MDLの効率的計算のための事前分布設定

豊橋技術科学大学　情報・知能工学系　渡辺一帆

概要

- 正規化最尤符号（NML）の計算：符号長，予測分布

NML

$$p_{NML}^{(n)}(x^n)$$

計算困難

Bayes

$$\int p(x^n \mid \theta)\underline{q(\theta)}d\theta$$

効率的に計算可
（近似，別表現）

□ 漸近的ミニマックスにする（精度良く近似する）にはどうしたらいいか？ [Watanabe, Roos, Myllymäki, ACML2013]
[Watanabe & Roos, J. Mach. Learn. Res.]

□ 別表現を用いて効率的計算が可能か？

[Barron, Roos, Watanabe, ISIT2014]

ユニバーサル符号化

- 系列の予測

$$x_1, \ldots, x_{i-1}, x_i, \ldots, x_n$$

$$\left\{ p(x^n \mid \theta); \theta \in \Theta \right\}$$:学習モデル

$$p(x^n) = \prod_{i=1}^{n} \underline{p(x_i \mid x^{i-1})}$$

予測分布

$$-\log p(x^n)$$:累積誤差（符号長）

$$-\log p(x^n \mid \underline{\hat{\theta}(x^n)})$$:理想的な符号長(x^nだけ符号化するなら)

最尤推定量

$$\boxed{\min_{\bar{p}} \max_{x^n} \log \frac{p(x^n \mid \hat{\theta}(x^n))}{\bar{p}(x^n)}}$$

最大リグレット最小化

どんなデータ系列であっても

ミニマックス解は

正規化最尤法(NML)

- 正規化最尤符号 [Shtarkov, 1987]

$$\boxed{p_{NML}^{(n)}(x^n) = \frac{p(x^n \mid \hat{\theta}(x^n))}{C_n}} \qquad C_n = \sum_{x^n} p(x^n \mid \hat{\theta}(x^n))$$

- いくつかのモデルでは計算可
- モデル選択への応用:
  多項モデル, マルコフモデル, ベイジアンネット …
- 非漸近(少サンプル)での有効性

[Kontkanen & Myllymäki, 2007]
[Silander, Roos, & Myllymäki, 2010]
[本セッションでの他の講演と参考文献]

正規化最尤法(2)

- 確率的コンプレキシティ(NMLの符号長)

$$-\log p_{NML}^{(n)}(x^n) = -\log p(x^n \mid \hat{\theta}(x^n)) + \log C_n$$

- パラメトリックコンプレキシティ(NMLのリグレット)

$$\log \frac{p(x^n \mid \hat{\theta}(x^n))}{p_{NML}^{(n)}(x^n)} = \log C_n$$　(すべての x^n で)

- 予測分布(逐次符号化で必要　例:算術符号)

$$p_{NML}^{(n)}(x_i \mid x^{i-1}) = \frac{p_{NML}^{(n)}(x^i)}{\sum_{x_i} p_{NML}^{(n)}(x^i)}$$

□ NMLの予測分布の計算はnの指数オーダー

ベイズ符号

$$p_B(x^n) = \int p(x^n \mid \theta)\underline{q(\theta)}d\theta$$ ：ベイズ混合(エビデンス)

事前分布

- 予測分布

$$p_B(x_{i+1} \mid x^i) = \frac{\int p(x_{i+1} \mid \theta)p(x^i \mid \theta)q(\theta)d\theta}{\int p(x^i \mid \theta)q(\theta)d\theta}$$

$$= \int p(x_{i+1} \mid \theta)\underline{q(\theta \mid x^i)}d\theta$$

$$q(\theta \mid x^i) = \frac{p(x_i \mid \theta)q(\theta \mid x^{i-1})}{p_B(x_i \mid x^{i-1})}$$ 事後分布

□ ベイズ符号の予測分布は逐次的に計算可

漸近的ミニマックス性

- NMLとの差がo(1)にできるか？　　g : 計算しやすい分布

$$\log \frac{p(x^n \mid \hat{\theta}(x^n))}{g(x^n)} \le \log C_n + o(1)$$　（すべての x^n で）

- サンプル数（ホライゾン）依存性
  - 「nに依存しない」 ⟺ $g(x^{\tilde{n}}) = \sum_{x^n_{\tilde{n}+1}} g(x^n) \quad (\tilde{n} < n)$
  - 例えば、$g(x^n) = \int p(x^n \mid \theta) \underbrace{q(\theta)}_{\text{事前分布}} d\theta$

