# Microprice, Bellman Values, and Semi-Markov Reinforcement Learning

*Paper II: companion note to* Microprice as a Centered Poisson Corrector

Lev Petersen *April 17, 2026*

**Abstract.** This note extends [1] in two ways. First, it recasts the move-time objects of the main paper – $r$, $N$, $G_1 = Nr$, $B = NT$, and the centered correction $g$ with $G^* = g + \mu\mathbf{1}$ – as value functions of an average-reward Markov chain, and derives exact event-time Bellman targets $V^{(H)}$ and $V_\gamma$ on the full chain $P$. Second, it writes down the online learning rules (differential TD and semi-Markov TD) these targets induce on price-move episodes. We preserve the notation of [1] throughout.

## 1 Discounted move-time Bellman equations

Let $P = Q + T$ and $r$ be as in [1], with $\rho(Q) < 1$, so $N = (I - Q)^{-1}$, $G_1 := Nr$, and $B := NT$. For $\gamma \in (0, 1)$, the discounted event-time Bellman value is

$$V_\gamma(x) := \mathbb{E}_x\left[\sum_{n=0}^{\infty} \gamma^n (p_{n+1} - p_n)\right]. \tag{1}$$

**Proposition 1.1** (Discounted move-time factorisation)**.** *For $\gamma \in (0, 1)$, set $G_{1,\gamma} := (I - \gamma Q)^{-1} r$ and $B_\gamma := \gamma(I - \gamma Q)^{-1} T$. Then*

$$V_\gamma = G_{1,\gamma} + B_\gamma V_\gamma = (I - B_\gamma)^{-1} G_{1,\gamma}, \tag{2}$$

*and $(I - \gamma Q)(I - B_\gamma) = I - \gamma P$, so both forms coincide with $V_\gamma = (I - \gamma P)^{-1} r$.*

*Proof.* Partition the discounted return at the first price move; summing over $n \geq 0$ no-move steps gives $V_\gamma = \sum_{n\geq 0} \gamma^n Q^n r + \sum_{n\geq 0} \gamma^{n+1} Q^n T V_\gamma = (I - \gamma Q)^{-1} r + \gamma(I - \gamma Q)^{-1} T V_\gamma$, which is (2). Since $(I - \gamma Q)(I - B_\gamma) = I - \gamma P$, multiplying the resolvent form by $I - \gamma P$ gives $V_\gamma = (I - \gamma P)^{-1} r$. □

(2) is the direct RL bridge from [1]: the only change from $(G_1, B)$ is that each no-move excursion is discounted before it is collapsed. Move-time SMDP-TD makes one update per price move instead of one per event, which on sparse-move panels compresses the per-pass work by an order of magnitude (Table 1).

*Remark* 1.2 (Event-time vs. move-time discounting). $(I - \beta B)^{-1} G_1$ of [1] discounts after the reduction, so every price-move episode carries weight $\beta^k$ regardless of its event-time duration. $V_\gamma$ discounts before the reduction, so $(G_{1,\gamma}, B_\gamma)$ carries episode-length-dependent factors. They agree only when every move occurs after a single event.

**Proposition 1.3** (Classical microprice as a move-time relative value)**.** *If $B$ is ergodic with stationary $\pi$, $\mu := \pi^\top G_1$ and $\Pi_B := \mathbf{1}\pi^\top$, then $g := (I - B + \Pi_B)^{-1}(G_1 - \mu\mathbf{1})$ satisfies*

$$g = G_1 - \mu\mathbf{1} + Bg, \qquad \pi^\top g = 0, \tag{3}$$

*and $G^* := (I - B + \Pi_B)^{-1} G_1 = g + \mu\mathbf{1}$. So the centered correction of [1] is the average-reward relative value of the move-time chain.*

*Proof.* $(I - B + \Pi_B)g = G_1 - \mu\mathbf{1}$ and $\Pi_B g = \mathbf{1}\pi^\top g = 0$ imply $(I - B)g = G_1 - \mu\mathbf{1}$. Conversely this, with $\pi^\top g = 0$, gives $\Pi_B g = 0$ and so the resolvent form. □

The undiscounted microprice correction is therefore already a value function; the discounted semi-Markov family replaces the average-reward boundary problem by a horizon-aware Bellman objective.

