Here we develop a Kalman filter model where the subject uses the pain perception on each trial, P_t, as the observation, rather than the sensory value N_t. Thus the subject assumes a latent random walk process (x_t) governed by

$$x_{t+1} \sim \mathcal{N}\left(x_t, \sigma_\eta^2\right),$$

with the current pain value governed by

$$P_t \sim \mathcal{N}\left(x_t, \sigma_\psi^2\right).$$

We assume also that the subject doesn't directly observe P_t but infers it using a sensory value N_t. (The sensory value can be modeled either as the true stimulus value or as a noisy version thereof.) The subject assumes N_t is an imperfect indicator of P_t, with

$$P_t - N_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right).$$

The difference $P_t - N_t$ represents other contributions to pain that are independent of sensory input and of pain expectation (i.e. belief about x). We can rewrite the relationship between N_t and P_t as

$$N_t \sim \mathcal{N}\left(P_t, \sigma_\varepsilon^2\right).$$

This system requires two levels of inference on the part of the subject. First, the subject must infer the true pain value on each trial, P_t, based on prior expectations combined with the current sensory input, N_t. Second, the subject must infer the current value of x_t based on P_t, in order to generate expectations for the next trial. Extending the standard Kalman filter, we assume the subject maintains a conjugate iterative prior on x_t conditioned on all previous observations, $\mathbf{N}_{t-1} = (N_1, \ldots, N_{t-1})$:

$$x_t|\mathbf{N}_{t-1} \sim \mathcal{N}\left(\mu_t, s_t^2\right).$$

Both levels of inference in the model are based on this iterative prior.

To derive the first level of inference, note that the prior on x_t also yields a prior on P_t:

$$P_t|\mathbf{N}_{t-1} \sim \mathcal{N}\left(\mu_t, \sigma_\psi^2 + s_t^2\right).$$

The mean of this prior is the subject's reported expectation at the beginning of a trial (i.e., after the cue has been observed):

$$E_t = \mu_t.$$

Once N_t is observed, it can be combined with the prior to derive a posterior for P_t:

$$P_t|\mathbf{N}_t \sim \mathcal{N}\left(\frac{\sigma_\varepsilon^2\mu_t + \left(\sigma_\psi^2 + s_t^2\right)N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}, \frac{\sigma_\varepsilon^2\left(\sigma_\psi^2 + s_t^2\right)}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}\right).$$

The mean of this posterior is the subject's pain report:

$$\hat{P}_t = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}\mu_t + \frac{\sigma_\psi^2 + s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}N_t.$$

Thus reported pain is a weighted average of expectations and experienced heat.

We can model the second level of inference in two ways. Model 1 assumes the subject treats $\hat{P}_t$ as veridical, discarding the uncertainty in the posterior $p\left(P_t|\mathbf{N}_t\right)$ and simply taking $P_t = \hat{P}_t$. The posterior on x_t will then be estimated by the subject as

$$\begin{aligned} x_t|\mathbf{N}_t &\sim \mathcal{N}\left(\frac{\sigma_\psi^2\mu_t + s_t^2\hat{P}_t}{\sigma_\psi^2 + s_t^2}, \frac{\sigma_\psi^2 s_t^2}{\sigma_\psi^2 + s_t^2}\right) \\ &= \mathcal{N}\left(\frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right)\mu_t + s_t^2 N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}, \frac{\sigma_\psi^2 s_t^2}{\sigma_\psi^2 + s_t^2}\right). \end{aligned}$$

The prior on the next trial is then obtained by adding the variance of the random walk:

$$x_{t+1}|\mathbf{N}_t \sim \mathcal{N}\left(\frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right)\mu_t + s_t^2 N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}, \frac{\sigma_\psi^2 s_t^2}{\sigma_\psi^2 + s_t^2} + \sigma_\eta^2\right).$$

By definition, this last distribution equals $\mathcal{N}\left(\mu_{t+1}, s_{t+1}^2\right)$, so we have the update equations

$$\mu_{t+1} = \frac{\sigma_\varepsilon^2 + \sigma_\psi^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}\mu_t + \frac{s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}N_t$$

and

$$s_{t+1}^2 = \frac{\sigma_\psi^2 s_t^2}{\sigma_\psi^2 + s_t^2} + \sigma_\eta^2.$$

Model 2 assumes the subject takes account of the uncertainty in $\hat{P}_t$ as an estimate of P_t, to obtain the true Bayesian posterior for x_t. The easiest way to see the result in this case is to derive the posterior directly from N_t, noting that

$$N_t \sim \mathcal{N}\left(x_t, \sigma_\varepsilon^2 + \sigma_\psi^2\right).$$

The posterior is then given by

$$x_t|\mathbf{N}_t \sim \mathcal{N}\left(\frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right)\mu_t + s_t^2 N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}, \frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right) s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}\right),$$

and the prior for the next trial by

$$x_{t+1}|\mathbf{N}_t \sim \mathcal{N}\left(\frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right)\mu_t + s_t^2 N_t}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}, \frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right) s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} + \sigma_\eta^2\right).$$

Therefore the update equation for the mean is the same as in Model 1,

$$\mu_{t+1} = \frac{\sigma_\varepsilon^2 + \sigma_\psi^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}\mu_t + \frac{s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} N_t,$$

and the update for the variance is given by

$$s_{t+1}^2 = \frac{\left(\sigma_\varepsilon^2 + \sigma_\psi^2\right) s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} + \sigma_\eta^2.$$

Therefore the two versions of the model differ only in the update for the uncertainty.

Finally, it is useful to rewrite the update equations in terms of the observed responses, E_t and $\hat{P}_t$. From above, we have

$$\hat{P}_t = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} E_t + \frac{\sigma_\psi^2 + s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} N_t.$$

$$\begin{aligned} E_{t+1} &= \frac{\sigma_\varepsilon^2 + \sigma_\psi^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} E_t + \frac{s_t^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2} N_t \\ &= \frac{\sigma_\psi^2}{\sigma_\psi^2 + s_t^2} E_t + \frac{s_t^2}{\sigma_\psi^2 + s_t^2} \hat{P}_t. \end{aligned}$$

It's important to note that this model differs from the standard Kalman filter with only two variables (true pain P_t and sensory input N_t, as in my previous writeup) instead of three (x_t , P_t, and N_t), because in the standard model the inferred pain on trial t ($\hat{P}_t$) and the expectation prior to trial $t+1$ (E_{t+1}) are identical. Furthermore, the present three-variable model is closely related to the RL model. Specifically, if we define learning rates

$$\gamma_t = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\psi^2 + s_t^2}$$

and

$$\alpha_t = \frac{s_t^2}{\sigma_\psi^2 + s_t^2},$$

then the learning rules can be written as

$$\hat{P}_t = \gamma_t E_t + \left(1 - \gamma_t\right) N_t$$

and

$$\begin{aligned} E_{t+1} &= \alpha_t \hat{P}_t + \left(1 - \alpha_t\right) E_t \\ &= \alpha_t \left(1 - \gamma_t\right) N_t + \left(1 - \alpha_t \left(1 - \gamma_t\right)\right) E_t. \end{aligned}$$

Other than the time-dependence of the learning rates, these equations are identical to the learning rules for the RL model, including the two formally equivalent expressions for E_{t+1}.