Last time: 1) More Fisher info.

2) Cramér-Rao bound

3) Differential correlations

HW: 1) How to estimate θ from x, y

2) Fisher information lower bound

Information theory

Def: The surprise of an outcome k w/ probability $P(k)$ is:

$$u_k = -\log_2 P(k) \quad \text{(in bits)}$$

Example 1: Weighted coin flip, prob. of heads $P(H)$

$P(H) = 1/2$, $u_H = -\log_2 1/2 = 1$ bit

$P(H) = 1$ $u_H = 0$ bits

$P(H) = 1/4$ $u_H = 2$ bits

Example 2: Two coins, $P(H) = 1/2$, $P(HH) = 1/4$

$u_{HH} = 2$ bits

Surprise is additive for independent events.

Def: The entropy of a r.v. X that takes values x_k w/ probabilities $P(x_k)$ is the avg. surprise:

$$H(x) = -\sum_k P(x_k) \log_2 P(x_k)$$

Example:

If $P(H) = P(T) = 1/2$

$$H = -P(H)\log_2 P(H) - P(T)\log_2 P(T) = 1 \text{ bit}$$

$P(H) = 1$, $P(T) = 0 \Longrightarrow H = 0$ bits

$P(H) = 3/4$, $P(T) = 1/4 \Longrightarrow H = 0.81$ bits

For discrete prob. dists., uniform dist. has highest entropy.

If n outcomes, then $P(x_k) = 1/n$, $k = 1 \ldots n$

$$H(x) = -\sum_{k=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = -\log_2 1/n = \log_2 n$$

Def: The mutual information between S and R is:

$$I(R,S) = H(R) - H(R|S) = H(S) - H(S|R)$$
$$= H(R) + H(S) - H(R,S)$$

where $H(R|S)$ is the conditional entropy

$$\sum_j P(S_j) H(R|S=S_j) = -\sum_j P(S_j) \sum_k P(R_k|S_j) \log_2 P(R_k|S_j)$$

(assuming that R takes values $\{R_k\}$ and S takes $\{S_j\}$)

Intuition using Kullback-Leibler divergence:

Given two prob dists. $P(x)$, $Q(x)$,

$$D_{KL}(P \| Q) = \sum_k P(x_k) \log_2 \frac{P(x_k)}{Q(x_k)}$$

(need $Q(x_k) = 0 \Rightarrow P(x_k) = 0$)

Properties: 1) $D_{KL}(P \| Q) \geq 0$ if $P(x_k) = Q(x_k) \ \forall k$

2) $D_{KL} \geq 0$

3) $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$

Claim: $I(r,s) = D_{KL}\left(P(r,s) \,\|\, P(r)P(s)\right)$

$$= \sum_{j,k} P(r_k, s_j) \log_2 \frac{P(r_k, s_j)}{P(r_k)P(s_j)}$$

Note: $P(r_k, s_j) = P(r_k|s_j)P(s_j)$

$$= \sum_j P(s_j) \sum_k P(r_k|s_j) \log_2 \frac{P(r_k|s_j)}{P(r_k)}$$

$$= \sum_j P(s_j) \sum_k P(r_k|s_j) \log_2 P(r_k|s_j) \qquad (= -H(r|s))$$

$$- \underbrace{\sum_j P(s_j) \sum_k P(r_k|s_j)}_{\sum_k P(r_k)} \log_2 P(r_k) \qquad (= H(r))$$

$$= H(r) - H(r|s) = I(r,s)$$

Same argument for $H(s) - H(s|r)$

Conclusion: Mutual info. is KL-divergence b/w joint & independent dists. of r, s.

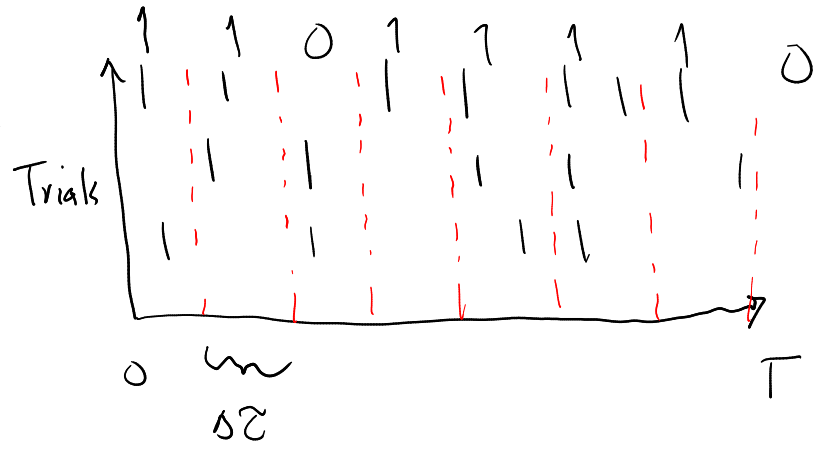
Application 1: Entropy of a spike train

Given spike times $P(\{t_k\}|s)$, how much info. about s is present?

