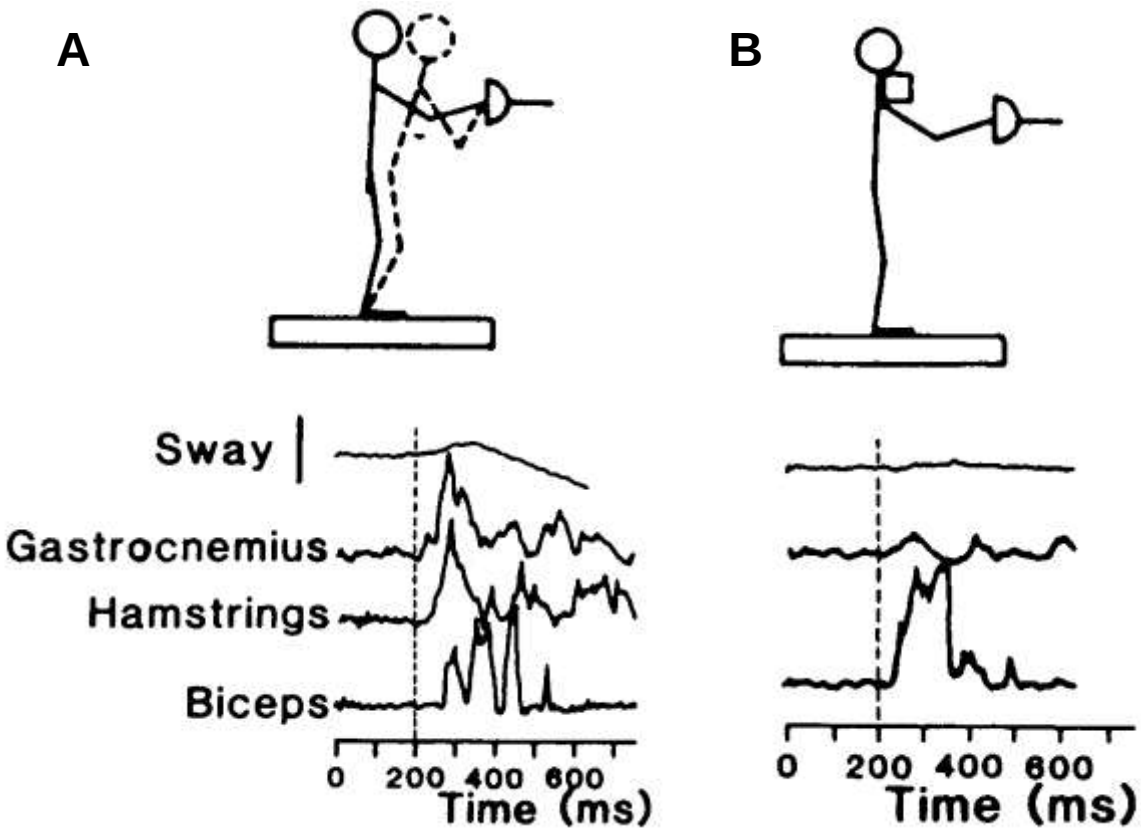
580.691 Learning Theory

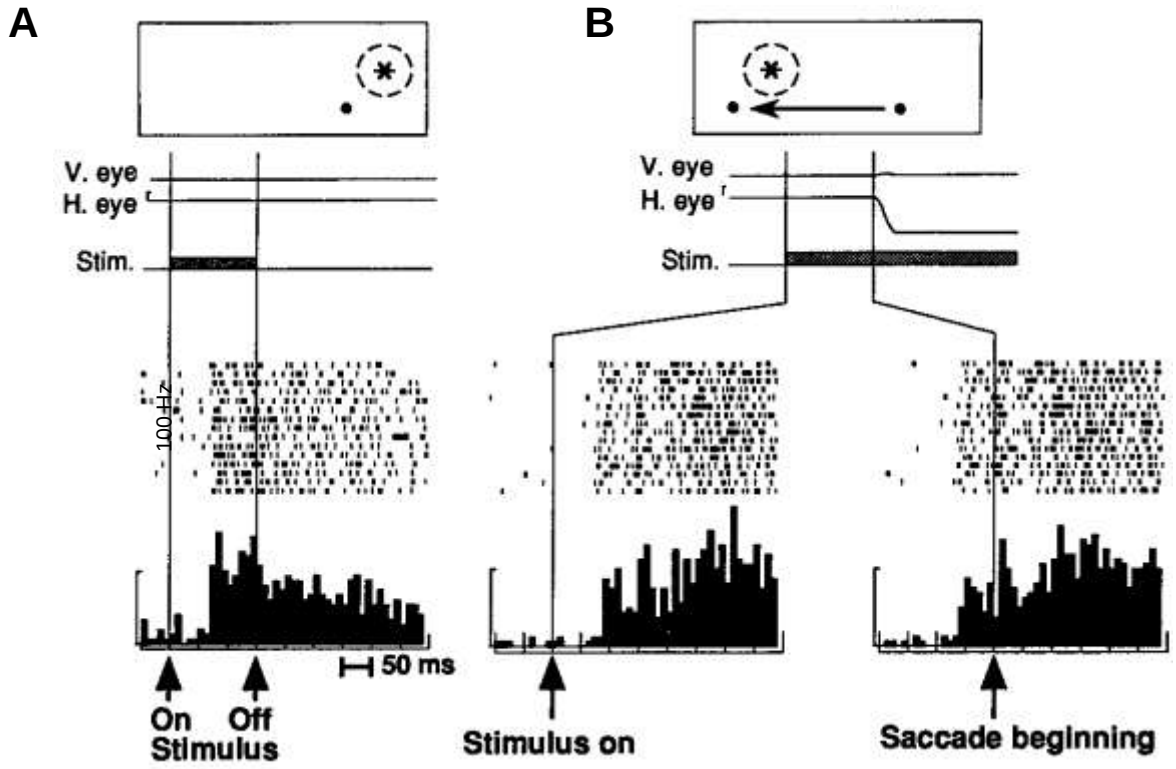
Reza Shadmehr

State estimation theory

1. Kalman Filter
2. Estimation with signal dependent noise

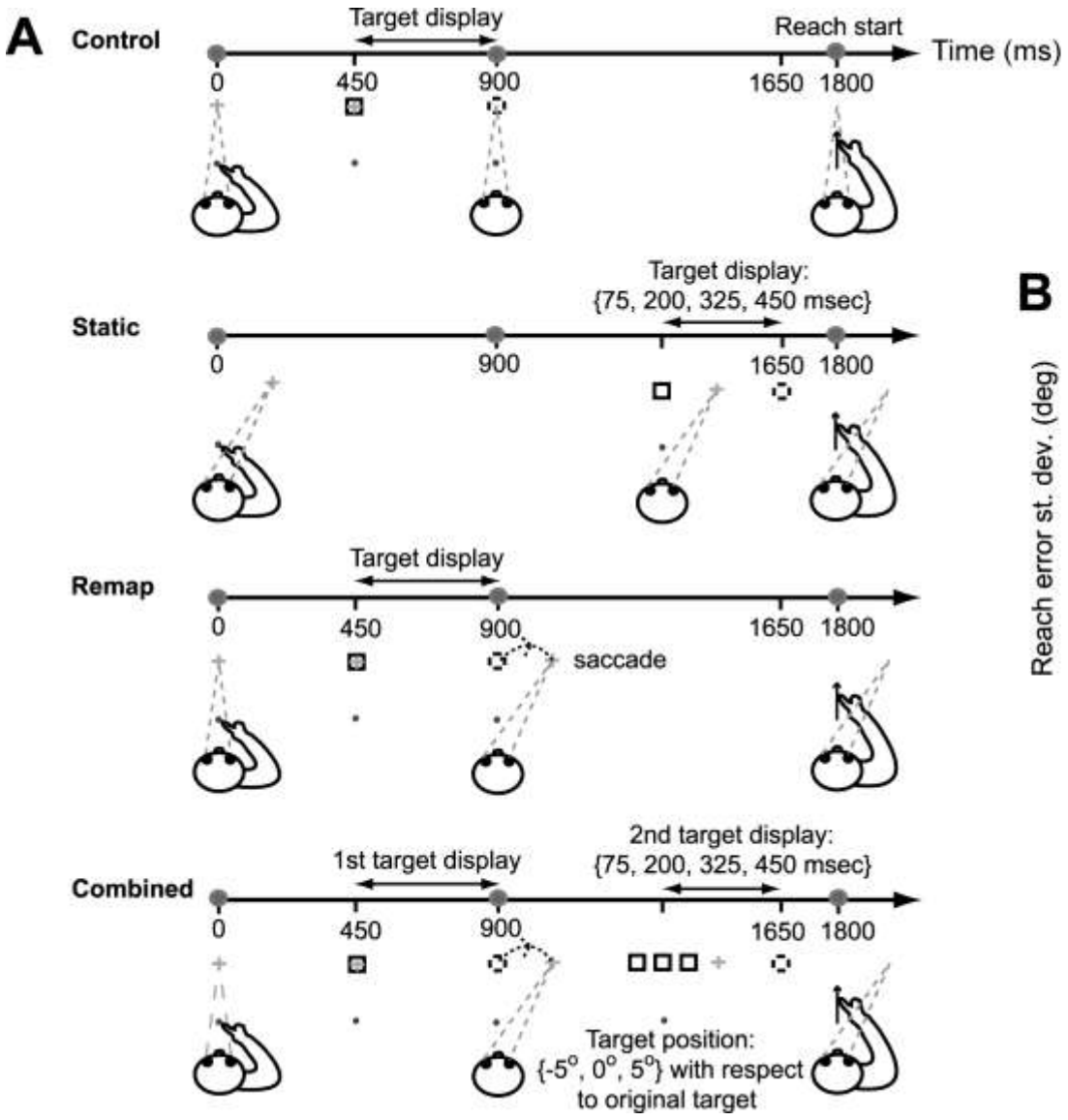


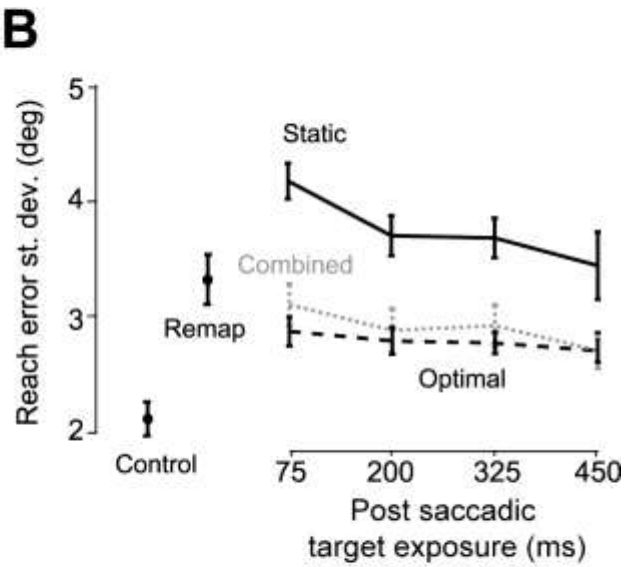
Subject was instructed to pull on a knob that was fixed on a rigid wall. A) EMG recordings from arm and leg muscles. Before biceps is activated, the brain activates the leg muscles to stabilize the lower body and prevent sway due to the anticipated pulling force on the upper body. B) When a rigid bar is placed on the upper body, the leg muscles are not activated when biceps is activated. (Cordo and Nashner, 1982)

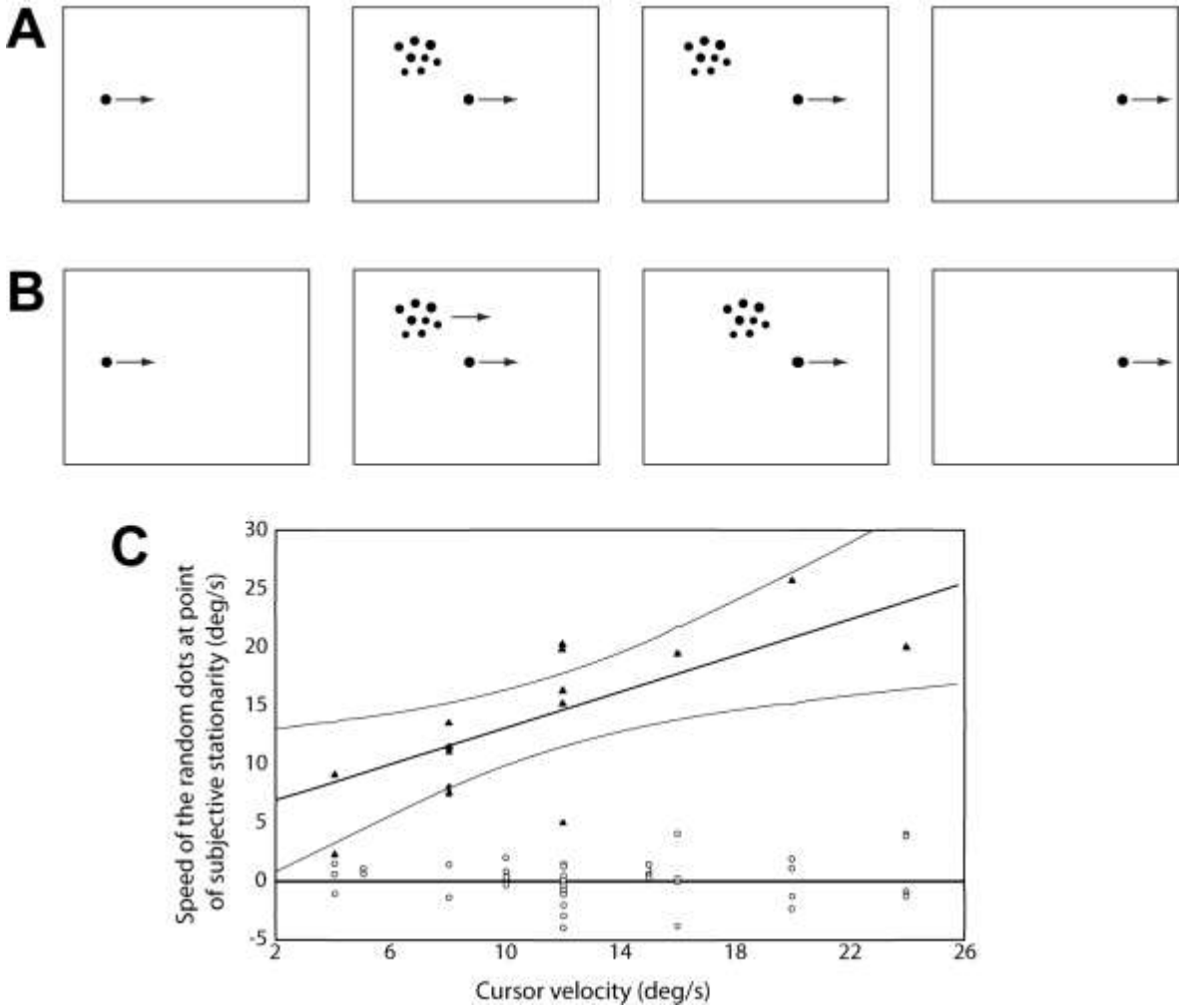


