Combining infinite sets of experts

Yoav Freund

January 20, 2025

Freund: Predicting a binary Sequence almost as well the the optimal biased coin.
Risannen: Fisher Information and Stochastic Complexity.

Outline

Outline

Outline

Outline

Review

Probabilities and codes

- $M_1, \ldots, M_n$ - possible messages

Probabilities and codes

- $M_1, \ldots, M_n$ - possible messages
- $P(M_i)$ - probability of message i

Probabilities and codes

- $M_1, \ldots, M_n$ - possible messages
- $P(M_i)$ - probability of message i
- Arithmetic coding defines a code of length $\lceil -\log_2 P(M_i) \rceil$ for message i

The online Bayes Algorithm

- Total loss of expert i

$$L_i^t = -\sum_{s=1}^{t} \log p_i^s(c^s); \quad L_i^0 = 0$$

The online Bayes Algorithm

- Total loss of expert i

$$L_i^t = -\sum_{s=1}^{t} \log p_i^s(c^s); \quad L_i^0 = 0$$

- Weight of expert i

$$w_i^t = w_i^1 e^{-L_i^{t-1}} = w_i^1 \prod_{s=1}^{t-1} p_i^s(c^s)$$

The online Bayes Algorithm

- Total loss of expert i

$$L_i^t = -\sum_{s=1}^{t} \log p_i^s(c^s); \quad L_i^0 = 0$$

- Weight of expert i

$$w_i^t = w_i^1 e^{-L_i^{t-1}} = w_i^1 \prod_{s=1}^{t-1} p_i^s(c^s)$$

- Freedom to choose initial weights.
  $w_t^1 \geq 0, \sum_{i=1}^{n} w_i^1 = 1$

The online Bayes Algorithm

- Total loss of expert i

$$L_i^t = -\sum_{s=1}^{t} \log p_i^s(c^s); \quad L_i^0 = 0$$

- Weight of expert i

$$w_i^t = w_i^1 e^{-L_i^{t-1}} = w_i^1 \prod_{s=1}^{t-1} p_i^s(c^s)$$

- Freedom to choose initial weights.
  $w_t^1 \geq 0$, $\sum_{i=1}^n w_i^1 = 1$
- Prediction of algorithm A

$$\mathbf{p}_A^t = \frac{\sum_{i=1}^N w_i^t \mathbf{p}_i^t}{\sum_{i=1}^N w_i^t}$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^{N} w_i^t$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^N w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^N w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^N w_i^t}$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^{N} w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^{N} w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^{N} w_i^t} = \frac{\sum_{i=1}^{N} w_i^t p_i^t(c^t)}{\sum_{i=1}^{N} w_i^t}$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^N w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^N w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^N w_i^t} = \frac{\sum_{i=1}^N w_i^t p_i^t(c^t)}{\sum_{i=1}^N w_i^t} = p_A^t(c^t)$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^N w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^N w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^N w_i^t} = \frac{\sum_{i=1}^N w_i^t p_i^t(c^t)}{\sum_{i=1}^N w_i^t} = p_A^t(c^t)$$

$$-\log \frac{W^{t+1}}{W^t} = -\log p_A^t(c^t)$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^N w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^N w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^N w_i^t} = \frac{\sum_{i=1}^N w_i^t p_i^t(c^t)}{\sum_{i=1}^N w_i^t} = p_A^t(c^t)$$

$$-\log \frac{W^{t+1}}{W^t} = -\log p_A^t(c^t)$$

$$-\log \frac{W^{T+1}}{W^1} = -\sum_{t=1}^T \log p_A^t(c^t)$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^{N} w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^{N} w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^{N} w_i^t} = \frac{\sum_{i=1}^{N} w_i^t p_i^t(c^t)}{\sum_{i=1}^{N} w_i^t} = p_A^t(c^t)$$

$$-\log \frac{W^{t+1}}{W^t} = -\log p_A^t(c^t)$$

$$-\log \frac{W^{T+1}}{W^1} = -\sum_{t=1}^{T} \log p_A^t(c^t) = L_A^T$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^N w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^N w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^N w_i^t} = \frac{\sum_{i=1}^N w_i^t p_i^t(c^t)}{\sum_{i=1}^N w_i^t} = p_A^t(c^t)$$

$$-\log \frac{W^{t+1}}{W^t} = -\log p_A^t(c^t)$$

$$-\log W^{T+1} = -\log \frac{W^{T+1}}{W^1} = -\sum_{t=1}^{T} \log p_A^t(c^t) = L_A^T$$

Cumulative loss vs. Final total weight

Total weight: $W^t \doteq \sum_{i=1}^N w_i^t$

$$\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^N w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^N w_i^t} = \frac{\sum_{i=1}^N w_i^t p_i^t(c^t)}{\sum_{i=1}^N w_i^t} = p_A^t(c^t)$$

$$-\log \frac{W^{t+1}}{W^t} = -\log p_A^t(c^t)$$

$$-\log W^{T+1} = -\log \frac{W^{T+1}}{W^1} = -\sum_{t=1}^{T} \log p_A^t(c^t) = L_A^T$$

EQUALITY not bound!

