Bayes Theory Essentials

Sumio Watanabe
Tokyo Institute of Technology

Notations

Datum : X in $\mathbf{R}^N$

Unknown True : q(x)

I.I.D. Sample

$D_n = \{X_1,\ldots,X_n\}$

Parameter w in $\mathbf{R}^d$

Model : p(x|w)

Prior : φ(w)

Posterior and Predictive

Log Loss : $L_n(w) = -(1/n) \sum_{i=1}^{n} \log p(X_i|w)$

Posterior $p(w|D_n) = (1/Z_n)\, \varphi(w) \exp(-nL_n(w))$

Predictive distribution $p(x|D_n) = \mathbf{E}_w\, [p(x|w)]$

$\mathbf{E}_w[\;]$, $\mathbf{V}_w[\;]$: posterior mean and variance

Random Variables

Let $p^*(x) = p(x|D_n)$

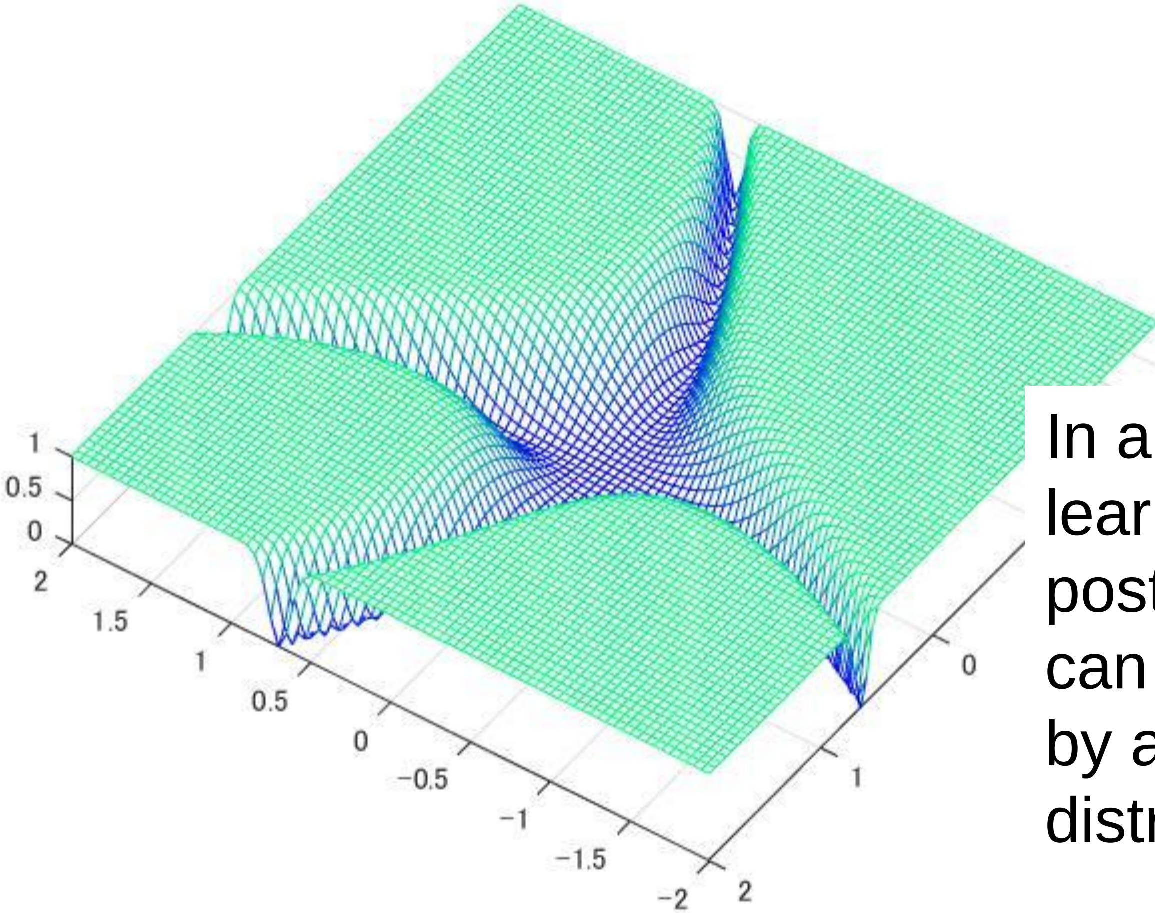
Training Loss: T $= -(1/n) \Sigma_{i=1}^{n} \log p^*(X_i)$

Generalization Loss: G $= -\mathbf{E}_{xy}[\log p^*(X)]$

Free energy : F $= -\log \int \varphi(w) \exp(-nL_n(w))\, dw$

Posterior Distribution

$$L(w) = -\ \mathbf{E}_x[\ \log p(X|w)\ ]$$



In almost all learning machines, posterior distributions can not be approximated by any normal distribution.

Birational Invariants

Real Log Canonical Threshold λ

Multiplicity m

Singular Fluctuation ν

Theorems

Theorem. Let w_0 be the parameter that minimizes L(w).

$$\mathbf{E}[\ G\ ] = L(w_0) + \lambda\ /n + o(1/n)$$

$$\mathbf{E}[\ T\ ] = L(w_0) + \{\lambda - \nu\ \}\ /\ n + o(1/n)$$

Theorem.

$$F\ =\ n\ L_n(w_0)\ \ + \lambda \log n - (m-1) \log \log n + O_p(1).$$

Theorems

(1) Definition : WAIC = T+ (1/n) $\sum_{i=1}^{n}$ $\mathbf{V}_w$[log p(X_i|w)].

Theorem **E**[G] = **E**[WAIC] +o(1/n).

(2) WAIC is asymptotically equivalent to LOOCV.

Theorem

Posterior $p(w|D_n,\beta) = (1/C)\ \varphi(w) \exp(-n\beta L_n(w))$

$\mathbf{E}_w^{(\beta)}[\ \ ]$, posterior mean using $p(w|D_n,\beta)$

(1) Definition : $\mathrm{WBIC} = \mathbf{E}_w^{(1/\log n)}[n L_n(w)]$.

Theorem. $F = \mathrm{WBIC} + o_p(\log n)$.