$q(\theta)$ をnに応じて設計 ➡ 漸近的ミニマックス
[Xie & Barron, 2000]
[Takeuchi & Barron, 1997]

多項モデル

- 有限アルファベット上の定常無記憶情報源

$$p(x^n \mid \theta) = \prod_{i=1}^{n} \theta_{x_i} \qquad \theta = (\theta_1, \theta_2, \ldots, \theta_m) \quad \sum_{j=1}^{m} \theta_j = 1$$

$x_i \in \{1, 2, \ldots, m\}$

- 共役事前分布（ディリクレ分布） $q(\theta) = \dfrac{\Gamma(m\alpha)}{\Gamma(\alpha)^m} \prod_{j=1}^{m} \theta_j^{\alpha-1}$
- ベイズ混合による予測

$$p(x_{l+1} = j \mid x^l) = \int p(x_{l+1} = j \mid \theta) \underbrace{q(\theta \mid x^l)}_{\text{ディリクレ分布}} d\theta$$

$$= \frac{l_j + \alpha}{l + m\alpha}$$

l_j ： x^l 中のjの数

α=1: Laplace

α=1/2: Krichevski-Trofimov

漸近的ミニマックス性とサンプル数依存性

- 定理 [Watanabe & Roos, J. Mach. Learn. Res.]:

多項モデルでは

漸近的ミニマックス ⇨ サンプル数nの知識が必要

（ベイズ混合ではq(θ)がnに依存）

証明の概略）

α=1/2のベイズ混合（ジェフリーズ混合）は、n→∞でNMLに十分に近くなるが漸近的ミニマックスではない。
あるサンプル数と系列におけるジェフリーズ混合との差を考えると、nに依存しない場合、（少なくとも一つの）その系列の延長では、差が残り続ける。

一般のモデルの場合

- 仮定　　　ある $x^{\tilde{n}}$ に対し、　　　　　　例: $\tilde{n} = \sqrt{n}$

$$\log \frac{p_{NML}^{(\tilde{n})}(x^{\tilde{n}})}{\sum_{x_{\tilde{n}+1}^{n}} p_{NML}^{(n)}(x^{n})} \to M > 0 \quad (n \to \infty)$$

意味: p_{NML} はnとともに変化する（その度合いがM）

- 定理 [Watanabe, Roos, Myllymäki, ACML2013]

仮定の下, g がnに依存しない

（すべての x^n で）

$$\log C_n - \underline{M} + o(1) \leq \log \frac{p(x^n \mid \hat{\theta}(x^n))}{g(x^n)} \leq \log C_n + o(1)$$

とはならない

$\underline{M}(< M)$ は任意の定数

多項モデルでのNML の変動（仮定のM）

- $$\log \frac{p_{NML}^{(\tilde{n})}(x^{\tilde{n}})}{\sum_{x_{\tilde{n}+1}^{n}} p_{NML}^{(n)}(x^{n})} = \sum_{j=1}^{m} \left\{ l_j \log l_j - \log\Gamma(l_j + 1/2) - l_j + \frac{\log 2\pi}{2} \right\} + o(1)$$

  l_j : $x^{\tilde{n}}$ 中のjの数

  - $\displaystyle\sum_{n_j \geq l_j} \binom{n-\tilde{n}}{n_j - l_j} \prod_{j=1}^{m} \left(\frac{n_j}{n}\right)^{n_j}$ の評価が必要

$$\boxed{M = \frac{m-1}{2}\log 2}$$ ととれる。

$l_1 = \tilde{n}$

$l_j = 0 \ (j = 2, \ldots, m)$

単純なn依存性と漸近的ミニマックス性

- $\alpha_n^* = \arg\min_{\alpha} \max_{x^n} \log \dfrac{p(x^n \mid \hat{\theta}(x^n))}{p_{B,\alpha}(x^n)}$ （数値的に計算）

