Topics in Prediction and Learning
Lecture 4:
Online Density Estimation

Peter Bartlett

Computer Science and Statistics
University of California at Berkeley

Mathematical Sciences
Queensland University of Technology

27 February–9 March, 2017
CREST, ENSAE

Online density estimation with log loss

Online Prediction as a Zero-Sum Game

Minimize *regret* wrt comparison $\mathcal{C}$:

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a}\in\mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t).$$

Online density estimation with log loss

Online Prediction as a Zero-Sum Game

Minimize *regret* wrt comparison $\mathcal{C}$:

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t).$$

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Online density estimation with log loss

Online Prediction as a Zero-Sum Game

Minimize *regret* wrt comparison $\mathcal{C}$:

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t).$$

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$,

Online density estimation with log loss

Online Prediction as a Zero-Sum Game

Minimize *regret* wrt comparison $\mathcal{C}$:

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a}\in\mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t).$$

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$,
where $p_\theta : \mathcal{Y}^n \to \mathbb{R}^+$ is a parameterized probability density with respect to the n-fold product of a fixed reference measure λ on $\mathcal{Y}$:

Online density estimation with log loss

Online Prediction as a Zero-Sum Game

Minimize *regret* wrt comparison $\mathcal{C}$:

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t).$$

Log loss

$$\ell(\hat{p}, y) = -\log \hat{p}(y).$$

Comparison class

Parametric family of densities: $\mathcal{C} = \{p_\theta : \theta \in \Theta\}$,
where $p_\theta : \mathcal{Y}^n \to \mathbb{R}^+$ is a parameterized probability density with respect to the n-fold product of a fixed reference measure λ on $\mathcal{Y}$:
For all $\theta \in \Theta$,

$$\int_{\mathcal{Y}^n} p_\theta(y_1, \ldots, y_n) \, d\lambda^n(y) = 1.$$

Online density estimation with log loss

Comparison class

For $p = p_\theta$ and $y \in \mathcal{Y}$, we write $p_t(y) = p(y|y_1, \ldots, y_{t-1})$. Thus,

$$\sum_{t=1}^{n} \log(p_t(y_t)) = \sum_{t=1}^{n} \log(p(y_t|y_1, \ldots, y_{t-1}) = \log(p(y_1^n)).$$

Online density estimation with log loss

Comparison class

For $p = p_\theta$ and $y \in \mathcal{Y}$, we write $p_t(y) = p(y|y_1, \ldots, y_{t-1})$. Thus,

$$\sum_{t=1}^{n} \log(p_t(y_t)) = \sum_{t=1}^{n} \log(p(y_t|y_1, \ldots, y_{t-1}) = \log(p(y_1^n)).$$

Regret

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t)$$

Online density estimation with log loss

Comparison class

For $p = p_\theta$ and $y \in \mathcal{Y}$, we write $p_t(y) = p(y|y_1, \dots, y_{t-1})$. Thus,

$$\sum_{t=1}^{n} \log(p_t(y_t)) = \sum_{t=1}^{n} \log(p(y_t|y_1, \dots, y_{t-1}) = \log(p(y_1^n)).$$

Regret

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t)$$

$$R(y_1^n, \hat{p}_1^n) = \sup_{p \in \mathcal{C}} \sum_{t=1}^{n} \log(p(y_t|y_1^{t-1})) - \sum_{t=1}^{n} \log(\hat{p}_t(y_t))$$

Online density estimation with log loss

Comparison class

For $p = p_\theta$ and $y \in \mathcal{Y}$, we write $p_t(y) = p(y|y_1, \ldots, y_{t-1})$. Thus,

$$\sum_{t=1}^{n} \log(p_t(y_t)) = \sum_{t=1}^{n} \log(p(y_t|y_1, \ldots, y_{t-1}) = \log(p(y_1^n)).$$

Regret

$$R(y_1^n, a_1^n) = \sum_{t=1}^{n} \ell(a_t, y_t) - \inf_{\hat{a} \in \mathcal{C}} \sum_{t=1}^{n} \ell(\hat{a}_t, y_t)$$

$$R(y_1^n, \hat{p}_1^n) = \sup_{p \in \mathcal{C}} \sum_{t=1}^{n} \log(p(y_t|y_1^{t-1})) - \sum_{t=1}^{n} \log(\hat{p}_t(y_t))$$

$$= \sup_{p \in \mathcal{C}} \log(p(y_1^n)) - \sum_{t=1}^{n} \log(\hat{p}_t(y_t)).$$

Online density estimation with log loss

Definition: Parametric constant model

Parametric family of i.i.d. densities on $\mathcal{Y}^n$:

Online density estimation with log loss

Definition: Parametric constant model

Parametric family of i.i.d. densities on $\mathcal{Y}^n$:
$\mathcal{C} = \{p_\theta^n : \theta \in \Theta\}$,
where p_θ^n is the n-fold product of the probability density $p_\theta : \mathcal{Y} \to \mathbb{R}^+$

Online density estimation with log loss

Definition: Parametric constant model

Parametric family of i.i.d. densities on $\mathcal{Y}^n$:
$\mathcal{C} = \{p_\theta^n : \theta \in \Theta\}$,
where p_θ^n is the n-fold product of the probability density $p_\theta : \mathcal{Y} \to \mathbb{R}^+$, which is a parameterized probability density with respect to the fixed reference measure λ on $\mathcal{Y}$:

$$\int_{\mathcal{Y}} p_\theta(y)\, d\lambda(y) = 1.$$

Online density estimation with log loss

Definition: Parametric constant model

Parametric family of i.i.d. densities on $\mathcal{Y}^n$:
$\mathcal{C} = \{p_\theta^n : \theta \in \Theta\}$,
where p_θ^n is the n-fold product of the probability density $p_\theta : \mathcal{Y} \to \mathbb{R}^+$, which is a parameterized probability density with respect to the fixed reference measure λ on $\mathcal{Y}$:

$$\int_{\mathcal{Y}} p_\theta(y)\, d\lambda(y) = 1.$$

For $p = p_\theta$ and $y_t \in \mathcal{Y}$, we have $p_t(y_t) = p(y_t | y_1, \ldots, y_{t-1}) = p(y)$.

