# TODAY: Maximum Entropy & MDL, S-Value Connection

1. Note: No More Homework
2. Test Kaltura for final examination
3. Maximum Entropy and Minimum Description Length
4. Wrap-Up, Feedback

# The Coding (or Log-Loss) Game

- Data-compression as a two-player zero-sum game
- *Nature* picks a distribution $P$
- *Statistician* only knows that $P \in \mathcal{P} = \{P: E_P[\phi(X)] = t\}$ but nothing else
- Statistician's goal is to minimize expected code-length in the worst-case, i.e. find $Q$ achieving

$$\min_{q} \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\log q(X)]$$

**Nature's choice**

**Statistician's choice: over all (incl defective) distrs**

# The Coding (or Log-Loss) Game

- Statistician's goal is to minimize expected code-length in the worst-case, i.e. find $Q$ achieving

$$\min_q \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\log q(X)]$$

- Nature's goal is to maximize expected code-length in the worst-case, i.e. find $P \in \mathcal{P}$ achieving

$$\max_{P \in \mathcal{P}} \min_q \mathbf{E}_{X \sim P}[-\log q(X)]$$

…it seems that Nature's goal is rather 'un-natural'.
However, we have:

$$\min_q \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\log q(X)] = \max_{P \in \mathcal{P}} \min_q \mathbf{E}_{X \sim P}[-\log q(X)]$$

*It does not matter who is allowed to move second!*

# The Coding (or Log-Loss) Game

$$\min_q \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\log q(X)] \;=\; \max_{P \in \mathcal{P}} \min_q \mathbf{E}_{X \sim P}[-\log q(X)]$$

- Instance of the celebrated minimax theorem of game-theory/convex analysis. Originally due to Von Neumann (1928), but only for finite sample spaces and functions with bounded range
- This form holds for (quite) general convex constraints and is due to Topsoe (1979)
- We will show it for linear constraints (proof is easy)

# Relation to Maximum Entropy

$$\min_q \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\log q(X)] = \max_{P \in \mathcal{P}} \min_q \mathbf{E}_{X \sim P}[-\log q(X)]$$
$$= \max_{P \in \mathcal{P}} H(P)$$

- Both minimum on left and maximum on right achieved for $P_{\mathrm{me}}$
  - …for the left-hand-side this is surprising: the solution satisfies the constraint, even though we did not impose it!
  - although the game is extremely asymmetric, the optimal move **for both players is the same**
- $P_{\mathrm{me}}$ can thus be thought of as **the worst-case optimal distribution to use for data-compression** when data comes from some distribution in $\mathcal{P}$ , but you have no idea which → motivation for use of MaxEnt in practice!

# Proof, Part 1

## (this part we already saw last week)

$$p_\beta(x) = \frac{1}{Z(\beta)} \cdot e^{\beta^T \phi(X)} \qquad Z(\beta) = \sum_{x \in \mathcal{X}} e^{\beta^T \phi(X)}$$

Theorem, Part 1: suppose there exists $\tilde{\beta}$ s.t. $P_{\tilde{\beta}} \in \mathcal{P}$ i.e. $E_{X \sim P_{\tilde{\beta}}}[\phi(X)] = t$ . Then: $P_{\tilde{\beta}} = P_{\text{me}} := \arg\max_{P \in \mathcal{P}} H(P)$

$$H(P_{\tilde{\beta}}) = \max_{P \in \mathcal{P}} \min_q \mathbf{E}_{X \sim P}[-\log q(X)] = \max_{P \in \mathcal{P}} H(P)$$

Proof: let $P \in \mathcal{P}$ . We have:

$$H(P) \leq \mathbf{E}_{X \sim P}[-\log p_{\tilde{\beta}}(X)] =$$

$$\mathbf{E}_{X \sim P}[-\tilde{\beta}^T \phi(X) + \log Z(\tilde{\beta})] = -\tilde{\beta}^T t + \log Z(\tilde{\beta}) =$$

$$\mathbf{E}_{X \sim P_{\tilde{\beta}}}[-\beta^T \phi(X) + \log Z(\tilde{\beta})] = H(P_{\tilde{\beta}})$$

# Proof, Part 2

Theorem, Part 2: suppose there exists $\tilde{\beta}$ s.t. $P_{\tilde{\beta}} \in \mathcal{P}$ i.e. $E_{X \sim P_{\tilde{\beta}}}[\phi(X)] = t$ . Then:

$$H(P_{\tilde{\beta}}) = \min_{q} \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[- \log q(X)]$$

$$p_{\tilde{\beta}} = p_{\text{me}} = \arg\min_{q \in \mathcal{Q}} \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[- \log q(X)]$$