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$
- Total Weight is at least the weight of the best expert.

$$L_A^T = -\log W^{T+1}$$

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$
- Total Weight is at least the weight of the best expert.

$$L_A^T = -\log W^{T+1}$$

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$
- Total Weight is at least the weight of the best expert.

$$L_A^T = -\log W^{T+1} = -\log \sum_{i=1}^{N} w_i^{T+1}$$

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$
- Total Weight is at least the weight of the best expert.

$$
\begin{aligned}
L_A^T &= -\log W^{T+1} = -\log \sum_{i=1}^{N} w_i^{T+1} \\
&= -\log \sum_{i=1}^{N} w_i^1 e^{-L_i^T}
\end{aligned}
$$

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$
- Total Weight is at least the weight of the best expert.

$$
\begin{aligned}
L_A^T &= -\log W^{T+1} = -\log \sum_{i=1}^{N} w_i^{T+1} \\
&= -\log \sum_{i=1}^{N} w_i^1 e^{-L_i^T} \le -\log \max_i \left(w_i^1 e^{-L_i^T} \right)
\end{aligned}
$$

Simple Bound

- Use non-uniform initial weights $\sum_i w_i^1 = 1$
- Total Weight is at least the weight of the best expert.

$$
\begin{aligned}
L_A^T &= -\log W^{T+1} = -\log \sum_{i=1}^{N} w_i^{T+1} \\
&= -\log \sum_{i=1}^{N} w_i^1 e^{-L_i^T} \le -\log \max_i \left(w_i^1 e^{-L_i^T} \right) \\
&= \min_i \left(L_i^T - \log w_i^1 \right)
\end{aligned}
$$

The Universal prediction machine

Standardizing online prediction algorithms

- Fix a universal Turing machine U.

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0,1\}^t$

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0, 1\}^t$
  - runs finite time and outputs

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0,1\}^t$
  - runs finite time and outputs
  - A prediction for the next bit $p(\vec{X}) \in [0,1]$.

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0,1\}^t$
  - runs finite time and outputs
  - A prediction for the next bit $p(\vec{X}) \in [0,1]$.
  - To ensure p has a finite description. Restrict to rational numbers n/m

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0,1\}^t$
  - runs finite time and outputs
  - A prediction for the next bit $p(\vec{X}) \in [0,1]$.
  - To ensure p has a finite description. Restrict to rational numbers n/m
- Any online prediction algorithm can be represented as code $\vec{b}(E)$ for U. The code length is $|\vec{b}(E)|$.

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0, 1\}^t$
  - runs finite time and outputs
  - A prediction for the next bit $p(\vec{X}) \in [0, 1]$.
  - To ensure p has a finite description. Restrict to rational numbers n/m
- Any online prediction algorithm can be represented as code $\vec{b}(E)$ for U. The code length is $|\vec{b}(E)|$.
- Most sequences do not correspond to valid prediction algorithms.

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0,1\}^t$
  - runs finite time and outputs
  - A prediction for the next bit $p(\vec{X}) \in [0,1]$.
  - To ensure p has a finite description. Restrict to rational numbers n/m
- Any online prediction algorithm can be represented as code $\vec{b}(E)$ for U. The code length is $|\vec{b}(E)|$.
- Most sequences do not correspond to valid prediction algorithms.
- $V(\vec{b}, \vec{X}, t) = 1$ if the program $\vec{b}$, given $\vec{X}$ as input, halts within t steps and outputs a well-formed prediction. Otherwise $V(\vec{b}, \vec{X}, t) = 0$

Standardizing online prediction algorithms

- Fix a universal Turing machine U.
- An online prediction algorithm E is a program that
  - given as input The past $\vec{X} \in \{0,1\}^t$
  - runs finite time and outputs
  - A prediction for the next bit $p(\vec{X}) \in [0,1]$.
  - To ensure p has a finite description. Restrict to rational numbers n/m
- Any online prediction algorithm can be represented as code $\vec{b}(E)$ for U. The code length is $|\vec{b}(E)|$.
- Most sequences do not correspond to valid prediction algorithms.
- $V(\vec{b},\vec{X},t) = 1$ if the program $\vec{b}$, given $\vec{X}$ as input, halts within t steps and outputs a well-formed prediction. Otherwise $V(\vec{b},\vec{X},t) = 0$
- $V(\vec{b},\vec{X},t)$ is computable (recursively enumerable).

A universal prediction machine

- Assign to the code $\vec{b}$ the initial weight $w^1_{\vec{b}} = 2^{-|\vec{b}|-\log_2 |\vec{b}|}$.

A universal prediction machine

- Assign to the code $\vec{b}$ the initial weight $w^1_{\vec{b}} = 2^{-|\vec{b}| - \log_2 |\vec{b}|}$.
- The total initial weight over all finite binary sequences is one.

A universal prediction machine

- Assign to the code $\vec{b}$ the initial weight $w^1_{\vec{b}} = 2^{-|\vec{b}|-\log_2|\vec{b}|}$.
- The total initial weight over all finite binary sequences is one.
- Run the Bayes algorithm over “all” prediction algorithms.