Online density estimation with log loss

Definition: Parametric constant model

Parametric family of i.i.d. densities on $\mathcal{Y}^n$:
$\mathcal{C} = \{p_\theta^n : \theta \in \Theta\}$,
where p_θ^n is the n-fold product of the probability density $p_\theta : \mathcal{Y} \to \mathbb{R}^+$, which is a parameterized probability density with respect to the fixed reference measure λ on $\mathcal{Y}$:

$$\int_{\mathcal{Y}} p_\theta(y)\, d\lambda(y) = 1.$$

For $p = p_\theta$ and $y_t \in \mathcal{Y}$, we have $p_t(y_t) = p(y_t | y_1, \ldots, y_{t-1}) = p(y)$. Thus,

$$\sum_{t=1}^{n} \log(p_t(y_t)) = \sum_{t=1}^{n} \log(p(y_t)).$$

Online density estimation with log loss

Strategies are joint densities

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) =$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1) \cdots \hat{p}(y_n|y_1^{n-1}).$$

Online density estimation with log loss

Strategies are joint densities

- A strategy $\hat{p}$ is a mapping from histories $y_1^t = (y_1, \ldots, y_t)$ to densities $\hat{p}(\cdot|y_1^t)$ on $\mathcal{Y}$.
- Every strategy is a joint density:

$$\hat{p}(y_1, \ldots, y_n) = \hat{p}(y_1)\hat{p}(y_2|y_1) \cdots \hat{p}(y_n|y_1^{n-1}).$$

- Every joint density $\hat{p}$ is a strategy, $\hat{p}_{t+1}(\cdot) = \hat{p}(\cdot|y_1^t)$.

Online density estimation with log loss

Regret

We abuse notation, and write:

$$p_\theta(y_1^n) = \prod_{t=1}^{n} p_\theta(y_t).$$

Online density estimation with log loss

Regret

We abuse notation, and write:

$$p_\theta(y_1^n) = \prod_{t=1}^{n} p_\theta(y_t).$$

Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is a log likelihood ratio,

$$R(y_1^n, \hat{p}) = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t)$$

Online density estimation with log loss

Regret

We abuse notation, and write:

$$p_\theta(y_1^n) = \prod_{t=1}^{n} p_\theta(y_t).$$

Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is a log likelihood ratio,

$$\begin{aligned} R(y_1^n, \hat{p}) &= \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t) \\ &= \sup_{\theta \in \Theta} \log p_\theta(y_1^n) - \log \hat{p}(y_1^n) \end{aligned}$$

Online density estimation with log loss

Regret

We abuse notation, and write:

$$p_\theta(y_1^n) = \prod_{t=1}^{n} p_\theta(y_t).$$

Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is a log likelihood ratio, which is a difference of KL-divergences:

$$\begin{aligned} R(y_1^n, \hat{p}) &= \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t) \\ &= \sup_{\theta \in \Theta} \log p_\theta(y_1^n) - \log \hat{p}(y_1^n) \end{aligned}$$

Online density estimation with log loss

Regret

We abuse notation, and write:

$$p_\theta(y_1^n) = \prod_{t=1}^{n} p_\theta(y_t).$$

Regret wrt comparison $\mathcal{C} = \{p_\theta\}$ is a log likelihood ratio, which is a difference of KL-divergences:

$$\begin{aligned} R(y_1^n, \hat{p}) &= \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t) \\ &= \sup_{\theta \in \Theta} \log p_\theta(y_1^n) - \log \hat{p}(y_1^n) \\ &= nKL(P_n \| \hat{p}) - \inf_{\theta \in \Theta} nKL(P_n \| p_\theta), \end{aligned}$$

where P_n is the empirical distribution, with mass $1/n$ on $y_1, \ldots, y_n$, and $KL(P_n \| p)$ is the Kullback-Leibler divergence of P_n with respect to p.

Online density estimation with log loss

Many interpretations of prediction with log loss

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.
- Sequential lossless data compression ("minimum description length")

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.
- Sequential lossless data compression ("minimum description length")
- Repeated gambling/investment.

Online density estimation with log loss

Many interpretations of prediction with log loss

- Sequential probability prediction.
- Sequential lossless data compression ("minimum description length")
- Repeated gambling/investment.

Long history in several communities.

[Kelly, 1956], [Solomonoff, 1964], [Kolmogorov, 1965], [Cover, 1974], [Rissanen, 1976, 1987, 1996], [Shtarkov, 1987], [Feder, Merhav and Gutman, 1992], [Freund, 1996], [Xie and Barron, 2000], [Cesa-Bianchi and Lugosi, 2001, 2006], [Grünwald, 2007]

Outline

- Normalized maximum likelihood
- Multinomials
- SNML: predicting like there's no tomorrow
- Bayesian strategies
- Optimality $=$ exchangeability

Outline

- **Normalized maximum likelihood**
- Multinomials
- SNML: predicting like there's no tomorrow
- Bayesian strategies
- Optimality = exchangeability

Normalized maximum likelihood

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_1^n) = \frac{\sup_{\theta \in \Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n)}$$

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_1^n) = \frac{\sup_{\theta \in \Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n)}$$

Integrability

We require that the *Shtarkov integral*,

$$\int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n)$$

is finite.

Normalized maximum likelihood

Example

Consider the Gaussian family of densities on $\mathbb{R}$ ($\lambda =$ Lebesgue measure):

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right),$$

for $\mu \in \mathbb{R}$.

Normalized maximum likelihood

Example

Consider the Gaussian family of densities on $\mathbb{R}$ ($\lambda =$ Lebesgue measure):

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right),$$

for $\mu \in \mathbb{R}$. Then the Shtarkov integral for $n = 1$ is

$$\int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1)\, dz_1 =$$

Normalized maximum likelihood

Example

Consider the Gaussian family of densities on $\mathbb{R}$ ($\lambda =$ Lebesgue measure):

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right),$$

for $\mu \in \mathbb{R}$. Then the Shtarkov integral for $n = 1$ is

$$\int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1)\, dz_1 = \frac{1}{\sqrt{2\pi}} \int_{\mathcal{Y}} dz_1$$

Normalized maximum likelihood

Example

Consider the Gaussian family of densities on $\mathbb{R}$ ($\lambda =$ Lebesgue measure):

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right),$$

for $\mu \in \mathbb{R}$. Then the Shtarkov integral for $n = 1$ is

$$\int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1)\, dz_1 = \frac{1}{\sqrt{2\pi}} \int_{\mathcal{Y}} dz_1 = \infty.$$

