MAP inference and regularization
1. What is the marginal $P(x)$?
2. Are $x$ and $y$ uncorrelated, positively or negatively correlated?
3. What is the marginal entropy $H(x)$?
4. What is the conditional entropy $H(x|y)$?
5. What is the mutual information $I(x,y)$?
1. What is the marginal $P(x)$?
2. Are $x$ and $y$ uncorrelated, positively or negatively correlated?
3. What is the marginal entropy $H(x)$?
4. What is the conditional entropy $H(x|y)$?
5. What is the mutual information $I(x,y)$?

\[
\begin{array}{ccc}
3 & 0 & 1/8 \\
2 & 0 & 1/8 \\
1 & 1/2 & 0 & 0 \\
\end{array}
\]

\[
P(x,y) = \begin{cases} 
0 & \text{if } x=1, y=2 \\
1/8 & \text{otherwise}
\end{cases}
\]

- $\log_2(1) = 0$
- $\log_2(1/2) = -1$
- $\log_2(1/4) = -2$
1. What is the marginal $P(x)$?

\[ [0.5 \ 0.25 \ 0.25] \]

2. Are $x$ and $y$ uncorrelated, positively or negatively correlated?

3. What is the marginal entropy $H(x)$?

\[ -\sum p(x) \log_2 p(x) \]

\[ 0.5(-\log(0.5)) + 2 \times 0.25(-\log(0.25)) \]
\[ = 0.5 \times 1 + 0.5 \times 2 \]
\[ = 1.5 \text{ bits} \]
4. What is the conditional entropy \( H(x|y) \)?

\[
-\sum p(y) \sum p(x|y) \log_2 p(x|y)
\]

\[
\begin{align*}
0.5 \text{ (0 bit)} + 0.25 \text{ (1 bit)} + 0.25 \text{ (1 bit)} \\
&= 0.5 \text{ bits}
\end{align*}
\]
4. What is the conditional entropy $H(x|y)$?

$-\sum p(y)\sum p(x|y) \log_2 p(x|y)$

$0.5 \times 0 \text{ bit} + 0.25 \times 1 \text{ bit} + 0.25 \times 1 \text{ bit} = 0.5 \text{ bits}$

5. What is the mutual information $I(x,y)$?

$H(x) - H(x|y)$

$1.5 \text{ bits} - 0.5 \text{ bits} = 1 \text{ bit}$
Review of ML estimation for GLMs.

\[ \hat{w}_{ML} = \arg \max_{\hat{w}} \log p(Y|X, \hat{w}) \]

Linear-Gaussian:

\[ L = -\frac{1}{2\sigma^2}(Y - X\hat{w})^T(Y - X\hat{w}) \]

\[ \hat{w}_{ML} = (X^T X)^{-1} X^T Y \]

- \[ \text{e.g.} \]
- \[ \frac{(y_i - x_i \cdot \hat{w})^2}{2 \sigma^2} \]

must be invertible.

Problem: \( X^T X \) may be non-invertible, or ill-conditioned.

Solution: add a "penalty" function.
Bayesian Approach

Posterior
\[ p(w | X, Y) = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{"marginal likelihood" or "evidence"}} \]

\[ = \int p(Y | X, w) p(w) \, dw \]

This ensures that \( p(w | X, Y) \) integrates to 1.

\[ \text{Maximum a posteriori (MAP) estimator} \]

\[ \hat{w}_{\text{MAP}} = \arg \max_w p(w | X, Y) \]

\[ = \arg \max_w \log p(w | X, Y) \]

\[ = \arg \max_w \left[ \log p(Y | X, w) + \log p(w) \right] \]

\[ \text{log prior or "penalty"} \]
• Maximiing the posterior $\iff$

maximizing $\text{log-likelihood} + \log \text{ prior}$. 

penalty term that penalizes “unlikely” solutions $w$. 

• Any time you see “penalized log likelihood” remember that this is always equivalent to a MAP estimator where “penalty” = “log prior” Bayesian.

---

**Common choices of prior**

1. “Ridge” prior.

   - also known as “$L_2$ penalty”, “$L_2$ shrinkage”, “Tikhonov Regression”, 

   penalty: $-\frac{1}{2} \lambda \|w\|^2 = -\frac{1}{2} \lambda \sum_{i=1}^{d} w_i^2 = -\frac{1}{2} \lambda \hat{w}^T \hat{w}$

   - Ridge Parameter, Squared ($L_2$) norm of $w$.
prior \( p(w) \propto e^{-\frac{1}{2}w^T C w} \)

\( C = \text{covariance} \)

\[ = N(0, \frac{1}{2}I) \]

\( N(m, C) \Rightarrow p(w) \propto e^{-\frac{1}{2}(w-m)^T C^{-1} (w-m)} \)

(2) \( L_2 \) smoothing prior.

\[
\text{penalty: } -\frac{1}{2} \lambda \sum_{i=2}^d (w_i - w_{i-1})^2
\]

\[
D = \begin{bmatrix}
-1 & 1 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 \\
0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & -1 & 1
\end{bmatrix}
\]

\[
\hat{w} = \begin{bmatrix}
w_2 - w_1 \\
w_3 - w_2 \\
\vdots \\
w_d - w_{d-1}
\end{bmatrix} = Dw
\]
\[ \| D w \|_2^2 = w^T D^T D w = \sum_{i=2}^{d} (w_i - w_{i-1})^2 \]

\[ \Pi = \text{graph Laplacian} \]

Equivalent prior:

\[ p(w) \propto e^{-\frac{1}{2} w^T \Pi w} \]

\[ N(0, \frac{1}{\lambda} \Pi^{-1}) \]

Such prior gives us the penalty:

\[ \log p(w) = -\frac{1}{2} \lambda w^T \Pi w \]

Note that we can define the graph Laplacian for other geometries of \( w \).

For example, if \( w \) is a 2D image we can enforce smoothness horizontally.
via a penalty of the form

$$-\frac{1}{2} \lambda \left[ \sum_{i=1}^{d_1} (v_{ij} - w_{i-1,j})^2 + \sum_{j=1}^{d_2} (w_{ij} - w_{ij-1})^2 \right]$$

sq. diffs along rows sq. diffs along columns.