A universal prediction machine

- Assign to the code $\vec{b}$ the initial weight $w^1_{\vec{b}} = 2^{-|\vec{b}|-\log_2 |\vec{b}|}$.
- The total initial weight over all finite binary sequences is one.
- Run the Bayes algorithm over "all" prediction algorithms.
- technical details: On iteration t, $|\vec{X}| = t$. Use the predictions of programs $\vec{b}$ such that $|\vec{b}| \leq t$ and for which $V(\vec{b}, \vec{X}, 2^t) = 1$.
  the unused algorithms predict $1/2$ (insuring a loss of 1)

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$
- Assume E is a prediction algorithm which generates the tth prediction in time smaller than 2^t

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$
- Assume E is a prediction algorithm which generates the tth prediction in time smaller than 2^t
- When $t \leq |\vec{b}(E)|$ the algorithm is not used and thus it's loss is 1

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$
- Assume E is a prediction algorithm which generates the tth prediction in time smaller than 2^t
- When $t \leq |\vec{b}(E)|$ the algorithm is not used and thus it's loss is 1
- We get that the loss of the Universal algorithm is at most $|\vec{b}(E)| + \log_2 |\vec{b}(E)| + L_E$

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$
- Assume E is a prediction algorithm which generates the tth prediction in time smaller than 2^t
- When $t \leq |\vec{b}(E)|$ the algorithm is not used and thus it's loss is 1
- We get that the loss of the Universal algorithm is at most $|\vec{b}(E)| + \log_2 |\vec{b}(E)| + L_E$
- More careful analysis can reduce to $|\vec{b}(E)| + L_E$

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$
- Assume E is a prediction algorithm which generates the tth prediction in time smaller than 2^t
- When $t \leq |\vec{b}(E)|$ the algorithm is not used and thus it's loss is 1
- We get that the loss of the Universal algorithm is at most $|\vec{b}(E)| + \log_2 |\vec{b}(E)| + L_E$
- More careful analysis can reduce to $|\vec{b}(E)| + L_E$
- How good is that?

Performance of the universal prediction algorithm

- Using $L_A \leq \min_i \left(L_i - \log w_i^1\right)$
- Assume E is a prediction algorithm which generates the tth prediction in time smaller than 2^t
- When $t \leq |\vec{b}(E)|$ the algorithm is not used and thus it's loss is 1
- We get that the loss of the Universal algorithm is at most $|\vec{b}(E)| + \log_2 |\vec{b}(E)| + L_E$
- More careful analysis can reduce to $|\vec{b}(E)| + L_E$
- How good is that?
- What is the two part code?

Bayes coding is better than two part codes

- Simple bound as good as bound for two part codes (MDL) but enables online compression

Bayes coding is better than two part codes

- Simple bound as good as bound for two part codes (MDL) but enables online compression
- Suppose we have K copies of each expert.

Bayes coding is better than two part codes

- Simple bound as good as bound for two part codes (MDL) but enables online compression
- Suppose we have K copies of each expert.
- Two part code has to point to one of the KN experts
  $L_A \leq \log NK + \min_i L_i^T = \log NK + \min_i L_i^T$

Bayes coding is better than two part codes

- Simple bound as good as bound for two part codes (MDL) but enables online compression
- Suppose we have K copies of each expert.
- Two part code has to point to one of the KN experts $L_A \leq \log NK + \min_i L_i^T = \log NK + \min_i L_i^T$
- If we use Bayes predictor + arithmetic coding we get:

$$L_A = -\log W^{T+1} \leq \log K \max_i \frac{1}{NK} e^{-L_i^T} = \log N + \min_i L_i^T$$

Bayes coding is better than two part codes

- Simple bound as good as bound for two part codes (MDL) but enables online compression
- Suppose we have K copies of each expert.
- Two part code has to point to one of the KN experts $L_A \leq \log NK + \min_i L_i^T = \log NK + \min_i L_i^T$
- If we use Bayes predictor + arithmetic coding we get:

$$L_A = -\log W^{T+1} \leq \log K \max_i \frac{1}{NK} e^{-L_i^T} = \log N + \min_i L_i^T$$

- We don't pay a penalty for copies.

Bayes coding is better than two part codes

- Simple bound as good as bound for two part codes (MDL) but enables online compression
- Suppose we have K copies of each expert.
- Two part code has to point to one of the KN experts $L_A \leq \log NK + \min_i L_i^T = \log NK + \min_i L_i^T$
- If we use Bayes predictor + arithmetic coding we get:

$$L_A = -\log W^{T+1} \leq \log K \max_i \frac{1}{NK} e^{-L_i^T} = \log N + \min_i L_i^T$$

- We don't pay a penalty for copies.
- More generally, the regret is smaller if many of the experts perform well.

The biased coins set of experts

- Each expert corresponds to a biased coin, predicts with a fixed $\theta \in [0, 1]$.

The biased coins set of experts

- Each expert corresponds to a biased coin, predicts with a fixed $\theta \in [0, 1]$.
- Set of experts is uncountably infinite.

The biased coins set of experts

- Each expert corresponds to a biased coin, predicts with a fixed $\theta \in [0, 1]$.
- Set of experts is uncountably infinite.
- Only countably many experts can be assigned non-zero weight.

The biased coins set of experts

- Each expert corresponds to a biased coin, predicts with a fixed $\theta \in [0, 1]$.
- Set of experts is uncountably infinite.
- Only countably many experts can be assigned non-zero weight.
- Instead, we assign the experts a Density Measure.