Effect of eye movement on the memory of a visual stimulus. In the top panel, the filled circle represents the fixation point, the asterisk indicates the location of the visual stimulus, and the dashed circle indicates the receptive field a cell in the LIP region of the parietal cortex. A) Discharge to the onset and offset of a visual stimulus in the cell's receptive field. Abbreviations: H. eye, horizontal eye position; Stim, stimulus; V. eye, vertical eye position. B) Discharge during the time period in which a saccade brings the stimulus into the cell's receptive field. The cell's discharge increased before the saccade brought the stimulus into the cell's receptive field. (From (Duhamel et al., 1992)

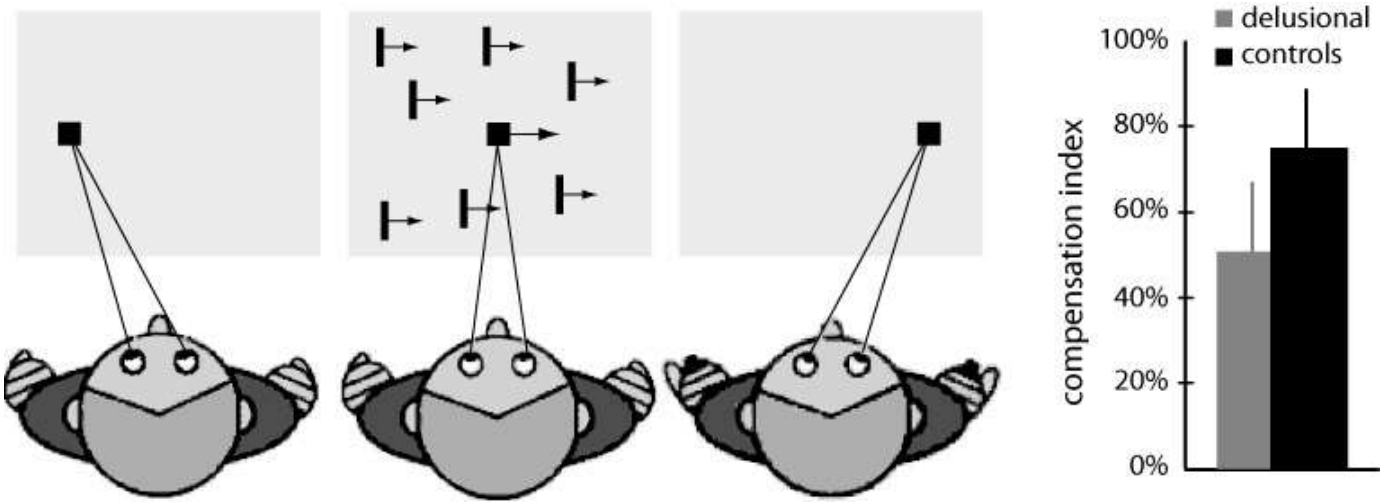
Why predict sensory consequences of motor commands?





