at an increasing rate. It is relevant to the discussion that C^* is not uniformly concave or convex; I mean to allow marginal cost, which is the slope of this curve, to be first decreasing and then increasing. It is important to remind students that total cost is not just some technological relation of cost to output, but that it depends on a certain kind of optimizing behavior by the firm. I usually think of the production function in this type of "black box" formulation: inputs march in and some output results. That is, no optimization is inherent in the traditional formulation of the production function; it could be just the way some firm arranges its production.

The total cost I plot in $C^*(y)$, however, is the *minimum* cost for achieving that level of output. In particular, the total, average, and marginal cost curves are all derived from (in two dimensions, for simplicity) the following:

Eugene Silberberg is a professor of economics at the University of Washington, Seattle (e-mail: silber@u.washington.edu).

FIGURE 1
Short- and Long-Run Cost Curves



$$\textit{minimize } C = w_L L + w_K K, \text{ with respect to } L \text{ and } K \qquad (1)$$
$$\textit{subject to } f(L,K) = y,$$

where L and K stand for labor and capital, and w_L and w_K are their respective factor prices. Output, y, is parametric in this discussion; that is, I derive the input combinations that minimize total cost for arbitrary output levels, without deciding what output will wind up being the one chosen by the firm, on the basis of, for example, profit maximization. The solution to equation (1) is the implied cost-minimizing levels of labor and capital.[2]

$$L = L^*(y) \tag{2}$$
$$K = K^*(y).$$

Substituting these cost-minimizing input levels into the objective function in equation (1) gives the total cost function $C^*(y)$. At each point along $C^*(y)$, some cost-minimizing levels of labor and capital are used, in accordance with equations (2). These levels change as y increases (although not predictably—it is possible for factors to become inferior, i.e., to be used less as output increases).

Consider now panel (b) of Figure 1, showing the behavior of total cost in a neighborhood of output level y^0. At that particular output, some cost-minimizing input levels of labor and capital are used by the firm; denote these as L^0 and K^0. In general, as the firm changes its output level in either direction, the cost-minimizing L and K will depart from these values. Suppose now capital is held fixed at K^0 as y changes. In that case, the cost of producing any output $y \neq y^0$ must be greater than the amount indicated by $C^*(y)$, because $C^*(y)$ is by definition the *minimum* cost. The cost curve incorporating this restriction on capital is $C^s(y)$ (s for short run). It is equal to $C^*(y)$ at y^0, because the cost-minimizing L and K are used there, but everywhere else, $C^s(y) \geq C^*(y)$. Assuming differentiability, it must be the case that $C^s(y)$ and $C^*(y)$ are tangent at y^0, and that $C^s(y)$ is locally relatively more convex or less concave than $C^*(y)$.

I now show how this tangency is related to the classic Viner–Wong diagram. Tangency means that $C^s(y)$ and $C^*(y)$ have the same slope at y^0, where the fixed capital input happens to be at the cost-minimizing level. But the slopes of these total cost functions at any point are the respective values of marginal cost in the short or long run. At the tangency point, short- and long-run marginal costs are thus obviously equal. Moreover, it is clear that because $C^s(y)$ must be relatively more convex or less concave than $C^*(y)$, the *second* derivative of $C^s(y)$ is greater than equal to the second derivative of $C^*(y)$: $C^s_{yy}(y) \geq C^*_{yy}(y)$. But this means that marginal cost either rises faster or falls slower in the short run than in the long run. Thus in panel (a), the short-run marginal cost curve (SRMC) cuts the long-run marginal cost curve (LRMC) from below.[3] It is interesting that even though one cannot say from cost minimization alone whether marginal cost will be rising or falling, one can nonetheless prove this relative relationship between the short- and long-run marginal cost curves.