The biased coins set of experts

- Each expert corresponds to a biased coin, predicts with a fixed $\theta \in [0, 1]$.
- Set of experts is uncountably infinite.
- Only countably many experts can be assigned non-zero weight.
- Instead, we assign the experts a Density Measure.
- $L_A \leq \min_i \left(L_i - \log w_i^1\right)$ is meaningless.

The biased coins set of experts

- Each expert corresponds to a biased coin, predicts with a fixed $\theta \in [0, 1]$.
- Set of experts is uncountably infinite.
- Only countably many experts can be assigned non-zero weight.
- Instead, we assign the experts a Density Measure.
- $L_A \leq \min_i \left(L_i - \log w_i^1\right)$ is meaningless.
- Can we still get a meaningful bound?

Bayes Algorithm for biased coins

- Replace the initial weight by a density measure
  $w(\theta) = w^1(\theta)$, $\int_0^1 w(\theta) d\theta = 1$

Bayes Algorithm for biased coins

- Replace the initial weight by a density measure $w(\theta) = w^1(\theta)$, $\int_0^1 w(\theta) d\theta = 1$
- Relationship between final total weight and total log loss remains unchanged:

$$L_A = \ln \int_0^1 w(\theta) e^{-L_\theta^{T+1}} d\theta$$

Bayes Algorithm for biased coins

- Replace the initial weight by a density measure $w(\theta) = w^1(\theta)$, $\int_0^1 w(\theta) d\theta = 1$
- Relationship between final total weight and total log loss remains unchanged:

$$L_A = \ln \int_0^1 w(\theta) e^{-L_\theta^{T+1}} d\theta$$

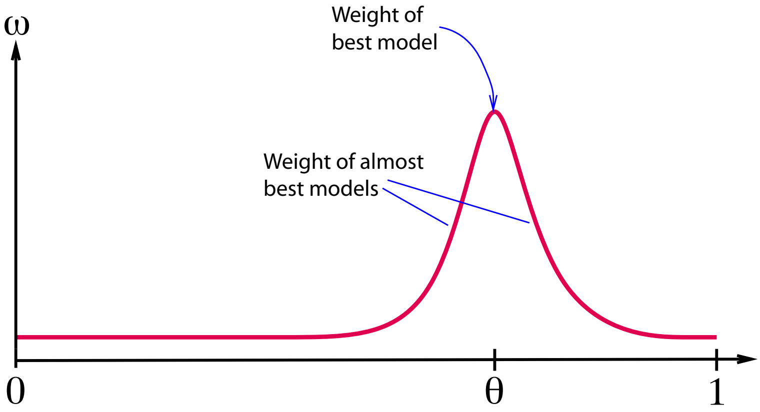
- We need a new lower bound on the final total weight

Main Idea

If $w^t(\theta)$ is large then $w^t(\theta + \epsilon)$ is also large.

Main Idea

If $w^t(\theta)$ is large then $w^t(\theta + \epsilon)$ is also large.



Expanding the exponent around the peak

- For log loss the best θ is empirical distribution of the seq.

$$\hat{\theta} = \frac{\#\{x^t = 1;\ \ 1 \leq t \leq T\}}{T}$$

Expanding the exponent around the peak

- For log loss the best θ is empirical distribution of the seq.

$$\hat{\theta} = \frac{\#\{x^t = 1;\ \ 1 \leq t \leq T\}}{T}$$

- The total loss scales with T

$$L_\theta = T \cdot (\hat{\theta}\ell(\theta, 1) + (1 - \hat{\theta})\ell(\theta, 0)) \doteq T \cdot g(\hat{\theta}, \theta)$$

Expanding the exponent around the peak

- For log loss the best θ is empirical distribution of the seq.

$$\hat{\theta} = \frac{\#\{x^t = 1; \;\; 1 \leq t \leq T\}}{T}$$

- The total loss scales with T

$$L_\theta = T \cdot (\hat{\theta}\ell(\theta, 1) + (1 - \hat{\theta})\ell(\theta, 0)) \doteq T \cdot g(\hat{\theta}, \theta)$$

Expanding the exponent around the peak

- For log loss the best θ is empirical distribution of the seq.

$$\hat{\theta} = \frac{\#\{x^t = 1; \;\; 1 \le t \le T\}}{T}$$

- The total loss scales with T

$$L_\theta = T \cdot (\hat{\theta}\ell(\theta, 1) + (1 - \hat{\theta})\ell(\theta, 0)) \doteq T \cdot g(\hat{\theta}, \theta)$$

$$L_A - L_{\min} \quad \le \quad \ln \int_0^1 w(\theta) e^{-L_\theta} d\theta - \ln e^{L_{\min}}$$

Expanding the exponent around the peak

- For log loss the best θ is empirical distribution of the seq.

$$\hat{\theta} = \frac{\#\{x^t = 1; \;\; 1 \le t \le T\}}{T}$$

- The total loss scales with T

$$L_\theta = T \cdot (\hat{\theta}\ell(\theta, 1) + (1 - \hat{\theta})\ell(\theta, 0)) \doteq T \cdot g(\hat{\theta}, \theta)$$

$$\begin{aligned} L_A - L_{\min} &\le \ln \int_0^1 w(\theta) e^{-L_\theta} d\theta - \ln e^{L_{\min}} \\ &= \ln \int_0^1 w(\theta) e^{-(L_\theta - L_{\min})} d\theta \end{aligned}$$

