Alternating Least Squares in Generalized Linear Models

Arshak Minasyan

Yerevan State University

arsh.minasyan@gmail.com

February 23, 2022

Abstract

We derived a convergence result for a sequential procedure known as alternating maximization (minimization) to the maximum likelihood estimator for a pretty large family of models - Generalized Linear Models (GLMs). Alternating procedure for linear regression becomes to the well-known algorithm of Alternating Least Squares (ALS), because of the quadraticity of log-likelihood function $L(\boldsymbol{v})$. In GLMs framework we lose quadraticity of $L(\boldsymbol{v})$, but still have concavity due to the fact that error-distribution is from exponential family (EF). Concentration property makes the Taylor approximation of $L(\boldsymbol{v})$ up to the second order accurate and makes possible the use of alternating minimization (maximization) technique. Examples and experiments confirm convergence result followed by the discussion of the importance of initial guess.

1 Introduction

Many statistical tasks can be viewed as problems of semi-parametric estimation when the unknown data distribution is described by a high or infinite dimensional parameter while the target is of low dimension. Typical examples are provided by functional estimation, estimation of a function at a point, or simply by estimating a given sub-vector of the parameter vector. The classical statistical theory provides a general solution to this problem: estimate the full parameter vector by the maximum likelihood method and project the obtained estimate onto the target subspace. This approach is known as *profile maximum likelihood* and it appears to be *semi-parametrically efficient* under some mild regularity conditions, which in case of Generalized Linear Models are satisfied. For more general case, for example, in M-estimation framework these technical conditions should be introduced separately and checked whether they are fulfilled or not. We refer to [8], [6] and the book of Kosorok [10] for a detailed presentation.

This study revisits the problem of profile semi-parametric estimation (see [4], [5] and references therein). One issue that is worth mentioning is the model mis-specification. In most of the cases of practical problems, it is unrealistic to expect that the model assumptions are exactly fulfilled, even if some rich non-parametric family is used. This means that the true data distribution $\mathbb{P}$ does not belong to the considered parametric family, in our case — exponential family. Applicability of the general semi-parametric theory in such cases is questionable. An important feature of the presented approach is that it equally applies under a possible model mis-specification.

Consider the following statistical model that assumes the unknown data distribution $\mathbb{P}$ belongs to a given parametric family $(\mathbb{P}_{\boldsymbol{v}})$:

$$\boldsymbol{\mathcal{Y}} \sim \mathbb{P} = \mathbb{P}_{\boldsymbol{v}^*} \in (\mathbb{P}_{\boldsymbol{v}}, \boldsymbol{v} \in \Theta), \tag{1.1}$$

where Θ is some high dimensional or even infinite dimensional parametric space.

The maximum likelihood approach in the parametric estimation suggests to estimate the whole parameter vector $\boldsymbol{v}$ by maximizing the corresponding log-likelihood

$$L(\boldsymbol{v}) = \log \frac{d\mathbb{P}_{\boldsymbol{v}}}{d\boldsymbol{\mu}_0}$$

for some dominating measure $\boldsymbol{\mu}_0$. Define the maximum likelihood estimator in the following way

$$\widetilde{\boldsymbol{v}} \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{v} \in \Theta} L(\boldsymbol{v}). \tag{1.2}$$

Our study admits a model specification $\mathbb{P} \notin (\mathbb{P}_{\boldsymbol{v}}, \boldsymbol{v} \in \Theta)$. Equivalently, one can say that $L(\boldsymbol{v})$ is the *quasi log-likelihood function* on Θ. The *target* value $\boldsymbol{v}^*$ of the parameter $\boldsymbol{v}$ can be defined by

$$\boldsymbol{v}^* \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{v} \in \Theta} \mathbb{E} L(\boldsymbol{v}). \tag{1.3}$$

Under model misspecification, $\boldsymbol{v}^*$ defines the best parametric fit to $\mathbb{P}$ by the considered family. For the results of similar kind see [1]. Kneip first started the work of this direction by introducing ordered linear functionals (see [9]). For general results of alternating maximization (minimization) we refer to [2] and references therein.

