Real-Time Electricity Markets

1.0 Two-settlement system

Material from this section was adapted from [1]. The electricity markets in the US are all two-settlement systems. That is, they are comprised of two interconnected markets, each of which results in a settlement:

- The day-ahead market and
- The real-time, or balancing market.

A third market, called the operating reserve market, is run in parallel.

These three markets together have four products:

- Energy: based on resource offers and demand bids
- Regulating reserve: for real-time balancing
- Spinning reserve: provides on-line energy to meet demand under contingency conditions
- Supplemental reserve: same as spinning, except can be from off-line resource

The day-ahead market utilitizes the security-constrained unit commitment (SCUC) to identify a unit commitment schedule and compute day-ahead locational marginal prices (LMPs).

The real-time market utilizes the security-constrained economic dispatch (SCED) to identify the real-time dispatch and compute real-time LMPs.

In these notes, we will show how to compute LMPs and examine those things which influence them.

2.0 Derivation of LMP

We will develop an optimization problem that characterizes what real-time markets do, and from this optimization problem, we will extract the definition for locational marginal prices (LMPs).

2.1 Objective function

We make two simplifications on the objective function:

- We assume that demand bids are not price sensitive, i.e., that demand is fixed, independent of the price. Therefore, the problem of maximizing the social surplus becomes one of maximizing the producer's surplus, which is the same as minimizing the producer's cost. So our objective function will be to minimize producers's cost.
- We assume that producers cost is expressed linearly as a function of their generation. You can think about this in two different ways.

- The producers represent their cost functions using a piecewise linear approximation, or
- The producers are simply making offers of certain quantities, at fixed prices.

In either case, we may represent the objective function as

$$G(\underline{P}_g) = \sum_{k=1}^{N} s_k P_{gk} \tag{1}$$

where s_k are the \$/MWhr offers being made on an amount of generation of P_{gk} over 1 hour, and there are N generators making such offers. If bus k is a pure load bus, then $P_{gk}=s_k=0$.

2.2 Power balance

If our analysis is based on linearized network representation, then it is implied that resistance has been assumed zero. Therefore, losses should be zero, and the power balance equation would be

$$\sum_{k=1}^{N} P_{gk} - P_{dk} = \sum_{k=1}^{N} P_k = 0 \tag{2}$$

That left-hand summation of (2), when placed equal to 0, says that the sum of generation is exactly equal to the sum of demand. The right-hand summation of (2), when placed equal to 0, says the sum of injections is exactly equal to zero.

However, we will, for the moment, give a more general relation that accounts for losses, i.e.,

$$\sum_{k=1}^{N} P_{gk} - P_{dk} = P_{loss} \tag{3}$$

2.3 Line flow constraints

Regarding network representation, we assume that we have a constraint sensitivity matrix $\underline{T}$ with elements t_{jk} that give the change in flow on circuit j to a change in real power injection at bus k, under a specified slack distribution, according to

$$t_{jk} = \frac{\Delta F_j}{\Delta P_k} \tag{4}$$

If the network is linear over its entire operating range, then (4) applies even when

$$\Delta F_j = F_j - 0, \quad \Delta P_k = P_k - 0 \tag{5}$$

so that

$$t_{jk} = \frac{F_j}{P_k} \tag{6}$$

or

$$F_j = t_{jk} P_k \tag{7}$$

In matrix form, (7) becomes:

$$\underline{F} = \underline{T}\,\underline{P} \tag{8}$$

It is important to note that the sensitivity factors of (6) are computed under a so-called “slack-bus” assumption which indicates how the change ΔP_k is assumed to be compensated. It could be compensated from another certain bus, or from several other buses, or from all other buses, and the elements of T will change depending on which of these is assumed. It is generally considered best to employ a so-called *distributed slack bus assumption* here where the compensation is assumed to come from all other generator buses.

Given that we know the "normal" flow constraints on every circuit, then

$$-\underline{F}_{\max} \le \underline{F} \le \underline{F}_{\max} \tag{9}$$

Substitution of (8) into (9) results in

$$-\underline{F}_{\max} \le \underline{T}\,\underline{P} \le \underline{F}_{\max} \tag{10}$$

We assume at this point that high flows in our network are unidirectional, i.e., we need not be concerned with high flows in both directions. This does not prevent bidirectional flows, it merely enables us to be concerned with reaching the upper bound in only one direction. Therefore, we may ignore the lower bound in (10) so that our circuit flow constraint is

$$\underline{T}\,\underline{P} \le \underline{F}_{\max} \tag{11}$$

In scalar form, (11) is

$$\sum_{k=1}^{N} t_{jk} P_k \le F_{j\max},\ j = 1,\ldots,M \tag{12}$$

and replacing injection with difference between generation and load, we obtain:

$$\sum_{k=1}^{N} t_{jk}(P_{gk} - P_{dk}) \le F_{j\max}, \; j = 1,\dots,M \qquad (13)$$