Normalized maximum likelihood

Example

The Shtarkov integral for $n = 2$ is

$$\int_{\mathcal{Y}} \int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1^2)\, dz_1\, dz_2$$

Normalized maximum likelihood

Example

The Shtarkov integral for $n = 2$ is

$$\int_{\mathcal{Y}} \int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1^2) \, dz_1 \, dz_2$$
$$= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(-\frac{(z_1 - (z_1 + z_2)/2)^2 + (z_2 - (z_1 + z_2)/2)^2}{2} \right) dz_1 \, dz_2$$

Normalized maximum likelihood

Example

The Shtarkov integral for $n = 2$ is

$$\int_{\mathcal{Y}} \int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1^2)\, dz_1\, dz_2$$

$$= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(-\frac{(z_1 - (z_1 + z_2)/2)^2 + (z_2 - (z_1 + z_2)/2)^2}{2} \right) dz_1\, dz_2$$

$$= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(-\frac{(z_1 - z_2)^2}{4} \right) dz_1\, dz_2$$

Normalized maximum likelihood

Example

The Shtarkov integral for $n = 2$ is

$$\int_{\mathcal{Y}} \int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1^2) \, dz_1 \, dz_2$$

$$= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(- \frac{(z_1 - (z_1 + z_2)/2)^2 + (z_2 - (z_1 + z_2)/2)^2}{2} \right) \, dz_1 \, dz_2$$

$$= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(- \frac{(z_1 - z_2)^2}{4} \right) \, dz_1 \, dz_2$$

$$= \frac{\sqrt{2}}{\sqrt{\pi}} \int_{\mathcal{Y}} dz_1$$

Normalized maximum likelihood

Example

The Shtarkov integral for $n = 2$ is

$$\begin{aligned}
&\int_{\mathcal{Y}} \int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(z_1^2)\, dz_1\, dz_2 \\
&= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(-\frac{(z_1 - (z_1 + z_2)/2)^2 + (z_2 - (z_1 + z_2)/2)^2}{2} \right) dz_1\, dz_2 \\
&= \frac{1}{2\pi} \int_{\mathcal{Y}} \int_{\mathcal{Y}} \exp\left(-\frac{(z_1 - z_2)^2}{4} \right) dz_1\, dz_2 \\
&= \frac{\sqrt{2}}{\sqrt{\pi}} \int_{\mathcal{Y}} dz_1 \\
&= \infty.
\end{aligned}$$

Normalized maximum likelihood

Definition

Given an initial sequence $y_1^m \in \mathcal{Y}^m$, define the *conditional Shtarkov integral*

$$\int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m, y_{m+1}^n) \, d\lambda^{n-m}(y_{m+1}^n).$$

Normalized maximum likelihood

Definition

Given an initial sequence $y_1^m \in \mathcal{Y}^m$, define the *conditional Shtarkov integral*

$$\int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m, y_{m+1}^n)\, d\lambda^{n-m}(y_{m+1}^n).$$

Example

For the Gaussian family of densities on $\mathbb{R}$ and $y_1 \in \mathbb{R}$, the conditional Shtarkov integral for $n = 2$ is

$$\int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(y_1, y_2)\, dy_2 = \frac{1}{2\pi} \int_{\mathcal{Y}} \exp\left(-\frac{(y_2 - y_1)^2}{4}\right) dy_2$$

Normalized maximum likelihood

Definition

Given an initial sequence $y_1^m \in \mathcal{Y}^m$, define the *conditional Shtarkov integral*

$$\int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m, y_{m+1}^n)\, d\lambda^{n-m}(y_{m+1}^n).$$

Example

For the Gaussian family of densities on $\mathbb{R}$ and $y_1 \in \mathbb{R}$, the conditional Shtarkov integral for $n = 2$ is

$$\int_{\mathcal{Y}} \sup_{\theta \in \Theta} p_\theta(y_1, y_2)\, dy_2 = \frac{1}{2\pi} \int_{\mathcal{Y}} \exp\left(-\frac{(y_2 - y_1)^2}{4}\right) dy_2 = \sqrt{\frac{2}{\pi}}.$$

Conditional regret

Definition

Fix $y_1^m \in \mathcal{Y}$.
The conditional regret given y_1^m wrt the comparison class $\mathcal{C} = \{p_\theta\}$ is

Conditional regret

Definition

Fix $y_1^m \in \mathcal{Y}$.
The conditional regret given y_1^m wrt the comparison class $\mathcal{C} = \{p_\theta\}$ is

$$R(y_{m+1}^n, \hat{p} | y_1^m) = \sum_{t=m+1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p \in \mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t)$$

Conditional regret

Definition

Fix $y_1^m \in \mathcal{Y}$.
The conditional regret given y_1^m wrt the comparison class $\mathcal{C} = \{p_\theta\}$ is

$$R(y_{m+1}^n, \hat{p}|y_1^m) = \sum_{t=m+1}^{n} \ell(\hat{p}_t, y_t) - \inf_{p\in\mathcal{C}} \sum_{t=1}^{n} \ell(p, y_t)$$
$$= \sup_{\theta\in\Theta} \log p_\theta(y_1^n) - \log \hat{p}(y_{m+1}^n).$$

Conditional normalized maximum likelihood

Conditional normalized maximum likelihood

Conditional NML

Given $y_1^m \in \mathcal{Y}^m$,

$$p_{nml}^{(n)}(y_{m+1}^n | y_1^m) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

Conditional normalized maximum likelihood

Conditional NML

Given $y_1^m \in \mathcal{Y}^m$,

$$p_{nml}^{(n)}(y_{m+1}^n | y_1^m) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_{m+1}^n | y_1^m) = \frac{\sup_{\theta \in \Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m, z_{m+1}^n)\, d\lambda^{n-m}(z_{m+1}^n)}$$

Conditional normalized maximum likelihood

Conditional NML

Given $y_1^m \in \mathcal{Y}^m$,

$$p_{nml}^{(n)}(y_{m+1}^n | y_1^m) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_{m+1}^n | y_1^m) = \frac{\sup_{\theta \in \Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m, z_{m+1}^n) \, d\lambda^{n-m}(z_{m+1}^n)}$$

Integrability

We require that the conditional Shtarkov integral given y_1^m is finite, that is,

$$\int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m, z_{m+1}^n) \, d\lambda^{n-m}(z_{m+1}^n) < \infty.$$

Normalized maximum likelihood

NML is optimal [Shtarkov, 1987]

Fix $n > 0$ and suppose that the Shtarkov integral is finite, so that NML is well defined.

