A decision Theoretic Formulation for Robust Automatic Speech Recognition

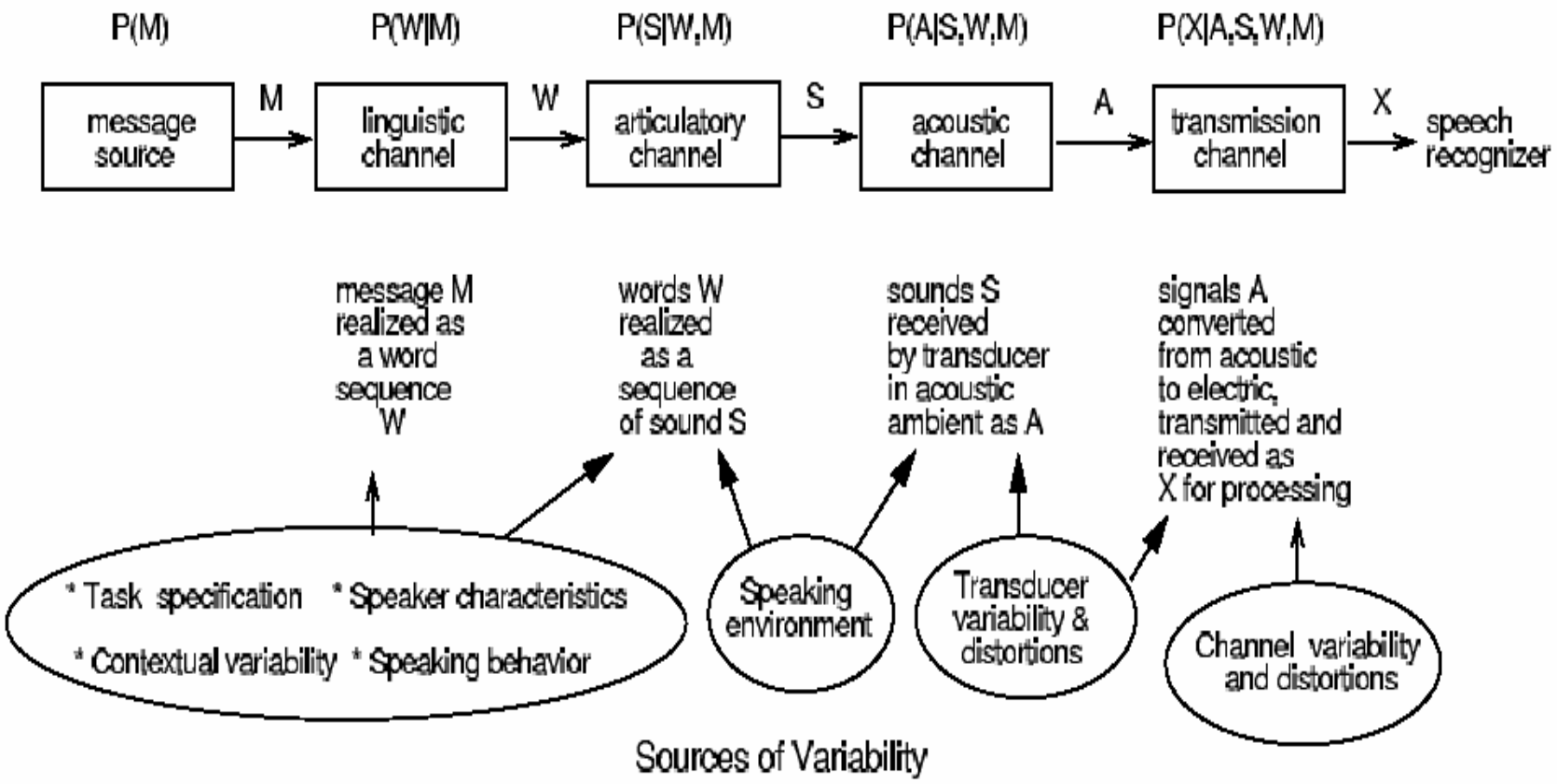
Author : Qiang Huo

Reporter : CHEN TZAN HWEI

Reference

- W. Chou and B.-H. Juang, “Pattern recognition in speech and language processing”, New York, CRC Press, 2003

Introduction (1/4)



Introduction (2/4)

- The goal of speech recognition can be viewed as a decision problem
  - i.e. based on the information of X, we attempted to make the best decision of the word sequence W that has been embedded in X
  - For the simplicity of discussion, we can view each W as a ***class***. So, speech recognition consists to find optimal decision rules for classification of the observation X into one of some fixed classes.

