The 34th Symposium on Information Theory
and its Applications (SITA2011)
Ousyuku, Iwate, Japan, Nov. 29–Dec. 2, 2011

予測のための min-max 問題
A min-max problem for prediction

池田思朗*
Shiro Ikeda

Abstract— We study a basic statistical problem to predict a new draw from a distribution, when i.i.d. samples are available. We consider a min-max problem and show that the optimal prior. The proposed method is practical for the prediction.

Keywords—predictive distribution, channel capacity, Bayesian statistics.

1 はじめに

本稿では簡単な予測の問題を考える．X は $p(x|\theta)$ という分布にしたがっているがパラメータ θ は未知だとする．もし θ に関して事前分布 $\pi(\theta)$ が得られているなら，$\pi(\theta)$ から構成される予測分布が Kullback-Leibler (KL) 情報量を損失とするベイズリスクを最小にすることが知られている [1]．

ベイズリスクは，事前分布のもとで損失の期待値をとったものである．本稿では，ベイズリスクのように期待値を基準にするのではなく，最悪評価をもとに最適な分布を求める問題を考える．

KL 情報量を損失とした min-max 問題は情報理論における通信路容量 [2] の定義と双対の関係にあることが知られている [3]．このときの最適なパラメータの分布はベイズ統計では reference 事前分布と呼ばれ，以前から研究されている [4, 5]．本稿ではこのような研究の背景を示し，新たな提案をおこなう．

2 Reference 事前分布と通信路容量

本節では reference 事前分布について説明する．Reference 事前分布の導出法は共役事前分布，Jefferyes の事前分布ほどは広く知られていないが，通信路容量と関係している．次節で提案する min-max の問題とも関係がある．

X を確率変数とし，それがしたがう確率分布を $p(x|\theta)$ とする．$p(x|\theta)$ はある族 $\mathcal{P}$ に属し $\theta \in \Theta$ は未知である．以下では簡単のため θ は 1 次元のパラメータとし，$\Theta = \{\theta | a \leq \theta \leq b\}$ とする．また，[3] にあるように $\mathcal{P}$ に属する $p(x|\theta)$ は全ての x および θ に対して正であり，有界だと仮定する．

$$\mathcal{P} = \{p = p(x|\theta) \mid x \in \mathcal{X},\ \theta \in \Theta\}.$$

* 〒 190-8562 東京都立川市緑町 10-3 統計数理研究所 The Institute of Statistical Mathematics, 10-4 Midoricho, Tachikawa-shi, Tokyo, 190-8562 Japan. E-mail: shiro@ism.ac.jp

以下では θ の分布は π であらわし，$\mathbb{E}_{\pi(\theta)}[\cdot]$ は $\pi(\theta)$ によって平均をとることを表す．$\pi(\theta)$ が連続分布のときには，

$$\mathbb{E}_{\pi(\theta)}[f(\theta)] = \int_{\Theta} \pi(\theta) f(\theta) d\theta,$$

また，離散分布のときには以下のようになる．

$$\mathbb{E}_{\pi(\theta)}[f(\theta)] = \sum_i \pi(\theta_i) f(\theta_i).$$

KL 情報量 $D(\cdot\|\cdot)$ と相互情報量 $I(\cdot;\cdot)$ の定義は次の通りである．

$$D(p_{X|\theta}\|q_X) = \int_{\mathcal{X}} p(x|\theta) \log \frac{p(x|\theta)}{q(x)} d\mu_X$$
$$I(X;\theta) = \mathbb{E}_{\pi(\theta)}[D(p_{X|\theta}\|p_{X;\pi})].$$