The main point of the work is that the Alternating Method gives only a little gain, if any, in the complexity of optimum point computation for Linear Models (see [12]). Meanwhile, in

non-linear models the gain is pretty sensible. In non-linear models in most of the cases the closed form solution can not be obtained, in some cases even the numerical solutions of first order conditions could be very hard to implement in full parameter dimension. The technique known as alternating maximization (minimization) [3] helps in these situations and gives the estimation of parameter vector with adequate time complexity.

The model that we consider has the parameter $\boldsymbol{v}$ which is of dimension $p+q$, where p is the dimension of *target* parameter and q is the dimension of *nuisance* parameter. Usually p is not large, because we also care about tractability and interpretability of our model, but q can be very large, although it is a *nuisance* parameter, we can not ignore it and exclude from considered model. Main problems with direct computations occur in high dimensions, i.e. in cases when $p+q$ is large enough to make it impossible to invert matrix of sizes $(p+q)\times(p+q)$.

The alternating maximization procedure can be understood as a special case of the Expectation Maximization algorithm (EM algorithm). The EM algorithm is a popular algorithm first derived in [7]. Further on a number of modifications and extensions of this algorithm come into playground. In [7] it is also described how EM algorithm can be implement in different fields and give fruitful results. We refer to the [11] for the brief introduction to the development of EM algorithm. We restrict ourselves to citing the well-known convergence result (see [13]), which is still state of the art in most settings. Unfortunately, the result from [13] - as most convergence results on these interative procedures - only ensures convergence to some set of local maximizers or fix-points of the procedure. In this work we consider one of the special cases where it is possible to show the actual convergence of the method.

The work has the following structure. In Section 2 contains the preliminaries about Generalized Linear Models (GLMs) and class of random variables called Exponential Family (EF). Section 3 contains the main results of the paper. Section 4 illustrates how the algorithm of alternating least squares (ALS) works and confirms above theoretical results using both real and simulated data.

2 Introduction to Generalized Linear Models (GLMs)

In this section we introduce to Generalized Linear Models (GLMs) from a slightly different angle. This section gives background for further research on Alternating Method for GLMs in Section 3.

Let Y_i be independent random vectors and regressors $X_i \in \mathbb{R}^p$ and $Y_i \sim P_i \in (\mathcal{P}_{\boldsymbol{v}})$, which means that $\exists v_i : P_i = P_{v_i}$, where $(\mathcal{P}_{\boldsymbol{v}})$ we assume to be an exponential family of distributions with canonical parameter. Exponential family will be discussed in details further in this section. Generalized linear model has the following representation $Y_i \sim P_{v(X_i)}$. In the case of Gaussian distribution we instead get the following form $Y_i = v(X_i) + \varepsilon_i$ with arbitrary function $\boldsymbol{v}(\cdot)$.

Function $\boldsymbol{v}(x)$ can be represented in the following way

$$v(x) = \sum_{j=1}^{\infty} \boldsymbol{\theta}_j \psi_j(x).$$

Then, our linear parametric assumption is

$$v(x) = \sum_{j=1}^{p+q} \boldsymbol{\theta}_j \psi_j(x) = \sum_{j=1}^{p} \boldsymbol{\theta}_j \psi_j(x) + \sum_{j=p+1}^{p+q} \boldsymbol{\theta}_j \psi_j(x) \text{ for given basis } \psi_j(\cdot). \tag{2.1}$$

Denote $\boldsymbol{\eta}_i = \boldsymbol{\theta}_{p+i}$ and column-vector $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_q)^T \in \mathbb{R}^q$. We will mostly discuss the case of finite p and q.