## 2 Event-time Bellman values and differential limits

**Proposition 2.1** (Finite-horizon Bellman values)**.** *For $H \geq 0$, with $V^{(0)} := 0$ and $V^{(h+1)} := r + PV^{(h)}$,*

$$V^{(H)} = \sum_{k=0}^{H-1} P^k r, \quad V^{(H)}(x) = \mathbb{E}_x\left[\sum_{n=0}^{H-1} (p_{n+1} - p_n)\right].$$

**Proposition 2.2** (Event-time differential value)**.** *If $P$ is ergodic with stationary $\pi_P$, $\Pi_P := \mathbf{1}\pi_P^\top$, and $\mu_P := \pi_P^\top r$, then*

$$g_P := (I - P + \Pi_P)^{-1}(r - \mu_P\mathbf{1}), \quad \pi_P^\top g_P = 0, \tag{4}$$

*and for every $H \geq 0$ and $\gamma \in (0, 1)$,*

$$V^{(H)} = H\mu_P\,\mathbf{1} + g_P - P^H g_P, \tag{5}$$

$$V_\gamma = \frac{\mu_P}{1 - \gamma}\,\mathbf{1} + (I - \gamma P)^{-1}(I - \Pi_P)r, \tag{6}$$

*so $V_\gamma - \mu_P\mathbf{1}/(1 - \gamma) \to g_P$ as $\gamma \uparrow 1$.*

*Proof.* From $(I - P)g_P = r - \mu_P\mathbf{1}$, $r = \mu_P\mathbf{1} + g_P - Pg_P$, and so $V^{(H)} = \sum_{k=0}^{H-1} P^k r = H\mu_P\mathbf{1} + g_P - P^H g_P$. For the discounted identity, decompose $r = \Pi_P r + (I - \Pi_P)r = \mu_P\mathbf{1} + (I - \Pi_P)r$. Since $P\mathbf{1} = \mathbf{1}$ and $P\Pi_P = \Pi_P$, $(I - \gamma P)^{-1}\Pi_P r = \mu_P\mathbf{1}/(1 - \gamma)$; on $(I - \Pi_P)\mathbb{R}^L$, $P^k \to 0$, so the Neumann series converges to $(I - P + \Pi_P)^{-1}(I - \Pi_P)r = g_P$. □

(5) is the event-time analogue of the main theorem of [1]: $G^*$ is the centered long-run value after the $Q/T$ reduction, while $g_P$ is the centered long-run value before it.

*Remark* 2.3 (Stationarity and horizon choice). $V^{(H)}$ requires $P$ to be approximately stationary over $H$ event-time steps. Intraday non-stationarity caps the useful range of $H$; in practice, refitting $P$ within rolling windows keeps $\|V^{(H)} - V^{(H)}_{\text{true}}\|$ bounded.

*Remark* 2.4 (Horizon-matched discounting). For a finite horizon $H$, the geometric discount $\gamma_H = 1 - 1/H$ has effective horizon $1/(1 - \gamma_H) = H$, matching the finite-horizon target in the expected number of discounted steps. This is the simplest one-shot surrogate for $V^{(H)}$ and the choice used below.