μ_X は X の測度である．また，$p_{X;\pi}$ は次の周辺分布 $p(x;\pi)$ を示す．

$$p(x;\pi) = \mathbb{E}_{\pi(\theta)}[p(x;\theta)]. \tag{1}$$

まず，Haussler によって示された結果を説明する [3]．次の 3 つの問題を考える．

問題 1. *[3]*

$$\inf_{q\in\mathcal{A}(X)} \sup_{\theta\in\Theta} D(p_{X|\theta}\|q_X), \tag{2}$$

$$\sup_{\pi(\theta)} \inf_{q\in\mathcal{A}(X)} \mathbb{E}_{\pi(\theta)}\big[D(p_{X|\theta}\|q_X)\big], \tag{3}$$

$$\sup_{\pi(\theta)} \mathbb{E}_{\pi(\theta)}\big[D(p_{X|\theta}\|p_{X;\pi})\big] = \sup_{\pi(\theta)} I(X;\theta). \tag{4}$$

ここで $\mathcal{A}(X)$ は X の任意の分布である．式 (2) は KL 情報量を損失とするゲームとして定式化できる [6]．[6] の定式化にしたがえば Nature が $\theta \in \Theta$ を選び，Statistician が X の任意の分布 $\mathcal{A}(X)$ から最適な $q(x)$ を選ぶ．Statistician にとっては最悪の損失を最小にするゲームである．

一方，式 (4) は，θ を入力，X を出力，$p(x|\theta)$ を通信路とする通信路容量の定義である．この問題の最適な分布は，通信路容量を達成する入力 θ の分布である．

このような双対な問題については Grünwald & Dawid [6] に詳しく述べられている．ここでは KL 情報

Copyright © 2011 IEICE

量の損失について知られている結果を示したが，Bregman 情報量を損失としても同様の結果を示すことができる [6].

Haussler は適当な仮定の下，3 つの問題が同じ最適値を持つことを示した．また，式 (2) の最適値を与える $q(x)$ は $q(x) = p(x;\pi)$ の形をしており，その $\pi(\theta)$ は式 (3) と式 (4) の最適値を与えることを示した．

式 (4) の最大値を与える $\pi(\theta)$ はベイズ統計では reference 事前分布として知られている．統計の問題はパラメータと変数の間の情報を扱うとも言えるため，その 2 つの間の相互情報量を最大にする事前分布は無情報事前分布として適当なものと考えることができる [4, 5].

ただし，通信路容量を達成する分布の多くが離散分布となることからわかるように [2, 7, 8]，reference 事前分布も最適な分布が離散分布となることが多いことが知られている [9].

3 予測のための min-max 問題

3.1 予測分布

Reference 事前分布は主に無情報事前分布として考えられたが，ここでは N 個のサンプル $x_1^N = \{x_1, \cdots, x_N\}$ が得られたときに，新たなデータ $x \sim p(x|\theta)$ を予測することを考える．

事前分布 $\pi(\theta)$ が与えられているとすると，パラメータの事後分布は以下のようになる．

$$\pi(\theta|x_1^N) = \frac{p(x_1^N|\theta)\pi(\theta)}{p(x_1^N;\pi)}, \quad p(x_1^N;\pi) = \mathbb{E}_{\pi(\theta)}\big[p(x_1^N|\theta)\big].$$

ベイズ統計の立場からは，新たな X の予測ためには次の予測分布を用いるのが自然である．

$$p(x|x_1^N;\pi) = \mathbb{E}_{\pi(\theta|x_1^N)}[p(x|\theta)],$$

予測分布は次のベイズリスクを最も小さくする．

問題 2 (ベイズリスク最小化 [1]).

$$\min_{q\in\mathcal{A}(X)} \mathbb{E}_{\pi(\theta|x_1^N)}\big[D(p_{X|\theta}\|q_X)\big],$$

Proof. $p(x;\pi)$ の最適性は以下のように示せる [1],

$$\begin{aligned}\mathbb{E}_{\pi(\theta|x_1^N)}\big[D(p_{X|\theta}\|q_X)\big] =& \mathbb{E}_{\pi(\theta|x_1^N)}\big[D(p_{X|\theta}\|p_{X|x_1^N;\pi})\big] \\ &+ D(p_{X|x_1^N;\pi}\|q_X), \end{aligned} \tag{5}$$

$p_{X|x_1^N;\pi}$ は $p(x|x_1^N;\pi)$ である．KL 情報量は非負であり，$q(x) = p(x|x_1^N;\pi)$ のとき最適である． □