2.4 Optimization problem

A linearized optimal power flow (OPF) problem minimizes (1) subject to (3) and (12), that is,

$$\min \quad G(\underline{P}) = \sum_{k=1}^{N} s_k P_{gk}$$

$$s.t.$$

$$\sum_{k=1}^{N} P_{gk} - P_{dk} = P_{loss}$$

$$\sum_{k=1}^{N} t_{jk}(P_{gk} - P_{dk}) \le F_{j\max}, \; j = 1,\dots, M \qquad (14)$$

We will call (14) LOPF-1.

The Lagrangian function for LOPF-1 is

$$L(\underline{P}_g, \lambda, \underline{\mu}) = \sum_{k=1}^{N} s_k P_{gk} - \lambda\left[\sum_{k=1}^{N} P_{gk} - P_{dk} - P_{loss}\right] - \sum_{j=1}^{M} \mu_j \left[\sum_{k=1}^{N} t_{jk}(P_{gk} - P_{dk}) - F_{j\max}\right] \qquad (15)$$

The first order conditions for finding the optimum to LOPF-1 include:

$$k \in gen: \qquad \frac{\partial L}{\partial P_{gk}} = s_k - \lambda(1 - \frac{\partial P_{loss}}{\partial P_{gk}}) - \sum_{j=1}^{M} \mu_j t_{jk} = 0 \qquad (16)$$

But we are more interested in the load buses. Consider

$$k \in load: \qquad \frac{\partial L}{\partial P_{dk}} = \lambda(1 + \frac{\partial P_{loss}}{\partial P_{dk}}) + \sum_{j=1}^{M} \mu_j t_{jk} \qquad (17)$$

Note carefully that P_{dk} is not a decision variable, and therefore we do not set it equal to 0.

Let's consider (17). What is this? To answer this question, we need to learn a theorem.

2.5 Envelope theorem

Consider the following optimization problem.

$$\begin{aligned} &\max_x f(x,\theta) \\ &s.t. \quad g(x,\theta) \le 0 \end{aligned} \qquad (18)$$

where x is the decision variable and θ is some parameter that is influential in the problem, but it is not a decision variable, i.e., we may not select its value. We desire to find how the optimal value of f changes with respect to θ.

Let's give a name to the optimal value of f. Let's call it V; it is a function of θ. That is,

$$V(\theta) = f(x^*(\theta), \theta) \qquad (19)$$

Then what we are trying to find is

$$\frac{\partial V(\theta)}{\partial \theta} \tag{20}$$

Note that V will change both because θ affects f and because it also affects the optimal choice of x.

The Lagrangian function is

$$L(x,\theta,\lambda) = f(x,\theta) - \lambda g(x,\theta) \tag{21}$$

Envelope theorem: The total rate of change in the optimal value of the objective function due to a small change in the parameter θ is the rate of change in the Lagrangian L evaluated at the optimal value of x. That is,

$$\frac{\partial V(\theta)}{\partial \theta} = \left.\frac{\partial L(x(\theta),\theta,\lambda)}{\partial \theta}\right|_{x=x^*} \tag{22}$$

The proof requires several lines of calculus that we omit here.

2.6 Locational marginal price

Armed with the envelope theorem, we may now identify the meaning to (17), which is repeated here for convenience:

$$k \in load: \qquad \frac{\partial L}{\partial P_{dk}} = \lambda\left(1 + \frac{\partial P_{loss}}{\partial P_{dk}}\right) + \sum_{j=1}^{M} \mu_j t_{jk} \tag{17}$$

Equation (17) gives the change in the optimal value of the objective function due to a small change in the parameter P_{dk}.

In other words, if we solve the optimization problem with $P_{dk}=P_{dk0}$, obtaining $G^*(P_{dk0})$, and then resolve the optimization problem with $P_{dk}=P_{dk0}+1$, obtaining $G^*(P_{dk0}+1)$, then

$$G^*(P_{dk0}+1)-G^*(P_{dk0})=\frac{\partial L}{\partial P_{dk}} \tag{23}$$

We call $\frac{\partial L}{\partial P_{dk}}$ the LMP for bus k, that is,

$$k \in load: \qquad LMP_k=\lambda(1+\frac{\partial P_{loss}}{\partial P_{dk}})+\sum_{j=1}^{M}\mu_j t_{jk} \tag{24}$$

Written slightly different, it is

$$k \in load: \qquad LMP_k=\lambda+\lambda\frac{\partial P_{loss}}{\partial P_{dk}}+\sum_{j=1}^{M}\mu_j t_{jk} \tag{25}$$

And (25) show us a very useful way to think about LMPs. They consist of three components:

$$\begin{aligned} k \in load: \quad LMP_k &= \lambda && \text{Energy component} \\ &+\lambda\frac{\partial P_{loss}}{\partial P_{dk}} && \text{Loss component} \\ &+\sum_{j=1}^{M}\mu_j t_{jk} && \text{Congestion component} \end{aligned} \tag{26}$$

We discuss each one of these terms in what follows.