## 3 Learning rules

### 3.1 Event-time TD

$V_\gamma = r + \gamma P V_\gamma$ is learnable online by temporal-difference updates [2]: with $X_t$ the discretised state and $R_t := p_{t+1} - p_t$,

$$V_{t+1}(X_t) \leftarrow V_t(X_t) + \eta_t\big(R_t + \gamma V_t(X_{t+1}) - V_t(X_t)\big). \quad (7)$$

With linear features $\phi$, least-squares TD [3] solves $A_T w_T = b_T$ with $A_T = \sum_{t<T} \phi_t(\phi_t - \gamma\phi_{t+1})^\top$, $b_T = \sum_{t<T} \phi_t R_t$; equivalently, the successor representation [4] $M_\gamma = (I - \gamma P)^{-1} = \sum_{k\geq 0} \gamma^k P^k$ gives $V_\gamma = M_\gamma r$.

### 3.2 Move-time differential TD for classical microprice

By Proposition 1.3, the centered correction of [1] is an average-reward Bellman value on move time. With $\Delta P_k := p_{\tau_{k+1}} - p_{\tau_k}$, $g(y) = \mathbb{E}[\Delta P_k - \mu + g(Y_{k+1}) \mid Y_k = y]$, and the tabular differential-TD recursion is

$$\delta_k = \Delta P_k - \bar{\mu}_k + h_k(Y_{k+1}) - h_k(Y_k), \quad (8)$$

$$h_{k+1}(Y_k) \leftarrow h_k(Y_k) + \alpha_k \delta_k, \qquad \bar{\mu}_{k+1} \leftarrow \bar{\mu}_k + \beta_k \delta_k.$$

Same move-time chain, same state space, same centered boundary object as classical microprice.

### 3.3 Move-time semi-Markov TD for discounted forecasting

For horizon-aware forecasting, replace the average-reward problem by the discounted semi-Markov value. With $D_k := \tau_{k+1} - \tau_k$ and $Y_k := X_{\tau_k}$,

$$W_\gamma(y) = \mathbb{E}\big[\gamma^{D_k - 1}\Delta P_k + \gamma^{D_k} W_\gamma(Y_{k+1}) \mid Y_k = y\big], \quad (9)$$

with $\gamma^{D_k - 1}$ because the price change lands on the last event of the episode. The SMDP-TD(0) update is

$$W_{k+1}(Y_k) \leftarrow W_k(Y_k) + \eta_k\big(\gamma^{D_k-1}\Delta P_k + \gamma^{D_k} W_k(Y_{k+1}) - W_k(Y_k)\big). \quad (10)$$

Differential TD learns the microprice boundary; SMDP-TD learns the horizon-aware discounted predictor on the same episodes. Model-based counterparts are the exact $V^{(H)}$ and its one-shot surrogate $V_{\gamma_H}$ with $\gamma_H = 1 - 1/H$ (Remark 2.4).

## 4 Empirical results

We evaluate the Bellman targets on the bundled public panels (Stoikov BAC/CVX, IEX BAC/SPY/GLD, WSELOB PEKAO/PZU/PKOBP/PKNORLEN/KGHM) using the grouped conditional-curve metric of [1].

**Table 1.** Bundled-data summary. Geo. ratio = RMSE/RMSE($G^*$); below one favours the RL extension. Events/move is the upper bound on per-pass SMDP-TD compression vs. event-time TD.

| *accuracy vs. $G^*$* | | |
|---|---|---|
| Predictor | Wins/19 | Geo. ratio |
| $V^{(H)}$ | 12/19 | 0.759 |
| $V_{\gamma_H}$ | 9/19 | 0.795 |
| $g_P$ | IEX 6/6; else ties | 0.999 |
| *events per move (geo. mean)* | | |
| Stoikov | 10.4× | |
| IEX | 5.2× | |
| WSELOB | 41.1× | |
| overall | 16.8× (median 30.4×) | |

Table 1 gives the main message. $V^{(H)}$ beats classical microprice in 12/19 cells with geometric-mean RMSE ratio 0.759 (≈ 24% reduction); $V_{\gamma_H}$ wins in 9/19 with ratio 0.795 (≈ 20% reduction). Figure 1 resolves the counts into per-cell texture: the short-horizon WSELOB cells are where the RL extension gains most, while IEX regresses in a handful of cells – the discount surrogate in particular struggles on BAC-60s and SPY-60s.