予測のためには上の予測分布を用いるが自然である．しかし，予測分布を求めるためには積分が必要である．この積分は解析的に求まらないことも多い．

3.2 予測に関する min-max 問題

Reference 事前分布の問題は KL 情報量を損失とする min-max 問題と同値であった．ここでは次の損失の min-max 問題を考える．

$$f(\theta)\ D(p_{X|\theta}\|q_X)$$

上式は KL 情報量に $f(\theta)$ でパラメータに関する重みをつけたものである．この損失は適当な Bregman 情報量 $D_B(\cdot||\cdot)$ を用いて $D_B(p_{X|\theta}||q_X)$ とは書けない．以下では特に $f(\theta)$ として $p(x_1^N|\theta)$ を選ぶ．この重みは θ の関数としてみると，最尤推定値のまわりに集中した関数で，最尤推定値から外れると 0 に近づくと考えられる．したがって，最尤推定値のまわりでの min-max 問題を考えていることになり，予測のための損失として適当だろう．

問題 3 (予測のための min-max 問題).

$$\inf_{q\in\mathcal{A}(X)} \sup_{\theta\in\Theta}\ p(x_1^N|\theta)\ D(p_{X|\theta}\|q_X).$$

上記の問題は Grünwald & Dawid の論文の枠組で考えれば，損失を KL 情報量から重みをつけた KL 情報量に変えたことになっている．

Haussler の結果と同様に次の 2 つの問題を考えることができる．

問題 4.

$$\sup_{\pi(\theta)} \inf_{q\in\mathcal{A}(X)}\ \mathbb{E}_{\pi(\theta)}[p(x_1^N|\theta)\ D(p_{X|\theta}\|q_X)]. \tag{6}$$

$$\sup_{\pi(\theta)}\ \mathbb{E}_{\pi(\theta)}\big[p(x_1^N|\theta)D(p_{X|\theta}\|p_{X|x_1^N;\pi})\big]. \tag{7}$$

ここで $p_{X|x_1^N;\pi}$ は $p(x|x_1^N;\pi)$ を表している．

式 (6) と式 (7) の最適値が同じであることは式 (5) から次のように示せる．

$$\begin{aligned}&\mathbb{E}_{\pi(\theta)}[p(x_1^N|\theta)\ D(p_{X|\theta}\|q_X)] \\ =&p(x_1^N;\pi)\ \mathbb{E}_{\pi(\theta|x_1^N)}[D(p_{X|\theta}\|q_X)] \\ =&p(x_1^N;\pi)\ \mathbb{E}_{\pi(\theta|x_1^N)}\big[D(p_{X|\theta}\|p_{X|x_1^N;\pi})\big] \\ &+ p(x_1^N;\pi)\ D(p_{X|x_1^N;\pi}\|q_X),\end{aligned}$$

したがって式 (6) の inf は $q(x) = p(x|x_1^N;\pi)$ のときに最小となり，式 (6) と式 (7) とは同じ最適値をもつ．また，共に同じ $\pi(\theta)$ が 2 つの問題の最適値を与える．

以下では問題 3 と 4 との最適値を比較する．結果として同じ最適値を持つことを示す．さらに問題 4 の最適な $p_{X|x_1^N;\pi}$ が問題 3 の最適な $q(x)$ となることも示す．

3.3 双対性について

ここではパラメータ θ は 1 次元とするので，$\pi(\theta)$ ではなく，累積密度関数 $F(\theta)$ を考える．累積密度関数は右連続な非減少関数である．全ての $F(\theta)$ の集合を $\mathcal{F}$ とする．

$$\mathcal{F} = \big\{F : \mathbb{R} \to [0,1] \mid F(\theta) = 0,\ (\forall \theta < a),\ F(\theta) = 1,\ (\forall \theta \geq b)\big\}. \tag{8}$$