Normalized maximum likelihood

NML is optimal [Shtarkov, 1987]

Fix $n > 0$ and suppose that the Shtarkov integral is finite, so that NML is well defined.

1. NML equalizes regret: for any y_1^n,

$$R(y_1^n, p_{nml}^{(n)}) = \log \int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n).$$

Normalized maximum likelihood

NML is optimal [Shtarkov, 1987]

Fix $n > 0$ and suppose that the Shtarkov integral is finite, so that NML is well defined.

1. NML equalizes regret: for any y_1^n,

$$R(y_1^n, p_{nml}^{(n)}) = \log \int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n).$$

2. Any strategy $\hat{p}$ that predicts differently from NML has strictly worse maximum regret.

Normalized maximum likelihood

NML is optimal [Shtarkov, 1987]

Fix $n > 0$ and suppose that the Shtarkov integral is finite, so that NML is well defined.

1. NML equalizes regret: for any y_1^n,

$$R(y_1^n, p_{nml}^{(n)}) = \log \int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n).$$

2. Any strategy $\hat{p}$ that predicts differently from NML has strictly worse maximum regret.
3. Thus, NML is the minimax optimal strategy:

$$\min_{\hat{p}} \max_{y_1^n} R(y_1^n, \hat{p}) = R(y_1^n, p_{nml}^{(n)}).$$

Normalized maximum likelihood

The regret,

$$\log \int_{\mathcal{Y}^n} \sup_{\theta \in \Theta} p_\theta(z_1^n) \, d\lambda^n(z_1^n)$$

is often called the *stochastic complexity* of $\{p_\theta : \theta \in \Theta\}$.

Conditional normalized maximum likelihood

Conditional NML is optimal

Fix $y_1^m \in \mathcal{Y}^m$ and $n > m$. Suppose that the conditional Shtarkov integral given y_1^m is finite, so that conditional NML is well defined.

1. Conditional NML equalizes conditional regret: for any y_{m+1}^n,

$$R(y_{m+1}^n, p_{nml}^{(n)}|y_1^m) = \log \int_{\mathcal{Y}^{n-m}} \sup_{\theta\in\Theta} p_\theta(y_1^m z_{m+1}^n)\, d\lambda^{n-m}(z_{m+1}^n).$$

2. Any conditional strategy $\hat{p}$ that predicts differently from conditional NML has strictly worse maximum conditional regret.
3. Thus, conditional NML is the minimax optimal strategy:

$$\min_{\hat{p}} \max_{y_{m+1}^n} R(y_{m+1}^n, \hat{p}|y_1^m) = R(y_{m+1}^n, p_{nml}^{(n)}|y_1^m).$$

Conditional normalized maximum likelihood

Call the regret,

$$\log \int_{\mathcal{Y}^{n-m}} \sup_{\theta \in \Theta} p_\theta(y_1^m z_{m+1}^n)\, d\lambda^{n-m}(z_{m+1}^n)$$

the *conditional stochastic complexity* of $\{p_\theta : \theta \in \Theta\}$, given y_1^m.

Optimality of NML

Proof

First, NML is an equalizer:

Optimality of NML

Proof

First, NML is an equalizer:

$$R(y_1^n, p_{nml}^{(n)})$$

Optimality of NML

Proof

First, NML is an equalizer:

$$R(y_1^n, p_{nml}^{(n)}) = \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(p_{nml}^{(n)}(y_1^n)\right)$$

Optimality of NML

Proof

First, NML is an equalizer:

$$R(y_1^n, p_{nml}^{(n)}) = \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(p_{nml}^{(n)}(y_1^n)\right)$$

$$= \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(\frac{\sup_{\theta\in\Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^n} \sup_{\theta\in\Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n)}\right)$$

Optimality of NML

Proof

First, NML is an equalizer:

$$
\begin{aligned}
R(y_1^n, p_{nml}^{(n)}) &= \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(p_{nml}^{(n)}(y_1^n)\right) \\
&= \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(\frac{\sup_{\theta\in\Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^n} \sup_{\theta\in\Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n)}\right) \\
&= \log \int_{\mathcal{Y}^n} \sup_{\theta\in\Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n),
\end{aligned}
$$

Optimality of NML

Proof

First, NML is an equalizer:

$$
\begin{aligned}
R(y_1^n, p_{nml}^{(n)}) &= \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(p_{nml}^{(n)}(y_1^n)\right) \\
&= \log\left(\sup_{\theta\in\Theta} p_\theta(y_1^n)\right) - \log\left(\frac{\sup_{\theta\in\Theta} p_\theta(y_1^n)}{\int_{\mathcal{Y}^n} \sup_{\theta\in\Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n)}\right) \\
&= \log\int_{\mathcal{Y}^n} \sup_{\theta\in\Theta} p_\theta(z_1^n)\, d\lambda^n(z_1^n),
\end{aligned}
$$

which is independent of y_1^n.

Optimality of NML

Proof

Second, for any other strategy, $\hat{p} \neq p_{nml}^{(n)}$, there is a sequence y_1^n with $\hat{p}(y_1^n) < p_{nml}^{(n)}(y_1^n)$.

Optimality of NML

Proof

Second, for any other strategy, $\hat{p} \neq p_{nml}^{(n)}$, there is a sequence y_1^n with $\hat{p}(y_1^n) < p_{nml}^{(n)}(y_1^n)$.

For this sequence,

$$R(y_1^n, \hat{p}) > R(y_1^n, p_{nml}^{(n)}).$$

Optimality of NML

Proof

Second, for any other strategy, $\hat{p} \neq p_{nml}^{(n)}$, there is a sequence y_1^n with $\hat{p}(y_1^n) < p_{nml}^{(n)}(y_1^n)$.

For this sequence,

$$R(y_1^n, \hat{p}) > R(y_1^n, p_{nml}^{(n)}).$$

So NML is the minimax optimal strategy.

Computing Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

Computing Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

- To predict, we compute conditional distributions, marginalize.

Computing Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1 \cdots y_{t-1}) = \frac{p_{nml}^{(n)}(y_1^t)}{p_{nml}^{(n)}(y_1^{t-1})}$$

- To predict, we compute conditional distributions, marginalize.