- 漸近的な近似

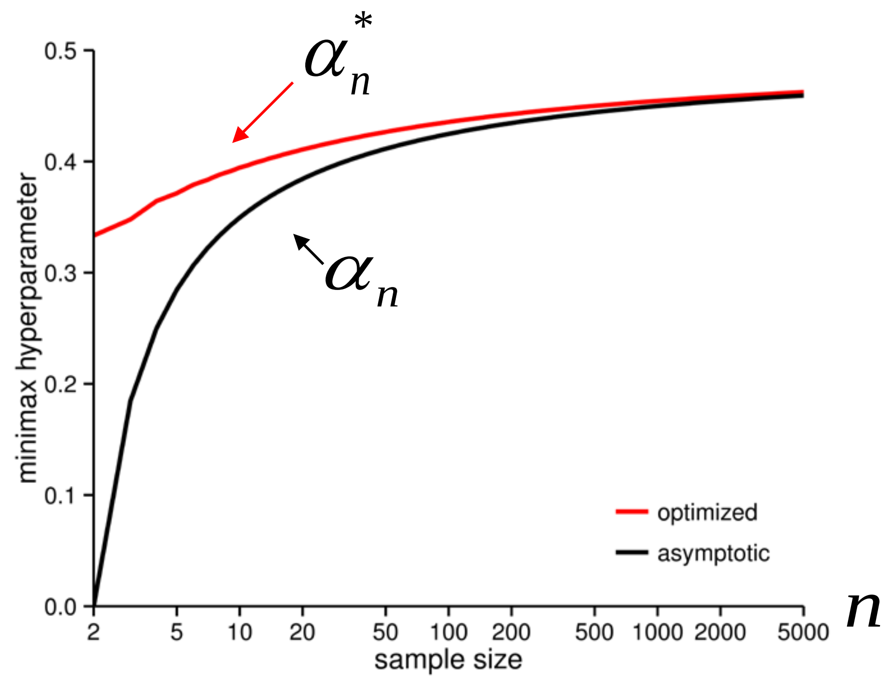
$$\boxed{\alpha_n = \frac{1}{2} - \frac{\log 2}{2} \frac{1}{\log n}}$$

導出:

○ $n_j = n$ (一つのjで)

○ $n_j = n / m$

⇒ リグレットを等しくする。



単純なn依存性と漸近的ミニマックス性(2)

- 定理 [Watanabe, Roos, Myllymäki, ACML2013]

α_n を用いたベイズ混合は漸近的ミニマックス。

しかも，

$$\log C_n - M + o(1) \le \log \frac{p(x^n \mid \hat{\theta}(x^n))}{p_{B,\alpha_n}(x^n)} \le \log C_n + o(1)$$

（すべての x^n で） となる。

$$\boxed{M = \frac{m-1}{2}\log 2}$$

証明の概略：スターリングの公式を適用し，定数オーダーの項を評価。

- 強い意味での漸近的ミニマックス性をほぼ達成する。

数値実験(ベルヌーイモデル：m=2)

- nに依存(ベイズ混合)
  - p_{B,α_n^*} p_{B,α_n}
  - 修正ジェフリーズ [Xie & Barron, 2000]

ジェフリーズ事前分布

$$q_{MJ}^{(n)}(\theta) = \frac{\varepsilon_n}{2}\{\delta(\theta - 1/n) + \delta(\theta - 1 + 1/n)\} + (1 - \varepsilon_n)b_{1/2}(\theta)$$

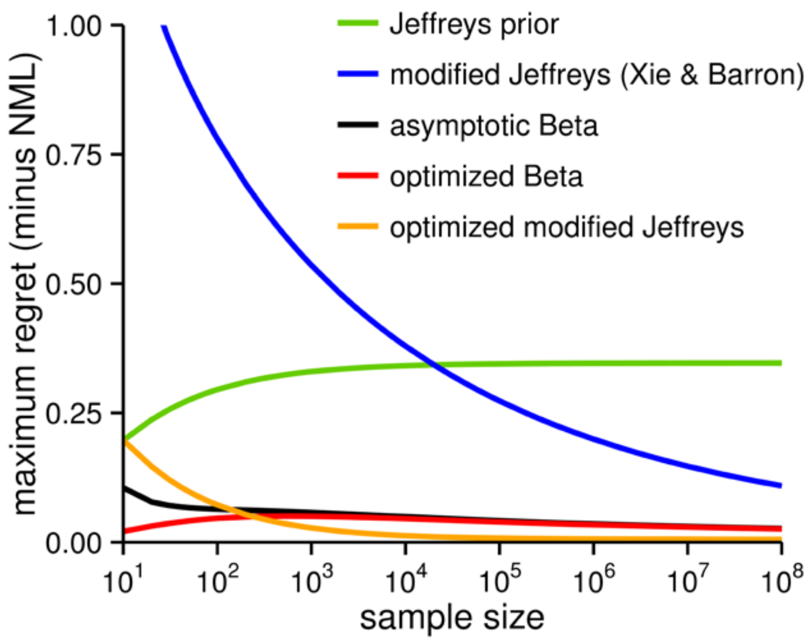
$\varepsilon_n = n^{-1/8}$ ε_n :最大リグレット最小化