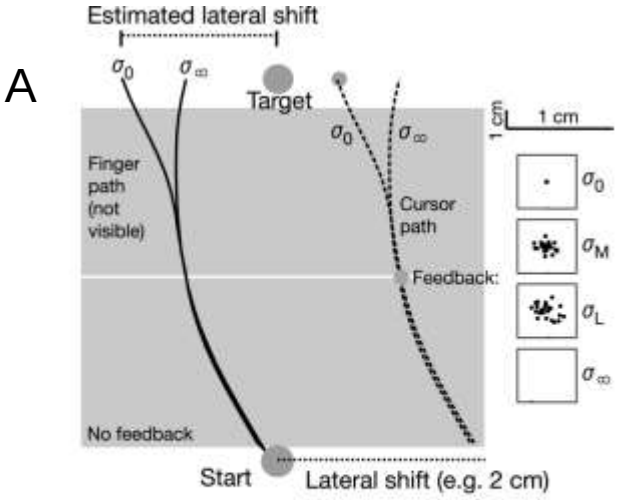


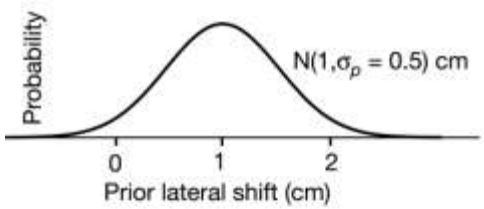
Subject looked at a moving cursor while a group of dots appeared on the screen for 300ms. In some trials the dots would remain still (A) while in other trials they would move together left or right with a constant speed (B). Subject indicated the direction of motion of the dots. From this result, the authors estimated the speed of subjective stationarity, i.e., the speed of dots for which the subject perceived them to be stationary. C) The unfilled circles represent performance of control subjects. Regardless of the speed of the cursor, they perceived the dots to be stationary only if their speed was near zero. The filled triangles represent performance of subject RW. As the speed of the cursor increased, RW perceived the dots to be stationary if their speed was near the speed of the cursor. (Haarmeier et al., 1997)

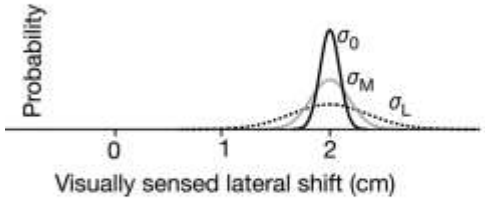


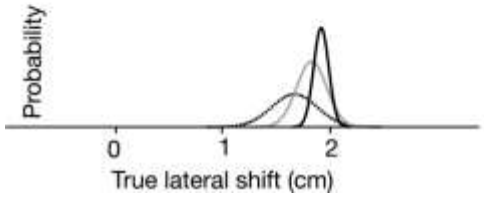
Disorders of agency in schizophrenia relate to an inability to compensate for sensory consequences of self-generated motor commands. In a paradigm similar to that shown in the last figure, volunteers estimated whether during motion of a cursor the background moved to the right or left. By varying the background speed, at each cursor speed the experimenters estimated the speed of perceptual stationarity, i.e., the speed of background motion for which the subject saw the background to be stationary. They then computed a compensation index as the difference between speed of eye movement and speed of background when perceived to be stationary, divided by speed of eye movement. The subset of schizophrenic patients who had delusional symptoms showed a greater deficit than control in their ability to compensate for sensory consequences of self-generated motor commands. (From (Lindner et al., 2005))

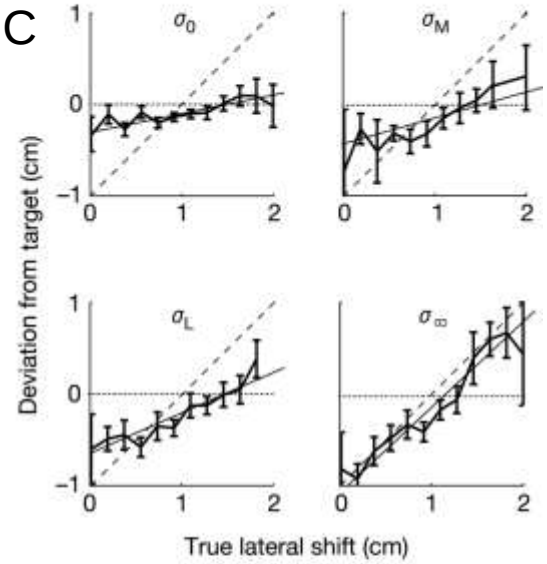
Combining predictions with observations

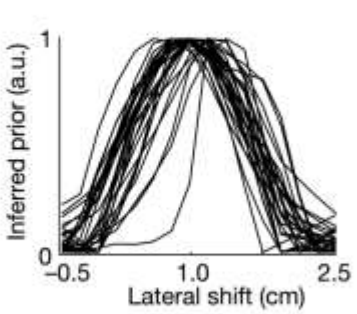




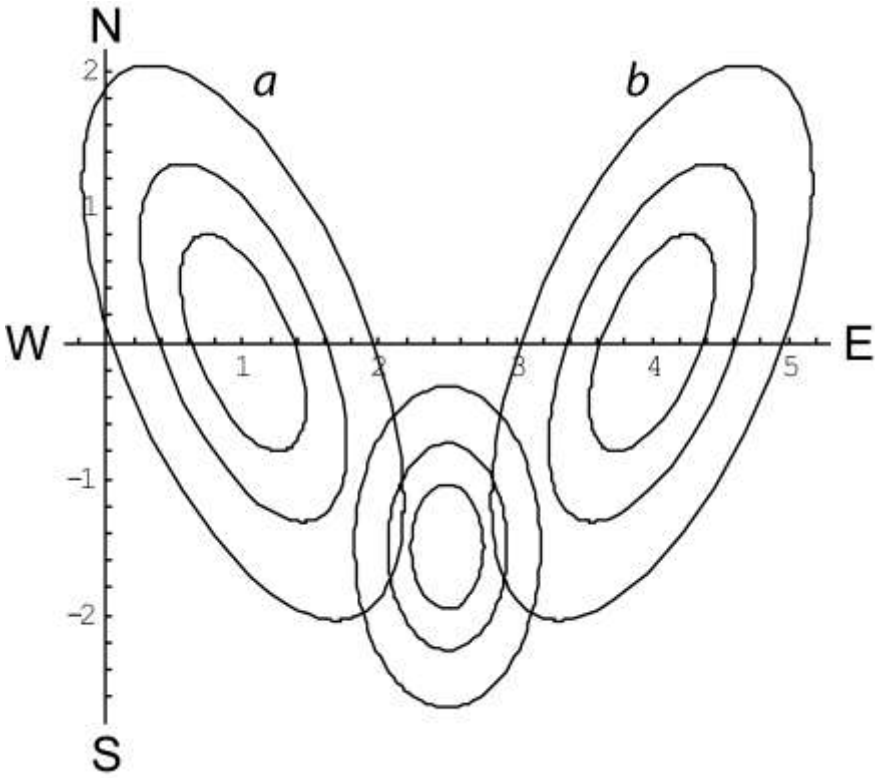








State estimation: the hiking in the woods problem



Device *a* and device *b* provide independent estimates of a hidden variable (position on a map). Each device has a Gaussian noise property. The ellipses describe the region centered on the mean of the distribution that contains 10%, 25%, and 50% of the data under the distribution. The maximum likelihood estimate of the hidden variable is marked by the distribution at the center.

$$\mathbf{y} = C\mathbf{x} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(0, R) \qquad R = \begin{bmatrix} R_a & 0 \\ 0 & R_b \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} I_{2x2} \\ I_{2x2} \end{bmatrix}$$

Maximum likelihood estimate for the hiking problem.

$$\mathbf{y} \sim N\left(C\mathbf{x}, C\,\mathrm{var}(\mathbf{x})C^T + R\right)$$

$$p(\mathbf{y}|\mathbf{x}) = N(C\mathbf{x}, R)$$

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^4 |R|}} \exp\left[-\frac{1}{2}(\mathbf{y} - C\mathbf{x})^T R^{-1}(\mathbf{y} - C\mathbf{x})\right]$$

$$\ln p(\mathbf{y}|\mathbf{x}) = -2\ln(2\pi) - \frac{1}{2}\ln|R| - \frac{1}{2}(\mathbf{y} - C\mathbf{x})^T R^{-1}(\mathbf{y} - C\mathbf{x})$$

$$\frac{d}{d\mathbf{x}} \ln p(\mathbf{y}|\mathbf{x}) = \left(C^T R^{-1}\mathbf{y} - C^T R^{-1} C\mathbf{x}\right)$$

$$\hat{\mathbf{x}} = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1}\mathbf{y} \longrightarrow$$

$$\mathrm{var}(\hat{\mathbf{x}}) = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1}\,\mathrm{var}(\mathbf{y})\,R^{-T} C\left(C^T R^{-1} C\right)^{-T}$$