Computing Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1 \cdots y_{t-1}) = \frac{p_{nml}^{(n)}(y_1^t)}{p_{nml}^{(n)}(y_1^{t-1})}$$

$$= \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n) \, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n) \, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.

Computing Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta\in\Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t|y_1 \cdots y_{t-1}) = \frac{p_{nml}^{(n)}(y_1^t)}{p_{nml}^{(n)}(y_1^{t-1})}$$

$$= \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta\in\Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta\in\Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!

Computing Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p^{(n)}_{nml}(y_t | y_1 \cdots y_{t-1}) = \frac{p^{(n)}_{nml}(y_1^t)}{p^{(n)}_{nml}(y_1^{t-1})}$$

$$= \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!
- When can we compute it cheaply?

Computing Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p^{(n)}_{nml}(y_t | y_1 \cdots y_{t-1}) = \frac{p^{(n)}_{nml}(y_1^t)}{p^{(n)}_{nml}(y_1^{t-1})}$$

$$= \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n) \, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n) \, d\lambda^{n-t+1}(z_t^n)}$$

- To predict, we compute conditional distributions, marginalize.
- All that conditioning is computationally expensive!
- When can we compute it cheaply?
- Multinomials.

[Kontkanen, Buntine, Myllymäki, Rissanen, Tirri, 2003]

Outline

- Normalized maximum likelihood
- **Multinomials**
- SNML: predicting like there's no tomorrow
- Bayesian strategies
- Optimality = exchangeability

Computing NML

Example

Consider $y \in \{1, \ldots, K\}$ and

$$p_\theta(y) = \theta_y, \qquad \theta \in \Delta^K.$$

Computing NML

Example

Consider $y \in \{1, \dots, K\}$ and

$$p_\theta(y) = \theta_y, \qquad \theta \in \Delta^K.$$

Then

$$p^{(n)}_{nml}(y_1^n) = \frac{\max_\theta p_\theta(y_1^n)}{\log \sum_{z_1^n} \max_\theta p_\theta(z_1^n)}.$$

Computing NML

Example

Consider $y \in \{1, \ldots, K\}$ and

$$p_\theta(y) = \theta_y, \qquad \theta \in \Delta^K.$$

Then

$$p_{nml}^{(n)}(y_1^n) = \frac{\max_\theta p_\theta(y_1^n)}{\log \sum_{z_1^n} \max_\theta p_\theta(z_1^n)}.$$

How do we compute the denominator (the stochastic complexity)?

Computing NML

Example

Consider $y \in \{1, \dots, K\}$ and

$$p_\theta(y) = \theta_y, \qquad \theta \in \Delta^K.$$

Then

$$p^{(n)}_{nml}(y_1^n) = \frac{\max_\theta p_\theta(y_1^n)}{\log \sum_{z_1^n} \max_\theta p_\theta(z_1^n)}.$$

How do we compute the denominator (the stochastic complexity)?
(The sums required to compute $p^{(n)}_{nml}(y_t | y_1 \cdots y_{t-1})$ are similar.)

Computing NML

Example

For $y_1^n \in \{1, \ldots, K\}^n$, define $h \in \{0, \ldots, n\}^K$ by

$$h_v = \sum_{t=1}^{n} 1[y_t = v].$$

Define the maximum likelihood estimator $\hat{\theta}(y_1^n) = h/n$.

Computing NML

Example

For $y_1^n \in \{1, \ldots, K\}^n$, define $h \in \{0, \ldots, n\}^K$ by

$$h_v = \sum_{t=1}^{n} 1[y_t = v].$$

Define the maximum likelihood estimator $\hat{\theta}(y_1^n) = h/n$. Then

$$\max_{\theta} p_\theta(y_1^n)$$

Computing NML

Example

For $y_1^n \in \{1, \ldots, K\}^n$, define $h \in \{0, \ldots, n\}^K$ by

$$h_v = \sum_{t=1}^{n} 1[y_t = v].$$

Define the maximum likelihood estimator $\hat{\theta}(y_1^n) = h/n$. Then

$$\max_{\theta} p_\theta(y_1^n) = \prod_{t=1}^{n} p_{\hat{\theta}(y_1^n)}(y_t)$$

Computing NML

Example

For $y_1^n \in \{1, \ldots, K\}^n$, define $h \in \{0, \ldots, n\}^K$ by

$$h_v = \sum_{t=1}^{n} 1[y_t = v].$$

Define the maximum likelihood estimator $\hat{\theta}(y_1^n) = h/n$. Then

$$\max_{\theta} p_\theta(y_1^n) = \prod_{t=1}^{n} p_{\hat{\theta}(y_1^n)}(y_t) = \prod_{v=1}^{K} \hat{\theta}_v^{h_v}$$

Computing NML

Example

For $y_1^n \in \{1, \ldots, K\}^n$, define $h \in \{0, \ldots, n\}^K$ by

$$h_v = \sum_{t=1}^{n} 1[y_t = v].$$

Define the maximum likelihood estimator $\hat{\theta}(y_1^n) = h/n$. Then

$$\max_{\theta} p_\theta(y_1^n) = \prod_{t=1}^{n} p_{\hat{\theta}(y_1^n)}(y_t) = \prod_{v=1}^{K} \hat{\theta}_v^{h_v} = \prod_{v=1}^{K} \left(\frac{h_v}{n}\right)^{h_v}.$$

Computing NML

Example

We can write

$$P_{K,n} := \sum_{z_1^n} p_{\hat{\theta}(z_1^n)}(z_1^n)$$

[Kontkanen, Buntine, Myllymäki, Rissanen, Tirri, 2003]

[Kontkanen, Myllymäki, 2005]

Computing NML

Example

We can write

$$P_{K,n} := \sum_{z_1^n} p_{\hat{\theta}(z_1^n)}(z_1^n) = \sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1!\cdots h_K!} \prod_{v=1}^{K} \left(\frac{h_v}{n}\right)^{h_v} .$$

[Kontkanen, Buntine, Myllymäki, Rissanen, Tirri, 2003]

[Kontkanen, Myllymäki, 2005]

Computing NML

Example

We can write

$$P_{K,n} := \sum_{z_1^n} p_{\hat{\theta}(z_1^n)}(z_1^n) = \sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1!\cdots h_K!} \prod_{v=1}^{K} \left(\frac{h_v}{n}\right)^{h_v}.$$