Turning to the average cost curves, consider in panel (b) that at outputs to the right and to the left of y^0, for example, at y^1 and y^2, $C^s(y) > C^*(y)$. Average cost is the slope of the chord connecting the origin and the total cost curve. Two such slopes are indicated in panel (b) at output level y^1. It is transparent that the average cost is greater when measured up to the short-run total cost curve C^s than to C^*. (To avoid clutter, I have drawn these chords only for y^1 but at y^2 the same relationship holds.) Obviously, average costs are equal at y^0. Thus the relationships between the short- and long-run average and marginal cost curves in the Viner–Wong diagram follow in an elementary way from the simple notion that a more-constrained minimum must always lie above a less-constrained minimum, but at the point where the added constraint is nonbinding, the two minima must be equal.

CONCLUDING REMARKS

Although I have defined the short run (as in Viner and the standard textbook treatments) as *holding one input fixed*—usually capital—the exhibited relationships between the short- and long-run average and marginal cost curves do not depend on this definition of the short run. Suppose, more generally, the short run means that some arbitrary relation $g(L, K) = 0$ holds between labor and capital, where $g(L^0,K^0) = 0$, so that this constraint is nonbinding at y^0, in the same sense that I held capital fixed at its cost-minimizing value at y^0. One can use such a more general constraint to contemplate more general restrictions on the firm; for example, although the firm need not hold capital literally fixed in the short run, it is not as free to adjust it as it is in the long run. Perhaps in the short run, capital can be varied only in the same direction as the change in labor and perhaps at some specified rate. Absolutely no change occurs in panel (b) for this more general constraint.[4] Whatever this more general constraint is, it remains the case that $C^s \geq C^*$ when $y \neq y^0$, and $C^s = C^*$ at $y = y^0$. Thus the same type of tangency occurs between C^s and C^* as in the simpler case, and the resulting relationships between short- and long-run average and marginal costs are the same as in the classic Viner analysis.

The above discussion assumes initially only two inputs, labor and capital, with the short run defined as *imposing an additional constraint*, for example, holding capital fixed. This is in keeping with Viner's original treatment and the standard textbook analysis. In fact, the two-factor cost-minimization problem degenerates when one imposes an additional constraint. In this model, I hold output fixed along some arbitrary isoquant and then one choose the labor and capital inputs that minimize cost. If, for example, the capital input is specified, then we can no longer vary labor at all—only one labor input will keep us on the same isoquant! The model stated in equation (1) is only a one-dimensional problem. If one uses the constraint to substitute for one variable, one is left with a single-variable minimization problem. If one proceeds to impose still another constraint, for example, holding one variable fixed, the model becomes a two-variable problem with two constraints, which then by themselves completely define the solution. In order to do the any kind of short-run vs. long-run analysis à la Viner–Wong, one must start out with at least three variable inputs. The analysis then goes through in the manner discussed above.

NOTES

1. Samuelson's proof of the envelope theorem was so opaque to me when I first encountered it in graduate school in a reading class with Jim Quirk that I promptly dropped the course!
2. These factor demands and total cost C depend on the factor prices w_K and w_L as well as output y, but because these variables are not germane to this discussion, I suppressed them and simply wrote, for example, $C = C^*(y)$.
3. Depending on the level of the class, one can include the following more formal exposition: because $C^s(y) \geq C^*(y)$ but $C^s(y) = C^*(y)$ at y^0, the function $F(y) = C^s(y) - C^*(y)$ has a minimum at $y = y^0$. The necessary first order condition for a minimum yields $C^s_y(y) - C^*_y(y) = 0$; the sufficient second-order condition yields $F''(y) = C^s_{yy}(y) > C^*_{yy}(y)$. See Silberberg (1990) for a more complete discussion of this and other LeChatelier effects.
4. An anonymous referee pointed out that the new constraint must allow production to expand, or the

Viner–Wong diagram will be truncated. For example, the linear constraint $aL + bK = C^0$, where C^0 is the isocost level at L^0, K^0, would constrain output to be no greater than the amount indicated by the isoquant tangent to that isocost line. If in fact $a = w_L$, $b = w_K$, then the Viner–Wong diagram would not exist beyond the original output level (y^0 in Figure 1).

REFERENCES

Samuelson, P. 1947. *Foundations of economic analysis*. Cambridge: Harvard University Press.