- nに依存しない
  - $p_{B,1/2}$ （ジェフリーズ事前分布）
  - k-Last-step minimax (Sequential NML) [Takimoto & Warmuth, 2000] [Rissanen & Roos, 2007]

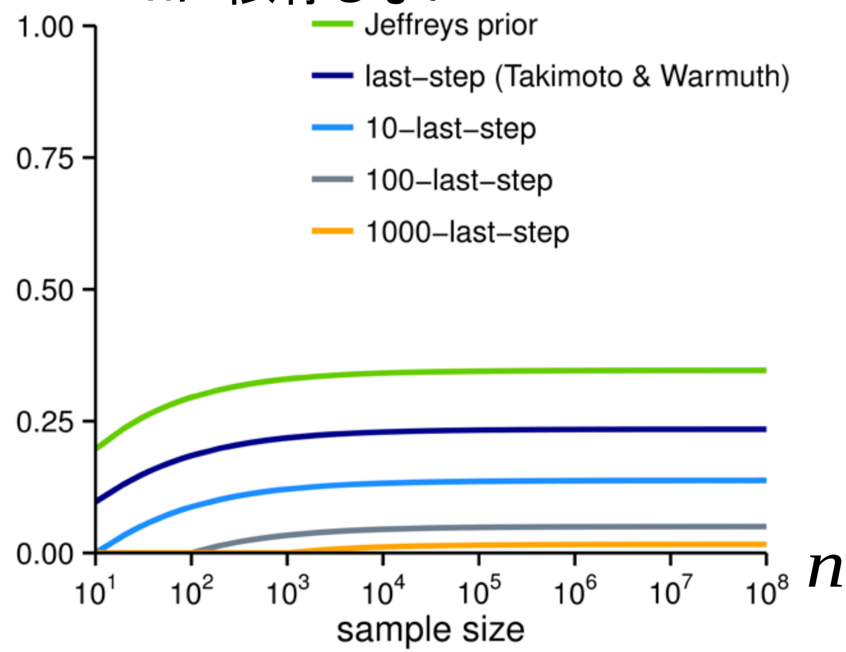
$$p_{kLS}(x_{t-k+1}^t \mid x^{t-k}) \propto p(x^t \mid \hat{\theta}(x^t))$$ （k個分を正規化）

数値実験(m=2)

- nに依存(ベイズ混合)



- nに依存しない



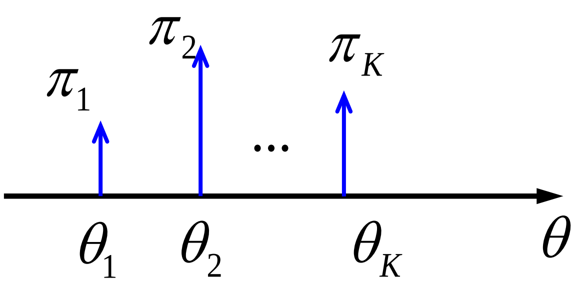
- 少サンプル数での有効性
- 漸近的ミニマックス(0に収束) ⇒ n依存性が必要

離散混合によるNML

- 事前分布を離散分布とする

$$q(\theta) = \sum_{j=1}^{K} \pi_j \delta(\theta - \theta_j)$$

$$p_B(x^n) = \sum_{j=1}^{K} \pi_j p(x^n \mid \theta_j)$$



- 定理 [Barron, Roos, Watanabe, ISIT2014]

$T(x^n)$ ：十分統計量（K通りの値をとる）

$\Theta_K = \{\theta_1, \ldots, \theta_K\}$ に対し、 $T(x^n)$ の分布が線形独立

$\Rightarrow$ $p_{NML}^{(n)}(x^n) = \sum_{j=1}^{K} \pi_j p(x^n \mid \theta_j)$ と表現可

