Lecture 3: Key Messages

- Recursive LS and LOOCV

$$\hat{\underline{w}}_N = \Sigma_N X_N^T \underline{z}_N = \Sigma_N \left( \sum_{i=1}^{N} \underline{x}_i z_i \right); \Sigma_N = \left( X^T X + \mu I \right)^{-1} = \left( \sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T + \mu I \right)^{-1}$$

$$\hat{\underline{w}}_{N+1} = \Sigma_{N+1} X_{N+1}^T \underline{z}_{N+1} = \left( \sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T + \mu I + \underline{x}_{N+1} \underline{x}_{N+1}^T \right)^{-1} \left( \sum_{i=1}^{N} \underline{x}_i z_i + \underline{x}_{N+1} z_{N+1} \right)$$

$$= \underbrace{\left( \Sigma_N - \frac{\Sigma_N \underline{x}_{N+1} \underline{x}_{N+1}^T \Sigma_N}{1 + \underline{x}_{N+1}^T \Sigma_N \underline{x}_{N+1}} \right)}_{\Sigma_{N+1}} \left( \sum_{i=1}^{N} \underline{x}_i z_i + \underline{x}_{N+1} z_{N+1} \right)$$

$$= \hat{\underline{w}}_N + \frac{\Sigma_N \underline{x}_{N+1}}{1 + \underline{x}_{N+1}^T \Sigma_N \underline{x}_{N+1}} (z_{N+1} - \underline{x}_{N+1}^T \hat{\underline{w}}_N) = \hat{\underline{w}}_N + \underbrace{\frac{\Sigma_N \underline{x}_{N+1}}{1 + \underline{x}_{N+1}^T \Sigma_N \underline{x}_{N+1}}}_{\text{Kalman gain}} (z_{N+1} - \underbrace{\hat{z}_{N+1|N}}_{\hat{z}_{-(N+1)}})$$

$$Note: \underline{x}_{N+1}^T \Sigma_{N+1} \underline{x}_{N+1} = \frac{\underline{x}_{N+1}^T \Sigma_N \underline{x}_{N+1}}{1 + \underline{x}_{N+1}^T \Sigma_N \underline{x}_{N+1}} \Rightarrow \underline{x}_{N+1}^T \Sigma_N \underline{x}_{N+1} = \frac{\underline{x}_{N+1}^T \Sigma_{N+1} \underline{x}_{N+1}}{1 - \underline{x}_{N+1}^T \Sigma_{N+1} \underline{x}_{N+1}}$$

$$So, \hat{z}_{N+1|N+1} == \hat{z}_{N+1|N} + \underline{x}_{N+1}^T \Sigma_{N+1} \underline{x}_{N+1} (z_{N+1} - \hat{z}_{N+1|N})$$

$$z_{N+1} - \hat{z}_{-(N+1)} = (1 - \underline{x}_{N+1}^T \Sigma_{N+1} \underline{x}_{N+1})(z_{N+1} - \hat{z}_{N+1|N})$$

$$(z_{N+1} - \hat{z}_{-(N+1)}) = \frac{z_{N+1} - \hat{z}_{N+1|N+1}}{1 - \underline{x}_{N+1}^T \Sigma_{N+1} \underline{x}_{N+1}}$$ ..... vaid for any measurement in LS... just replace $N+1$ by $i$

Not true for dynamic systems.

$$J_{LOOCV} = \frac{1}{N+1} \sum_{i=1}^{N+1} (z_i - \hat{z}_{-i})^2 = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{z_i - \hat{z}_{i|N+1}}{1 - \underline{x}_i^T \Sigma_{N+1} \underline{x}_i} \right)^2$$

- Bernoulli and Gaussian

$$p(x) = p^x (1-p)^{1-x} = \exp\left[ x \ln p + (1-x) \ln(1-p) \right] = \exp\left[ x \ln \frac{p}{1-p} + \ln(1-p) \right]$$

$$\text{Let} \quad T(x) = x; \; w = \ln \frac{p}{1-p} \Rightarrow p = \frac{1}{1+e^{-w}} = \sigma(w)$$

$$\Rightarrow 1 - p = \frac{1}{1+e^{w}} = \sigma(-w) = 1 - \sigma(w)$$

$$\text{so, } p(x) = \exp[x \ln \frac{p}{1-p} - \ln(1+e^{w})]$$

$$A(w) = \ln(1+e^{w}); \frac{dA(w)}{dw} = \frac{e^{w}}{1+e^{w}} = \frac{1}{1+e^{-w}} = p = E(x) = E[T(x)]$$

$$\frac{d^2 A(w)}{dw^2} = p(1-p) = Var(x) = Var[T(x)]$$

- XOR:
  - 

| B | C | P(A=1\|B,C) | P(A=0\|B,C) |
|---|---|---|---|
| 0 | 0 | 0.10 | 0.90 |
| 0 | 1 | 0.99 | 0.01 |
| 1 | 0 | 0.80 | 0.20 |
| 1 | 1 | 0.25 | 0.75 |