Proof: let $q$ be a (defective) prob. mass fn. We have

$$\max_{P \in \mathcal{P}} \mathbf{E}_{P}[- \log q(X)] \geq \mathbf{E}_{X \sim P_{\tilde{\beta}}}[- \log q(X)] \geq H(P_{\tilde{\beta}})$$ …yet

$$\max_{P \in \mathcal{P}} \mathbf{E}_{P}[- \log p_{\tilde{\beta}}(X)] =$$

$$\max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\tilde{\beta}^T \phi(X) + \log Z(\tilde{\beta})] = -\tilde{\beta}^T t + \log Z(\tilde{\beta}) =$$

$$\mathbf{E}_{X \sim P_{\tilde{\beta}}}[-\beta^T \phi(X) + \log Z(\tilde{\beta})] = H(P_{\tilde{\beta}})$$

# Equalizer Property

- In fact we proved something stronger than
  $p_{\tilde{\beta}} = p_{\text{me}} = \arg\min_{q \in \mathcal{Q}} \max_{P \in \mathcal{P}} \mathbf{E}_{X \sim P}[-\log q(X)]$
- Namely, we showed that for all $P \in \mathcal{P}$,
  $\mathbf{E}_{X \sim P}[-\log p_{\tilde{\beta}}(X)] = \mathbf{E}_{X \sim P_{\tilde{\beta}}}[-\log p_{\tilde{\beta}}(X)] = H(P_{\tilde{\beta}}).$
- So not only is $p_{\tilde{\beta}}$ worst-case optimal for coding, you also have a guarantee how well you will do in expectation!
- Data behaves as if $P_{\tilde{\beta}}$ were the true distribution, even though it isn't!
  - weird property. Called "robustness" in book
  - have already seen this e.g. for Bernoulli

# MaxEnt vs MDL

- So the maximum entropy distribution minimizes worst-case expected codelength
- Can MaxEnt therefore be seen as ‘a form of’ MDL?

Not really: with MDL model selection

  - we restrict the models we look at beforehand (e.g. all polynomials)
  - we then pick the model minimizing actual codelength on the data…where the code we use minimizes maximum regret.

With MaxEnt

  - we don't pick any model beforehand; we just observe a constraint.
  - We then pick distribution minimizing maximum codelength of the data

# MaxEnt vs MDL, II

- Also, the MaxEnt distribution is a solution to a minimax absolute codelength problem
  - Solution in set of distributions under consideration (constraint)
- ….whereas the NML distribution is a solution to a minimax relative codelength problem
  - Solution not in set of distributions under consideration (model); leads to ‘learning’ (predictive distributions pick up on patterns in past data)

Usually the first is taken in-expectation and the second for individual sequences, but that is a less fundamental difference

# From MaxEnt to MinRelEnt

- We can extend the story from MaxEnt to general exponential families (with nonuniform carrier $r_0(x)$ ):
- Let $L_{r_0}(P, q) \coloneqq E_{X \sim P}\,[-\log\, q(X) - [-\log r_0\,(X)]]$ be '$P$-expected codelength achieved by $q$ relative to $r_0$'
- Let $p_\beta(x) = \frac{1}{Z(\beta)} \cdot e^{\beta^T \phi(X)} \cdot r_0(x)$ $\quad Z(\beta) = \sum_{x \in \mathcal{X}} e^{\beta^T \phi(X)} r_0(x)$
- Theorem: fix arbitrary $r_0$ s.t. there exists $\tilde{\beta}$ s.t. $p_{\tilde{\beta}} \in \mathcal{P}$ i.e. $E_{X \sim P_{\tilde{\beta}}}[\phi(X)] = t$ . Then

$$\min_q \max_{P \in \mathcal{P}} L_{r_0}(P, q) = \max_{P \in \mathcal{P}} \min_q L_{r_0}(P, q)$$

…both min on left and max on right achieved by $P_{\tilde{\beta}}$

# From MaxEnt to MinRelEnt

Theorem: fix arbitrary $r_0$ s.t. there exists $\tilde{\beta}$ s.t. $p_{\tilde{\beta}} \in \mathcal{P}$ i.e. $E_{X\sim P_{\tilde{\beta}}}[\phi(X)] = t$ . Then

$$\min_{q} \max_{P\in\mathcal{P}} L_{r_0}(P,q) \quad = \max_{P\in\mathcal{P}} \min_{q} L_{r_0}(P,q)$$

…both min on left and max on right achieved by $P_{\tilde{\beta}}$

**$P_{\tilde{\beta}}$ can now be thought of as minimum relative entropy distribution 'the closest to $R_0$ satisfying constraint':**

$$\begin{aligned}
P_{\tilde{\beta}} &= \arg\max_{P\in\mathcal{P}} \min_{q} \mathbf{E}_{X\sim P}[-\log q(X) + \log r_0(X)] \\
&= \arg\max_{P\in\mathcal{P}} \mathbf{E}_{X\sim P}[-\log p(X) + \log r_0(X)] \\
&= \arg\min_{P\in\mathcal{P}} \mathbf{E}_{X\sim P}[\log p(X) - \log r_0(X)] \\
&= \arg\min_{P\in\mathcal{P}} D(P\|R_0).
\end{aligned}$$

