Sequential design and maximum likelihood recursion

Fritjof Freise
Otto-von-Guericke-Universität Magdeburg
e-mail: fritjof.freise@ovgu.de

1 Introduction

The logistic model is widely used for modeling binary or binomial response experiments. Let $Y(x)$ be a binary random variable depending on the "dose" x, then the three parameter logistic model is of the shape

$$P(Y(x)=1) := \alpha + (1-\alpha)\,F(\beta\,(x-\mu)) = 1 - P(Y(x)=0) \tag{1}$$

$$F(t) := \frac{1}{1+\exp(-t)}\ . \tag{2}$$

While $\mu \in \mathbb{R}$ can be interpreted as the difficulty of an item in a psychological test, $\beta > 0$ is the discriminating effect and $\alpha \in [0,1)$ describes the probability to answer the item correctly by guessing. The model reduces to the two parameter case for $\alpha = 0$ and with $\beta = 1$ in addition to the so called Rasch model.

One of the problems in optimal experimental design concerning these models is, as for nonlinear regression in general, that the information and hence the optimal design depends on the parameters. The (locally) optimal design to estimate μ in the two parameter model for example is a one point design at μ itself. Sequential designs are one way to cope with this situation.

An example for such an approach is the algorithm introduced by Robbins and Monro[1](1951), a procedure to find the root of a strictly increasing function, which is only observable via noisy observations $\widetilde{Y}$. The algorithm is defined in the following way:

$$X_{n+1} = X_n - a_n\widetilde{Y}_n\ , \quad \widetilde{Y}_n := \widetilde{Y}(X_n)\,,\ a_n > 0,\ a_n \to 0\ .$$

To adopt this procedure to estimate μ from the logistic model we set

$$\widetilde{Y}_n = Y_n - P(Y(\mu)=1) = Y_n - \frac{1+\alpha}{2}\ , \text{ with } Y_n := Y(X_n)\ .$$

[1] In the following referred to as RM.

In a more general setting we are interested in the argument x_p corresponding to $E(Y(x_p)) = p$ and $\widetilde{Y}_n = Y_n - p$.

There is an extensive literature about the asymptotic properties of this procedure. See Lai (2003) for a comprehensive review on this topic.

The present report first presents the results of Ying and Wu (1997) on recursive maximum likelihood and illustrates them using the logistic model. After that some related results from the literature are presented briefly, including some notes on the choice of initial designs, ensuring the existence of the maximum likelihood estimate.

2 Maximum likelihood recursion

To estimate the location parameter μ in the two parameter logistic model, Wu (1985) investigated a sequential scheme based on the maximum likelihood estimator. The next design point x_{n+1} in the approach is the ML-estimate $\widehat{\mu}_n$ of the location parameter, which is based on all previous observations.

To motivate this idea let us assume that the model we are interested in is a linear regression model, i.e.

$$Y(x) = \beta_0 + \beta x + \varepsilon = \beta\,(x + \beta_0/\beta) + \varepsilon = \beta\,(x - \mu) + \varepsilon\ ,$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ with $\sigma^2 > 0$.

Let β be known, then the maximum likelihood estimate for μ is given by $\widehat{\mu} = -\widehat{\beta_0}/\beta$. For normal errors, like in our case, the maximum likelihood estimator for β_0 is the same as the least square estimator, namely

$$\widehat{\beta_0} = \overline{Y} - \beta\overline{X}\ .$$

The maximum likelihood estimator for μ based on the first n observations is then given by

$$\widehat{\mu}_n = \overline{X}_n - \frac{\overline{Y}_n}{\beta}\ .$$

Using the fact that $\widehat{\mu}_n = X_{n+1}$, this can be rewritten in a recursive shape as

$$\widehat{\mu}_n = \widehat{\mu}_{n-1} - \frac{1}{n\beta} Y_{n-1}$$

or alternatively as

$$X_{n+1} = X_n - \frac{1}{n\beta} Y_{n-1}\ .$$

This is a special case of the Robbins-Monro recursion with $a_n = (n\beta)^{-1}$. (cf. Lai and Robbins, 1979) In the case of unknown slope, where β is estimated by the maximum likelihood method, Wu (1986) established a connection to an adaptive RM procedure. The shape of the step length a_n is more complicated here. E.g. it is based on the previous observations.