$$p(x_A, x_B, x_C) = (0.1)^{x_A(1-x_B)(1-x_C)}(0.9)^{(1-x_A)(1-x_B)(1-x_C)}(0.99)^{x_A(1-x_B)x_C}(0.01)^{(1-x_A)(1-x_B)x_C}$$
$$(0.8)^{x_A x_B(1-x_C)}(0.2)^{(1-x_A)x_B(1-x_C)}(.25)^{x_A x_B x_C}(.75)^{(1-x_A)x_B x_C}$$
$$\ln p(x_A, x_B, x_C) = (1-x_B)(1-x_C)\ln(0.9) + x_A(1-x_B)(1-x_C)\ln(\frac{1}{9})$$
$$+(1-x_B)x_C\ln(0.01) + x_A(1-x_B)(1-x_C)\ln(99) + x_B(1-x_C)\ln(.2) + x_A x_B(1-x_C)\ln(4)$$
$$+x_B x_C\ln(0.75) + x_A x_B x_C\ln(\frac{1}{3})$$
$$\ln(0, x_B, x_C) = (1-x_B)(1-x_C)\ln(0.9) + (1-x_B)x_C\ln(0.01) + x_B(1-x_C)\ln(.2)$$
$$MAP: p(x_B, x_C \mid x_A = 0) = \max_{x_B, x_C}[(1-x_B)(1-x_C)\ln(0.9) + (1-x_B)x_C\ln(0.01) + x_B(1-x_C)\ln(.2)]$$
$$Evidently, x_B = 1; x_C = 1$$

- Beta distribution: HW 1 problem
- Gaussian: Use Least squares problem
  Z=Xw+v; v ~ N(0, diag(R)); w ~N(0,$\Sigma_0$)…. $\Sigma_0$ = 1/μ. I … Ridge Regression
  p(z|w,X)=N(Xw,diag(R))
  p(w|z,X)=p(z|w,X)p(w)/p(z|X)
  Discuss MAP versus ML versus Bayesian MMSE
  ML: max p(z|w,X) wrt w => max ln p(z|w,x) => min –ln p(z|w,X)…. Weighted least squares
  MAP: max p(w|z,X) => min -ln ln p(z|w,x) –ln p(w)
  MMSE: E(w|z,X) …. Need posterior density …. In this case, it is MAP

$$ML: J = \frac{1}{2}\| \underline{z} - X\,\underline{w} \|^2_{R^{-1}} \Rightarrow \underline{\hat{w}}_N = \underbrace{(X^T R^{-1} X)^{-1}}_{\Sigma_N} X^T R^{-1} \underline{z}$$

$$\underline{\hat{w}}_{N+1} = \underline{\hat{w}}_N + \underbrace{\frac{\Sigma_N \underline{x}_{N+1}}{r_{N+1} + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}}}_{Kalman\ gain}(z_{N+1} - \hat{z}_{N+1|N});\ \hat{z}_{N+1|N} = \underline{x}^T_{N+1}\underline{\hat{w}}_N$$

$$MAP: J = \frac{1}{2}\| \underline{z} - X\,\underline{w} \|^2_{R^{-1}} + \frac{1}{2}\| \underline{w} - \underline{w}_0 \|^2_{\Sigma_0^{-1}}$$

$$\nabla_w J = (X^T R^{-1} X + \Sigma_0^{-1})w - X^T R^{-1}\underline{z} - \Sigma_0^{-1}\underline{w}_0 \Rightarrow \underline{\hat{w}}_N = \underbrace{(X^T R^{-1} X + \Sigma_0^{-1})^{-1}}_{\Sigma_N}[X^T R^{-1}\underline{z} + \Sigma_0^{-1}\underline{w}_0]$$

$$\Sigma^{-1}_{N+1} = \Sigma^{-1}_N + \frac{\underline{x}_{N+1}\underline{x}^T_{N+1}}{r_{N+1}} \Rightarrow \Sigma_{N+1} = \Sigma_N - \frac{\Sigma_N \underline{x}_{N+1}\underline{x}^T_{N+1}\Sigma_N}{r_{N+1} + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}}$$

$$\underline{\hat{w}}_{N+1} = \underline{\hat{w}}_N + \underbrace{\frac{\Sigma_N \underline{x}_{N+1}}{r_{N+1} + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}}}_{Kalman\ gain}\underbrace{(z_{N+1} - \hat{z}_{N+1|N})}_{\nu_{N+1}};\ \hat{z}_{N+1|N} = \underline{x}^T_{N+1}\underline{\hat{w}}_N$$