$$Y_i \sim P_{v_i}, \;\; v_i = \Psi_i^T \boldsymbol{\theta} + \Phi_i^T \boldsymbol{\eta}, \tag{2.2}$$

where $\Psi_i = (\psi_1(x), \dots, \psi_p(x))^T \in \mathbb{R}^p$, $\Phi_i = (\psi_{p+1}(x), \dots, \psi_{p+q}(x))^T \in \mathbb{R}^q$ and $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\eta} \in \mathbb{R}^q$. Log-likelihood

$$\log \frac{dP_{\boldsymbol{v}}}{d\mu_0^n}(\mathcal{Y}) = \sum_{i=1}^{n} (v_i Y_i - g(\nu_i)) = \sum_{i=1}^{n} \left(\Psi_i^T \boldsymbol{\theta} Y_i + \Phi_i^T \boldsymbol{\eta} Y_i - g(\Psi_i^T \boldsymbol{\theta} + \Phi_i^T \boldsymbol{\eta}) \right),$$

which follows from the properties of exponential family, see (2.10). Function $g(\cdot)$ is derived from (2.8) when the distribution is fixed. Equivalently, one writes

$$L(\boldsymbol{\theta}, \boldsymbol{\eta}) \stackrel{\text{def}}{=} S^T \boldsymbol{\theta} + R^T \boldsymbol{\eta} - A(\boldsymbol{\theta}, \boldsymbol{\eta}), \tag{2.3}$$

where

$$S \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_i \Psi_i \in \mathbb{R}^p, \;\; R \stackrel{\text{def}}{=} \sum_{i=1}^{n} Y_i \Phi_i \in \mathbb{R}^q, \;\; A(\boldsymbol{\theta}, \boldsymbol{\eta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} g(\Psi_i^T \boldsymbol{\theta} + \Phi_i^T \boldsymbol{\eta}).$$

Denote log-likelihood function as $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ or $L(\boldsymbol{v})$, where $\boldsymbol{v} \stackrel{\text{def}}{=} (\boldsymbol{\theta}, \boldsymbol{\eta})$. For the ease of representation sometimes we use $\boldsymbol{v}$ instead of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ and vice versa.

For instance, in terms of $\boldsymbol{v}$ log-likelihood can be written in the following way:

$$L(\boldsymbol{v}) = \Upsilon^T \boldsymbol{v} - A(\boldsymbol{v}),$$

where $\Upsilon = \begin{pmatrix} S \\ R \end{pmatrix}^T \in \mathbb{R}^{p+q}$, the concatenation of above defined S and R. Fisher information matrix is defined by $\mathcal{D}^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{v}^*) = F(\boldsymbol{v}^*)$, where $\boldsymbol{v} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\eta} \end{pmatrix}^T \in \mathbb{R}^{p+q}$ and the Hessian matrix

$$\mathbb{F}(\boldsymbol{v}) = -\nabla^2 \mathbb{E} L(\boldsymbol{v}) = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{v}) & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{v}) \\ \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\theta}}(\boldsymbol{v}) & \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{v}) \end{pmatrix}.$$

Further we will prove that function $g(\cdot)$ is convex which gives us positively defined Fisher information matrix $\mathbb{F}$.

Also define score vector $\nabla \stackrel{\text{def}}{=} \nabla L(\boldsymbol{v}^*)$ as well as standardized score vector $\breve{\xi}$ as follows:

$$\breve{\xi} = \mathcal{D}^{-1} \nabla$$

Rewrite the definition (1.2) and (1.3) as follows

$$\widetilde{\boldsymbol{v}} = (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta}), \;\; \boldsymbol{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\eta}} \mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\eta}). \tag{2.4}$$

Parameters $\widetilde{\boldsymbol{v}} = (\widetilde{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\eta}})$ are data-dependent and change by the change of Y, while $\boldsymbol{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$ is not random (i.e. data-dependent), but are unavailable under real world conditions. In real world we know nothing about distribution of Y, nevertheless we make parametric assumptions on Y and always keep in mind that parametric assumption is probably wrong. From the definition of $\boldsymbol{v}^*$ follows that $\nabla \mathbb{E} L(\boldsymbol{v}^*) = 0$ which yields

$$\mathbb{E} \begin{pmatrix} S \\ R \end{pmatrix}^T = \nabla A(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$$