- π_j は負でもよい

負の重みが必要な例

- $X \in \{1, 2, 3\}$

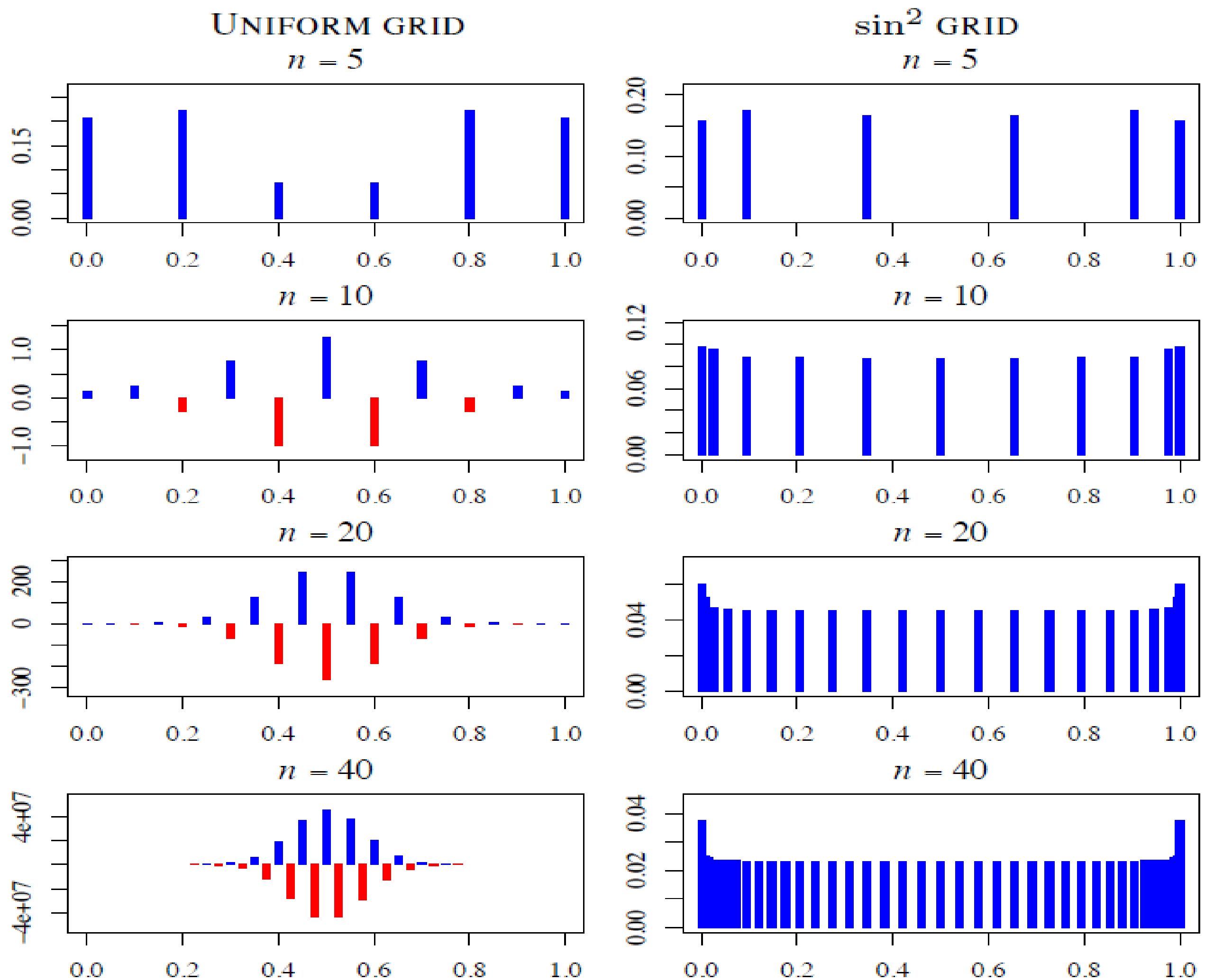
モデル：
$$p_1 = \left(\frac{1}{2}, \frac{1}{2}, 0\right) \quad p_2 = \left(0, \frac{1}{2}, \frac{1}{2}\right) \quad p_3 = \left(\frac{2}{7}, \frac{3}{7}, \frac{2}{7}\right)$$