Because of the relationship to the RM procedure, the convergence of the recursive maximum likelihood scheme follows from the convergence of the Robbins-Monro algorithm. (cf. Lai and Robbins, 1979)

In the two parameter logistic model simulations showed, that the performance of Wu's ML-scheme (Wu, 1985) turned out to be better with respect to the root mean squared error than the one of the RM-procedure, if the initial design assures the existence of the estimate. The simulations suggested, that the ML-recursion converges as well.

A proof of (almost sure) convergence of the sequential ML design for a class of location-scale models containing the two parameter logistic model was published by Ying and Wu (1997).

Denote the mean function of $Y(x)$ by $H(x-\mu)$ and the variance function by $V(x-\mu)$. Let $E(Y(\mu)) = H(0) = p$.

The sequential design is defined via the following equations

$$\sum_{i=1}^{n} \psi(X_i)\,(Y_i - H(X_i - \widehat{\mu}_n)) = 0 \tag{3}$$

$$X_{n+1} = \widehat{\mu}_n \tag{4}$$

where $\psi \geq 0$ is a weight function. Equation (3) is an estimating equation based on the first n observations, which becomes the maximum likelihood estimating equation if the ψ is chosen correctly. The second step defines the new design point as the estimate of the location parameter.

Ying and Wu (1997) introduced four sets of conditions to prove theorem 1, which is the main result of their article's second section:

(C1) H continuous, strictly increasing, and with probability one; the recursion defined by (3) and (4) is well defined for large n

(C2) for every $K > 0$,

$$\infty > \sup_{|t|<K} \psi(t+\mu)\,V(t) \geq \inf_{|t|<K} \psi(t+\mu)\,V(t) > 0$$

(C3) $\liminf_{|t|\to\infty} \frac{|H(t)-p|}{V(t)\psi(t+\mu)} > 0$ and $\liminf_{|t|\to\infty} \frac{|H(2t)-H(t)|}{V(t)\psi(t+\mu)} > 0$

(C4) V is continuous at 0, H is continuously differentiable in a neighborhood of 0 and $H'(0) > 0$.

For pairs of models and estimating equations satisfying (C1)-(C3) the sequential procedure yields an almost sure converging sequence of design points or the sequence diverges to infinity. The condition (C4) is needed for asymptotic normality.

Theorem 1 (cf. Ying and Wu, 1997) *Let X_i be the ith design point generated by* (3) *and* (4) *with $Y_i := Y(x_i)$. Suppose the corresponding mean, variance and weight functions H, V and ψ satisfy conditions* (C1)-(C3). *Define the disjoint events $A_\mu = \{X_n \to \mu\}$, $A_\infty = \{X_n \to \infty\}$ and $A_{-\infty} = \{X_n \to -\infty\}$.*

(i) $P(A_\mu \cup A_\infty \cup A_{-\infty}) = 1$. *In fact*

$$P\left(A_\mu | \sum_{n=1}^{\infty} \psi^2(X_n)\, V(X_n - \mu) = \infty\right) = 1$$

and $P(A_\infty \cup A_{-\infty} | \sum_{n=1}^{\infty} \psi^2(X_n)\, V(X_n - \mu) < \infty) = 1$,

(ii) If $\liminf_{t\to\infty} \psi^2(t)\,[1 + V(t-\mu)] > 0$ *then* $P(A_\infty) = 0$. *Similar for* $t \to -\infty$ *and* $P(A_{-\infty})$,

(iii) Suppose V and ψ satisfy the tail growth condition

$$\liminf_{|t|\to\infty} \psi^2\,(t)\,[1 + V(t-\mu)] > 0.$$

Then $X_n \longrightarrow \mu$ with probability one,

(iv) If $X_n \longrightarrow \mu$ with probability one and assumption (C4) *is also satisfied, then*

$$\sqrt{n}\,(X_n - \mu) \xrightarrow{\mathcal{D}} N\Big(0, (H'(0))^{-2}\, V(0)\Big)\,.$$

For the maximum likelihood equation in the two parameter logistic model with known β we have $\psi(t) \equiv 1$, $H(t) = F(t\beta)$, $V(t) = F(t\beta)\,(1 - F(t\beta))$. If we are interested in the location parameter μ then $p = 0.5$. While (C4) and the first part of (C1) are consequences of the definition of F, it follows directly, that (C2) is fulfilled, too, because ψ is constant and

$$0 < V(K) < V(t) < 1$$

for all $0 \le |t| < K$. For (C3) we note, that

$$\lim_{|t|\to\infty} |F(t\beta) - p| \ge \min\{p, 1-p\}$$

and $\lim_{|t|\to\infty} V(t) = 0$ which yields the first inequality.

