Choosing an Adaptation Mechanism and Gain

The most logical approach to the adaptation problem is to assume a certain model for how the "true" parameters θ_0 change. A typical choice is to describe these parameters as a random walk.

$$\theta_0(t) = \theta_0(t-1) + w(t) \tag{3-58}$$

Here $w(t)$ is assumed to be white Gaussian noise with covariance matrix

$$Ew(t)w^T(t) = R_1 \tag{3-59}$$

Suppose that the underlying description of the observations is a linear regression (3-57). An optimal choice of $Q(t)$ in (3-55)-(3-56) can then be computed from the Kalman filter, and the complete algorithm becomes

$$\begin{aligned}
\hat{\theta}(t) &= \hat{\theta}(t-1) + K(t)(y(t) - \hat{y}(t)) \\
\hat{y}(t) &= \psi^T(t)\hat{\theta}(t-1) \\
K(t) &= Q(t)\psi(t) \\
Q(t) &= \frac{P(t-1)}{R_2 + \psi(t)^T P(t-1)\psi(t)} \\
P(t) &= P(t-1) + R_1 - \frac{P(t-1)\psi(t)\psi(t)^T P(t-1)}{R_2 + \psi(t)^T P(t-1)\psi(t)}
\end{aligned} \tag{3-60}$$

Here R_2 is the variance of the innovations $e(t)$ in (3-57): $R_2 = Ee^2(t)$ (a scalar). The algorithm (3-60) will be called the **Kalman filter (KF) approach** to adaptation, with *drift matrix* R_1. See eq (11.66)-(11.67) in Ljung (1999). The algorithm is entirely specified by R_1, R_2, $P(0)$, $\theta(0)$, and the sequence of data $y(t)$, $\psi(t)$, $t = 1$, *2*,.... Even though the algorithm was motivated for a linear regression model structure, it can also be applied in the general case where $\hat{y}(t)$ is computed in a different way from (3-60b).

Another approach is to discount old measurements exponentially, so that an observation that is τ samples old carries a weight that is λ^τ of the weight of the most recent observation. This means that the following function is minimized rather than (3-39)

$$\sum_{k=1}^{t} \lambda^{t-k} e^2(k) \tag{3-61}$$

at time *t*. Here λ is a positive number (slightly) less than 1. The measurements that are older than $\tau = 1/(1-\lambda)$ carry a weight in the expression above that is less than about 0.3. Think of $\tau = 1/(1-\lambda)$ as the *memory horizon* of the approach. Typical values of λ are in the range 0.97- 0.995.

The criterion (3-61) can be minimized exactly in the linear regression case giving the algorithm (3-60abc) with the following choice of $Q(t)$.

$$Q(t) = P(t) = \frac{P(t-1)}{\lambda + \psi(t)^T P(t-1)\psi(t)} \qquad \textbf{(3-62)}$$

$$P(t) = \frac{1}{\lambda}\left(P(t-1) - \frac{P(t-1)\psi(t)\psi(t)^T P(t-1)}{\lambda + \psi(t)^T P(t-1)\psi(t)}\right)$$

This algorithm will be called the **Forgetting Factor (FF) approach** to adaptation, with the *forgetting factor* λ. See eq (11.63) in Ljung (1999). The algorithm is also known as *recursive least squares* (RLS) in the linear regression case. Note that $\lambda = 1$ in this approach gives the same algorithm as $R_1 = 0, R_2 = 1$ in the Kalman filter approach.

A third approach is to allow the matrix $Q(t)$ to be a multiple of the identity matrix.

$$Q(t) = \gamma I \qquad \textbf{(3-63)}$$

It can also be normalized with respect to the size of ψ.

$$Q(t) = \frac{\gamma}{|\psi(t)|^2} I \qquad \textbf{(3-64)}$$

See eqs (11.45) and (11.46), respectively in Ljung (1999). These choices of $Q(t)$ move the updates of $\hat{\theta}$ in (3-55) in the negative gradient direction (with respect to θ) of the criterion (3-39). Therefore, (3-63) will be called the **Unnormalized Gradient (UG) approach** and (3-64) the **Normalized Gradient (NG) approach** to adaptation, with gain γ. The gradient methods are also known as *least mean squares* (LMS) in the linear regression case.