But we can split this sum: for any $k_1 + k_2 = K$,

$$P_{K,n} = \sum_{h_1+h_2=n} \frac{n!}{h_1!h_2!} \left(\frac{h_1}{n}\right)^{h_1} \left(\frac{h_2}{n}\right)^{h_2} P_{k_1,h_1} P_{k_2,h_2}.$$

[Kontkanen, Buntine, Myllymäki, Rissanen, Tirri, 2003]

[Kontkanen, Myllymäki, 2005]

Computing NML

Example

We can write

$$P_{K,n} := \sum_{z_1^n} p_{\hat{\theta}(z_1^n)}(z_1^n) = \sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1!\cdots h_K!} \prod_{v=1}^{K} \left(\frac{h_v}{n}\right)^{h_v}.$$

But we can split this sum: for any $k_1 + k_2 = K$,

$$P_{K,n} = \sum_{h_1+h_2=n} \frac{n!}{h_1!h_2!} \left(\frac{h_1}{n}\right)^{h_1} \left(\frac{h_2}{n}\right)^{h_2} P_{k_1,h_1} P_{k_2,h_2}.$$

So we can build up a table of these values, with a suitable geometric sequence of k_1s and all values of h_1, to compute $P_{K,n}$ in $O(n^2 \log K)$ time.

[Kontkanen, Buntine, Myllymäki, Rissanen, Tirri, 2003]

[Kontkanen, Myllymäki, 2005]

Computing NML

Example

We can write

$$P_{K,n} := \sum_{z_1^n} p_{\hat{\theta}(z_1^n)}(z_1^n) = \sum_{h_1+\cdots+h_K=n} \frac{n!}{h_1!\cdots h_K!} \prod_{v=1}^{K} \left(\frac{h_v}{n}\right)^{h_v}.$$

But we can split this sum: for any $k_1 + k_2 = K$,

$$P_{K,n} = \sum_{h_1+h_2=n} \frac{n!}{h_1!h_2!} \left(\frac{h_1}{n}\right)^{h_1} \left(\frac{h_2}{n}\right)^{h_2} P_{k_1,h_1} P_{k_2,h_2}.$$

So we can build up a table of these values, with a suitable geometric sequence of k_1s and all values of h_1, to compute $P_{K,n}$ in $O(n^2 \log K)$ time.

[Kontkanen, Buntine, Myllymäki, Rissanen, Tirri, 2003]

Also conditional multinomial models on $\{1, \ldots, K\}^d$.

[Kontkanen, Myllymäki, 2005]

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1 \cdots y_{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n) \, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n) \, d\lambda^{n-t+1}(z_t^n)}$$

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1 \cdots y_{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- Computationally cheaper strategies:

Normalized maximum likelihood

NML

$$p_{nml}^{(n)}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p_{nml}^{(n)}(y_t | y_1 \cdots y_{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- Computationally cheaper strategies:
  - Horizon-independent NML ("Sequential NML")

Normalized maximum likelihood

NML

$$p^{(n)}_{nml}(y_1 \cdots y_n) \propto \sup_{\theta \in \Theta} p_\theta(y_1^n)$$

$$p^{(n)}_{nml}(y_t | y_1 \cdots y_{t-1}) = \frac{\int_{\mathcal{Y}^{n-t}} \sup_{\theta \in \Theta} p_\theta(y_1^t z_{t+1}^n)\, d\lambda^{n-t}(z_{t+1}^n)}{\int_{\mathcal{Y}^{n-t+1}} \sup_{\theta \in \Theta} p_\theta(y_1^{t-1} z_t^n)\, d\lambda^{n-t+1}(z_t^n)}$$

- Computationally cheaper strategies:
  - Horizon-independent NML ("Sequential NML")
  - Bayesian prediction

Outline

- Normalized maximum likelihood.
- Multinomials
- **SNML: predicting like there's no tomorrow.**
- Bayesian strategies.
- Optimality = exchangeability.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

- Pretend that this is the last prediction we'll ever make.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) := p_{nml}^{(t)}(y_t|y_1^{t-1})$$

- Pretend that this is the last prediction we'll ever make.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) := p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

- Pretend that this is the last prediction we'll ever make.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) := p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

- Pretend that this is the last prediction we'll ever make.
- Simpler conditional calculation.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) := p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

- Pretend that this is the last prediction we'll ever make.
- Simpler conditional calculation.
- Has asymptotically optimal regret.

[Roos and Rissanen, 2008], [Kotłowski and Grünwald, 2011]

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

Theorem

SNML is optimal iff p_{snml} is exchangeable.

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta\in\Theta} p_\theta(y_1^t)$$

Theorem

SNML is optimal iff p_{snml} is exchangeable.

[Hedayati and B., 2016]

- p_{snml} is exchangeable means:
  for any n, any y_1^n, and any permutation σ on $\{1,\ldots,n\}$,
  $p_{snml}(y_1,\ldots,y_n) = p_{snml}(y_{\sigma(1)},\ldots,y_{\sigma(n)})$.

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

1. SNML's regret doesn't depend on the last observation.

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

1. SNML's regret doesn't depend on the last observation.

$$R(y_1^n, p_{snml})$$

Predicting like there's no tomorrow: Sequential NML

Proof (⇐)

1. SNML's regret doesn't depend on the last observation.

$$R(y_1^n, p_{snml}) = \log \frac{p_{\hat{\theta}}(y_1^n)}{p_{snml}(y_1^n)}, \qquad (\hat{\theta} \text{ is maximum likelihood})$$

Proof ($\Leftarrow$)

1. SNML's regret doesn't depend on the last observation.

$$R(y_1^n, p_{snml}) = \log \frac{p_{\hat{\theta}}(y_1^n)}{p_{snml}(y_1^n)}, \qquad (\hat{\theta} \text{ is maximum likelihood})$$

$$p_{snml}(y_1^n) = p_{snml}(y_n|y_1^{n-1})p_{snml}(y_1^{n-1})$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