$$p(\underline{w} \,|\, \underline{z}, X) \sim N(\underline{\hat{w}}_N, \Sigma_N);\ p(\underline{z} \,|\, X) = p(\underline{\nu} \,|\, X) \sim N(X\,\underline{\hat{w}}_N, X\Sigma_N X^T + R)$$

$$\underline{\hat{w}}_N = \Sigma_N X^T_N \underline{z}_N = \Sigma_N\left(\sum_{i=1}^N \underline{x}_i z_i\right);\ \Sigma_N = \left(X^T X + \mu I\right)^{-1} = \left(\sum_{i=1}^N \underline{x}_i \underline{x}^T_i + \mu I\right)^{-1}$$

$$\underline{\hat{w}}_{N+1} = \Sigma_{N+1} X^T_{N+1}\underline{z}_{N+1} = \left(\sum_{i=1}^N \underline{x}_i \underline{x}^T_i + \mu I + \underline{x}_{N+1}\underline{x}^T_{N+1}\right)^{-1}\left(\sum_{i=1}^N \underline{x}_i z_i + \underline{x}_{N+1} z_{N+1}\right)$$

$$= \underbrace{(\Sigma_N - \frac{\Sigma_N \underline{x}_{N+1}\underline{x}^T_{N+1}\Sigma_N}{1 + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}})}_{\Sigma_{N+1}}\left(\sum_{i=1}^N \underline{x}_i z_i + \underline{x}_{N+1} z_{N+1}\right)$$

$$= \underline{\hat{w}}_N + \frac{\Sigma_N \underline{x}_{N+1}}{1 + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}}(z_{N+1} - \underline{x}^T_{N+1}\underline{\hat{w}}_N) = \underline{\hat{w}}_N + \underbrace{\frac{\Sigma_N \underline{x}_{N+1}}{1 + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}}}_{Kalman\ gain}(z_{N+1} - \underbrace{\hat{z}_{N+1|N}}_{\hat{z}_{-(N+1)}})$$

$$Note: \underline{x}^T_{N+1}\Sigma_{N+1}\underline{x}_{N+1} = \frac{\underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}}{1 + \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1}} \Rightarrow \underline{x}^T_{N+1}\Sigma_N \underline{x}_{N+1} = \frac{\underline{x}^T_{N+1}\Sigma_{N+1}\underline{x}_{N+1}}{1 - \underline{x}^T_{N+1}\Sigma_{N+1}\underline{x}_{N+1}}$$

$$So, \hat{z}_{N+1|N+1} == \hat{z}_{N+1|N} + \underline{x}^T_{N+1}\Sigma_{N+1}\underline{x}_{N+1}(z_{N+1} - \hat{z}_{N+1|N})$$

$$z_{N+1} - \hat{z}_{-(N+1)} = (1 - \underline{x}^T_{N+1}\Sigma_{N+1}\underline{x}_{N+1})(z_{N+1} - \hat{z}_{N+1|N})$$

$(z_{N+1} - \hat{z}_{-(N+1)}) = \dfrac{z_{N+1} - \hat{z}_{N+1|N+1}}{1 - \underline{x}^T_{N+1}\Sigma_{N+1}\underline{x}_{N+1}}$..... vaid for any measurement in LS... just replace $N+1$ by $i$

Not true for dynamic systems.

$$J_{LOOCV} = \frac{1}{N+1}\sum_{i=1}^{N+1}(z_i - \hat{z}_{-i})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{z_i - \hat{z}_{i|N+1}}{1 - \underline{x}^T_i \Sigma_{N+1}\underline{x}_i}\right)^2$$

- Information Theory: Entropy, KL-divergence, Mutual Information, use Gaussian example
- ML versus MAP
- Gradient and Hessian

- $$L(x) = L(f(h(g(x)))$$
  $$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial f}\frac{\partial f}{\partial h}\frac{\partial h}{\partial g}\frac{\partial g}{\partial x} = \lambda_g \frac{\partial g}{\partial x}$$
  $$\lambda_f = \frac{\partial L}{\partial f}; \lambda_h = \frac{\partial L}{\partial h} = \lambda_f \frac{\partial f}{\partial h};$$
  $$\lambda_g = \frac{\partial L}{\partial g} = \lambda_h \frac{\partial h}{\partial g}$$
  $$x \to g \to h \to f \to L$$
  $$\frac{\partial L}{\partial x} = \lambda_g \frac{\partial g}{\partial x} \leftarrow \lambda_g = \frac{\partial L}{\partial g} = \lambda_h \frac{\partial h}{\partial g} \leftarrow \lambda_h = \frac{\partial L}{\partial h} = \lambda_f \frac{\partial f}{\partial h} \leftarrow \lambda_f = \frac{\partial L}{\partial f}$$
  Back propagation!!!!
- Optimization algorithms