$$\mathrm{var}(\mathbf{y}) = R$$

$$\mathrm{var}(\hat{\mathbf{x}}) = \left(C^T R^{-1} C\right)^{-1}$$

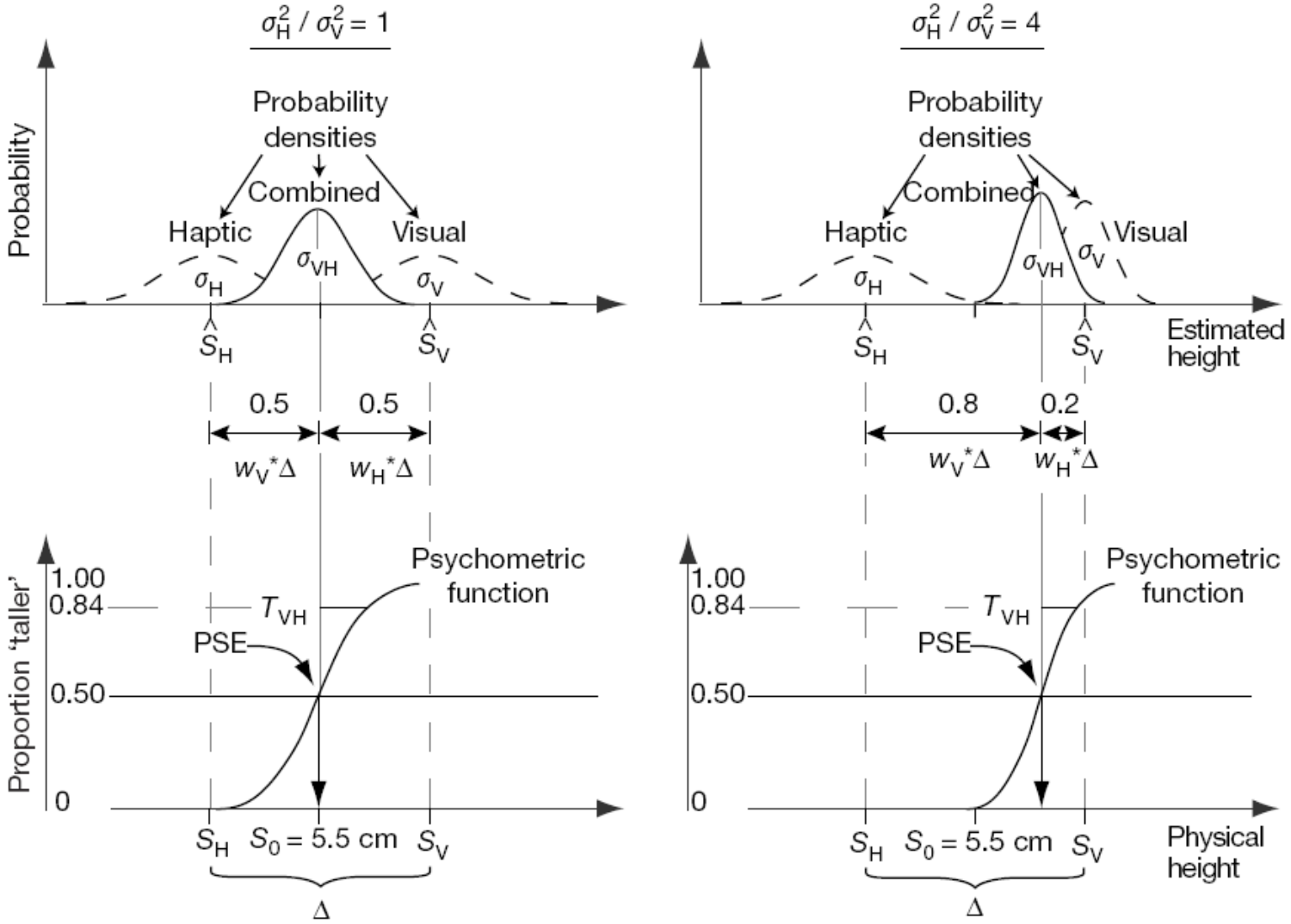
$$R^{-1} = \begin{bmatrix} R_a^{-1} & 0 \\ 0 & R_b^{-1} \end{bmatrix}$$

$$C^T R^{-1} = \begin{bmatrix} R_a^{-1} & R_b^{-1} \end{bmatrix}$$

$$\mathrm{var}(\hat{\mathbf{x}}) = \left(R_a^{-1} + R_b^{-1}\right)^{-1}$$

$$\hat{\mathbf{x}} = \left(R_a^{-1} + R_b^{-1}\right)^{-1}\left(R_a^{-1}\mathbf{y}_a + R_b^{-1}\mathbf{y}_b\right)$$

Optimal integration of sensory information by the brain



Parameter variance depends only on input selection and noise

A noisy process produces n data points and we form an ML estimate of w.

$$y^{*(i)} = \mathbf{w}^{*T}\mathbf{x}^{(i)}$$

$$y^{(i)} = y^{*(i)} + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

$$D^{(1)} = \left(\left\{\mathbf{x}^{(1)}, y^{(1,1)}\right\}, \left\{\mathbf{x}^{(2)}, y^{(1,2)}\right\}, \ldots, \left\{\mathbf{x}^{(n)}, y^{(1,n)}\right\}\right)$$

$$\mathbf{w}_{ML} = \left(X^T X\right)^{-1} X^T \mathbf{y}^{(1)}$$

We run the noisy process again with the same sequence of x's and re-estimate w:

$$D^{(2)} = \left(\left\{\mathbf{x}^{(1)}, y^{(2,1)}\right\}, \left\{\mathbf{x}^{(2)}, y^{(2,2)}\right\}, \ldots, \left\{\mathbf{x}^{(n)}, y^{(2,n)}\right\}\right)$$

$$\mathbf{w}_{ML} = \left(X^T X\right)^{-1} X^T \mathbf{y}^{(2)}$$

$$\vdots$$

The distribution of the resulting w will have a var-cov that depends only on the sequence of inputs, the bases that encode those inputs, and the noise sigma.

$$\mathbf{w}_{ML} \sim N\left(\mathbf{w}^*, \sigma^2\left(X^T X\right)^{-1}\right)$$

The Gaussian distribution and its var-cov matrix

A 1-D Gaussian distribution is defined as $$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$$

In *n* dimensions, it generalizes to $$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T C^{-1} (\mathbf{x}-\bar{\mathbf{x}}) \right]$$

When **x** is a vector, the variance is expressed in terms of a *covariance matrix* ***C***, where *ρij* corresponds to the degree of correlation between variables *xi* and *xj*

$$c_{ij} = E\left[(x_i - \bar{x}_i)(x_j - \bar{x}_j) \right] = E\left[x_i x_j \right] - \bar{x}_i \bar{x}_j$$

$$C = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n}\sigma_1\sigma_n & \rho_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\rho = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2} \sqrt{\sum_i (y_i - \mu_y)^2}} \equiv \frac{C_{xy}}{\sqrt{C_{xx}}\sqrt{C_{yy}}} \equiv \frac{C_{xy}}{\sigma_x \sigma_y}$$

$$\mathbf{x} \sim N(\boldsymbol{\mu}, C)$$

$$C = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

x1 and x2 are positively correlated

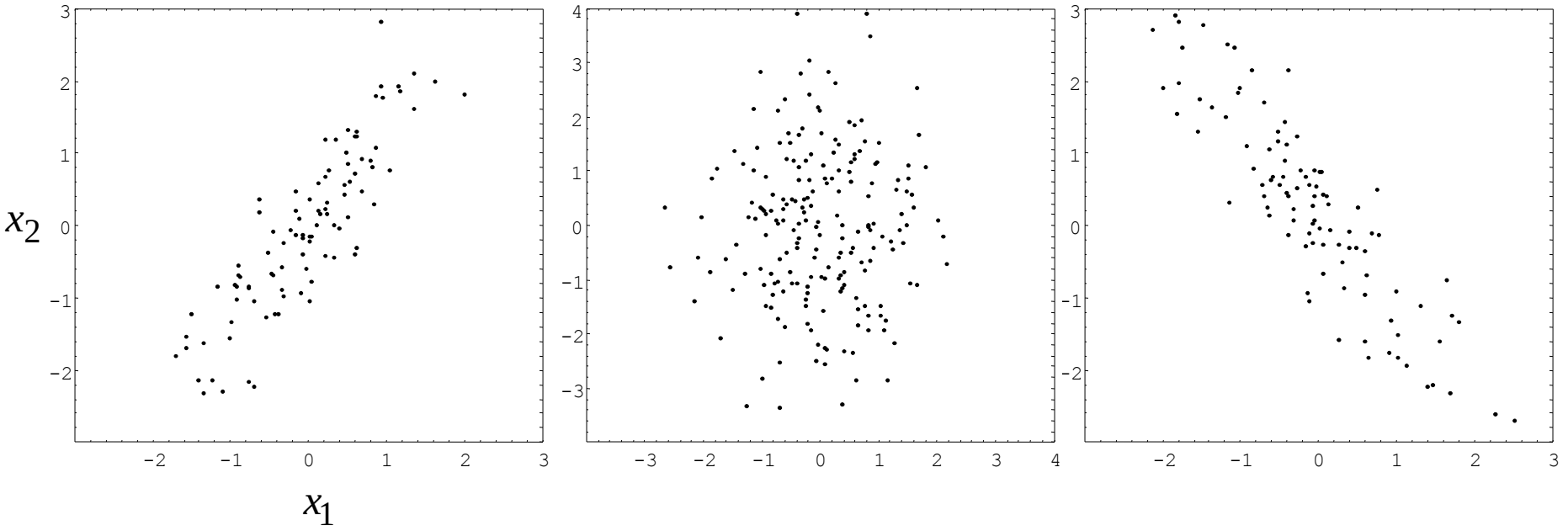
$$\mathbf{x} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9\sqrt{2} \\ 0.9\sqrt{2} & 2 \end{bmatrix}\right)$$

x1 and x2 are not correlated

$$\mathbf{x} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.1\sqrt{2} \\ 0.1\sqrt{2} & 2 \end{bmatrix}\right)$$

x1 and x2 are negatively correlated

$$\mathbf{x} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.9\sqrt{2} \\ -0.9\sqrt{2} & 2 \end{bmatrix}\right)$$