Problem: Infinite-dimensional space of $\{t_k\}$.

Possibilities: Consider statistics (e.g. spike count)

Binning:



$N = T/\delta\tau$ bins

If 0/1, 2^N possible binary "words"

Hard to estimate for large T.

Depends on $\delta\tau$.

"Direct method" (Strong, Koberle, de Ruyter van Steveninck, Bialek, PRL 1997)

Example (Strong et al. 1997):

If $P(x_k)$ is uniform over n outcomes, $H = \log_2 n$

and 2 random observations are the same w/ prob. $1/n$.

Can use the prob. of "coincidence" P_c to

estimate $H \approx -\log_2 P_c$

In general, $P_c = \sum_k P(x_k)^2 = E\left[P(x_k)\right]$

$$H = -\sum_k P(x_k) \log_2 P(x_k) = -E\left[\log_2 P(x_k)\right]$$

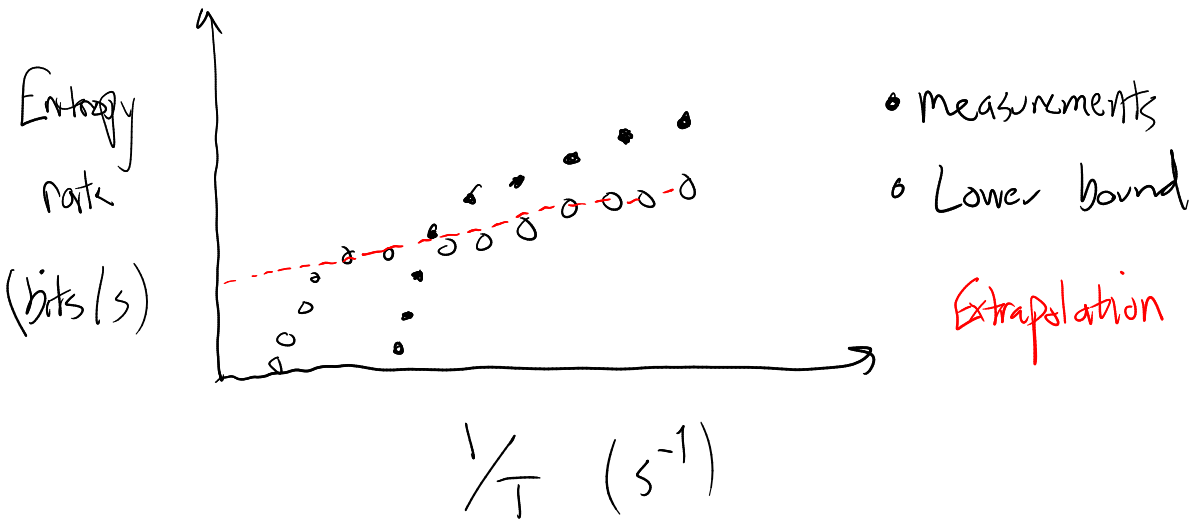
$$\geq -\log_2 E\left[P(x_k)\right] = -\log_2 P_c$$

For spike trains, assume words w/ same # of spikes S equally probable.

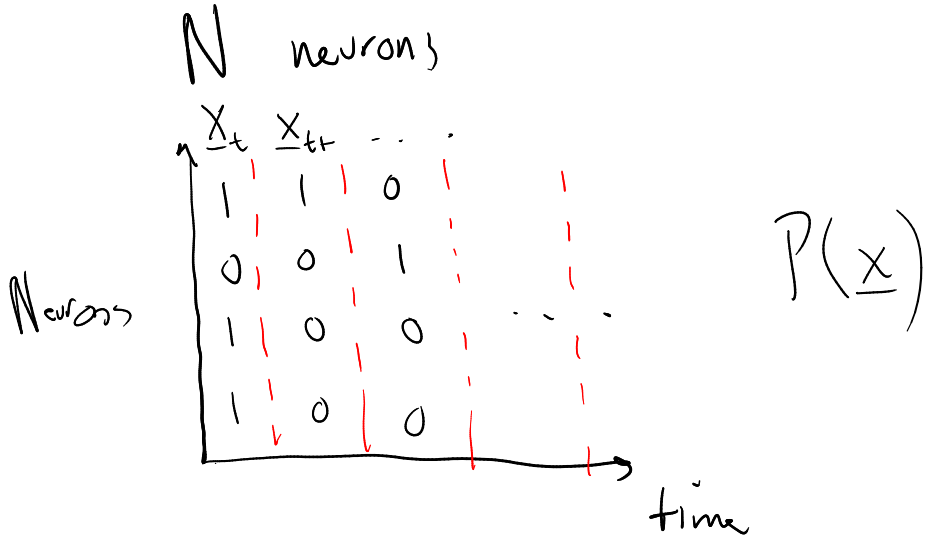
$$H_{LB} = -\sum_S P(S) \log_2 \left[P(S) \frac{2N_c(S)}{N_{obs}(S)\left[N_{obs}(S) - 1\right]} \right]$$