or equivalently

$$\mathbb{E} \Upsilon = \nabla A(\boldsymbol{v}^*),$$

An important feature of exponential family is that the stochastic component $\zeta(\boldsymbol{\theta}, \boldsymbol{\eta})$ of (log) likelihood function is linear in $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Put $\varepsilon_i = Y_i - \mathbb{E} Y_i$ and $\zeta = L - \mathbb{E} L$ then

$$\zeta(\boldsymbol{\theta}, \boldsymbol{\eta}) = (S^T - \mathbb{E} S^T)\boldsymbol{\theta} + (R^T - \mathbb{E} R^T)\boldsymbol{\eta} = \sum_{i=1}^n \varepsilon_i (\Psi_i^T \boldsymbol{\theta} + \Phi_i^T \boldsymbol{\eta}),$$

$$\nabla \zeta(\boldsymbol{\theta}, \boldsymbol{\eta}) = \begin{pmatrix} S - \mathbb{E} S & R - \mathbb{E} R \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \varepsilon_i \Psi_i & \sum_{i=1}^n \varepsilon_i \Phi_i \end{pmatrix}.$$

Now consider the following elliptic set

$$\Omega_\circ(r) \stackrel{\text{def}}{=} \{\boldsymbol{v} : \|\mathcal{D}(\boldsymbol{v} - \boldsymbol{v}^*)\| \le r\}. \tag{2.5}$$

Set $\Omega_\circ(r)$ is called local vicinity of $\boldsymbol{v}^*$ for $\mathcal{D}^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{v}^*) = \mathbb{F}(\boldsymbol{v}^*)$ and $\mathcal{V}^2 \stackrel{\text{def}}{=} \mathbb{C}ov(\nabla L(\boldsymbol{v}^*))$. Consider the covariance matrix in the block form:

$$\mathcal{V}^2 = \begin{pmatrix} V^2 & E \\ E^T & Q^2 \end{pmatrix}. \tag{2.6}$$

Now we want to understand the what these results yield for the estimator $\widetilde{\boldsymbol{\theta}}$ of the target parameter $\boldsymbol{\theta}^*$. Recall the block representation of the Fisher information matrix $\mathbb{F}(\boldsymbol{v}^*) = -\nabla^2 \mathbb{E} L(\boldsymbol{v}^*)$:

$$\mathbb{F}(\boldsymbol{v}) = \begin{pmatrix} \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{v}) & \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{v}) \\ \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\theta}}(\boldsymbol{v}) & \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{v}) \end{pmatrix}. \tag{2.7}$$

For the central point $\boldsymbol{v}^*$ the decomposition can be written in the following form

$$\mathcal{D}^2 = \mathbb{F}(\boldsymbol{v}^*) = \begin{pmatrix} D^2 & A \\ A^T & H^2 \end{pmatrix},$$

where $D^2 = \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{v}^*)$, $A = \mathbb{F}_{\boldsymbol{\theta}\boldsymbol{\eta}}(\boldsymbol{v}^*)$ and $H^2 = \mathbb{F}_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{v}^*)$. Decompose also the score vector $\nabla \stackrel{\text{def}}{=} \nabla L(\boldsymbol{v}^*)$ as follows

$$\nabla = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix}.$$

2.1 Exponential Family with Canonical Parameter (EFc)

In this part of our work we extend the scope of our modeling toolbox to accommodate a variety of additional data types, including counts and rates. We introduce the exponential family of distributions. It is worth mentioning that exponential family is pretty wide class of random variables and covers the vast majority of possible reasonable error distributions. It includes the Gaussian, binomial, Poisson, gamma, multinomial and many others. The simplest examples of distributions which are not from exponential family are Uniform distribution and Student's t distribution. The beauty and elegance of exponential family is the fact that the (log) likelihood function has a simple form and can be written explicitly.