Introduction (3/4)

- This chapter explains :
  - The decision theoretic formulation for the ASR problem and the optimal decision rule that can be constructed if everything about the problem is known
  - How to construct the adaptive decision rules when learning from a training sample set

Introduction (4/4)

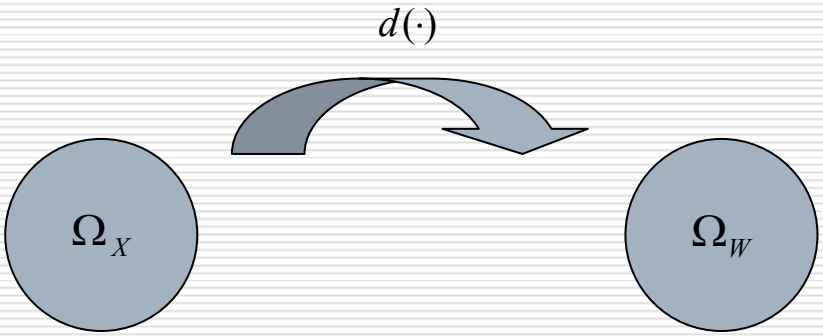
- This chapter explains (cont):
  - The classification of possible distortions of hypothetical models and data
  - Some of the recent parameter adaptation techniques for improving adaptive decision rules

Optimal Bayes' Decision Rule for ASR (1/8)

- Let us assume :
  - speech observation X belongs to a suitable space Ω_X
  - A class $W \in \Omega_W$, where $\Omega_W = \{W_1, W_2, \cdots, W_M\}$
- The problem of constructing a speech recognizer is the equivalent to find a decision rules $d(\cdot)$ in a set of possible decision rules D , such that

$$W = d(X), \text{ for } X \in \Omega_x, W \in \Omega_W, \text{ and } d(\cdot) \in D \qquad (1)$$

Optimal Bayes' Decision Rule for ASR (2/8)



There may exist an infinite set of decison rules for the same given classification problem.

To determine whether a decision rule is "good" one has to agree on a reasonable set of criteria for assessing the "goodness"

Optimal Bayes' Decision Rule for ASR (3/8)

One possible formulation by using the classical statistical decision theory pioneered by Wald:

let us view W and X as a jointly distributed random pair (W, X), whose joint PDF is denoted by $p(W, X)$.

$\ell(W, d(X))$ be the loss associated with making a decision, $d(X)$, if the true class is W

We would like the loss function to have the following property

$$0 \le \ell(W, d(X) = W) \le \ell(W, d(X) \ne W)$$

Optimal Bayes' Decision Rule for ASR (4/8)

- If we assume the true distribution $P(W,X)$ is known. Then we can define the total risk, $r(d(\cdot))$,for a decision rule $d(\cdot)$ as expected loss function, i.e. ,

$$
\begin{aligned}
r(d(\cdot)) &= E_{(W,X)}\left[\ell(W,d(X))\right] \\
&= \sum_{W\in\Omega_W}\int_{X\in\Omega_X}\ell(W,d(X))p(W,X)dX \qquad (2)\\
&= \int_{X\in\Omega_X}p(X)\left[\sum_{W\in\Omega_W}\ell(W,d(X))p(W\mid X)\right]dX \qquad (3)\\
&= \sum_{W\in\Omega_W}P(W)\int_{X\in\Omega_X}\ell(W,d(X))p(X\mid W)dX \qquad (4)
\end{aligned}
$$

Optimal Bayes' Decision Rule for ASR (5/8)

- In this framework, the issue of constructing an optimal decision rule becomes the following loss minimization problem :

$$\min_{d(\cdot)\in D} r(d(\cdot)) = \min_{d(\cdot)\in D} \int_{X\in\Omega_X} p(X)\left[\sum_{W\in\Omega_W} \ell(W, d(X))\, p(W \mid X)\right] dX \qquad (5)$$

- The optimization can be solved by minimizing the expression in the square brackets

$$d_o(X) = \arg\min_{d(X)\in\Omega_W} \sum_{W\in\Omega_W} \ell(W, d(X))\, p(W \mid X) \qquad (6)$$

Bayes' decision rule

Optimal Bayes' Decision Rule for ASR (6/8)