Parameter uncertainty: Example 1

- Input history:

x_1	x_2	y^*
1	0	0.5
1	0	0.5
1	0	0.5
1	0	0.5
0	1	0.5

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

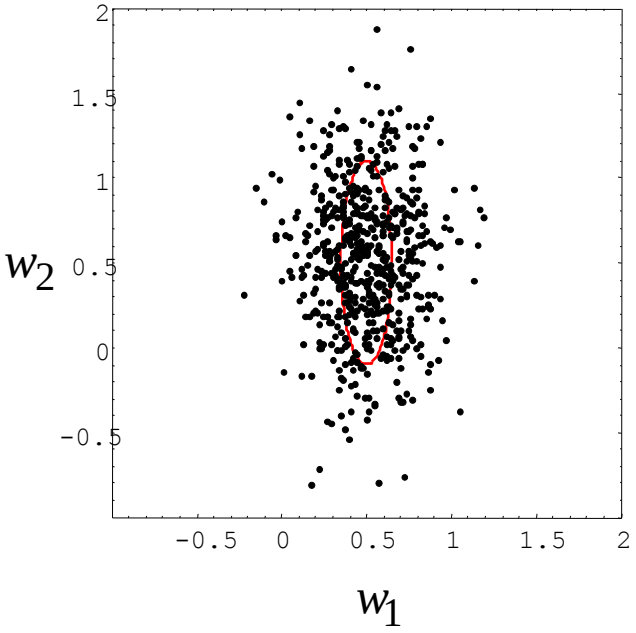
$$\hat{y} = w_1 x_1 + w_2 x_2 = \mathbf{x}^T \mathbf{w}$$

$$\mathbf{w}_{ML} \sim N\left(E[\mathbf{w}], \begin{bmatrix} \mathrm{var}[w_1] & \mathrm{cov}[w_1, w_2] \\ \mathrm{cov}[w_2, w_1] & \mathrm{var}[w_2] \end{bmatrix} \right)$$

$$\sim N\left(\mathbf{w}^*, \sigma^2 \left(X^T X \right)^{-1} \right)$$

$$\sim N\left(\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \sigma^2 \begin{bmatrix} 0.25 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

x1 was “on” most of the time. I’m pretty certain about w1. However, x2 was “on” only once, so I’m uncertain about w2.



Parameter uncertainty: Example 2

- Input history:

x_1	x_2	y^*
1	1	1
1	1	1
1	1	1
1	1	1
1	0	0.5

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\mathbf{w}_{ML} \sim N\left(E[\mathbf{w}], \begin{bmatrix} \mathrm{var}[w_1] & \mathrm{cov}[w_1, w_2] \\ \mathrm{cov}[w_2, w_1] & \mathrm{var}[w_2] \end{bmatrix} \right)$$

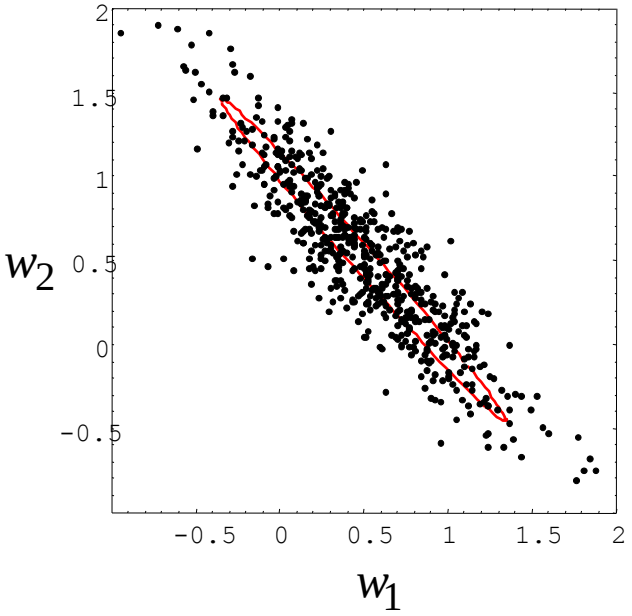
$$\sim N\left(\mathbf{w}^*, \sigma^2 \left(X^T X \right)^{-1} \right)$$

$$\sim N\left(\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & -1 \\ -1 & 1.25 \end{bmatrix} \right)$$

x1 and x2 were "on" mostly together. The weight var-cov matrix shows that what I learned is that: $w_1 + w_2 = 1$

I do not know individual values of w1 and w2 with much certainty.

x1 appeared slightly more often than x2, so I'm a little more certain about the value of w1.



Parameter uncertainty: Example 3

- Input history:

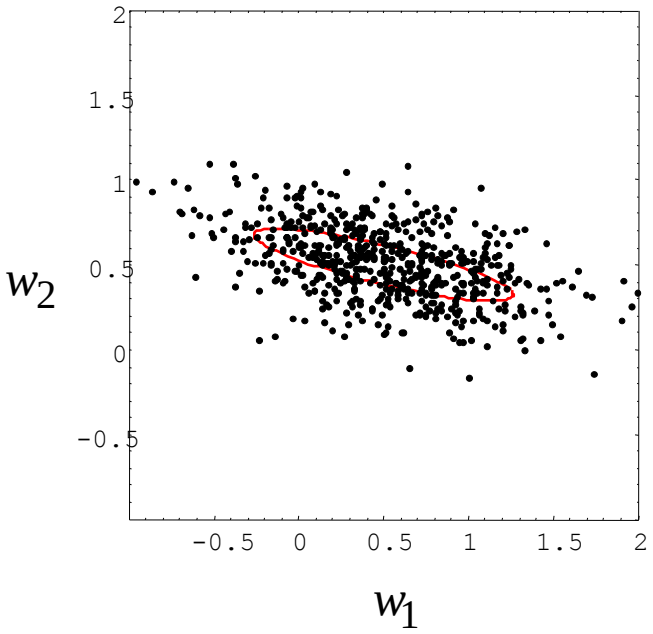
x_1	x_2	y^*
0	1	0.5
0	1	0.5
0	1	0.5
0	1	0.5
1	1	1

$$\mathbf{w}_{ML} \sim N\left(E[\mathbf{w}], \begin{bmatrix} \mathrm{var}[w_1] & \mathrm{cov}[w_1, w_2] \\ \mathrm{cov}[w_2, w_1] & \mathrm{var}[w_2] \end{bmatrix}\right)$$

$$\sim N\left(\mathbf{w}^*, \sigma^2 \left(X^T X\right)^{-1}\right)$$

$$\sim N\left(\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1.25 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}\right)$$

x2 was mostly “on”. I’m pretty certain about w2, but I am very uncertain about w1. Occasionally x1 and x2 were on together, so I have some reason to believe that: $w_1 + w_2 = 1$



Effect of uncertainty on learning rate

- When you observe an error in trial n, the amount that you should change w should depend on how certain you are about w. The more certain you are, the less you should be influenced by the error. The less certain you are, the more you should "pay attention" to the error.

mx1 mx1 error

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n)}\right)$$

Kalman gain

Rudolph E. Kalman (1960) A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering, 82 (Series D): 35-45.

Research Institute for Advanced Study
7212 Bellona Ave, Baltimore, MD

Example of the a variable learning gain: running estimate of average

$$x^{(i)} = 1$$

$$y^{*(i)} = w^{*}; \quad y^{(i)} = y^{*(i)} + \varepsilon \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

$$X^T = \overbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}}^{n}$$

$$w^{(n)} = \left(X^T X\right)^{-1} X^T \mathbf{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$$

w(n) is the online estimate of the mean of y

$$w^{(n-1)} = \frac{1}{n-1}\sum_{i=1}^{n-1} y^{(i)}$$

Past estimate

New measure

$$w^{(n)} = \frac{1}{n}\left(\sum_{i=1}^{n-1} y^{(i)} + y^{(n)}\right) = \frac{1}{n}\left((n-1)w^{(n-1)} + y^{(n)}\right) = \left(1 - \frac{1}{n}\right)w^{(n-1)} + \frac{1}{n}y^{(n)}$$

$$\boxed{w^{(n)} = w^{(n-1)} + \frac{1}{n}\left(y^{(n)} - w^{(n-1)}\right)}$$

As n increases, we trust our past estimate w(n-1) a lot more than the new observation y(n)

Kalman gain: learning rate decreases as the number of samples increase

Example of a variable learning gain: running estimate of variance

sigma_hat is the online estimate of the var of y

$$\hat{\sigma}^2_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - E[y]\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - w^{(n)}\right)^2$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n-1}\left(y^{(i)} - w^{(n)}\right)^2 + \left(y^{(n)} - w^{(n)}\right)^2\right]$$

$$= \frac{1}{n}\left[(n-1)\hat{\sigma}^2_{(n-1)} + \left(y^{(n)} - w^{(n)}\right)^2\right]$$

$$= \hat{\sigma}^2_{(n-1)} - \frac{1}{n}\hat{\sigma}^2_{(n-1)} + \frac{1}{n}\left(y^{(n)} - w^{(n)}\right)^2$$

$$\boxed{\hat{\sigma}^2_{(n)} = \hat{\sigma}^2_{(n-1)} + \frac{1}{n}\left[\left(y^{(n)} - w^{(n)}\right)^2 - \hat{\sigma}^2_{(n-1)}\right]}$$

Some observations about variance of model parameters

$$y^{(n)} = \mathbf{x}^{(n)T}\mathbf{w}^{*} + \varepsilon^{(n)} \qquad \varepsilon \sim N\left(0, \sigma^{2}\right)$$

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n)}\right)$$

$$P^{(n)} \equiv \mathrm{var}\left[\mathbf{w}^{(n)}\right]$$

$$= E\left[\left(\mathbf{w}^{(n)} - E\left[\mathbf{w}^{(n)}\right]\right)\left(\mathbf{w}^{(n)} - E\left[\mathbf{w}^{(n)}\right]\right)^{T}\right]$$

$$= E\left[\left(\mathbf{w}^{(n)} - \mathbf{w}^{*}\right)\left(\mathbf{w}^{(n)} - \mathbf{w}^{*}\right)^{T}\right]$$

$$= E\left[\tilde{\mathbf{w}}^{(n)}\tilde{\mathbf{w}}^{(n)T}\right]$$

$$trace\left[P^{(n)}\right] = E\left[\tilde{\mathbf{w}}^{(n)T}\tilde{\mathbf{w}}^{(n)}\right]$$

We note that P is simply the var-cov matrix of our model weights. It represents the uncertainty in our estimates of the model parameters.