In general we say that the random variable with probability density function $f(\cdot)$ is from EFc, if the probability density function could be expressed in the following way

$$f(x|\boldsymbol{\eta}) = h(x)\exp\{\boldsymbol{\eta}^T T(x) - A(\boldsymbol{\eta})\} \tag{2.8}$$

where $\boldsymbol{\eta}$ from l.h.s. is a given measure and the vector parameter $\boldsymbol{\eta}$ from r.h.s. is referred to as the canonical parameter. $T(X)$ is a *joint sufficient statistics* and we call function $A(\cdot)$ as *link function*.

As stated above there is a huge number of famous distributions the probability density functions of which could be expressed in the form (2.8).

2.1.1 GLM with error distribution from EFc

Let Y be response vector of two sets of factors: Ψ and Φ. Consider the following model

$$Y \sim P \in (\mathcal{P}_{\boldsymbol{v}})_{v\in\mathbb{R}} \ll \mu_0, \tag{2.9}$$

where $(\mathcal{P}_{\boldsymbol{v}})_{v\in\mathbb{R}}$ is an exponential family with canonical parameter (EFc) and μ_0 is some dominating measure. Hence,

$$\log\frac{dP_{\boldsymbol{v}}}{d\mu_0} = y\cdot v - g(\boldsymbol{v}) \text{ for some function } g. \tag{2.10}$$

Lemma 2.1 (see []). *Let $(\mathcal{P}_\nu)$ be EFc. Then, it holds.*

$$\mathbb{E}_\nu Y = g'(\nu),\ \ \mathbb{V}ar_\nu(Y) = \mathbb{E}_\nu[Y - g'(\nu)]^2 = g''(\nu) \tag{2.11}$$

Hence, function $g(\cdot)$ is convex.

Remark 2.1. The above proof is provided for univariate functions $g(\cdot)$, but generally it is also true for multivariate functions. The only difference will be in the equation (**??**), instead of variance there will be a covariance matrix, which will be positive-defined.

2.2 Concentration

Recall that the following properties hold true for GLMs: the stochastic component $\zeta(\boldsymbol{\theta},\boldsymbol{\eta})$ is linear in both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, and the deterministic part $\mathbb{E}L(\boldsymbol{v})$ is concave in $\boldsymbol{v}$. Consider the following elliptic set defined in (2.5). In [12] it is proved that there is a elliptic set $\Omega_\circ(\mathrm{r})$ around oracle $\boldsymbol{v}$ such that

$\Omega_\circ(\mathrm{r})$ contains the estimator $\widetilde{\boldsymbol{v}}$ with high probability. Further on we assume a sufficiently large value of $\boldsymbol{x}$ to be fixed. It determines the level of overwhelming probability: a generic random set $\Omega_0(\boldsymbol{x})$ is of dominating probability if

$$\mathbb{P}\left(\Omega_0(\boldsymbol{x})\right) \geq 1 - \boldsymbol{\mathcal{C}} e^{-\boldsymbol{x}},$$

for some absolute constant $\boldsymbol{\mathcal{C}} > 0$.

The value of $\boldsymbol{x}$ may depend on sample size n and grows to infinity with n. The possible choices of $\boldsymbol{x}$ are $\boldsymbol{x} \asymp n^{1/2}$ and $\boldsymbol{x} \asymp \log n$, which would entail that $\mathbb{P}\left(\Omega_0(\boldsymbol{x})\right) \geq 1 - \boldsymbol{\mathcal{C}}/n$. We only require that sequence $\{x_n\}$ is not too large, more precisely, $\boldsymbol{x} \leq \boldsymbol{x}_c \stackrel{\text{def}}{\asymp} n^{1/2}$.

The way of construction and the exact probability of desired elliptic set is a very technical issue, hence omitted. All the results obtained below are considered to be true on the random set $\Omega_\circ(\boldsymbol{x})$ of high probability.