Expanding the exponent around the peak

- For log loss the best θ is empirical distribution of the seq.

$$\hat{\theta} = \frac{\#\{x^t = 1;\ \ 1 \leq t \leq T\}}{T}$$

- The total loss scales with T

$$L_\theta = T \cdot (\hat{\theta}\ell(\theta, 1) + (1 - \hat{\theta})\ell(\theta, 0)) \doteq T \cdot g(\hat{\theta}, \theta)$$

$$\begin{aligned}
L_A - L_{\min} &\leq \ln \int_0^1 w(\theta) e^{-L_\theta} d\theta - \ln e^{L_{\min}} \\
&= \ln \int_0^1 w(\theta) e^{-(L_\theta - L_{\min})} d\theta \\
&= \ln \int_0^1 w(\theta) e^{T(g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))} d\theta
\end{aligned}$$

Laplace approximation (idea)

- Taylor expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\theta = \hat{\theta}$.

Laplace approximation (idea)

- Taylor expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\theta = \hat{\theta}$.
- First and second terms in the expansion are zero.

Laplace approximation (idea)

- Taylor expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\theta = \hat{\theta}$.
- First and second terms in the expansion are zero.
- Third term gives a quadratic expression in the exponent

Laplace approximation (idea)

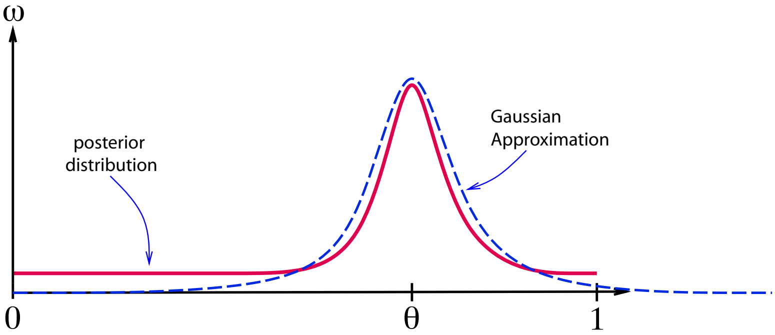
- Taylor expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\theta = \hat{\theta}$.
- First and second terms in the expansion are zero.
- Third term gives a quadratic expression in the exponent
- $\Rightarrow$ a gaussian approximation of the posterior.

Laplace approximation (idea)

- Taylor expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\theta = \hat{\theta}$.
- First and second terms in the expansion are zero.
- Third term gives a quadratic expression in the exponent
- $\Rightarrow$ a gaussian approximation of the posterior.

Laplace approximation (idea)

- Taylor expansion of $g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})$ around $\theta = \hat{\theta}$.
- First and second terms in the expansion are zero.
- Third term gives a quadratic expression in the exponent
- $\Rightarrow$ a gaussian approximation of the posterior.



Laplace Approximation, Watson's lemma

$$\int_0^1 w(\theta) e^{T(g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))} d\theta$$

Laplace Approximation, Watson's lemma

$$\int_0^1 w(\theta) e^{T(g(\hat{\theta},\theta)-g(\hat{\theta},\hat{\theta}))} d\theta$$

$$= \quad w(\hat{\theta}) \sqrt{\frac{-2\pi}{T \left. \frac{d^2}{d\theta^2} \right|_{\theta=\hat{\theta}} (g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))}} + O(T^{-3/2})$$

Choosing the optimal prior

- Choose $w(\theta)$ to maximize the worst-case final total weight

$$\min_{\hat{\theta}} w(\hat{\theta}) \sqrt{\frac{-2\pi}{T \left. \frac{d^2}{d\theta^2} \right|_{\theta=\hat{\theta}} \left(g(\hat{\theta}, \theta) - g(\hat{\theta}, \hat{\theta})\right)}}$$

Choosing the optimal prior

- Choose $w(\theta)$ to maximize the worst-case final total weight

$$\min_{\hat{\theta}} w(\hat{\theta}) \sqrt{\frac{-2\pi}{T \left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} (g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))}}$$

- Make bound equal for all $\hat{\theta} \in [0,1]$ by choosing

$$w^*(\hat{\theta}) = \frac{1}{Z} \sqrt{\frac{\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} (g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))}{-2\pi}} \; ,$$

where Z is the normalization factor:

$$Z = \sqrt{\frac{1}{2\pi}} \int_0^1 \sqrt{\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} (g(\hat{\theta},\hat{\theta}) - g(\hat{\theta},\theta))} \; d\hat{\theta}$$

The bound for the optimal prior

- We get that the regret is

$$
\begin{aligned}
L_A - L_{\min} &\leq \ln \int_0^1 w^*(\theta) e^{T(g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}))} d\theta \\
&= \ln \left(\sqrt{\frac{2\pi Z}{T}} + O(T^{-3/2}) \right) \\
&= \frac{1}{2} \ln \frac{T}{2\pi} - \frac{1}{2} \ln Z + O(1/T) \, .
\end{aligned}
$$