We want to update the weights in such a way as to minimize the trace of this variance. The trace is the sum of the squared errors between our estimates of w and the true estimates.

Trace of parameter var-cov matrix is the sum of squared parameter errors

$$P \equiv E\left[\tilde{\mathbf{w}}\tilde{\mathbf{w}}^T\right]$$

$$\tilde{\mathbf{w}} = \begin{bmatrix}\tilde{w}_1 \\ \tilde{w}_2\end{bmatrix} \sim N\left(\mathbf{0}, P\right) = N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\mathrm{var}\left(\tilde{w}_1\right) & \mathrm{cov}\left(\tilde{w}_1, \tilde{w}_2\right) \\ \mathrm{cov}\left(\tilde{w}_2, \tilde{w}_1\right) & \mathrm{var}\left(\tilde{w}_2\right)\end{bmatrix}\right)$$

$$\mathrm{var}\left(\tilde{w}_1\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{w}_1^{(i)} - E\left[\tilde{w}_1\right]\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{w}_1^{(i)2}$$

$$trace\left(P\right) = \mathrm{var}\left(\tilde{w}_1\right) + \mathrm{var}\left(\tilde{w}_2\right) = \frac{1}{n}\sum_{i=1}^{n}\tilde{w}_1^{(i)2} + \tilde{w}_2^{(i)2}$$

Our objective is to find learning rate **k** (Kalman gain) such that we minimize the sum of the squared error in our parameter estimates. This sum is the trace of the P matrix. Therefore, given observation y(n), we want to find k such that we minimize the variance of our estimate w.

$$\mathbf{w}^{\left(n|n\right)} = \mathbf{w}^{\left(n|n-1\right)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{\left(n|n-1\right)}\right)$$

Objective: adjust learning gain in order to minimize model uncertainty

Hypothesis about data observation in trial n $y^{(n)} = \mathbf{x}^{(n)T}\mathbf{w}^{*} + \varepsilon^{(n)} \qquad \varepsilon \sim N\left(0, \sigma^{2}\right)$

my estimate of w* before I see y in trial n, given that I have seen y up to n-1 $\mathbf{w}^{\left(n|n-1\right)}$

error in trial n $y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{\left(n|n-1\right)}$

my estimate after I see y in trial n $\mathbf{w}^{\left(n|n\right)} = \mathbf{w}^{\left(n|n-1\right)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{\left(n|n-1\right)}\right)$

$P^{\left(n|n-1\right)} \equiv \mathrm{var}\left[\mathbf{w}^{\left(n|n-1\right)}\right]$ a prior variance of parameters

$= E\left[\left(\mathbf{w}^{\left(n|n-1\right)} - E\left[\mathbf{w}^{\left(n|n-1\right)}\right]\right)\left(\mathbf{w}^{\left(n|n-1\right)} - E\left[\mathbf{w}^{\left(n|n-1\right)}\right]\right)^{T}\right]$

$P^{\left(n|n\right)} \equiv \mathrm{var}\left[\mathbf{w}^{\left(n|n\right)}\right]$ a posterior variance of parameters

$= E\left[\left(\mathbf{w}^{\left(n|n\right)} - E\left[\mathbf{w}^{\left(n|n\right)}\right]\right)\left(\mathbf{w}^{\left(n|n\right)} - E\left[\mathbf{w}^{\left(n|n\right)}\right]\right)^{T}\right]$

Evolution of parameter uncertainty

$$\mathbf{w}^{(n|n)} = \mathbf{w}^{(n|n-1)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n|n-1)} \right)$$

$$\mathbf{w}^{(n|n)} = \mathbf{w}^{(n|n-1)} + \mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}\mathbf{w}^{*} + \varepsilon^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n|n-1)} \right)$$

$$\mathbf{w}^{*} - \mathbf{w}^{(n|n)} = \mathbf{w}^{*} - \mathbf{w}^{(n|n-1)} + \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\mathbf{w}^{*} + \mathbf{k}^{(n)}\varepsilon^{(n)} - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\mathbf{w}^{(n|n-1)}$$

$$= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n-1)} \right) + \mathbf{k}^{(n)}\varepsilon^{(n)}$$

$$P^{(n|n)} = E\left[\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n)} \right)\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n)} \right)^{T} \right]$$

$$= E\left[\left(\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n-1)} \right) + \mathbf{k}^{(n)}\varepsilon^{(n)} \right)\left(\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n-1)} \right) + \mathbf{k}^{(n)}\varepsilon^{(n)} \right)^{T} \right]$$

$$= E\left[\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n-1)} \right)\left(\mathbf{w}^{*} - \mathbf{w}^{(n|n-1)} \right)^{T}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)^{T} + \mathbf{k}^{(n)}\varepsilon^{(n)}\varepsilon^{(n)T}\mathbf{k}^{(n)T} \right]$$

$$= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)P^{(n|n-1)}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)^{T} + E\left[\mathbf{k}^{(n)}\varepsilon^{(n)}\varepsilon^{(n)T}\mathbf{k}^{(n)T} \right]$$

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)P^{(n|n-1)}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)^{T} + \mathbf{k}^{(n)}\sigma^{2}\mathbf{k}^{(n)T}$$

Find K to minimize trace of uncertainty

$$\mathbf{w}^{(n|n)} = \mathbf{w}^{(n|n-1)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n|n-1)} \right)$$

$$\mathbf{w}^{(n|n)} = \mathbf{w}^{(n|n-1)} + \mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}\mathbf{w}^{*} + \varepsilon^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n|n-1)} \right)$$

$$\mathbf{w}^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)\mathbf{w}^{(n|n-1)} + \mathbf{k}^{(n)}\varepsilon^{(n)} + \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\mathbf{w}^{*}$$

$$P^{(n|n-1)} = \mathrm{var}\left[\mathbf{w}^{(n|n-1)} \right]$$

$$P^{(n|n)} = \mathrm{var}\left[\mathbf{w}^{(n|n)} \right]$$

$$= \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right) P^{(n|n-1)} \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)^{T} + \mathbf{k}^{(n)}\,\mathrm{var}\left[\varepsilon^{(n)} \right]\mathbf{k}^{(n)T}$$

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right) P^{(n|n-1)} \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T} \right)^{T} + \mathbf{k}^{(n)}\sigma^{2}\mathbf{k}^{(n)T}$$

Find K to minimize trace of uncertainty

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}\left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)^T + \mathbf{k}^{(n)}\sigma^2\mathbf{k}^{(n)T}$$

$$= P^{(n|n-1)} - P^{(n|n-1)}\mathbf{x}^{(n)}\mathbf{k}^{(n)T} - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} + \mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)}\mathbf{k}^{(n)T} + \mathbf{k}^{(n)}\sigma^2\mathbf{k}^{(n)T}$$

$$tr\left[P^{(n|n)}\right] = tr\left[P^{(n|n-1)}\right] - tr\left[P^{(n|n-1)}\mathbf{x}^{(n)}\mathbf{k}^{(n)T}\right] - tr\left[\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\right]$$

$$+tr\left[\mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\right]$$

$$= tr\left[P^{(n|n-1)}\right] - 2tr\left[\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\right] + tr\left[\mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\right]$$

$$tr[A] = tr\left[A^T\right]$$
$$P = P^T$$

scalar

$$tr\left[\mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\right] = tr\left[\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)}\mathbf{k}^{(n)T}\right]$$

$$= \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)tr\left[\mathbf{k}^{(n)}\mathbf{k}^{(n)T}\right]$$

$$= \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\mathbf{k}^{(n)}$$

$$tr[aB] = atr[B]$$

The Kalman gain

$$tr\left[P^{(n|n)}\right] = tr\left[P^{(n|n-1)}\right] - 2tr\left[\mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)}\right] + \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}\mathbf{k}^{(n)}$$

$$\frac{d}{d\mathbf{k}^{(n)}}tr\left[P^{(n|n)}\right] = -2P^{(n|n-1)}\mathbf{x}^{(n)} + \left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\left(2\mathbf{k}^{(n)}\right) = 0$$

$$\frac{d}{dA}tr\left[AB\right] = B^T$$

$$\mathbf{k}^{(n)} = \frac{P^{(n|n-1)}\mathbf{x}^{(n)}}{\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)}$$

If I have a lot of uncertainty about my model, P is large compared to sigma. I will learn a lot from the current error.

If I am pretty certain about my model, P is small compared to sigma. I will tend to ignore the current error.