$\pi(\theta)$ によって周辺化した $p(*;\pi)$ については以下では $p(*;F)$ と書く．また，積分は Riemann-Stieltjes 積分の表記とするため $\mathbb{E}_{\pi(\theta)}[g(\theta)]$ は $\int g(\theta)dF(\theta)$ とかく．次にベイズリスク $\mathsf{R}(F)$ を以下のように定義する．

$$\mathsf{R}(F) = \int_\Theta L(\theta, F)dF(\theta), \tag{9}$$

ただし次の損失を考える．

$$L(\theta, F) = p(x_1^N|\theta)\ D(p_{X|\theta} \| p_{X|x_1^N;F}).$$

問題 4 は次のように書き直せる．

$$\sup_{F\in\mathcal{F}} \mathsf{R}(F). \tag{10}$$

以下では [2] の手順にしたがって議論する．まず，$\mathcal{F}$ がコンパクトであることを示すが，文献にある通りなので省略する．$\mathsf{R}(F)$ が F の汎関数として $\mathcal{F}$ 上で連続で上に凸であれば，式 (7) の最適化の問題は $\mathcal{F}$ 上の最適な関数 $\hat{F}$ が最適値を達成する．$\mathsf{R}(F)$ の凸性については別途調べる必要がある．以下では $\mathsf{R}(F)$ が上の凸だとして進める．このとき上の問題は以下のように書き換えられる．

$$\max_{F\in\mathcal{F}} \mathsf{R}(F). \tag{11}$$

最大値を与える $\hat{F}$ の満たす Karush Kuhn Tucker (KKT) 条件を導く．そのために，次の方向微分を考える．

$$\mathsf{R}'_{F_0}(F) = \lim_{\eta\downarrow 0} \frac{\mathsf{R}((1-\eta)F_0 + \eta F) - \mathsf{R}(F_0)}{\eta}, \qquad F \in \mathcal{F}.$$

ここで上の方向微分が全ての $F, F_0 \in \mathcal{F}$ に対して存在すると仮定する．KKT 条件は以下の通りである．

Proposition 1. $\mathsf{R}(F)$ が $\mathcal{F}$ 上の全ての F で上の凸であり，全ての $F, F_0 \in \mathcal{F}$ に対して方向微分 $\mathsf{R}'_{F_0}(F)$ が存在するとする．$\hat{F}$ が $\mathcal{F}$ 上で $\mathsf{R}(F)$ の最大値を与えるとすると，全ての $F \in \mathcal{F}$ に対して以下の方向微分が負，または *0* となる．

$$\mathsf{R}'_{\hat{F}}(F) \leq 0.$$

上の結果は，最大値を与える $\hat{F}$ からどの方向に F を動かしても $\mathsf{R}(F)$ が増えることがないということである．また，$\mathsf{R}'_{F_0}(F)$ は一般に以下のように書ける．

$$\mathsf{R}'_{F_0}(F) = \int_\Theta L(\theta, F_0)dF(\theta) - \mathsf{R}(F_0).$$

したがって Proposition 1 は最適な $\hat{F}$ が以下の関係を満たすことを示している．

$$\int_\Theta L(\theta, \hat{F})dF(\theta) \leq \mathsf{R}(\hat{F}), \quad \forall F \in \mathcal{F}. \tag{12}$$

ベイズリスクが損失の平均であることから，次の結果が導ける．

Corollary 1. E_0 を Θ における $\hat{F}(\theta)$ の増加点の集合とする．最適な $\hat{F}$ は次の関係を満たす．

$$\begin{aligned} L(\theta;\hat{F}) &\leq \mathsf{R}(\hat{F}), \quad \forall\theta \in \Theta \\ L(\theta;\hat{F}) &= \mathsf{R}(\hat{F}), \quad \forall\theta \in E_0. \end{aligned} \tag{13}$$

Proof. もし $\theta \in E_0$ のある点において $L(\theta;\hat{F}) > \mathsf{R}(\hat{F})$ であるならば式 (12) の左辺は全ての確率をその点に集中した F を選べば不等式を満たさない．したがって最適な $\hat{F}$ は式 (13) を満たさなければならない． □