The second inequality follows from the fact, that

$$\begin{aligned}\frac{|F(2t\beta) - F(t\beta)|}{V(t)\,\psi(t+\mu)} &= \frac{|1-\exp(-t\beta)|\,(1+\exp(-t\beta))}{1+\exp(-2t\beta)}\\ &= \frac{|1-\exp(-2t\beta)|}{1+\exp(-2t\beta)} = \frac{|\exp(2t\beta)-1|}{\exp(2t\beta)+1}\end{aligned}$$

$$\begin{aligned}&\Longrightarrow \lim_{t\to\infty} \frac{|F(2t\beta) - F(t\beta)|}{V(t)\,\psi\,(t+\mu)} = \lim_{t\to\infty} \frac{|1-\exp(-2t\beta)|}{1+\exp(-2t\beta)} = 1\\ \text{and} \quad &\lim_{t\to-\infty} \frac{|F(2t\beta) - F(t\beta)|}{V(t)\,\psi\,(t+\mu)} = \lim_{t\to-\infty} \frac{|\exp(2t\beta)-1|}{\exp(2t\beta)+1} = 1\ .\end{aligned}$$

Now that we have verified (C1) through (C4) for the two parameter logistic model, we can apply the theorem. From (C1) - (C3) and

$$\psi^2\,(t)\,[1+V(t-\mu)] = 1+V(t-\mu) > 1$$

we get the almost sure convergence of the sequence X_n. It follows that

$$\begin{aligned}\sqrt{n}\,(X_n-\mu) \xrightarrow{\mathcal{D}} &N\Big(0, (F'\,(0))^{-2}\,V(0)\Big)\\ &= N\left(0, \frac{F(0)\,(1-F(0))}{\beta^2 F(0)^2\,(1-F(0))^2}\right)\\ &= N\left(0, \frac{1}{\beta^2 p\,(1-p)}\right) = N\left(0, \frac{4}{\beta^2}\right)\ .\end{aligned}$$

Ying and Wu even showed that a constrained sequence of the estimates converges with probability one to the location parameter μ if the model is misspecified, in the sense that the mean function of the location model used for estimation is not the true one. The constraint mentioned above is, that one knows an interval $[a, b]$ containing μ. The estimates are restricted to this interval by changing (4) to

$$X_{n+1} = \max\{a, \min(\widehat{\mu}_n, b)\}\ . \tag{5}$$

To prove convergence Ying and Wu's approach is to transform the corresponding estimating equation into a Robbins-Monro shaped recursion. After that they applied the results of Robbins and Siegmund (1971) for convergence with probability one and results of Lai and Robbins (1979) for the asymptotic normality.

In addition to (5) they needed conditions for the misspecified mean function $\widetilde{H}$ to assure convergence with probability one ((C5)-(C7)) and asymptotic normality (C8).

(C5) $\widetilde{H}$ is twice continuously differentiable, strictly increasing and $\widetilde{H}(0) = p$,

(C6) There exists a $C > 0$ such that $\widetilde{H}(-C) < H(x-\mu) < \widetilde{H}(C)$ for all $x \in [a, b]$,

(C7) $\sup_{t\in[a,b]} E\left(|Y(t)|^4\right) < \infty$,

(C8) ψ is continuous at μ and $\widetilde{H}'(0) < 2H'(0)$.

If in our example, the two parameter logistic model, the β is known but it is impossible to adjust the experiment, e.g. the item, to the parameter or the β is unknown, and it is substituted by a constant b instead, then $\widetilde{H}(t) = F(tb)$ and (C5) is satisfied. Because x is restricted to a finite interval and C is not, (C6) is true. Condition (C7) follows because Y is a binary random variable. Hence almost sure convergence holds. Demanding that the slope of the "estimated" function $\widetilde{H}$ is not to large compared with the true one, (C8) reduces to a condition on the constant b:

$$\widetilde{H}'(0) = b\widetilde{H}(0)\left(1 - \widetilde{H}(0)\right) < 2H'(0) = 2\beta H(0)\left(1 - H(0)\right) \iff b < 2\beta\,.$$

Whenever b is small enough asymptotic normality holds. More precisely

$$\begin{aligned}\sqrt{n}\left(X_n - \mu\right) \xrightarrow{\mathcal{D}} N&\left(0, \frac{V(0)}{\widetilde{H}'(0)\left(2H'(0) - \widetilde{H}'(0)\right)}\right)\\ = N&\left(0, \frac{4}{b\left(2\beta - b\right)}\right)\,,\end{aligned}$$

i.e. the asymptotic variance is smallest for $b = \beta$.