$N_c(S)$: # of coincidences w/ S spikes

$N_{obs}(S)$: # total occurrences of words w/ S spikes



Application 2: Maximum entropy models



$P(\underline{x})$

How much structure in this population code can be explained by constraints?

Idea: Find maximum entropy distribution consistent w/ a set of constraints. Which constraints?

0) $\sum_{k=1}^{2^N} P(\underline{x}_k) = 1$

1) $\sum_{k} P(\underline{x}_k)\, x_i = \mu_i = E[x_i]$

(avg. rate of neuron i).

$$\sum_k P(\underline{x}_k)\, x_i x_j = E[x_i x_j] = c_{ij} \quad \text{(correlations)}$$

Let $P(\underline{x}_k) = P_k$

Maximize $-\sum_k P_k \log_2 P_k$ subject to these constraints

Method of Lagrange multipliers:

$$\frac{\partial}{\partial P_k}\left[-\sum_k P_k \log_2 P_k + \lambda\left(\sum_k P_k - 1\right) + \sum_{i=1}^{N} h_i \left(\sum_k P_k x_i - \mu_i\right) + \sum_{i,j=1}^{N} J_{ij}\left(\sum_k P_k x_i x_j - c_{ij}\right)\right] = 0$$

$$= -\log P_k - 1 + \lambda + \sum_i h_i x_i + \sum_{ij} J_{ij} x_i x_j$$

$$P(\underline{x}_k) = \exp\left(\lambda - 1 + \sum_i h_i x_i + \sum_{ij} J_{ij} x_i x_j\right)$$

$$= \frac{1}{Z} \exp\left(\sum_i h_i x_i + \sum_{ij} J_{ij} x_i x_j \right)$$

normalization

first-order constraints

second-order

"Boltzmann distribution"

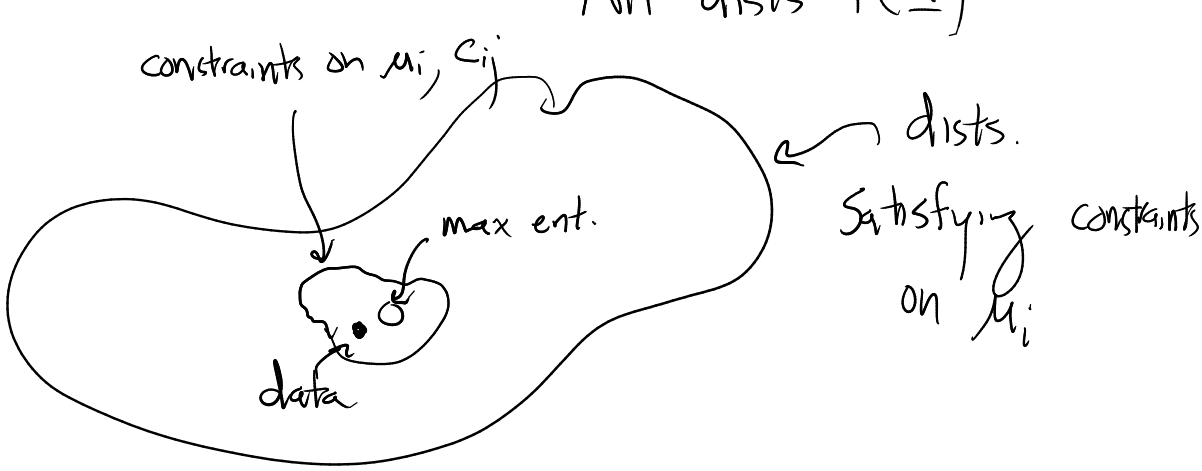
If no constraints on μ_i, c_{ij}

$P(\underline{x}_k) = 1/Z$, uniform dist.

If only on μ_i

$$P(\underline{x}_k) = \frac{1}{Z} \exp\left(\sum_i h_i x_i \right) = \frac{1}{Z} \prod_i e^{h_i x_i}$$

(independ)



If H_2 is entropy of max-ent. dist satisfying k^{th} order constraints

$$H_1 > H_2 > \dots H_N$$

Schneidman et al Nature 2006