問題 3 と 4 との弱双対性は [3] と同様に証明できる．双対ギャップが 0 となることを以下で簡単に示す．

Corollary 2. 問題 *3* の最適値を与える $q(X)$ はある $\hat{F}$ を用いて $p(x|x_1^N;\hat{F})$ と表される．また，その $\hat{F}$ は問題 *4* の最適値を与え，*2* つの最適値は一致する．

$$\begin{aligned} \mathsf{R}(\hat{F}) &= \inf_{q\in\mathcal{A}(X)} \sup_{\theta\in\Theta}\ p(x_1^N|\theta) D(p_{X|\theta}\|q_X), \\ &= L(\hat{\theta}, \hat{F}), \quad \hat{\theta} \in E_0. \end{aligned}$$

Proof. まず Corollary 1 と式 (13) から次の関係が示せる．

$$\max_{\theta\in\Theta} L(\theta;\hat{F}) = L(\hat{\theta}, \hat{F}), \quad \hat{\theta} \in E_0. \tag{14}$$

また，$F \neq \hat{F}$ でない分布に対しては次の関係がなりたつ．

$$\max_{\theta\in\Theta} L(\theta;F) > \mathsf{R}(F), \tag{15}$$

$F \neq \hat{F}$ に対して $\max_{\theta\in\Theta} L(\theta;F) - \mathsf{R}(\hat{F})$ を考えると以下のようになる．

$$\begin{aligned} \max_{\theta\in\Theta} L(\theta;F) - \mathsf{R}(\hat{F}) &= \max_{\theta\in\Theta} L(\theta;F) - \int_\Theta L(\theta;\hat{F})d\hat{F}(\theta) \\ &\geq \int_\Theta \Big[L(\theta;F) - L(\theta;\hat{F})\Big]d\hat{F}(\theta) \qquad (16) \\ &= \int_\Theta p(x_1^N|\theta)\Big[D(p_{X|\theta}\|p_{X|x_1^N;F}) \\ &\qquad - D(p_{X|\theta}\|p_{X|x_1^N;\hat{F}})\Big]d\hat{F}(\theta) \\ &= p(x_1^N;\hat{F})D(p_{X|x_1^N;\hat{F}}\|p_{X|x_1^N;F}) \geq 0. \end{aligned} \tag{17}$$

式 (16) は $\max_{\theta\in\Theta} L(\theta;F) \geq L(\hat{\theta};F)$ $(\hat{\theta}\in E_0)$ から導かれる．式 (17) は KL 情報量の非負性から成り立つ．$F=\hat{F}$ のとき，等号が成り立ち，2 つの問題の最適値が等しいことがわかる．弱双対性が成り立つことから 2 つの最適化問題は等しい最適値を持ち，問題 4 の最適値を与える $\hat{F}$ によって構成される $p(x|x_1^N;\hat{F})$ は問題 3 の最適値を与える □

4　数値実験

本節では正規分布を例にとり，前に提案した min-max 問題で得られる分布の性質を考える．

正規分布の例

平均が未知で分散が 1 の正規分布を考える．ただし，正規分布の平均は絶対値はある正の実数 a 以下だとする．

$$\mathcal{P}=\Big\{p(x|\theta)=\frac{1}{\sqrt{2\pi}}\exp\Big[-\frac{(x-\theta)^2}{2}\Big]\Big|\ |\theta|\leq a, x\in\Re\Big\}. \tag{18}$$

この正規分布から N 点の独立なサンプル x_1^N が得られたとする．次の問題を考える．

$$\sup_{F}\ \int_{|\theta|\leq a} p(x_1^N|\theta)D(p_{X|\theta}\|p_{X|x_1^N;F})\ dF(\theta). \tag{19}$$

このベイズリスクは F に関して上の凸である．したがってこの問題の最適値は 次の問題の最適値と同じであり，最適値を与える F を $\hat{F}$ とすると，次式の最適値を与える $q(x)$ は $q(x)=p(x|x_1^N;\hat{F})$ と書ける．

$$\inf_{q\in\mathcal{A}(X)}\ \sup_{|\theta|\leq a}\ p(x_1^N|\theta)\ D(p_{X|\theta}\|q_X),.$$