# Relation to S-Values

- And now for something completely different…

**Hypothesis Testing**

**with S-Values**

…but then again, maybe not so different…

# Recall Definition of S-Values

- Let $H_0 = \{ P_\theta | \theta \in \Theta_0 \}$ represent the null hypothesis
  - Assume data $X_1, X_2, \ldots$ are i.i.d. under all $P \in H_0$ .
- Let $H_1 = \{ P_\theta | \theta \in \Theta_1 \}$ represent alternative hypothesis

- An **S-value** for sample size $n$ is a function $S : \mathcal{X}^n \to \mathbb{R}_0^+$ such that for **all** $P_0 \in H_0$ , we have

$$\mathbf{E}_{X^n \sim P_0} [S(X^n)] \leq 1$$

# Safe Tests

- The Safe Test against $H_0$ at level $\alpha$ based on S-value $S$ is defined as the test which rejects $H_0$ if $S(X^n) \geq \frac{1}{\alpha}$
- Since for all $P \in H_0$, all $0 \leq \alpha \leq 1$ ,

$$P\left(\frac{1}{S(X^n)} \leq \alpha\right) \leq \alpha$$

- ....the safe test which rejects $H_0$ iff $S(X^n) \geq 20$ , i.e. $S^{-1}(X^n) \leq 0.05$ , has **Type-I Error** Bound of 0.05

# How to design S-Values?

- Suppose we are willing to admit that we'll only be able to tell $H_0$ and $H_1$ apart if $P \in H_0 \cup H_1'$ for some $H_1' \subset H_1$ that excludes points that are 'too close' to $H_0$ e.g.

$$H_1' = \{P_\theta : \theta \in \Theta_1'\}, \Theta_1' = \{\theta \in \Theta_1 : \inf_{\theta_0 \in \Theta_0} \|\theta - \theta_0\|_2 \geq \delta\}$$

- We can then look for the GROW (growth-optimal in worst-case) S-value achieving

$$\sup_S \inf_{\theta \in \Theta_1'} \mathbf{E}_{X^n \sim P_\theta}[\log S]$$

# GROW: an analogue of Power

- The GROW (growth-optimal in worst-case) S-value relative to $H_{1,\delta}$ is the S-value achieving

$$\sup_{S} \inf_{\theta \in \Theta'_1} \mathbf{E}_{X^n \sim P_\theta}[\log S]$$

where the supremum is over all $S$-values relative to $H_0$

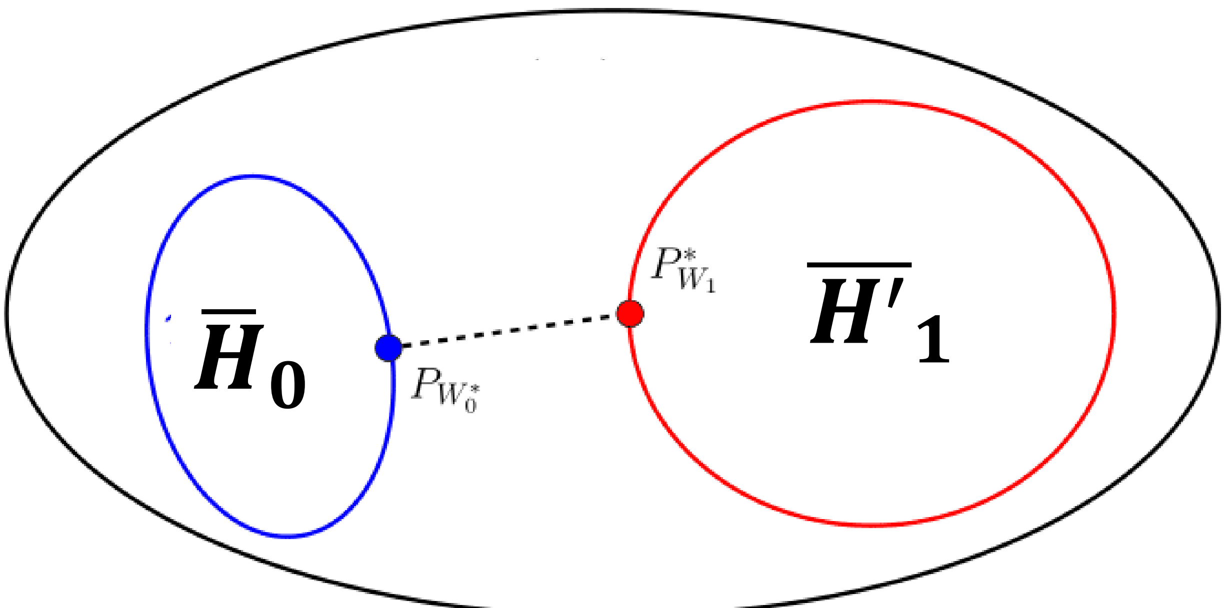
- ...so we don't expect to gain anything when investing in $S$ under $H_0$
- ...but among all such $S$ we pick the one(s) that make us rich fastest if we keep reinvesting in new gambles under $H_1$