Update of model uncertainty

$$P^{(n|n)} = P^{(n|n-1)} - P^{(n|n-1)}\mathbf{x}^{(n)}\mathbf{k}^{(n)T} - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}P^{(n|n-1)} + \mathbf{k}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)\mathbf{k}^{(n)T}$$

$$\mathbf{k}^{(n)} = P^{(n|n-1)}\mathbf{x}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)^{-1}$$

$$P^{(n|n)} = P^{(n|n-1)} - P^{(n|n-1)}\mathbf{x}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)^{-T}\mathbf{x}^{(n)T}P^{(n|n-1)}$$

$$-P^{(n|n-1)}\mathbf{x}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)^{-1}\mathbf{x}^{(n)T}P^{(n|n-1)}$$

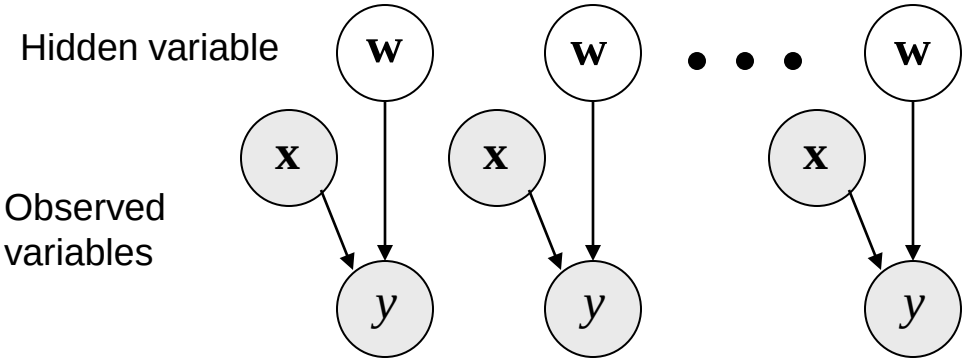
$$+P^{(n|n-1)}\mathbf{x}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)^{-1}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)$$

$$\times\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)^{-T}\mathbf{x}^{(n)T}P^{(n|n-1)}$$

$$= P^{(n|n-1)} - P^{(n|n-1)}\mathbf{x}^{(n)}\left(\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2\right)^{-1}\mathbf{x}^{(n)T}P^{(n|n-1)}$$

$$\boxed{P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}}$$

Model uncertainty decreases with every data point that you observe.



In this model, we hypothesize that the hidden variables, i.e., the "true" weights, do not change from trial to trial.

$$\mathbf{w}_{(n+1)} = \mathbf{w}_{(n)}$$

$$y^{(n)} = \mathbf{x}^{(n)T}\mathbf{w} + \varepsilon^{(n)} \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

A priori estimate of mean and variance of the hidden variable before I observe the first data point

$$\hat{\mathbf{w}}\left(1|0\right), P\left(1|0\right)$$

Update of the estimate of the hidden variable after I observed the data point

$$\hat{\mathbf{w}}\left(n|n\right) = \hat{\mathbf{w}}\left(n|n-1\right) + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\hat{\mathbf{w}}\left(n|n-1\right)\right)$$

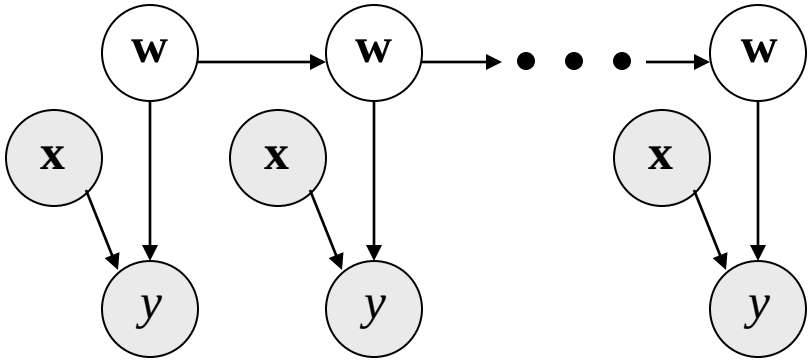
$$\mathbf{k}^{(n)} = \frac{P\left(n|n-1\right)\mathbf{x}^{(n)}}{\mathbf{x}^{(n)T}P\left(n|n-1\right)\mathbf{x}^{(n)} + \sigma^2}$$

$$P\left(n|n\right) = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P\left(n|n-1\right)$$

Forward projection of the estimate to the next trial

$$\hat{\mathbf{w}}\left(n+1|n\right) = \hat{\mathbf{w}}\left(n|n\right)$$

$$P\left(n+1|n\right) = P\left(n|n\right)$$



In this model, we hypothesize that the hidden variables change from trial to trial.

$$\mathbf{w}^{(n+1)} = A\mathbf{w}^{(n)} + \boldsymbol{\varepsilon}_w^{(n)} \qquad \boldsymbol{\varepsilon}_w \sim N(0, Q)$$

$$y^{(n)} = \mathbf{x}^{(n)T}\mathbf{w}^{(n)} + \varepsilon_y^{(n)} \qquad \varepsilon_y \sim N\left(0, \sigma^2\right)$$

A priori estimate of mean and variance of the hidden variable before I observe the first data point

$$\hat{\mathbf{w}}^{(1|0)}, P^{(1|0)}$$

Update of the estimate of the hidden variable after I observed the data point

$$\mathbf{k}^{(n)} = \frac{P^{(n|n-1)}\mathbf{x}^{(n)}}{\mathbf{x}^{(n)T}P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2}$$

$$\hat{\mathbf{w}}^{(n|n)} = \hat{\mathbf{w}}^{(n|n-1)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\hat{\mathbf{w}}^{(n|n-1)}\right)$$

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right)P^{(n|n-1)}$$

Forward projection of the estimate to the next trial

$$\hat{\mathbf{w}}^{(n+1|n)} = A\hat{\mathbf{w}}^{(n|n)}$$

$$P^{(n+1|n)} = AP^{(n|n)}A^T + Q$$

Uncertainty about my model parameters

$$\mathbf{k}^{(n)} = \frac{P^{(n|n-1)}\mathbf{x}^{(n)}}{\underbrace{\mathbf{x}^{(n)T} P^{(n|n-1)}\mathbf{x}^{(n)} + \sigma^2}}$$

Uncertainty about my measurement

- Learning rate is proportional to the ratio between two uncertainties: my model vs. my measurement.
- After we observe an input x, the uncertainty associated with the weight of that input decreases.

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}\mathbf{x}^{(n)T}\right) P^{(n|n-1)}$$

- Because of state update noise Q, uncertainty increases as we form the prior for the next trial.

$$P^{(n+1|n)} = AP^{(n|n)}A^T + Q$$

Comparison of Kalman gain to LMS

$$\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)}$$

$$y^{(n)} = \mathbf{x}^{(n)T}\mathbf{w} + \varepsilon^{(n)} \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

See derivation of this in homework

$$\mathbf{k}^{(n)} = \frac{P^{(n|n)}\mathbf{x}^{(n)}}{\sigma^2}$$

$$\hat{\mathbf{w}}^{(n)} = \hat{\mathbf{w}}^{(n-1)} + \mathbf{k}^{(n)}\left(y^{(n)} - \mathbf{x}^{(n)T}\hat{\mathbf{w}}^{(n-1)}\right)$$

$$= \hat{\mathbf{w}}^{(n-1)} + \frac{P^{(n|n)}}{\sigma^2}\left(y^{(n)} - \mathbf{x}^{(n)T}\hat{\mathbf{w}}^{(n-1)}\right)\mathbf{x}^{(n)}$$

In the Kalman gain approach, the P matrix depends on the history of all previous and current inputs. In LMS, the learning rate is simply a constant that does not depend on past history.

$$\mathbf{w}^{(n)} = \mathbf{w}^{(n-1)} + \eta\left(y^{(n)} - \mathbf{x}^{(n)T}\mathbf{w}^{(n-1)}\right)\mathbf{x}^{(n)}$$

With the Kalman gain, our estimate converges on a single pass over the data set. In LMS, we don't estimate the var-cov matrix P on each trial, but we will need multiple passes before our estimate converges.

How to set the initial var-cov matrix

$$\mathbf{x}^{(n)} = A\mathbf{x}^{(n)} + \boldsymbol{\varepsilon}_w^{(n)} \qquad \boldsymbol{\varepsilon}_w \sim N(0, Q)$$

$$\mathbf{y}^{(n)} = C\mathbf{x}^{(n)} + \boldsymbol{\varepsilon}_y^{(n)} \qquad \boldsymbol{\varepsilon}_y \sim N(0, R)$$

$$P^{(1|0)} = ?$$

$$P^{(n|n)} = P^{(n|n-1)} - \mathbf{k}^{(n)} C P^{(n|n-1)}$$

$$= P^{(n|n-1)} - P^{(n|n-1)} C^T \left(C P^{(n|n-1)} C^T + R \right)^{-1} C P^{(n|n-1)}$$

Matrix inversion lemma $\left(Z - XY^{-1}X^T \right)^{-1} = Z^{-1} + Z^{-1}X\left(Y - X^T Z^{-1} X \right)^{-1} X^T Z^{-1}$