Assuming, that a consistent estimator $\widehat{\beta}$ for β exists, that is plugged in into the estimating equation, results for location scale models with unknown parameter β are proven in the last part of Ying and Wu (1997). The Theorems and conditions are analog to the part concerning location models.

Even though the results from above yield consistency for the maximum likelihood estimate of μ for many typical models, in some cases of interest it does not work. Consider for example the three parameter logistic model with known α and β. Remember that

$$p = P(Y(\mu) = 1) = \frac{1+\alpha}{2} \; .$$

For the maximum likelihood equation we have

$$\psi(t) = \frac{1}{1 + \alpha \exp(-\beta\,(t-\mu))} \; .$$

As can be seen, the weighting function depends on μ, the parameter of interest. Hence we cannot apply the results from above directly, because the maximum likelihood estimating equation is not of the shape of (3). The estimating equation has to be altered, i.e. some kind of quasi likelihood can be used. If we consider ψ to be constant, say $(1+\alpha)^{-1}$, then (C1), (C2) and (C4) still hold, but the second part of (C3) does not. The problem is the negative half axis, where H tends to α and hence the variance to $\alpha\,(1-\alpha)$.

$$\Longrightarrow \frac{|H(2t) - H(t)|}{V(t)} \overset{t\to -\infty}{\longrightarrow} 0 \; .$$

Truncating, i.e. $\psi(t) := 0$ for all t smaller than some constant $c < \mu$, would violate (C2). But using

$$\psi(t) := \begin{cases} (1+\alpha)^{-1} \exp((t-c)\,\beta) & ,t < c \\ (1+\alpha)^{-1} & ,t \geq c \end{cases}$$

all conditions are fulfilled. Then by theorem 1 *(i)* and *(ii)* we get at least, that $\widehat{\mu}_n$ either converges to μ or tends to $-\infty$.

3 Extensions and related results

Joseph, Tian and Wu (2007) extended the approach of Ying and Wu (1997). They introduced a random error to the scale parameter β in the model, which originated from a Gaussian process. I.e. they used $(\beta + \epsilon(x))$ instead of β. Starting with the linear regression model from the beginning of section 2 the new model becomes

$$Y(x) = (\beta + \epsilon(x))\,(x-\mu) + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$. Additionally we assume $\text{cov}(\varepsilon, \epsilon(x)) = 0$ for all x and $\text{cov}(\epsilon(x_i), \epsilon(x_j)) = \tau^2 R(x_i, x_j)$ with covariance function R. The explanation for such a model is, that the variance

$$\text{Var}(Y(x)) = \tau^2 (x - \mu) + \sigma^2$$

becomes smaller for values next to μ. This results in more weight on values near the true value of μ. Even though this approach works quite nice in the simulations given in the article, there are some problems. First of all because of the computational issues, e.g. additional parameters have to estimated, this approach is recommended in situations only, where the computational costs are "low" compared to the costs for observations. The second is, that most of the paper is concerned with regression models using normal errors only. Because of the even more complex situation non-normal models are left "as a topic for future research". (Joseph, Tian and Wu (2007), p. 1556)

Starting with the one parameter logistic model Chang and Ying (2009) propose sequential designs, which are based on the maximum likelihood estimate as well. After an initial phase, they estimate μ. This first result is a direct consequence of Ying and Wu (1997). For the two parameter model they introduce the usual restriction, that β has to be in some known interval and bounded away from 0. In the three parameter case they approximate the maximum likelihood estimating equation and use this to simplify the estimation. Without the simplification they point out, that the maximum likelihood equation might have multiple roots. Additionally the (sequence of) guessing parameter(s) is assumed to be smaller then some positive constant $c < 1$. Up to the constraints from above the choice of the sequences of "estimates" for β and α is arbitrary. But it is proposed to start with relatively small values of β and to increase them during the trial. (cf. Chang and Ying, 2008)

All results from above, Joseph, Tian and Wu (2007) as well as Chang and Ying (2009), are concerned with almost sure convergence and asymptotic normality of $\widehat{\mu}_n$ or X_n, respectively.