まず，次のことを証明する．

Lemma 1. 式 (19) に示された問題の最適値を与える $\hat{F}$ は離散分布となる．

Proof. 証明の手順は [2] と同じである．

式 (13) より $L(\theta;\hat{F})$ は $\hat{F}$ の増加点において定数となる．

$$L(\theta,\hat{F})=p(x_1^N|\theta)D(p_{X|\theta}\|p_{X|x_1^N;\hat{F}})=\mathsf{R}(\hat{F}),\quad \theta\in\Theta.$$

実数のパラメータ θ を複素数 z に拡張する．$L(\theta;\hat{F})$ は複素関数 $L(z,\hat{F})$ となり，この関数は実数軸を含む領域で解析的となる．

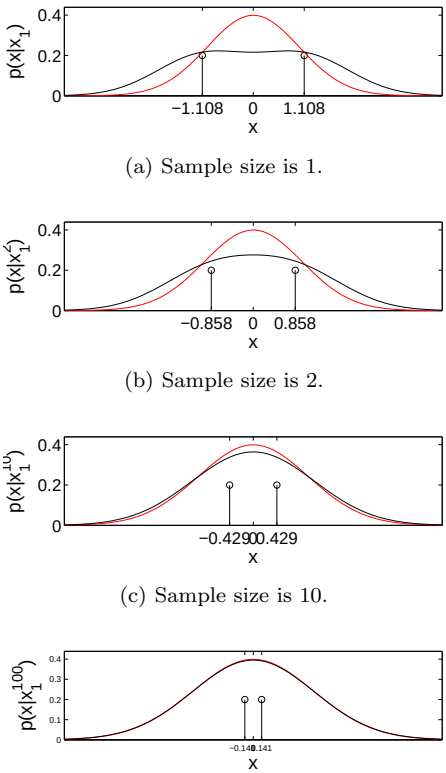
次に E_0 の点の数が無限にあると仮定する．Θ は閉区間であることから E_0 収束点がある．一致の定理から $L(\theta;\hat{F})$ は $L(z,\hat{F})$ は実数軸を含む領域で定数となる．したがって，全ての実数 $\theta\in\Re$ において次式を満たさなければならない．

$$-\int_{\mathcal{X}} p(x|\theta)\log p(x|x_1^N;\hat{F})d\mu_X=\frac{\mathsf{R}(\hat{F})}{p(x_1^N|\theta)}+H(X|\theta)$$
$$=\frac{\mathsf{R}(\hat{F})}{p(x_1^N|\theta)}+\frac{1}{2}\log 2\pi e,$$
$$=C_1(\hat{F})\exp\Big[\frac{N}{2}(\theta-\bar{x})^2\Big]+C_2.$$

ただし $\bar{x}=\sum_i \hat{x}_i/N$ であり $C_1(\hat{F})$ は $\hat{F}$ の関数，C_2 は定数である．C_1 と C_2 は θ を含まない．

この等式を満たす $\hat{F}$ が存在するかを考える．$p(x|x_1^N;\hat{F})$ は $\hat{F}$ と尤度を用いた $p(x|\theta)$ の周辺分布である．したがってどのような F を用いても左辺のように $\exp[\theta^2/2]$ の項を構成できず，この等号は成り立たない．$\hat{F}$ の増加点は有限個の孤立点からなり，$p(x|x_1^N;\hat{F})$ は有限個の正規分布の混合分布となる． □

この結果から最適化問題を解いて得られる事前分布 $\hat{F}$ は有限個の点からなる離散分布となる．しかし，それぞれの点の位置と確率は解析的には求まらないため，数値



(a) Sample size is 1.

(b) Sample size is 2.

(c) Sample size is 10.

(d) Sample size is 100.