Set: $-Z^{-1} = P^{(n|n-1)} \quad X = C^T \quad Y = R$

$$P^{(n|n)} = -Z^{-1} - Z^{-1}X\left(Y - X^T Z^{-1} X \right)^{-1} X^T Z^{-1}$$

$$-P^{(n|n)} = \left(Z - XY^{-1}X^T \right)^{-1} = \left(-\left(P^{(n|n-1)} \right)^{-1} - C^T R^{-1} C \right)^{-1}$$

$$\left(P^{(n|n)} \right)^{-1} = \left(P^{(n|n-1)} \right)^{-1} + C^T R^{-1} C$$

$$\left(P^{(n|n)} \right)^{-1} = \left(P^{(n|n-1)} \right)^{-1} + C^T R^{-1} C$$

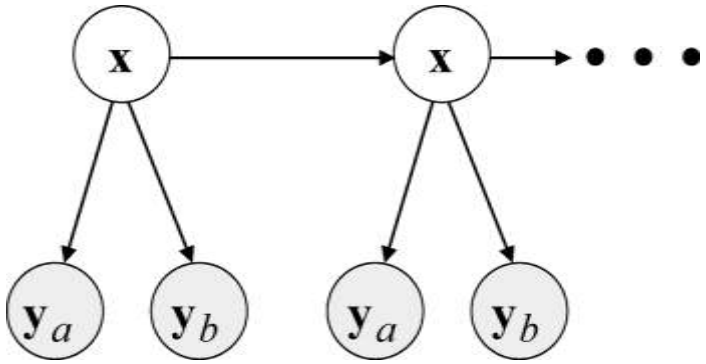
Now if we have absolutely no prior information on w, then before we see the first data point P(1|0) is infinity, and therefore its inverse in zero. After we see the first data point, we will be using the above equation to update our estimate. The updated estimate will become:

$$\left(P^{(1|1)} \right)^{-1} = C^T R^{-1} C$$

$$P^{(1|1)} = \left(C^T R^{-1} C \right)^{-1}$$

A reasonable and conservative estimate of the initial value of P would be to set it to the above value. That is, set:

$$P^{(1|0)} = \left(C^T R^{-1} C \right)^{-1}$$



$$\mathbf{x}^{(n+1)} = A\mathbf{x}^{(n)} + \boldsymbol{\varepsilon}_x^{(n)} \qquad \boldsymbol{\varepsilon}_x \sim N(\mathbf{0}, Q)$$

$$\mathbf{y}^{(n)} = C\mathbf{x}^{(n)} + \boldsymbol{\varepsilon}_y^{(n)} \qquad \boldsymbol{\varepsilon}_y \sim N(\mathbf{0}, R)$$

$$\mathbf{k}^{(n)} = P^{(n|n-1)}C^T\left(CP^{(n|n-1)}C^T + R\right)^{-1} \longrightarrow \mathbf{k}^{(n)}\left(CP^{(n|n-1)}C^T + R\right) = P^{(n|n-1)}C^T$$

$$P^{(n|n)} = \left(I - \mathbf{k}^{(n)}C\right)P^{(n|n-1)} \qquad \mathbf{k}^{(n)}\left(CP^{(n|n-1)}C^T + R\right)R^{-1} = P^{(n|n-1)}C^TR^{-1}$$

$$P^{(1|0)} = \infty \qquad \mathbf{k}^{(n)}CP^{(n|n-1)}C^TR^{-1} + \mathbf{k}^{(n)} = P^{(n|n-1)}C^TR^{-1}$$

$$\mathbf{x}^{(1|0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{k}^{(n)} = \left(P^{(n|n-1)} - \mathbf{k}^{(n)}CP^{(n|n-1)}\right)C^TR^{-1}$$

$$\mathbf{k}^{(n)} = P^{(n|n)}C^TR^{-1}$$

$$\left(P^{(n|n)}\right)^{-1} = \left(P^{(n|n-1)}\right)^{-1} + C^T R^{-1} C$$

$$P^{(1|0)} = \infty,\ \mathbf{x}^{(1|0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$P^{(1|1)} = \left(C^T R^{-1} C\right)^{-1}$$

$$\mathbf{k}^{(n)} = P^{(n|n)} C^T R^{-1}$$

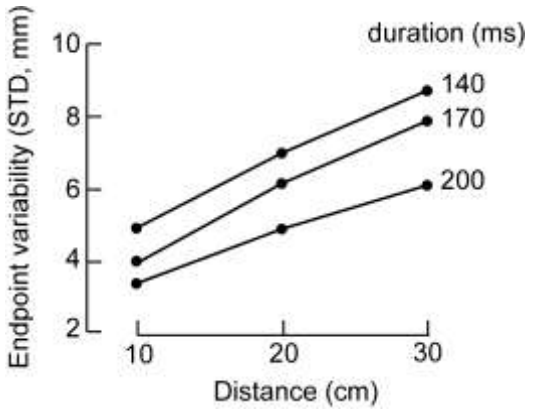
$$\mathbf{k}^{(1)} = P^{(1|1)} C^T R^{-1} = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1}$$

$$\mathbf{x}^{(1|1)} = \mathbf{x}^{(1|0)} + \mathbf{k}^{(1)} \left(\mathbf{y}^{(1)} - C\mathbf{x}^{(1|0)}\right)$$

$$\mathbf{x}^{(1|1)} = \left(C^T R^{-1} C\right)^{-1} C^T R^{-1} \mathbf{y}^{(1)}$$

This expression is our maximum likelihood estimate that we got earlier. Furthermore, the variance of our estimate is the variance of our maximum likelihood estimate. Therefore, if we are naïve in the sense that we have no prior knowledge about the state that we wish to estimate, then a weighted combination of the two sources of information is the optimal solution. On the other hand, if we also have a prior, e.g., we have hiked this path before and have some idea of where we might be, then the Kalman framework gives us the tools to weigh in this additional piece of information.

State estimation with signal dependent noise



$$f = u\left(1 + c\phi\right) \qquad \phi \sim N\left(0,1\right)$$

$$\mathrm{var}\left[\, f \,\right] = c^2 u^2$$

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} + B\left(\mathbf{u}^{(k)} + \boldsymbol{\varepsilon}_u^{(k)}\right) + \boldsymbol{\varepsilon}_x^{(k)} \qquad \boldsymbol{\varepsilon}_x \sim N\left(\mathbf{0}, Q_x\right)$$

$$\mathbf{y}^{(k)} = H\left(\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_s^{(k)}\right) + \boldsymbol{\varepsilon}_y^{(k)} \qquad \boldsymbol{\varepsilon}_y \sim N\left(\mathbf{0}, Q_y\right)$$

$$\boldsymbol{\varepsilon}_u^{(k)} \equiv \begin{bmatrix} c_1 u_1^{(k)} \phi_1^{(k)} \\ c_2 u_2^{(k)} \phi_2^{(k)} \\ \vdots \\ c_n u_n^{(k)} \phi_n^{(k)} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_s^{(k)} \equiv \begin{bmatrix} d_1 x_1^{(k)} \mu_1^{(k)} \\ d_2 x_2^{(k)} \mu_2^{(k)} \\ \vdots \\ d_m x_m^{(k)} \mu_m^{(k)} \end{bmatrix}$$

$$\phi \sim N\left(0,1\right) \qquad \mu \sim N\left(0,1\right)$$

$$C_1 \equiv \begin{bmatrix} c_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \ddots \end{bmatrix} \qquad C_2 \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & \ddots \end{bmatrix}$$

$$D_1 \equiv \begin{bmatrix} d_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \ddots \end{bmatrix} \qquad D_2 \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & \ddots \end{bmatrix}$$

$$\boldsymbol{\varepsilon}_u^{(k)} = \sum_{i=1}^{n} C_i \mathbf{u}^{(k)} \phi_i^{(k)}$$

$$\boldsymbol{\varepsilon}_s^{(k)} = \sum_{i=1}^{m} D_i \mathbf{x}^{(k)} \mu_i^{(k)}$$

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)} + B\mathbf{u}^{(k)} + \boldsymbol{\varepsilon}_x^{(k)} + B\sum_i C_i \mathbf{u}^{(k)} \phi_i^{(k)}$$

$$\mathbf{y}^{(k)} = H\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_y^{(k)} + H\sum_i D_i \mathbf{x}^{(k)} \mu_i^{(k)}$$

$$\hat{\mathbf{x}}^{(k|k)} = \hat{\mathbf{x}}^{(k|k-1)} + K^{(k)}\left(\mathbf{y}^{(k)} - H\hat{\mathbf{x}}^{(k|k-1)}\right)$$

$$\hat{\mathbf{x}}^{(k|k)} = \hat{\mathbf{x}}^{(k|k-1)} + K^{(k)}\left(H\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_y^{(k)} + H\sum_i D_i \mathbf{x}^{(k)} \mu_i^{(k)} - H\hat{\mathbf{x}}^{(k|k-1)}\right)$$

$$= \left(I - K^{(k)}H\right)\hat{\mathbf{x}}^{(k|k-1)} + K^{(k)}\left(H\mathbf{x}^{(k)} + \boldsymbol{\varepsilon}_y^{(k)} + H\sum_i D_i \mathbf{x}^{(k)} \mu_i^{(k)}\right)$$

$$P^{(k|k)} = \left(I - K^{(k)}H\right)P^{(k|k-1)}\left(I - K^{(k)}H\right)^T + K^{(k)}Q_y K^{(k)T}$$

$$+\sum_i KHD_i \mathbf{x}^{(k)}\mathbf{x}^{(k)T} D_i^T H^T K^T$$

$$= P^{k|k-1} - 2P^{k|k-1}H^T K^{(k)T}$$

$$+K^{(k)}\left(HP^{k|k-1}H^T + Q_y + \sum_i HD_i \mathbf{x}^{(k)}\mathbf{x}^{(k)T} D_i^T H^T\right)K^{(k)T}$$

$$\frac{d}{dK^{(k)}} tr\left[P^{(k|k)}\right] = -2P^{(k|k-1)}H^T$$

$$+2K^{(k)}\left(HP^{(k|k-1)}H^T + Q_y + \sum_i HD_i \mathbf{x}^{(k)}\mathbf{x}^{(k)T} D_i^T H^T\right)$$

$$K^{(k)} = P^{(k|k-1)}H^T\left(HP^{(k|k-1)}H^T + Q_y + \sum_i HD_i \mathbf{x}^{(k)}\mathbf{x}^{(k)T} D_i^T H^T\right)^{-1}$$

$$P^{k|k} = P^{k|k-1}\left(I - H^T K^{(k)T}\right)$$

$$P^{(k+1|k)} = AP^{(k|k)}A^T + Q_x + \sum_i BC_i \mathbf{u}^{(k)}\mathbf{u}^{(k)T} C_i^T B^T$$

the state uncertainty increases with the size of the motor commands, and the Kalman gain decreases with the size of the state