2.3 Local Quadratic Approximation of the Log-likelihood

Recall that function $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ could be rewritten in terms of the whole parameter vector $\boldsymbol{v}$ as $L(\boldsymbol{v})$. In this part of our work we will approximate our log-likelihood function $L(\boldsymbol{v})$ with quadratic form in a local elliptic set of optimum $\boldsymbol{v}^*$. Put $L(\boldsymbol{v}_1, \boldsymbol{v}_2) = L(\boldsymbol{v}_1) - L(\boldsymbol{v}_2)$ and recall that $\widetilde{\boldsymbol{v}}$ is data-dependent maximum of the likelihood function $L(\cdot)$, more precisely,

$$\widetilde{\boldsymbol{v}} = \arg\max_{\boldsymbol{v}} L(\boldsymbol{v}), \quad \boldsymbol{v}^* = \arg\max_{\boldsymbol{v}} \mathbb{E} L(\boldsymbol{v}).$$

Then, using Taylor expansion for function $L(\cdot)$ up to second order around the point $\boldsymbol{v}^*$ yields

$$L(\boldsymbol{v}, \boldsymbol{v}^*) = \nabla L(\boldsymbol{v}^*)(\boldsymbol{v} - \boldsymbol{v}^*) - \frac{1}{2}\|\mathcal{D}^2(\boldsymbol{v} - \boldsymbol{v}^*)\|^2 + \alpha'(\boldsymbol{v}, \boldsymbol{v}^*), \tag{2.12}$$

where $\alpha'(\cdot)$ is defined in (2.12).

Analogically, we approximate our function $L(\boldsymbol{v})$ around $\widetilde{\boldsymbol{v}}$ together with the fact that $\nabla L(\widetilde{\boldsymbol{v}}) = 0$, which gives us the following expression:

$$L(\boldsymbol{v}, \widetilde{\boldsymbol{v}}) = -\frac{1}{2}\|\mathcal{D}^2(\boldsymbol{v} - \widetilde{\boldsymbol{v}})\|^2 + \alpha(\boldsymbol{v}, \widetilde{\boldsymbol{v}}). \tag{2.13}$$

Remark 2.2. In order to apply Taylor expansion we use the concentration property, which is crucial, because it helps us to claim that this expansion adequately represents the original function $L(\cdot)$. The idea comes from a well-known fact that every concave function is quadratic around its maximizer, hence it can be expanded in the Taylor series up to the second order without any significant loses.

3 Alternating Least Squares for GLMs

This chapter specifies the previously obtained general results in linear models to the case of a Generealized Linear Models (GLMs). The question is non-trivial because we can not obtain direct estimates and closed forms and we loose the quadraticity of log-likelihood function. Instead we will approximate our likelihood by a quadratic form in a set where the measure concentration is

observed. Generalized Linear Models are frequently used in many areas and applications including categorical data analysis, classification problem, Poisson and Binary regressions, statistical learning and density estimation. GLMs are popular in economics, because of its flexibility to various types of models. In economics sometimes linear regression is not enough to fully explain the phenomenon and deeply analyze the situation. Using GLMs partially solves this problem, but it makes problem from the semiparametric family, which is the payment for more reliable and accurate results. The restrictions of GLMs are not very strict, so that it can be easily adopted in various models.

Here we will discuss alternating method for the (quasi) likelihood function obtained in previous section. In general the alternating maximization (minimization) procedure is used in cases when the direct full dimension computations are not feasible or simply very difficult to implement.

Recall the log-likelihood function $\mathcal{L}(\boldsymbol{v})$, where vector $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta})$ can be decomposed to the *target* parameter $\boldsymbol{\theta}$ and *nuisance* parameter $\boldsymbol{\eta}$. Alternating maximization is an iterative procedure starting from initial value $\boldsymbol{v}^\circ \in \mathbb{R}^{p+q}$ and updating iteratively in the way shown below

$$\begin{aligned} \widetilde{\boldsymbol{v}}_{k,k} &\stackrel{\text{def}}{=} (\widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\eta}}_k) = \left(\widehat{\boldsymbol{\theta}}_k, \operatorname*{argmax}_{\boldsymbol{\eta} \in \mathbb{R}^q} L(\widehat{\boldsymbol{\theta}}_k, \boldsymbol{\eta}) \right), \\ \widetilde{\boldsymbol{v}}_{k+1,k} &\stackrel{\text{def}}{=} (\widehat{\boldsymbol{\theta}}_{k+1}, \widehat{\boldsymbol{\eta}}_k) = \left(\operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^p} L(\boldsymbol{\theta}, \widehat{\boldsymbol{\eta}}_k), \widehat{\boldsymbol{\eta}}_k \right). \end{aligned} \tag{3.1}$$