図 1: 問題 3 の最適化問題で得られる分布 $\hat{F}$(棒グラフ) とそれによって構成される $p(x|x_1^N;\hat{F})$(黒の実線)，ひかく のために $\mathcal{N}(0,1)$ を示してある (赤の実線)．分散は 1 とした．観測データの平均は 0 にそろえてある．サンプルの数は (a)∼(d) のそれぞれで 1, 2, 10, 100 とした．また，$|\theta|\leq 10$ である．

的に求めることになる．それぞれの点の位置と確率を以下のように定義する．

$$\theta_1, \cdots, \theta_M, \quad -a \le \theta_1 < \cdots, < \theta_M \le a,$$
$$\pi_1, \cdots, \pi_M, \quad \sum_{j=1}^{M} \pi_j = 1, \quad \pi_j > 0,$$

点の数は M とする．M も未知であるため M を 2 に初期値とする．M を固定し，$L(\theta; F)$ を最大にするように $\{\theta_i\}$，$\{\pi_i\}$ を最急降下法によって最適化する．求まった F が最適かどうかは式 (13) の KKT 条件が成り立つかどうかによって確認する．最適でないときには M を増やし，再度最適化を行なう．KKT 条件が満たされればそれが最適な $\hat{F}$ となる．

最適な事前分布とそれから構成される予測分布を図 1 に示す．図では得られたサンプルの数 N を 1 から 100 まで変化させている．$\mathsf{R}(F)$ を求める際に，混合正規分布を数値的に評価する必要があるが，ここでは 30 点の Gauss-Hermite 積分を行なった．図では最適な事前分布を棒状に表現し，それによって構成される min-max 予測分布を実線で示している．予測分布は常に 2 点からなる分布で，それぞれ 1/2 という等しい確率をもっている．2 点間の距離はサンプルの数が増えるにしたがい狭まる．これはサンプルによる予測という意味では適切である．

サンプルは確率的にふるまう．したがってサンプルによって得られる $p(x|x_1^N; \hat{F})$ も確率的に変化する．図 2 ではデータを生成した正規分布から $p(x|x_1^N; \hat{F})$ までの KL 情報量

$$D(p_{X|\theta} \| p_{X|x_1^N;\hat{F}}), \tag{20}$$

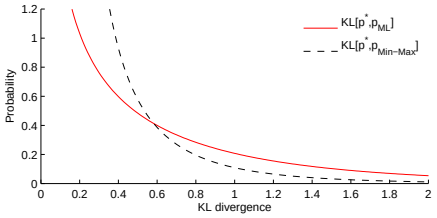
がどのように分布するかを示してある．点線で示したのがその結果である．赤い実戦で示したのは真の分布から最尤推定値 $\hat{\theta}(x_1^N) = \overline{x}_1^N = \sum_i x_i/N$ をプラグインした分布への KL 情報量

$$D(p_{X|\theta} \| p_{X|\hat{\theta}(x_1^N)}). \tag{21}$$

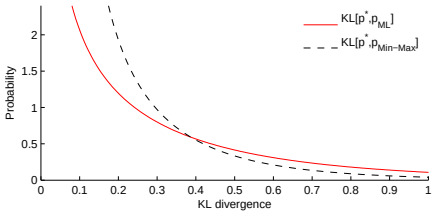
の分布である．

提案する方法で求めた分布は 2 つの正規分布の混合である．この分布を用いた場合と最尤推定をプラグインしたものを比べると，提案するものは KL 情報量が大きくなる確率が小さくなっていることがわかる．また，サンプルの数が増えるにしたがって，2 つの曲線の差は小さくなっている．

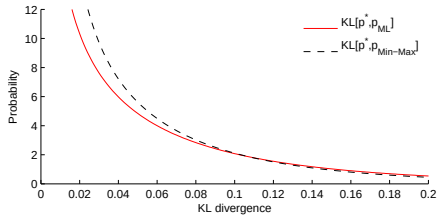
以上の結果から，提案する手法によって予測のための分布を構成すると，この場合には 2 つの正規分布の混合となり簡単に構成できること，そしてそれらが予測の意味では望ましい性質をもっていることがわかった．



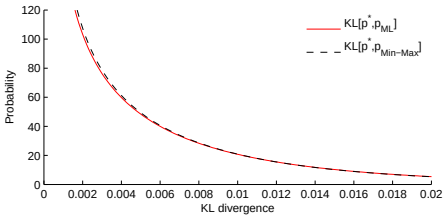
(a) Sample size is 1.