2.7 Energy component

We are considering the components of the LMP at a particular bus *k*. The first component is the energy component, represented by λ.

To gain better understanding of exactly what this is, we will neglect losses in our original formulation (14), resulting in

$$
\begin{aligned}
&\min \quad G(\underline{P}) = \sum_{k=1}^{N} s_k P_{gk} \\
&s.t. \\
&\sum_{k=1}^{N} P_{gk} - P_{dk} = 0 \\
&\sum_{k=1}^{N} t_{jk} (P_{gk} - P_{dk}) \le F_{j\max},\ j = 1,\ldots,M
\end{aligned}
\tag{27}
$$

Rewriting the equality constraint in (27) so that the function of decision variables is on the left-hand-side and constants on the right-hand-side, we have

$$
\begin{aligned}
&\min \quad G(\underline{P}) = \sum_{k=1}^{N} s_k P_{gk} \\
&s.t. \\
&\sum_{k=1}^{N} P_{gk} = \sum_{k=1}^{N} P_{dk} = P_{D,tot} \\
&\sum_{k=1}^{N} t_{jk}(P_{gk} - P_{dk}) \le F_{j\max},\ j = 1,\ldots,M
\end{aligned}
\tag{28}
$$

Now write the Lagrangian function:

$$
L(\underline{P}_g, \lambda, \underline{\mu}) = \sum_{k=1}^{N} s_k P_{gk} - \lambda \left[\sum_{k=1}^{N} P_{gk} - \sum_{k=1}^{N} P_{dk} \right] - \sum_{j=1}^{M} \mu_j \left[\sum_{k=1}^{N} t_{jk}(P_{gk} - P_{dk}) - F_{j\max} \right]
\tag{29}
$$

or

$$
L(\underline{P}_g, \lambda, \underline{\mu}) = \sum_{k=1}^{N} s_k P_{gk} - \lambda \left[\sum_{k=1}^{N} P_{gk} - P_{D,tot} \right] - \sum_{j=1}^{M} \mu_j \left[\sum_{k=1}^{N} t_{jk}(P_{gk} - P_{dk}) - F_{j\max} \right]
\tag{30}
$$

Notice that λ is the Lagrange multiplier (or dual variable) on the power balance equality constraint. This immediately gives us an interpretation of λ.

> ➔The energy component λ of the LMP is the increase in the objective function (in this case, cost per hour) if demand $P_{D,tot}$ increases by 1 unit.

Without losses, the LMP expression becomes (from (29)):

$$k \in load: \qquad LMP_k = \frac{\partial L}{\partial P_{dk}} = \lambda + \sum_{j=1}^{M} \mu_j t_{jk} \qquad (31)$$

The summation is the congestion component. If there is no congestion, then

$$k \in load: \qquad LMP_k = \lambda \qquad (32)$$

Equation (32) makes the interesting point that, if we ignore losses, and if there is no congestion, then the LMP will equal to λ, *and this will be true for every load bus in the network*.

One last comment here. It is worthwhile to identify what determines λ. We may gain insight to this via the first order condition (16) which, without losses, becomes

$$k \in gen: \qquad \frac{\partial L}{\partial P_{gk}} = s_k - \lambda - \sum_{j=1}^{M} \mu_j t_{jk} = 0 \qquad (33)$$

Solving for λ, we obtain:

$$k \in gen: \quad \lambda = s_k - \sum_{j=1}^{M} \mu_j t_{jk} \qquad (34)$$

Under the condition of no congestion, then

$$k \in gen: \quad \lambda = s_k \qquad (35)$$

What does this mean?...

To understand what this means, it is important to understand that P_{gk} for which we differentiate to obtain (35) must be "regulating," i.e., it cannot be at its limit. We could have exposed this idea more clearly by including constraints on P_{gk} in the optimization problem formulation, in which case we would have obtained corresponding terms in the objective that would have vanished for regulating units and would have contributed for non-regulating units.

Now consider how an electricity market works. Each generation owner offers in their s_k with a corresponding range. The algorithm selects the lowest offer, and takes the full range of that offer, and then selects the next lowest offer, and then the next, and so on until the demand is met. Figure 1 illustrates.