Solving for log-loss

- The exponent in the integral is

$$g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}) = \hat{\theta}\ln\frac{\hat{\theta}}{\theta} + (1-\hat{\theta})\ln\frac{1-\hat{\theta}}{1-\theta} = D_{KL}(\hat{\theta}||\theta)$$

Solving for log-loss

- The exponent in the integral is

$$g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}) = \hat{\theta}\ln\frac{\hat{\theta}}{\theta} + (1-\hat{\theta})\ln\frac{1-\hat{\theta}}{1-\theta} = D_{KL}(\hat{\theta}||\theta)$$

- The second derivative

$$\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} D_{KL}(\hat{\theta}||\theta) = \frac{1}{\hat{\theta}(1-\hat{\theta})}$$

Is called the Fisher information

Solving for log-loss

- The exponent in the integral is

$$g(\hat{\theta},\theta) - g(\hat{\theta},\hat{\theta}) = \hat{\theta} \ln \frac{\hat{\theta}}{\theta} + (1-\hat{\theta}) \ln \frac{1-\hat{\theta}}{1-\theta} = D_{KL}(\hat{\theta}||\theta)$$

- The second derivative

$$\left.\frac{d^2}{d\theta^2}\right|_{\theta=\hat{\theta}} D_{KL}(\hat{\theta}||\theta) = \frac{1}{\hat{\theta}(1-\hat{\theta})}$$

  Is called the Fisher information

- The optimal prior:

$$w^*(\hat{\theta}) = \frac{1}{\pi\sqrt{\hat{\theta}(1-\hat{\theta})}}$$

  Known in as Jeffrey's prior. And, in this case, the Dirichlet-$(1/2, 1/2)$ prior.

The regret of Bayes using Jeffrey's prior

- The regret (for coding: redundancy) is

$$L_A - L_{\min} \leq \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\frac{\pi}{2} + O(1/T)$$

But what is the prediction rule?

- As luck would have it the Dirichlet prior is the conjugate prior for the Binomial distribution.

But what is the prediction rule?

- As luck would have it the Dirichlet prior is the conjugate prior for the Binomial distribution.
- Observed t bits, n of which were 1. The posterior is:

$$\frac{1}{Z\sqrt{\theta(1-\theta)}}\theta^{n}(1-\theta)^{t-n} = \frac{1}{Z}\theta^{n-1/2}(1-\theta)^{t-n-1/2}$$

But what is the prediction rule?

- As luck would have it the Dirichlet prior is the conjugate prior for the Binomial distribution.
- Observed t bits, n of which were 1. The posterior is:

$$\frac{1}{Z\sqrt{\theta(1-\theta)}}\theta^n(1-\theta)^{t-n} = \frac{1}{Z}\theta^{n-1/2}(1-\theta)^{t-n-1/2}$$

- The posterior average is:

$$\frac{\int_0^1 \theta^{n+1/2}(1-\theta)^{t-n-1/2}d\theta}{\int_0^1 \theta^{n-1/2}(1-\theta)^{t-n-1/2}d\theta} = \frac{n+1/2}{t+1}$$

But what is the prediction rule?

- As luck would have it the Dirichlet prior is the conjugate prior for the Binomial distribution.
- Observed t bits, n of which were 1. The posterior is:

$$\frac{1}{Z\sqrt{\theta(1-\theta)}}\theta^n(1-\theta)^{t-n} = \frac{1}{Z}\theta^{n-1/2}(1-\theta)^{t-n-1/2}$$

- The posterior average is:

$$\frac{\int_0^1 \theta^{n+1/2}(1-\theta)^{t-n-1/2}d\theta}{\int_0^1 \theta^{n-1/2}(1-\theta)^{t-n-1/2}d\theta} = \frac{n+1/2}{t+1}$$

- This is called the Trichevsky Trofimov prediction rule.

Laplace Rule of Succession

- Laplace suggested using the uniform prior, which is also a conjugate prior.

Laplace Rule of Succession

- Laplace suggested using the uniform prior, which is also a conjugate prior.
- In this case the posterior average is:

$$\frac{\int_0^1 \theta^{n+1}(1-\theta)^{t-n} d\theta}{\int_0^1 \theta^{n}(1-\theta)^{t-n} d\theta} = \frac{n+1}{t+2}$$

Laplace Rule of Succession

- Laplace suggested using the uniform prior, which is also a conjugate prior.
- In this case the posterior average is:

$$\frac{\int_0^1 \theta^{n+1}(1-\theta)^{t-n} d\theta}{\int_0^1 \theta^{n}(1-\theta)^{t-n} d\theta} = \frac{n+1}{t+2}$$

- The bound on the cumulative log loss is worse:

$$L_A - L_{\min} = \ln T + O(1)$$

Laplace Rule of Succession

- Laplace suggested using the uniform prior, which is also a conjugate prior.
- In this case the posterior average is:

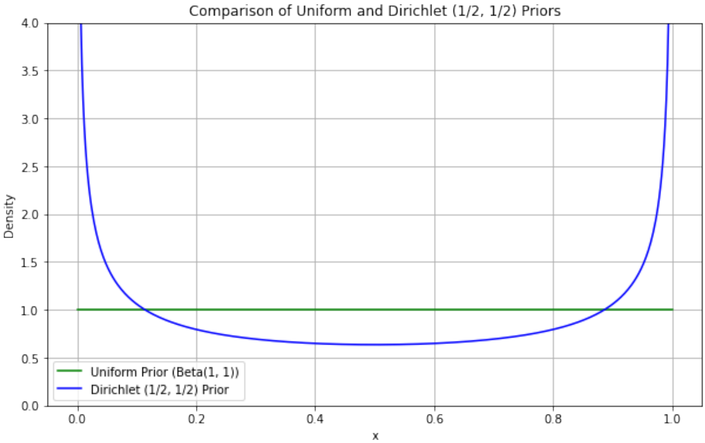
$$\frac{\int_0^1 \theta^{n+1}(1-\theta)^{t-n} d\theta}{\int_0^1 \theta^{n}(1-\theta)^{t-n} d\theta} = \frac{n+1}{t+2}$$

- The bound on the cumulative log loss is worse:

$$L_A - L_{\min} = \ln T + O(1)$$

- Suffers larger regret when $\hat{\theta}$ is far from $1/2$

Comparing the priors



Shtarkov's Normalized Maximal Likelihood (NML)

- What is the optimal prediction when T is know in advance?

Shtarkov's Normalized Maximal Likelihood (NML)

- What is the optimal prediction when T is know in advance?
- Equalize the reducndancy over all sequences.

Shtarkov's Normalized Maximal Likelihood (NML)

- What is the optimal prediction when T is know in advance?
- Equalize the reducndancy over all sequences.
- Assign to each **sequence** the **Maximal** likelihood:
  $P(x) = \frac{1}{Z} \max_p P_p(x)$

Shtarkov's Normalized Maximal Likelihood (NML)

- What is the optimal prediction when T is know in advance?
- Equalize the reducndancy over all sequences.
- Assign to each **sequence** the **Maximal** likelihood: $P(x) = \frac{1}{Z} \max_p P_p(x)$
- $\log Z$ is the regret (redundancy).

Shtarkov's Normalized Maximal Likelihood (NML)

- What is the optimal prediction when T is know in advance?
- Equalize the reducndancy over all sequences.
- Assign to each **sequence** the **Maximal** likelihood: $P(x) = \frac{1}{Z} \max_p P_p(x)$
- $\log Z$ is the regret (redundancy).
- For Bernoulli distributions:

$$L_*^T - \min_\theta L_\theta^T \geq \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\frac{\pi}{2} - O(\frac{1}{\sqrt{T}})$$

Shtarkov's Normalized Maximal Likelihood (NML)

- What is the optimal prediction when T is know in advance?
- Equalize the reducndancy over all sequences.
- Assign to each **sequence** the **Maximal** likelihood: $P(x) = \frac{1}{Z} \max_p P_p(x)$
- $\log Z$ is the regret (redundancy).
- For Bernoulli distributions:

$$L_*^T - \min_\theta L_\theta^T \geq \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\frac{\pi}{2} - O(\frac{1}{\sqrt{T}})$$

- The regret (for coding: redundancy) is

$$L_A - L_{\min} \leq \frac{1}{2}\ln(T+1) + \frac{1}{2}\ln\frac{\pi}{2} + O(1/T)$$

Generalization to larger sets of distributions

Multinomial Distributions

- For a distribution over k elements (Multinomial) [Xie and Barron]

Multinomial Distributions

- For a distribution over k elements (Multinomial) [Xie and Barron]
- Use the add 1/2 rule (KT).

$$p(i) = \frac{n_i + 1/2}{t + k/2}$$

Multinomial Distributions

- For a distribution over k elements (Multinomial) [Xie and Barron]
- Use the add 1/2 rule (KT).

$$p(i) = \frac{n_i + 1/2}{t + k/2}$$

- Bound is

$$L_A - L_{\min} \leq \frac{k-1}{2} \ln T + C + o(1)$$

Multinomial Distributions

- For a distribution over k elements (Multinomial) [Xie and Barron]
- Use the add 1/2 rule (KT).

$$p(i) = \frac{n_i + 1/2}{t + k/2}$$

- Bound is

$$L_A - L_{\min} \leq \frac{k-1}{2} \ln T + C + o(1)$$

- The constant C is optimal.

The Fisher Information Matrix

- $$\mathbf{I}(\theta) = \nabla^2_{\theta'} D_{\mathrm{KL}}(p(x;\theta) \| p(x;\theta'))\Big|_{\theta'=\theta}$$

The Fisher Information Matrix

- $$\mathbf{I}(\theta) = \nabla^2_{\theta'} D_{\mathrm{KL}}(p(x;\theta)\|p(x;\theta'))\Big|_{\theta'=\theta}$$

- $$\mathbf{I}(\theta) = \begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2} D_{\mathrm{KL}} & \frac{\partial^2}{\partial\theta_1\partial\theta_2} D_{\mathrm{KL}} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_n} D_{\mathrm{KL}} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} D_{\mathrm{KL}} & \frac{\partial^2}{\partial\theta_2^2} D_{\mathrm{KL}} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_n} D_{\mathrm{KL}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_n\partial\theta_1} D_{\mathrm{KL}} & \frac{\partial^2}{\partial\theta_n\partial\theta_2} D_{\mathrm{KL}} & \cdots & \frac{\partial^2}{\partial\theta_n^2} D_{\mathrm{KL}} \end{bmatrix}_{\theta'=\theta}$$

