

Battery SOC: Coulomb Counting vs. Kalman Estimation

The equivalent-circuit state space, the full EKF recursion, what sets the gain, and dual estimation for state of health.

1. Coulomb counting and why it drifts

With discharge current positive, coulombic efficiency η (about 1 for discharge), and capacity Q in ampere-seconds ($3600 \times \text{Ah}$):

$$\text{SOC}(t) = \text{SOC}(t_0) - \frac{1}{Q} \int_{t_0}^t \eta i(\tau) d\tau.$$

Three error sources, all uncorrected: a current-sensor bias b integrates to $b t / Q$, growing without bound; the initial condition $\text{SOC}(t_0)$ may be unknown; and Q itself fades with age and varies with temperature.

How fast does bias drift?

A 50 mA bias on a 100 Ah cell is 0.05 Ah per hour, or 0.05% SOC per hour. After 240 operating hours without a re-anchor the estimate is off by 12%.

2. Equivalent-circuit model as a state space

A first-order Thevenin model is an OCV source, a series resistance R_0 , and one $R_1 C_1$ pair. Discretize with step Δt and $\alpha = e^{-\Delta t / (R_1 C_1)}$:

$$\underbrace{\begin{bmatrix} \text{SOC} \\ v_1 \end{bmatrix}}_{\mathbf{x}_{k+1}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix}}_{\mathbf{A}} \mathbf{x}_k + \underbrace{\begin{bmatrix} -\eta \Delta t / Q \\ R_1 (1 - \alpha) \end{bmatrix}}_{\mathbf{B}} i_k + \mathbf{w}_k,$$

$$y_k = \text{OCV}(\text{SOC}_k) - v_{1,k} - R_0 i_k + v_k.$$

The state equation is coulomb counting plus RC relaxation. The output equation ties the hidden SOC to the measured terminal voltage through the OCV curve. $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{Q}_w)$ and $v_k \sim \mathcal{N}(0, R_v)$ model process and sensor noise. Many BMSs use two RC pairs (fast and slow), which simply adds another state.

3. The EKF recursion

The output is nonlinear in SOC, so linearize it at the predicted state:

$$\mathbf{C}_k = \left. \frac{\partial y}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}_k^-} = [s_k \quad -1], \quad s_k = \left. \frac{d \text{OCV}}{d \text{SOC}} \right|_{\widehat{\text{SOC}}_k^-}.$$

Predict

$$\hat{\mathbf{x}}_k^- = \mathbf{A}\hat{\mathbf{x}}_{k-1}^+ + \mathbf{B}i_{k-1}, \quad \mathbf{P}_k^- = \mathbf{A}\mathbf{P}_{k-1}^+\mathbf{A}^\top + \mathbf{Q}_w.$$

Update

$$\begin{aligned} \mathbf{K}_k &= \mathbf{P}_k^- \mathbf{C}_k^\top (\mathbf{C}_k \mathbf{P}_k^- \mathbf{C}_k^\top + R_v)^{-1}, \\ \hat{\mathbf{x}}_k^+ &= \hat{\mathbf{x}}_k^- + \mathbf{K}_k (y_k - \hat{y}_k), \quad \hat{y}_k = \text{OCV}(\widehat{\text{SOC}}_k^-) - \hat{v}_{1,k} - R_0 i_k, \\ \mathbf{P}_k^+ &= (\mathbf{I} - \mathbf{K}_k \mathbf{C}_k) \mathbf{P}_k^- (\mathbf{I} - \mathbf{K}_k \mathbf{C}_k)^\top + \mathbf{K}_k R_v \mathbf{K}_k^\top \quad (\text{Joseph form, numerically robust}). \end{aligned}$$

The SOC entry of $\mathbf{Q}_w$ is where coulomb-counting drift lives: it makes $\mathbf{P}$ grow every prediction, until a voltage update shrinks it again.