- The resulting minimum total risk

$$r(d_o(\cdot)) = \int_{X \in \Omega_X} p(X) \left[\sum_{W \in \Omega_W} \ell(W, d_o(X)) p(W \mid X) \right] dX \qquad (7)$$

Bayes' risk

- In speech recognition, a reasonable option is to assume that every misclassification of X is equally serious :

$$\ell(W, d(X)) = \begin{cases} 0 \text{ if } W = d(X) \\ 1 \text{ if } W \neq d(X) \end{cases} \qquad (8)$$

Optimal Bayes' Decision Rule for ASR (7/8)

- Substituting (8) into (6), we obtain

$$r^{01}(d(\cdot)) = \sum_{W \in \Omega_W} P(W) \int_{X \notin \Omega_X(w)} p(X \mid W) dX \qquad (9)$$

$$= 1 - \sum_{W \in \Omega_W} \int_{X \in \Omega_X(w)} P(W) p(X \mid W) dX \qquad (10)$$

- The optimal decision rule $d_o^{01}(X)$ is then solved as $d_{MAP}(X) = W'$ such that

$$W' = 1 - \arg\min_W \sum_{W'' \in \Omega_W, W'' \neq W} P(W) P(X \mid W)$$

$$= \arg\max_W P(W) P(X \mid W) = \arg\max_W P(W \mid X) \qquad (11)$$

Optimal Bayes' Decision Rule for ASR (8/8)

- In summary, in constructing these optimal decision rules, it was assumed that complete prior information about the classes is known
  - The observation space Ω_X is given
  - The loss function $\ell(W, d(X))$ is given
  - The true PDF $p(W, X)$ or $p(X \mid W)$ and $p(W)$ are given

Adaptive Decision Rule Constructed from Training samples (1/2)

- In practice, we don't know the true parametric form of the joint distribution $p(W, X)$
- We shall say that we have *prior uncertainty*
- If we have some labeled independent training sample set, $\chi = \{(W^i, X^i); i = 1,2,\cdots,n\}$, obtained by a series of independent experiments such that $(W^i, X^i) \sim p(W, X)$

Adaptive Decision Rule Constructed from Training samples (2/2)

- The decision rule $d(\cdot) = d(X; \chi)$ based on the training set χ and used to classify a random observation X that is independent of χ is called an *adaptive decision rule*.
  - Plug-in Bayes' decision Rules with Maximum-likelihood Density Estimate
  - Maximum-Discriminant Decision Rules Minimizing the Empirical Classification Error

Plug-in Bayes' decision Rules (1/10)

- It might be the most popular family of adaptive decision rules.
- For this approach, let $\{\hat{p}(W), \hat{p}(X|W)\}$ be any statistical estimators of true distributions $\{p(W), p(X|W)\}$
- The plug-in risk $\hat{r}(d(\cdot))$

$$\hat{r}(d(\cdot)) = \sum_{W \in \Omega_W} \hat{p}(W) \int_{X \in \Omega_X} \ell(W, d(X))\, \hat{p}(X|W) dX$$

Plug-in Bayes' decision Rules (2/10)

- The minimum plug-in risk is then $\hat{r}(\hat{d}(\cdot))$ where,

$$\hat{d}_o(\cdot) = \arg\min_{d(\cdot)\in D} \hat{r}(d(\cdot))$$

- Why it works?
  - Property : if the estimators $\{\hat{p}(W), \hat{p}(X \mid W)\}$ are pointwise unbiased, then

$$\hat{r}(\hat{d}_o(\cdot)) \geq r(d_o(\cdot))$$

Plug-in Bayes' decision Rules (3/10)

- why it works (cont):
  - Theorem (Bayes Risk Consistency) :

    If the estimators $\{\hat{p}(W), \hat{p}(X|W)\}$ are strongly consistent, i.e. :

$$p(W) \xrightarrow{\text{as(n->}\infty\text{)}} p(W), \hat{p}(X|W) \xrightarrow{\text{as(n->}\infty\text{)}} P(X|W), \text{for W} \in \Omega_{\text{W}} \text{ and } X \in \Omega_{\text{X}}$$

    then

$$\hat{r}(\hat{d}_o(\cdot)) \xrightarrow{\text{a.s.(n->}\infty\text{)}} r(d_o(\cdot))$$

Plug-in Bayes' decision Rules (4/10)

- Because of the constraints of the limited training data, we always have to assume some parameter form for $p(W)$ and $p(X|W)$, e.g. $p_\Gamma(W)$ And $p_\Lambda(X|W)$

Plug-in Bayes' decision Rules (5/10)

- Some estimate method of HMM parameter :
  - Maximum likelihood (ML)
  - Discriminative training
    - Corrective training
    - minimum empirical classification error (known as MCE)

Plug-in Bayes' decision Rules (6/10)

- The problem of ML :
  - If the wrong model is used, there can be no true parameter.
  - ML doesn't minimize the recognition error rate.
  - Seeking statistic which maximize some function that is loosely associated with speech recognition performance

Plug-in Bayes' decision Rules (7/10)

- Corrective training
  - A simple example :

Let $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$ denote linearly separable sets of vectors.