$$\max_{\theta} p(x \mid \theta) = \frac{1}{2} \qquad (x = 1, 2, 3)$$

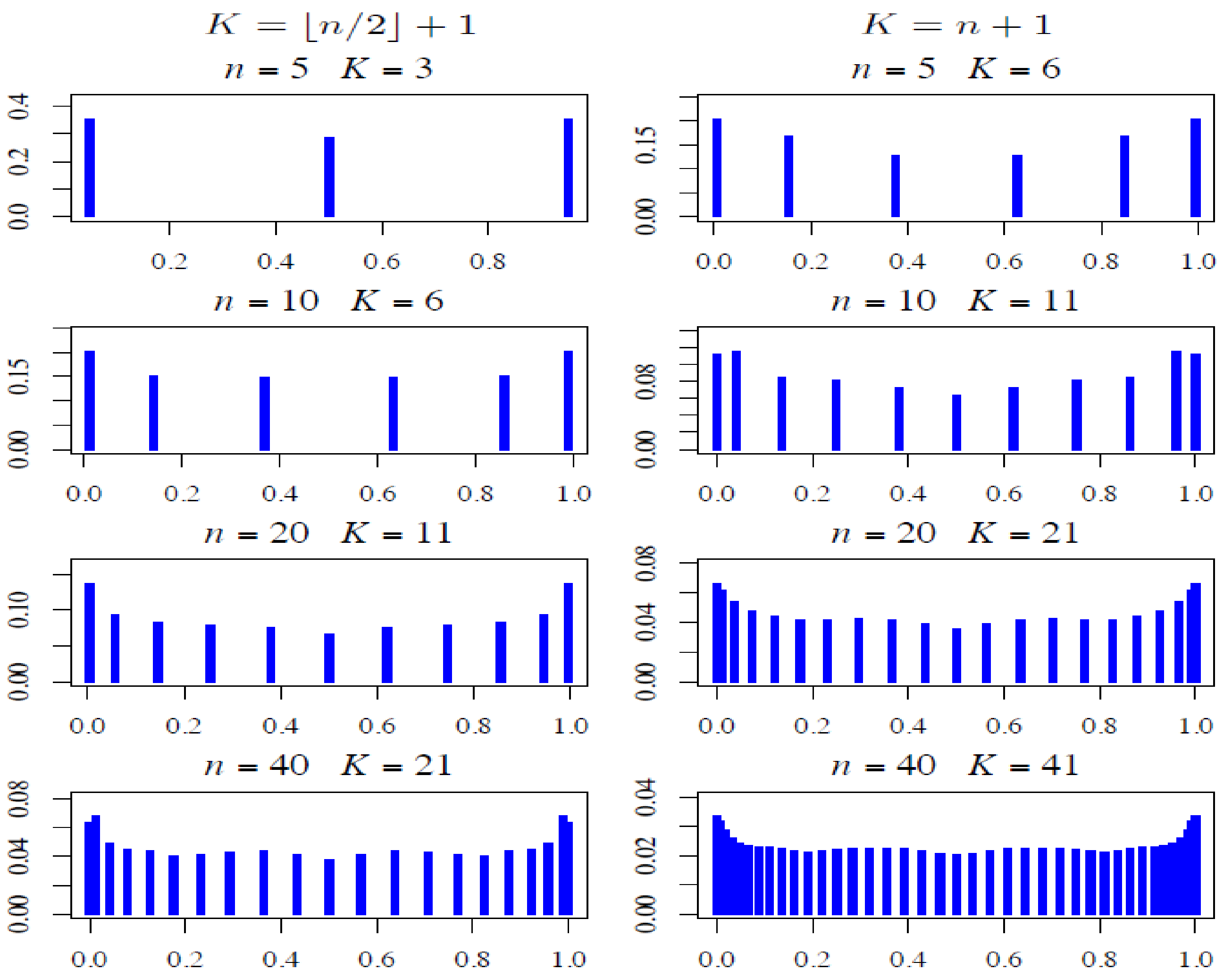
$$p_{NML} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

$$= -\frac{2}{3} p_1 - \frac{2}{3} p_2 + \frac{7}{3} p_3 \qquad \text{（一意）}$$

ベルヌーイモデル：サポートを固定



ベルヌーイモデル：サポートを最適化



まとめ

- 正規化最尤法の効率的な計算について
  - 漸近的ミニマックス性とサンプル数依存性の関係
  - 離散混合を用いた正確計算法

を紹介した。