4. What actually sets the gain

Ignoring the RC state for intuition, the SOC gain and posterior variance are

$$K = \frac{P s}{s^2 P + R_v}, \quad P^+ = P (1 - K s) = \frac{P R_v}{s^2 P + R_v}.$$

The gain depends on prior uncertainty P , sensor noise R_v , and **slope** s , and not directly on load current. A good model predicts the $R_0 i$ and RC sag, so voltage under load stays informative. BMSs back the gain off during transients only by inflating R_v when they expect the model to be wrong (adaptive filtering).

Steep curve vs. flat plateau (numbers)

Prior SOC uncertainty $\sigma = 5\%$ ($P = 0.0025$), voltage noise 5 mV ($R_v = 2.5 \times 10^{-5} \text{ V}^2$):

Region	Slope s (V per unit SOC)	Ks	Posterior σ
NMC, mid-range	≈ 0.7	0.98	0.7%
LFP, flat plateau	≈ 0.03	0.08	4.8%

On the steep curve a single measurement collapses the uncertainty. On the LFP plateau it barely moves it, because the voltage simply contains little SOC information there. No filter can fix that; LFP packs re-anchor at the steep ends of the curve or at rest.

5. The unscented alternative

Instead of a Jacobian, the UKF propagates $2n + 1$ **sigma points** through the true nonlinear functions:

$$\begin{aligned} \mathbf{x}_0 &= \hat{\mathbf{x}}, \quad \mathbf{x}_i = \hat{\mathbf{x}} \pm \left(\sqrt{(n + \lambda) \mathbf{P}} \right)_i, \quad \lambda = \alpha^2 (n + \kappa) - n, \\ W_0^{(m)} &= \frac{\lambda}{n + \lambda}, \quad W_0^{(c)} = W_0^{(m)} + (1 - \alpha^2 + \beta), \quad W_i = \frac{1}{2(n + \lambda)}. \end{aligned}$$

The predicted mean and covariance are weighted sums over the transformed points, accurate to second order for any nonlinearity versus first order for the EKF. That matters where the OCV curve bends sharply (near empty and near full). The cost is about $2n + 1$ model evaluations per step, which is trivial for $n = 2-4$.

6. Dual estimation: tracking state of health

Treat slow parameters as random-walk states, $\theta_{k+1} = \theta_k + \mathbf{r}_k$, with $\theta = [Q^{-1}, R_0]$, and either augment the main filter (joint estimation) or run a second, slower filter fed by the first (dual estimation). Then

$$\text{SOH}_Q = \frac{Q_{\text{now}}}{Q_{\text{new}}}, \quad \text{SOH}_R = \frac{R_{0,\text{new}}}{R_{0,\text{now}}} \text{ (or the resistance-growth ratio).}$$

Observability caveat: capacity is only identifiable when SOC moves substantially between good voltage anchors, and R_0 is best seen at current steps. Real BMSs gate parameter updates on those conditions.

7. Self-test, worked

Exercise 1: reason the gain at two operating points

(a) At rest, steep region: s is large and the model's RC states have relaxed, so $Ks \rightarrow 1$. The filter trusts the voltage, pulls SOC to the OCV-implied value, and erases accumulated coulomb-counting drift. **(b)**

Heavy current on a flat plateau: the gain is small, and the cause is the *flat curve*, not the current: $s \approx 0$ makes $K \approx 0$ regardless of load. The filter coasts on the model's prediction, with P growing, until the next informative region. Load matters only if the model is poor at high current, which is what adaptive R_v handles.

Exercise 2: synthesis

Coulomb counting is smooth and needs no model of the voltage, but it is an open-loop integrator: bias, unknown initial SOC, and capacity error all accumulate. Voltage lookup is absolute (no drift) but valid only at rest and ill-conditioned wherever the OCV curve is flat. The Kalman filter uses coulomb counting as its *prediction*, tracks the drift as growing covariance, and uses the voltage *measurement* to correct the prediction in proportion to how informative the measurement is right now. Each method covers the other's failure mode.