If necessary, to arrange this. We wish to find a vector z statisfying

$$z^t x > 0 \qquad \text{for each } x \in X$$

$$z^t y < 0 \qquad \text{for each } y \in Y$$

A solution vector z may be obtained via the following simple procedure

1) Choose any nonzero value for z

2) For each $x \in X$ compute $z^t x$. if $z^t x \le 0$, the replace z by $z + cx$

3)For each $x \in Y$ compute $z^t y$. if $z^t y \ge 0$, the replace z by $z - cy$

4)if any adjustments to z were made, return step 2

Plug-in Bayes' decision Rules (8/10)

- Corrective training (cont)
  - Basically, the procedure works by adjusting the vector z incrementally so as to make the classification errors go away.

$$\text{if } z^t x \le 0, \text{ for example, the replacing } z \text{ by } \bar{z} = z + cx, \text{ ensure that}$$

$$z \text{ is moved in the right direction: } \bar{z}^t x = z^t x + cx^t x > z^t x$$

Plug-in Bayes' decision Rules (9/10)

- Corrective training (cont) :
  - Applying to discrete HMM

$$p(y_k \mid a_i) = \frac{c(y_k, s_i)}{\sum_{j=1}^{L} c(y_j, s_i)}$$

$$p(a_k \mid s_k) = \frac{\sum_{j=1}^{L} c(y_j, a_i)}{\sum_{k:a_i \in O_k} \sum_{j=1}^{L} c(y_j, a_k)}$$

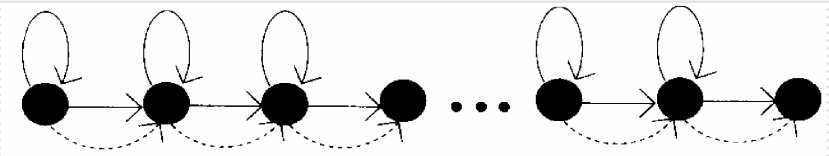


Fig. 2 Topology of a Markov word model.

O_k denotes the set of all arcs which originate in state s_k

Plug-in Bayes' decision Rules (10/10)

- Corrective training (cont) :
  - Applying to discrete HMM

1) using some labeled training data, apply the forward - backward algorithm to compute the approximate frequency $c(y,a)$ for each label y and each arc a

2) For each utterance u in the training data, compute the probability $p(u \mid w)$ for the correct word w and the probability $p(u \mid w_i)$ for each incorrect word w_i on the corresponding short list of acoustically confusable words.Perform step3 if $\log p(u \mid w_i) > \log p(u \mid w) - \delta$

3)Estimating $c_w^u(y,a)$ and $c_{w_i}^u(y,a)$ by using the Markov model w and w_i. for each label y and arc a repleace $c(y,a)$ by $c(y,a) + \gamma(c_w^u(y,a) - c_{w_i}^u(y,a))$

4)If any adjustments were made in Step 3, replace any negative counts $c(y,a)$ by 0 and return to step 2.

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (1/10)

- Suppose one can define a discriminant function $g_i(X;\Lambda)$ for each class C_i that characterize the similarity between an observation X and the class C_i

- Naturally, the following maximum-discriminant decision rule $d(\cdot)$

$$C' = \arg\max_i g_i(X;\Lambda)$$

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (2/10)

- The obvious criterion is to minimize the empirical classification error :

$$\bar{r}(d(\cdot)) = 1 - \frac{\text{number of correct classfication by } d(\cdot)}{\text{total number of sample observation on } \chi}$$

- A sample-based discriminant decision rule

$$\bar{r}(\bar{d}(\cdot)) = \min_{d(\cdot)\in D} \bar{r}(d(\cdot))$$

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (3/10)

- The choice of discriminant functions and the practical training algorithms
  - How to define an optimal form for the discriminant functions remains largely an open research problem
  - The smooth MCE objective function proposed can approximate the empirical error rate.