| | | $V^{(H)}$ | $V_{\gamma_H}$ |
|---|---|---|---|
| *Stoikov* | BAC, 60s | 0.974 | 0.515 |
| | BAC, 180s | 1.001 | 0.790 |
| | CVX, 10s | 1.030 | 1.234 |
| *IEX* | BAC, 10s | 0.997 | 0.999 |
| | BAC, 60s | 1.000 | 1.937 |
| | GLD, 10s | 1.018 | 1.011 |
| | GLD, 60s | 0.985 | 0.982 |
| | SPY, 10s | 1.040 | 1.041 |
| | SPY, 60s | 1.240 | 1.253 |
| *WSELOB* | KGHM, 10s | 0.784 | 0.789 |
| | KGHM, 60s | 1.007 | 1.038 |
| | PEKAO, 10s | 0.220 | 0.249 |
| | PEKAO, 60s | 0.935 | 1.051 |
| | PKNORLEN, 10s | 0.495 | 0.525 |
| | PKNORLEN, 60s | 1.019 | 1.088 |
| | PKOBP, 10s | 0.172 | 0.192 |
| | PKOBP, 60s | 0.980 | 1.072 |
| | PZU, 10s | 0.261 | 0.298 |
| | PZU, 60s | 1.130 | 1.340 |
| | *geo. mean* | **0.759** | **0.795** |

**Figure 1.** Per-cell RMSE ratio of $V^{(H)}$ and $V_{\gamma_H}$ against classical microprice $G^*$. Moss cells are improvements (ratio < 1), clay cells regressions. Dataset groups are separated by horizontal rules; the footer strip shows the column geometric means.

> *Remark* 4.1 (Wilson interval on win counts). A null of "no difference" gives 9.5/19 expected wins; the observed 12/19 has a Wilson 95% confidence interval of roughly [41%, 81%] on the win rate, so the improvement is indicative but not overwhelming on a per-cell basis. The geometric-mean ratios are the more robust headline.

The event-time boundary value $g_P$ does what the theory predicts. On Stoikov and WSELOB panels, already close to the centered regime of [1], $g_P$ and $G^*$ coincide to numerical tolerance. On the six IEX cells, where event and move time are more visibly separated, $g_P$ improves every cell with geometric-mean ratio 0.998. It is the correct event-time boundary object, though not a replacement for the objective-matched finite-horizon value.

Whether the per-pass saving translates into fewer samples to convergence is an empirical question left to a learning-curve study (range 3.7×–56.9× across cells). The discounted linear solve remains the cleanest model-based object: solving $(I - \gamma P)V = r$ is 19.9× cheaper on average than the classical microprice solve stage.

## 5 Conclusion

The move-time reduction of [1] admits exact event-time Bellman objectives $V^{(H)}$, $V_\gamma$, and $g_P$ together with differential and SMDP TD learning rules on price-move episodes: $V^{(H)}$ matches the fixed-horizon objective, $V_{\gamma_H}$ is the cleanest one-shot surrogate, $g_P$ is the correct event-time boundary object, and move-time TD delivers an order-of-magnitude per-pass saving on sparse-move panels.

## References

[1] L. Petersen, *Microprice as a Centered Poisson Corrector*, preprint, 2025.

[2] R. S. Sutton, *Learning to Predict by the Methods of Temporal Differences*, Machine Learning, 3(1):9–44, 1988.

[3] S. J. Bradtke and A. G. Barto, *Linear Least-Squares Algorithms for Temporal Difference Learning*, Machine Learning, 22:33–57, 1996.

[4] P. Dayan, *Improving Generalization for Temporal Difference Learning: The Successor Representation*, Neural Computation, 5(4):613–624, 1993.