(b) Sample size is 2.



(c) Sample size is 10.



(d) Sample size is 100.

図 2: サンプルの出方による $D(p_{X|\theta} \| p_{X|x_1^N;\hat{F}})$ (黒の点線) と $D(p_{X|\theta} \| p_{X|\hat{\theta}(x_1^N)})$ (赤の実線) の分布．サンプルの数は (a)~(d) のそれぞれで 1, 2, 10, 100 である．

5 まとめ

本稿では，まず KL 情報量を損失とするベイズリスクの最適化問題として定式化される reference 事前分布 [5] の問題を示し，それと双対な関係にある min-max 問題を紹介した．この双対性は Haussler [3] によって示されており， Grünwald & Dawid [6] が示した枠組に含まれてる．

本稿では損失を尤度で重みつけをした KL 情報量に変え，予測のための min-max 問題を考えた．この損失も KL 情報量と同様に双対となる問題があり，それぞれの最適値は一致することを示した．本稿の結果は重みをつけた KL 情報量も Grünwald & Dawid [6] の枠組に含まれることを示したことになる．

Reference 事前分布を求める問題は Shannon の定義

した通信路容量 [10] の問題で最適な入力分布を求める問題と一致する．多くの通信路容量の問題で容量を達成する入力の分布が有限個の点からなる離散分布となることは広くしられており [2, 11]，ここで提案した予測のための min-max の問題でも最適な分布は有限個の点からなる離散分布が最適な事前分布となることが予想される．有限個の点からなる離散分布によって予測分布を構成した場合，その分布は有限個の要素からなる混合分布となる．

例として平均値の値が分らない正規分布を例に数値実験を行なった．サンプルの確率的振舞いによって KL 情報量により損失がどのように分布するかを調べてみると，最尤推定値をプラグインする場合に比べ大きい損失をとる確率が小さくなることが分った．これは予測の観点からは良い振舞いである．今後は広いクラスの問題で提案する方法の有効性を検証していきたい．

参考文献

[1] J. Aitchison, “Goodness of prediction fit,” Biometrika, vol.62, no.3, pp.547-554, 1975.

[2] J.G. Smith, “The information capacity of amplitude- and variance-constrained scalar Gaussian channels,” Information and Control, vol.18, pp.203-219, 1971.

[3] D. Haussler, “A general minimax result for relative entropy,” IEEE Trans. Inf. Theory, vol.43, no.4, pp.1276-1280, 1997.

[4] J.M. Bernardo, “Reference posterior distributions for Bayesian inference,” J. R. Statistical Society, Series B, vol.41, no.2, pp.113-147, 1979.

[5] J.O. Berger, and J.M. Bernardo, “On the development of reference priors,” Bayesian Statistics, vol.4, pp.35-60, 1992.

[6] P.D. Grünwald, and A.P. Dawid, “Game thoery, maximum entropy, minimum discrepancy and robust Bayesian decision theory,” The Annals of Statistics, vol.32, no.4, pp.1367-1433, 2004.

[7] 竹内純一，池田思朗，“通信路容量に関する情報幾何学的考察,” SITA 2010 予稿集第 33 回情報理論とその応用シンポジウム，December 2010.

[8] 池田思朗，林和則，田中利幸，“最大出力制約下での通信路容量と変調方式,” SITA 2010 予稿集第 33 回情報理論とその応用シンポジウム，December 2010.

[9] Z. Zhang, Discrete Noninformative Priors, Ph.D thesis, Yale University, Nov. 1994.

[10] C.E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol.27, pp.379-423 and 623-656, July and October 1948.

[11] T.H. Chan, S. Hranilovic, and F.R. Kschischang, “Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs,” IEEE Trans. Inf. Theory, vol.51, no.6, pp.2073-2088, June 2005.

[12] F. Komaki, “On asymptotic properties of predictive distributions,” Biometrika, vol.83, no.2, pp.299-313, 1996.