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (4/10)

- MCE : three step procedure to derive the object function
  - First, the discriminant function $g_i(X;\Lambda) = p(C_i \mid X)$ are described
  - Second, introduce a misclassification measure. One proposal is

$$d_k(X) = \sum_{i \in S_k} \frac{1}{m_k} [g_i(X;\Lambda) - g_k(X;\Lambda)]$$

S_k is not a fixed set !

where $S_k = \{i \mid g_i(x;\Lambda) > g_k(x;\Lambda)\}$, the set of "confusing classes"

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (5/10)

- MCE : three step procedure to derive the object function (cont)
  - One reasonable misclassification measure

$$d(X) = -g_k(X;\Lambda) + \left[\frac{1}{M-1}\sum_{j,j\neq k} g_j(X;\Lambda)^{\eta}\right]^{\frac{1}{\eta}}$$

  - The measure is continuous and offer a fair amount of flexibility. Ex. : when η approach ∞ it becomes

$$d_k(X) = -g_k(X;\Lambda) + g_i(X;\Lambda)$$

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (6/10)

proof :

we could find $g_i(X;\Lambda) = \max_{j, j \neq k} g_j(X;\Lambda)$

right – hand become

$$\lim_{\eta \to \infty} \sqrt[\eta]{\frac{\left(\frac{g_1(X;\Lambda)}{g_i(X;\Lambda)}\right)^{\eta} + \left(\frac{g_2(X;\Lambda)}{g_i(X;\Lambda)}\right)^{\eta} + \cdots \left(\frac{g_i(X;\Lambda)}{g_i(X;\Lambda)}\right)^{\eta} + \cdots + \left(\frac{a_M}{g_i(X;\Lambda)}\right)^{\eta}}{\frac{M-1}{\left(g_i(X;\Lambda)\right)^{\eta}}}}$$

$$= \lim_{\eta \to \infty} \sqrt[\eta]{\frac{\left(g_i(X;\Lambda)\right)^{\eta}}{M-1}}$$

$$= g_i(X \mid \Lambda)$$

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (7/10)

- MCE : three step procedure to derive the object function (cont) :
  - the above misclassification measure is used in the third step where the minimum error objective $\ell_k(X;\Lambda)=\ell_k(d_k)$ is formulated :

$$a)\quad \text{exponential}: \ell_k(d_k)=\begin{cases}(d_k)^{\varsigma}, & d_k>0\\ 0 & d_k\le 0\end{cases} \qquad \text{where } \varsigma>0 \text{ and } \varsigma\to 0$$

$$b)\quad \text{translated sigmoid}: \ell_k(d_k)=\frac{1}{1+e^{-\xi(d_k+\alpha)}}, \qquad \xi>0$$

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (8/10)

- MCE : three step procedure to derive the object function (cont) :
  - Finally, for any unknown X the classifier performance is measured by

$$\ell(X;\Lambda) = \sum_{k=1}^{M} \ell_k(X;\Lambda)\delta(X \in C_k)$$

$$\delta(\Re) = \begin{cases} 1, & \text{if } \Re \text{ is true} \\ 0, & \text{if } \Re \text{ is false} \end{cases}$$

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (9/10)

- MCE : three step procedure to derive the object function (cont) :
  - How this formulation relates to the minimum classification?

The minimum classification error (Bayes minimum risk) :

$$\mathrm{E} = \sum_{k=1}^{M} \int_{X \in \Omega_X(k)} p(X, C_i)\delta(X \in C_k)dX$$

where

$$\Omega_X(k) = \left\{ X \in \Omega_X \mid p(C_k \mid X) \neq \max_i p(C_i \mid X) \right\}$$

the classification error can be rewriteen as

$$\mathrm{E} = \sum_{k=1}^{M} \int_{X \in \Omega_X} p(X, C_i)\delta(X \in C_k)\delta\left[X \in \Omega_X \mid p(C_k \mid X) \neq \max_i p(C_i \mid X) \right] dX$$

$$\cong \sum_{k=1}^{M} \int_{X \in \Omega_X} p(X, C_i)\delta(X \in C_k)\ell_k(d_k(X))dX$$