# The best S-Value is given by the **Joint Information Projection (JIPr)**

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$\mathcal{W}_1$ set of all priors (prob distrs) on $\Theta'_1$

$$(W_1^*, W_0^*) := \text{arg} \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$



# Towards Main Theorem on S-Values

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Here $D$ is the relative entropy or Kullback-Leibler divergence, the central divergence measure in information theory and large deviations

$$D(P\|Q) := \mathbf{E}_{X^n \sim P} \left[ \log \frac{p(X^n)}{q(X^n)} \right]$$

(can give measure-theoretic definition making it well-defined even if $P$ and $Q$ not abs. cont.)

# Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \text{arg} \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose $(W_1^*, W_0^*)$ exists. Then $S^* := \frac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-value relative to $H_0$. (b)….

# Main Theorem

$$p_W(X^n) := \int p_\theta(X^n)\, dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose $(W_1^*, W_0^*)$ exists. Then $S^* := \dfrac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-value. (b) In fact it is the **GROW** S-value, i.e.

$$\inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S^*] = \sup_S \inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}}[\log S]$$

# Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \arg \min_{W_1 \in \mathcal{W}_1} \min_{W_0 : \text{distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

Suppose $(W_1^*, W_0^*)$ exists. Then $S^* := \frac{p_{W_1^*}(X^n)}{p_{W_0^*}(X^n)}$

is (a) an S-value. (b) In fact it is the **GROW** S-value, i.e.

$$\inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}} [\log S^*] = \sup_S \inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}} [\log S]$$

and (c) ,

$$= \min_{W_1 \in \mathcal{W}_1} \min_{W_0} D(P_{W_1} \| P_{W_0})$$

# Main Theorem

$$p_W(X^n) := \int p_\theta(X^n) dW(\theta)$$

$$(W_1^*, W_0^*) := \text{arg} \min_{W_1 \in \mathcal{W}_1} \min_{W_0 \text{:distr on } \Theta_0} D(P_{W_1} \| P_{W_0})$$

**This is really an extension of the previous minimum-relative-entropy minimax theorem!**

**(nobody knows this ☺ )**

is (a) an S-value. (b) In fact it is the **GROW** S-value, i.e.

$$\inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}} [\log S^*] = \sup_S \inf_{\theta_1 \in \Theta_1'} \mathbf{E}_{X^n \sim P_{\theta_1}} [\log S]$$

and (c) ,

$$= \min_{W_1 \in \mathcal{W}_1} \min_{W_0} D(P_{W_1} \| P_{W_0})$$

# Wrap-Up: What I hope you take away from this course and why

- Basics of Data Compression
    - Because it's highly important by itself, and needed for rest
- Material:

1. Kraft inequality
2. Entropy as expected codelength; KL as expected CL difference; Fisher information as 'correction' in approximation to KL by squared Euclidean distance

- Homework mainly intended to get a feel for basic properties of entropy such as concavity, upper bounds)

# Wrap-Up: What I hope you take away from this course and why

- Some observations about likelihood
  - Because it's highly important if you do statistics and too much of it is taken for granted usually (I think)
  - maximizing over data vs over parameters, a little bit about sufficient statistics
- Exponential Families
  - because they're highly important in statistics
  - Because all our important theorems hold for general exponential families
  - Some homework was to give you a feel for this; some (e.g. uniform distribution) to show that properties of exp fams are quite special
-

# Wrap-Up: What I hope you take away from this course and why

- Basics of Bayesian statistics.
  - Generally important (30% of all statistics papers)
  - Relation to Data Compression/Sequential Prediction (underappreciated!)
- Relation between MaxEnt and MDL
  - Takes away the magic from MaxEnt

# Wrap-Up: What I hope you take away from this course and why

- **Universal Coding/MDL Model Selection**
  - Highly important in Information Theory; should also be important in machine learning/statistics, but somewhat neglected there. Even if you can't use this, there was enough other stuff you will be able to use
- S-Values/Hypothesis Testing: the future of MDL based methods?
- **General: the interaction between information theory (data compression, gambling) and learning from data**

- Questions?