A proof of convergence in probability was published recently by Tian (2009) for the two parameters of the 2-PL model. The starting point of his article is the consistency of the maximum likelihood estimator based on the observations from a sequential design plan. Even though the author describes two examples, the "Up-and-Down method" and a design based on the determinant of the estimated information matrix, the goal is to present more general sequential designs. I.e. while the articles from above are concerned with designs generated by maximum likelihood recursion, this one leaves unanswered how exactly the design is or should be generated. Beside conditions assuring the existence of the maximum likelihood estimator (cf. Silvapulle, 1981) Tian uses restrictions on the design space, the range of $\beta\,(x-\mu)$ and the sequence of design points X_i generated by the sequential design. (cf. Datta, 1995) His restrictions on the design sequence are, that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq c \quad \text{and} \quad \min_{w_1^2+w_2^2=1}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(w_1 X_i + w_2\right)^2\right\} \geq c^*$$

for $c, c^* > 0$ and all n larger than some $m \in \mathbb{N}$. The first condition assures, that the design points are not spreading to fast. This would be easily satisfied, if the design region is bounded. The second part seems to assure that region containing the design sequence is not to narrow.

4 A note on initial designs

One of the problems occurring in the above setting is the existence of the maximum likelihood estimate. In the case of the two parameter model Silvapulle (1981) gave conditions for the existence. Efficient initial designs are import, i.e. designs assuring the existence of the estimates after a relatively small number of observations.

With the goal to estimate μ, Wu (1985) used fixed sequences and the Robbins-Monro procedure as initial designs. In the latter case maximum likelihood estimation started when the estimate existed. After comparing the results a RM-procedure with "large“ step length is recommended. Especially in the case where little prior knowledge is given.

The initial design described by Chang and Ying (2009) uses a sequence of doses increasing, if the first observation is "negative“, i.e. $Y_1 = 0$. It ends as soon as a "positive“ response occurs for the first time. If the first observation is already "positive“, a decreasing sequence is used ending with first the "negative“ response. After that, maximum likelihood estimation starts. It is not specified how to determine the doses in the initial design. An approach like in Wu (1985) using Robbins-Monro to choose μ is possible.

Karvanen (2008) proposed a starting design related to bisection, like in numerical root finding. Starting with an interval, observations are taken at the middle point. If only non-responses ($Y = 0$) are observed, this is the lower bound of the new interval. If only responses are observed, it becomes the new upper bound. As soon as both 1's and 0's are observed at an interval's middle point additional observations are taken near it until the maximum likelihood estimate exists. This was compared using simulations to initial designs using fixed number of design points. Wu and Karvanen both observed situations where the estimate did not exist after the fixed design.

References

[1] Chang, H. and Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika 73*, 441-450

[2] Chang, H. and Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Annals of Statistics 37*, 1466-1488

[3] Datta, S. (1995). Consistency of the mle for a general sequential design problem. *Sankhyā 57*, 88-98

[4] Joseph, V.R., Tian, Y. and Wu, C.F.J. (2007). Adaptive designs for stochastic root-finding. *Statistica Sinica 17*, 1549-1565

[5] Karvanen, J. (2008). Efficient initial designs for binary response data. *Statistical Methodology 5*, 462-473

[6] Lai, T.L. (2003). Stochastic approximation. *Annals of Statistics 31*, 391-406.

[7] Lai, T.L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *Annals of Statistics 7*, 1196-1221.

[8] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics 22*, 400-407.

[9] Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In Rustagi, J.S. (ed.): Optimizing methods in statistics. *Academic Press*, 233-257.

[10] Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society, B 43*, 310-313.

[11] Tian, Y. (2009). Consistency of the maximum likelihood estimator and Bayesian estimator based on sequential sensitivity experiments. *Statistics and Probability Letters 79*, 728-732

[12] Wu, C.F.J. (1985). Efficient sequential designs with binary data. *J. Amer. Statist. Assoc. 80*, 974-984.

[13] Wu, C.F.J. (1986). Maximum likelihood recursion and stochastic approximation in sequential designs. In: van Ryzin, J. (ed.):Adaptive statistical procedures and related topics. *IMS Monograph Series* , 298-313.

[14] Ying, Z. and Wu, C.F.J. (1997). An asymptotic theory of sequential designs based on maximum likelihood recursions. *Statistica Sinica 7*, 75-91.