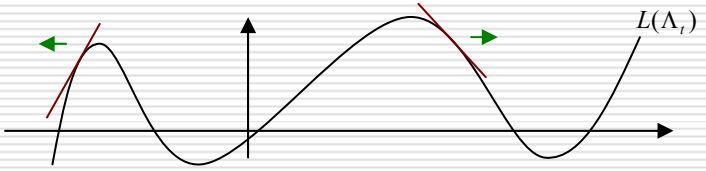
Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (10/10)

- MCE : minimum the average cost
  - Given a set of design observation $\aleph = \{X_1, X_2, \cdots, X_N\}$, we can define an empirical average cost as

$$L_0(\Lambda) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M} \ell_k(X_i;\Lambda)\delta(X_i \in C_k)$$

  - This well-defined cost function can be minimized by Gradient Descent algorithm

$$\Lambda_{t+1} = \Lambda_t - \varepsilon \nabla L_0(\Lambda_t)$$



Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (11/10)

- MCE : minimum the expected cost
  - Probabilistic Descent Algorithm :

The expected cost can be expressed as

$$L(\Lambda) = E[\ell(X;\Lambda)] = \sum_{k=1}^{M} P(C_k) \int \ell_k(X;\Lambda) p(X \mid C_k) dX$$

The adjustment of Λ is again according to

$$\Lambda_{t+1} = \Lambda_t + \delta\Lambda_t$$

where the "correction" term $\delta\Lambda_t$ is a function : $\delta(X, C_k, \Lambda_t)$

the manitude of the correction term is samm such that

$$L(\Lambda_{t+1}) \cong L(\Lambda_t) + \delta\Lambda_t \nabla L(\Lambda)\mid_{\Lambda=\Lambda_t}$$

hold

Maximum-discriminant Decision Rules minimizing the Empirical Classification Error (12/10)

- MCE : minimum the expected cost (cont)
  - Probabilistic Descent Algorithm :

the goal is to find an adaptation rule such that

$$E[L(\Lambda_{t+1}) - L(\Lambda_t)] = E[\delta L(\Lambda_t)] = E[\delta(X, C_k, \Lambda_t)]\nabla L(\Lambda_t) < 0$$

if we choose $\delta(X, C_k, \Lambda_t) = -\varepsilon U \nabla \ell_k(X, \Lambda)$

then

$$E[\delta L(\Lambda_t)] = -\varepsilon U \left[\sum_{k=1}^{M} P(C_k) \int \nabla \ell_k(X; \Lambda) p(X \mid C_k) dX \right] \nabla L(\Lambda_t)$$

$$= -\varepsilon U \bullet \nabla L(\Lambda_t) \bullet \nabla L(\Lambda_t) \le 0$$

Discussion about adaptive decision rules (1/2)

- Using plug-in MAP as a decision rule for recognition
  - ML as a criterion for the estimation of decision parameters
  - The asymptotic behavior will depend on that appropriateness of the parametric forms of the assumed distributions.

Discussion about adaptive decision rules (2/2)

- Using Maximum discriminant as a decision rule
  - The MCE as a criterion for the estimation of decision parameters
  - The asymptotic behavior will depend on the choice of the discriminant function.

Violations of modeling assumption in ASR (1/5)

- Three main distortion types
  - Distortion causing by small-sample effects
  - Distortion of models or discriminant functions for training samples
  - Distortion of trained model or discrimininat functions for observation to be classified.

Violations of modeling assumption in ASR (2/5)

- Distortion causing by small-sample effects
  - They arise from the noncoincidence of the statistical estimates $\{\hat{p}(W), \hat{p}(X|W)\}$ and $\{p(W), p(X|W)\}$
  - The design and/or collection of the training sample become very critical
    - To make the samples in χ follow the intended distribution $\{p(W), p(X|W)\}$ as closely as possible
    - Some more intelligent ways of using the available training data must be developed

Violations of modeling assumption in ASR (3/5)

- Distortion of models or discriminant functions for training samples can be caused by:
  - the wrong assumptions and/or inflexible parametric forms of the model or discriminant function.
  - The misclassifiaction of training samples
  - Outlier in training samples.

Violations of modeling assumption in ASR (4/5)

- Distortion of trained model or discrimininat functions for observation to be classified.
  - Might be the biggest problem for ASR.
  - There always exist some form of mismatch which causes a distortion between the trained models or discriminant function and test data

Violations of modeling assumption in ASR (5/5)

- Toward adaptive and robust ASR :
  - Find invariant features so as to minimize the observation variability.
  - Adapting recognizer parameters to new operating conditions using adaptation and/or testing data.
  - Using robust decision strategies
  - Possible combinations of the above techniques.