1. SNML's regret doesn't depend on the last observation.

$$R(y_1^n, p_{snml}) = \log \frac{p_{\hat{\theta}}(y_1^n)}{p_{snml}(y_1^n)}, \qquad (\hat{\theta} \text{ is maximum likelihood})$$

$$p_{snml}(y_1^n) = p_{snml}(y_n | y_1^{n-1}) p_{snml}(y_1^{n-1})$$

$$= \frac{p_{\hat{\theta}}(y_1^n)}{\int_{\mathcal{Y}} \sup_\theta p_\theta(y_1^{n-1}, z)\, d\lambda(z)} p_{snml}(y_1^{n-1}),$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

1. SNML's regret doesn't depend on the last observation.

$$R(y_1^n, p_{snml}) = \log \frac{p_{\hat{\theta}}(y_1^n)}{p_{snml}(y_1^n)}, \qquad (\hat{\theta} \text{ is maximum likelihood})$$

$$p_{snml}(y_1^n) = p_{snml}(y_n | y_1^{n-1}) p_{snml}(y_1^{n-1})$$

$$= \frac{p_{\hat{\theta}}(y_1^n)}{\int_{\mathcal{Y}} \sup_\theta p_\theta(y_1^{n-1}, z)\, d\lambda(z)} p_{snml}(y_1^{n-1}),$$

so

$$R(y_1^n, p_{snml}) = \log \frac{p_{snml}(y_1^{n-1})}{\int_{\mathcal{Y}} \sup_\theta p_\theta(y_1^{n-1}, z)\, d\lambda(z)}.$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

2. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^n p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

3. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^n p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

In that case, SNML's regret is independent of observations:

$$R(y_1, \ldots, y_{n-1}, y_n; p_{snml})$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

4. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^n p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

In that case, SNML's regret is independent of observations:

$$R(y_1, \ldots, y_{n-1}, y_n; p_{snml}) = R(y_1, \ldots, y_{n-1}, \tilde{y}_1; p_{snml})$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

5. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^n p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

In that case, SNML's regret is independent of observations:

$$\begin{aligned} R(y_1, \ldots, y_{n-1}, y_n; p_{snml}) &= R(y_1, \ldots, y_{n-1}, \tilde{y}_1; p_{snml}) \\ &= R(\tilde{y}_1, \ldots, y_{n-1}, y_1; p_{snml}) \end{aligned}$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

6. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^{n} p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

In that case, SNML's regret is independent of observations:

$$\begin{aligned} R(y_1, \ldots, y_{n-1}, y_n; p_{snml}) &= R(y_1, \ldots, y_{n-1}, \tilde{y}_1; p_{snml}) \\ &= R(\tilde{y}_1, \ldots, y_{n-1}, y_1; p_{snml}) \\ &\vdots \\ &= R(\tilde{y}_1, \ldots, \tilde{y}_{n-1}, \tilde{y}_n; p_{snml}). \end{aligned}$$

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

7. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^n p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

In that case, SNML's regret is independent of observations:

$$\begin{aligned} R(y_1, \ldots, y_{n-1}, y_n; p_{snml}) &= R(y_1, \ldots, y_{n-1}, \tilde{y}_1; p_{snml}) \\ &= R(\tilde{y}_1, \ldots, y_{n-1}, y_1; p_{snml}) \\ &\vdots \\ &= R(\tilde{y}_1, \ldots, \tilde{y}_{n-1}, \tilde{y}_n; p_{snml}). \end{aligned}$$

So if SNML is exchangeable, then it is an equalizer,

Predicting like there's no tomorrow: Sequential NML

Proof ($\Leftarrow$)

8. If SNML is exchangeable, then its regret is permutation-invariant:

$$R(y_1^n, p_{snml}) = \log \frac{\prod_{t=1}^n p_{\hat{\theta}}(y_t)}{p_{snml}(y_1^n)}.$$

In that case, SNML's regret is independent of observations:

$$\begin{aligned} R(y_1, \dots, y_{n-1}, y_n; p_{snml}) &= R(y_1, \dots, y_{n-1}, \tilde{y}_1; p_{snml}) \\ &= R(\tilde{y}_1, \dots, y_{n-1}, y_1; p_{snml}) \\ &\vdots \\ &= R(\tilde{y}_1, \dots, \tilde{y}_{n-1}, \tilde{y}_n; p_{snml}). \end{aligned}$$

So if SNML is exchangeable, then it is an equalizer, and so it is the same as NML.

Predicting like there's no tomorrow: Sequential NML

Proof ($\Rightarrow$)

Predicting like there's no tomorrow: Sequential NML

Proof ($\Rightarrow$)

1. $p_{nml}^{(n)}(y_1^n)$ is permutation-invariant:

$$p_{nml}^{(n)}(y_1^n) \propto \sup_{\theta \in \Theta} \prod_{t=1}^{n} p_\theta(y_t).$$

Predicting like there's no tomorrow: Sequential NML

Sequential Normalized Maximum Likelihood (SNML)

$$p_{snml}(y_t|y_1^{t-1}) = p_{nml}^{(t)}(y_t|y_1^{t-1}) \propto \sup_{\theta \in \Theta} p_\theta(y_1^t)$$

Theorem

SNML is optimal iff p_{snml} is exchangeable.

Outline

- Normalized maximum likelihood.
- Multinomials
- SNML: predicting like there's no tomorrow.
- **Bayesian strategies.**
- Optimality = exchangeability.

Bayesian strategies

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(y_1^t) = \int_{\theta \in \Theta} p_\theta(y_1^t)\, d\pi(\theta)$$

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(y_1^t) = \int_{\theta \in \Theta} p_\theta(y_1^t)\, d\pi(\theta)$$

- Sequential update to prior.

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(y_1^t) = \int_{\theta\in\Theta} p_\theta(y_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta|y_1^t) \propto p_\pi(\theta|y_1^{t-1}) p_\theta(y_t).$$

- Sequential update to prior.

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(y_1^t) = \int_{\theta\in\Theta} p_\theta(y_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta|y_1^t) \propto p_\pi(\theta|y_1^{t-1})p_\theta(y_t).$$

- Sequential update to prior.
- Consider Jeffreys prior:

$$\pi(\theta) \propto \sqrt{|I(\theta)|},$$

$$I(\theta) = \mathrm{Cov}\left(\nabla_\theta \ln p_\theta(X)\right). \qquad (X \sim p_\theta)$$

Bayesian strategies

For prior π on Θ:

$$p_\pi(y_1^t) = \int_{\theta\in\Theta} p_\theta(y_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta|y_1^t) \propto p_\pi(\theta|y_1^{t-1}) p_\theta(y_t).$$

- Sequential update to prior.
- Consider Jeffreys prior:

$$\pi(\theta) \propto \sqrt{|I(\theta)|},$$

$$I(\theta) = \mathrm{Cov}\left(\nabla_\theta \ln p_\theta(X)\right). \qquad (X \sim p_\theta)$$

- Attractive properties (e.g., invariant to parameterization).