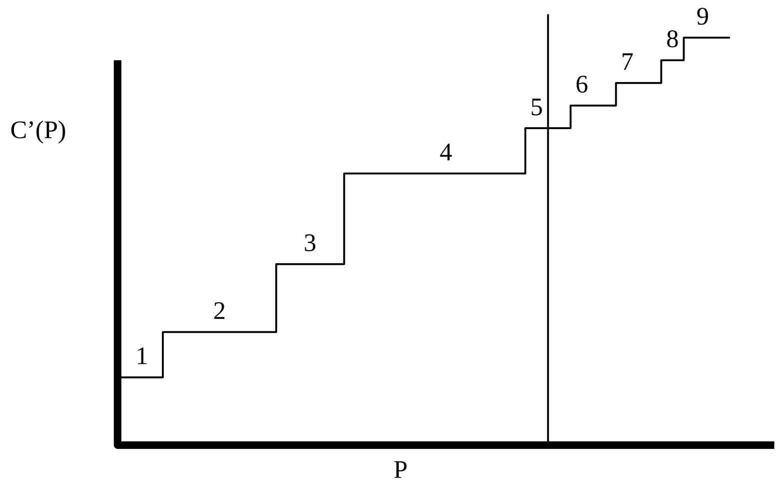

Fig. 1

The only unit that is selected, and is regulating, is unit 5. This is the unit for which $\lambda = s_k$. It is the unit that will pick up the extra demand when the demand is increased by 1 unit. We say that unit 5 is "on the margin."

2.8 Loss component

Consider the expression for LMP again, from (25)

$$k \in load: \qquad LMP_k = \lambda + \lambda \frac{\partial P_{loss}}{\partial P_{dk}} + \sum_{j=1}^{M} \mu_j t_{jk} \qquad (25)$$

Assuming no congestion, we have

$$k \in load: \qquad LMP_k = \lambda + \lambda \frac{\partial P_{loss}}{\partial P_{dk}} \qquad (36)$$

When we increase the demand at bus *k* by one unit, the losses will increase due to more current flowing through the network. Therefore the term $\frac{\partial P_{loss}}{\partial P_{dk}}$ will be positive. This results in each bus seeing a higher LMP than that set by the energy component λ.

For a particular bus *k*, the increase in LMP_k beyond λ will depend on how an increase in that bus's demand P_{dk} would be compensated. The way it would *really* be compensated is that the marginal unit would increase its generation. This would require $\frac{\partial P_{loss}}{\partial P_{dk}}$ to be recomputed each time the marginal unit changes which is very frequent. Alternatively, $\frac{\partial P_{loss}}{\partial P_{dk}}$ can be computed for each bus under an assumed compensation strategy. For example, reference [2][1] shows how to compute $\frac{\partial P_{loss}}{\partial P_{dk}}$ relative to a distributed slack bus reference. We will not cover this but will simply assume the availability of $\frac{\partial P_{loss}}{\partial P_{dk}}$.

2.9 Congestion component

Finally, we reconsider the expression for LMP once again, from (25)

$$k \in load: \qquad LMP_k = \lambda + \lambda \frac{\partial P_{loss}}{\partial P_{dk}} + \sum_{j=1}^{M} \mu_j t_{jk} \tag{25}$$

At this point, our interest is the last term. Let's ignore the losses, resulting in

$$k \in load: \qquad LMP_k = \lambda + \sum_{j=1}^{M} \mu_j t_{jk} \tag{37}$$

The summation in (37) will contain zero terms for all circuits *j* for which flow is not at the rating, i.e., the only non-zero terms in the summation will be for circuits that are at their rating, i.e., that are *congested*. Let's consider that there is only one such circuit in the network, circuit 5. Then

$$k \in load: \qquad LMP_k = \lambda + \mu_5 t_{5k} \tag{38}$$

[1] I have placed this reference on the web page. It is an excellent paper on LMP calculation.

The Lagrange multiplier (dual variable) μ_5 is on the flow constraint for circuit 5, and it will always be nonnegative. On the other hand, t_{5k}, the generation shift factor, representing the change in flow on circuit 5 for an increase in injection at bus k, may be positive or negative. Thus we see that congestion, although usually increasing LMPs for most buses, can also decrease LMPs under certain conditions.

We will study the effects of congestion on LMPs in some depth in the next set of notes.

---

[1] Midwest ISO Training Materials, "Bids and Offers," 2008, available at http://www.midwestiso.org/publish/Folder/10b1ff_101f945f78e_-7b9a0a48324a?rev=1.
[2] Eugene Litvinov, Tongxin Zheng, Gary Rosenwald, and Payman Shamsollahi, "Marginal Loss Modeling in LMP Calculation," IEEE Transactions On Power Systems, Vol. 19, No. 2, May 2004.