The Fisher Information Matrix

- $$\mathbf{I}(\theta) = \nabla^2_{\theta'} D_{\mathrm{KL}}(p(x;\theta) \| p(x;\theta'))\Big|_{\theta'=\theta}$$

- $$\mathbf{I}(\theta) = \begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2} D_{\mathrm{KL}} & \frac{\partial^2}{\partial\theta_1\partial\theta_2} D_{\mathrm{KL}} & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_n} D_{\mathrm{KL}} \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1} D_{\mathrm{KL}} & \frac{\partial^2}{\partial\theta_2^2} D_{\mathrm{KL}} & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_n} D_{\mathrm{KL}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_n\partial\theta_1} D_{\mathrm{KL}} & \frac{\partial^2}{\partial\theta_n\partial\theta_2} D_{\mathrm{KL}} & \cdots & \frac{\partial^2}{\partial\theta_n^2} D_{\mathrm{KL}} \end{bmatrix}_{\theta'=\theta}$$

- Jeffrey's prior
$$\pi_J(\boldsymbol{\theta}) \propto \sqrt{\det(\mathbf{I}(\boldsymbol{\theta}))}$$

Properties of Jeffrey's prior

- Known as "least informative prior" in Bayesian statistics.

Properties of Jeffrey's prior

- Known as "least informative prior" in Bayesian statistics.
- Min/max: Equalizes the risk for all parameter setting

Properties of Jeffrey's prior

- Known as "least informative prior" in Bayesian statistics.
- Min/max: Equalizes the risk for all parameter setting
- invariant under re-parametrization.

Properties of Jeffrey's prior

- Known as "least informative prior" in Bayesian statistics.
- Min/max: Equalizes the risk for all parameter setting
- invariant under re-parametrization.
- Often improper (integral = ∞).

Exponential Distributions

- The canonical form of an exponential distribution is

$$p(x|\theta) = \exp\left[\eta(\theta) \cdot T(x)\right]$$

for some fixed functions η, T

Exponential Distributions

- The canonical form of an exponential distribution is

$$p(x|\theta) = \exp\left[\eta(\theta) \cdot T(x)\right]$$

  for some fixed functions η, T

- Multinomial: $p(i| < p_1, \ldots, p_k >) = p_i$
  $\eta(\theta) =< \log p_1, \ldots, \log p_k >$
  $T : i \to (0, \ldots, 1, 0, \ldots, 0)$

Exponential Distributions

- The canonical form of an exponential distribution is

$$p(x|\theta) = \exp\left[\eta(\theta) \cdot T(x)\right]$$

  for some fixed functions η, T
- Multinomial: $p(i| < p_1, \ldots, p_k >) = p_i$
  $\eta(\theta) =< \log p_1, \ldots, \log p_k >$
  $T : i \rightarrow (0, \ldots, 1, 0, \ldots, 0)$
- Normal: $p(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$
  $\eta(\mu, \sigma^2) = \left[\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2, \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]$
  $T(y) = (1, y, y^2)$

Exponential Distributions

- The canonical form of an exponential distribution is

$$p(x|\theta) = \exp\left[\eta(\theta) \cdot T(x)\right]$$

  for some fixed functions η, T
- Multinomial: $p(i| < p_1, \ldots, p_k >) = p_i$
  $\eta(\theta) =< \log p_1, \ldots, \log p_k >$
  $T : i \to (0, \ldots, 1, 0, \ldots, 0)$
- Normal: $p(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)$
  $\eta(\mu, \sigma^2) = \left[\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2, \frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right]$
  $T(y) = (1, y, y^2)$
- Many more: 1D: Poisson, Exponential, Gamma ...
  Multi-Variate: Gaussian, Dirichlet, Multivariate t-distribution

Online learning for Exponential Families

- For any set of distributions from the exponential family defined by k parameters, the observations and the parameters come from compact sets.

Online learning for Exponential Families

- For any set of distributions from the exponential family defined by k parameters, the observations and the parameters come from compact sets.
- Use Bayes Algorithm with Jeffrey's prior:

$$w^*(\theta) = \frac{1}{Z}\sqrt{\det\left(I(\theta)\right)}$$

$I(\theta), Z$ do not depend on T.

Online learning for Exponential Families

- For any set of distributions from the exponential family defined by k parameters, the observations and the parameters come from compact sets.
- Use Bayes Algorithm with Jeffrey's prior:

$$w^*(\theta) = \frac{1}{Z}\sqrt{\det\left(I(\theta)\right)}$$

$I(\theta), Z$ do not depend on T.

- Redundancy bound

$$L_A - L_{\min} \leq \frac{k-1}{2} \ln T - \ln Z + o(1)$$

next Class

- Variable-length markov models - a set of distributions with increasing number of parameters.

next Class

- Variable-length markov models - a set of distributions with increasing number of parameters.
- The context algorithm: An efficient implementation of the Bayes algorithm which achieves close-to-optimal worst case bounds.