Bayesian strategies

Bayesian strategies

For prior π on Θ:

$$p_\pi(y_1^t) = \int_{\theta \in \Theta} p_\theta(y_1^t)\, d\pi(\theta)$$

$$p_\pi(\theta | y_1^t) \propto p_\pi(\theta | y_1^{t-1}) p_\theta(y_t).$$

- Sequential update to prior.
- Consider Jeffreys prior:

$$\pi(\theta) \propto \sqrt{|I(\theta)|},$$

$$I(\theta) = \mathrm{Cov}\left(\nabla_\theta \ln p_\theta(X)\right). \qquad (X \sim p_\theta)$$

- Attractive properties (e.g., invariant to parameterization).
- Asymptotically optimal regret for exponential families.

[Clarke and Barron, 1990, 1994]

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.
- Why?

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.
- Why? If NML=SNML, then we can consider long time horizons, so the asymptotics emerge.

Sequential NML and Bayesian strategies

Optimality

[Hedayati and B., 2016]

For regular p_θ (asymptotically normal maximum likelihood estimator, Fisher information well-behaved, integrals exist), the following are equivalent:

1. NML = SNML.
2. p_{snml} exchangeable.
3. NML = Bayesian.
4. NML = Bayesian with Jeffreys prior.
5. SNML = Bayesian.
6. SNML = Bayesian with Jeffreys prior.

- If we can ignore the time horizon and be optimal, that's the same as Bayesian prediction with Jeffreys prior.
- If any Bayesian strategy is optimal, it uses Jeffreys prior.
- Why? If NML=SNML, then we can consider long time horizons, so the asymptotics emerge. Asymptotic normality of the MLE implies Jeffreys prior is the only candidate.

Online density estimation with log loss

Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

Online density estimation with log loss

Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:
$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$
- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$

Online density estimation with log loss

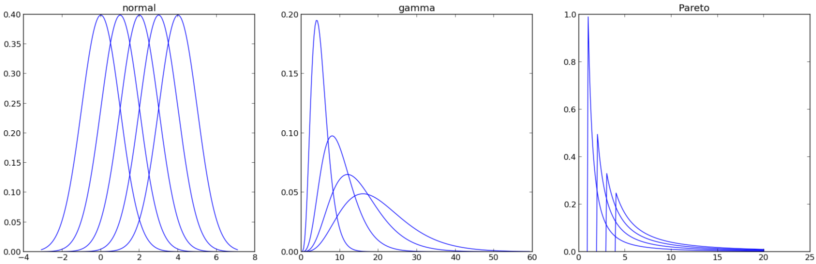
Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,



Online density estimation with log loss

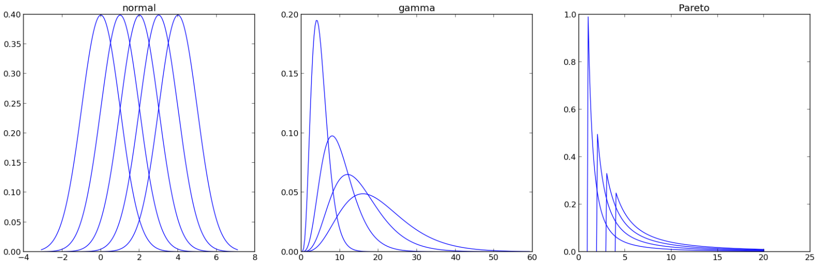
Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y)\exp\left(\theta y - A(\theta)\right).$$

- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,
  2. gamma distributions with fixed shape $k > 0$,



Online density estimation with log loss

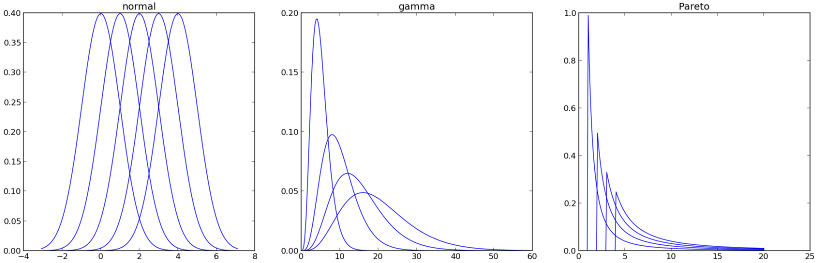
Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,
  2. gamma distributions with fixed shape $k > 0$,
  3. Tweedie exponential family of order $3/2$,



Online density estimation with log loss

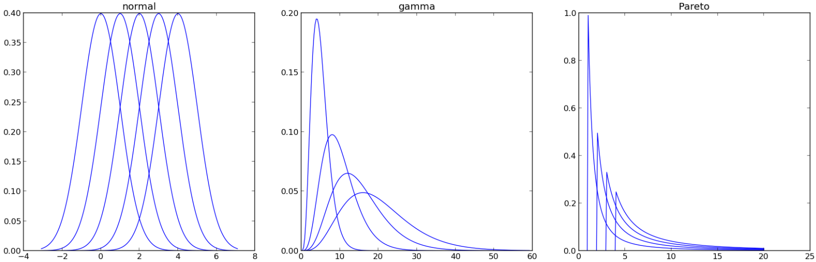
Examples

[B., Grünwald, Harremoës, Hedayati, Kotłowski, 2013]

- One-dimensional exponential families:

$$p_\theta(y) = h(y) \exp\left(\theta y - A(\theta)\right).$$

- p_{SNML} is exchangeable (i.e., SNML optimal, Bayesian optimal) $\Leftrightarrow$
  1. Gaussian distributions with fixed variance $\sigma^2 > 0$,
  2. gamma distributions with fixed shape $k > 0$,
  3. Tweedie exponential family of order $3/2$,
  4. Or smooth transformations.



Outline

- Normalized maximum likelihood.
- Multinomials
- SNML: predicting like there's no tomorrow.
- Bayesian strategies.
- Optimality = exchangeability.