In this section we will try to answer to some natural questions that arise with the iterative procedure described above: Does the sequence $\left(\widehat{\boldsymbol{\theta}}_k\right)$ converge? And if the answer is yes, what is the convergence rate? What are the conditions under which the sequence actually converges to the global maximizer $\widetilde{\boldsymbol{v}}$?

3.1 Convergence to Likelihood Estimator

The following theorem is one of the main results about the convergence of alternating least squares for generalized linear models.

Theorem 3.1. *Suppose a model given by (2.2) and put $\boldsymbol{v} = (\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathbb{R}^{p+q}$. $L(\boldsymbol{v})$ is defined in (2.3). Let $\mathcal{D}^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{v}^*)$ be the Hessian matrix for the log-likelihood function $L(\boldsymbol{v})$ with the following block-matrix representation $\begin{pmatrix} D^2 & A \\ A^T & H^2 \end{pmatrix}$ at point $\widetilde{\boldsymbol{v}}$. Assume that $\rho \stackrel{\text{def}}{=} \|D^{-1} A H^{-1}\|^2 < 1$ (here and afterwards the norm $\|\cdot\|$ with matrix argument denotes as spectral norm of matrix) and the condition $\|\mathcal{D}^{-1} \nabla^2 \mathbb{E} L(\boldsymbol{v}) \mathcal{D}^{-1} - I_{p+q}\| \le \delta(r) \le \boldsymbol{\delta}$ holds, where I_{p+q} is an identity matrix of size $p+q$.*

Then the sequence of estimators $\left(\widehat{\boldsymbol{\theta}}_k\right)$ obtained by alternating least squares converges to $\widetilde{\boldsymbol{\theta}} = \Pi_{\boldsymbol{\theta}} \widetilde{\boldsymbol{v}} \stackrel{\text{def}}{=} \Pi_{\boldsymbol{\theta}} \arg\max_{\boldsymbol{v}} L(\boldsymbol{v})$. $\Pi_{\boldsymbol{\theta}}$ is the projection of full vector to its sub-vector $\boldsymbol{\theta}$.

Remark 3.1. The notation $\widetilde{v}_{k(+1),k}$ is used in cases when the results are true for both $\widetilde{v}_{k,k}$ and $\widetilde{v}_{k+1,k}$.

Proof. Writing (2.13) in terms of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ we get

$$L(\boldsymbol{v}, \widetilde{\boldsymbol{v}}) =$$

$$-\frac{1}{2}\|D^2(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})\|^2-\frac{1}{2}\|H^2(\boldsymbol{\eta}-\widetilde{\boldsymbol{\eta}})\|^2-(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}})^TA(\boldsymbol{\eta}-\widetilde{\boldsymbol{\eta}})+\alpha(\boldsymbol{v},\widetilde{\boldsymbol{v}}). \tag{3.2}$$

Starting with an initial guess $\boldsymbol{\theta}^\circ$ and using the method described in (3.1) we get

$$\widehat{\boldsymbol{\eta}}_0=\widetilde{\boldsymbol{\eta}}(\boldsymbol{\theta}^\circ)=$$

$$\operatorname*{argmin}_{\boldsymbol{\eta}}\left[\frac{1}{2}\|D^2(\boldsymbol{\theta}^\circ-\widetilde{\boldsymbol{\theta}})\|^2+\frac{1}{2}\|H^2(\boldsymbol{\eta}-\widetilde{\boldsymbol{\eta}})\|^2+(\boldsymbol{\theta}^\circ-\widetilde{\boldsymbol{\theta}})^TA(\boldsymbol{\eta}-\widetilde{\boldsymbol{\eta}})+\alpha(\boldsymbol{v},\widetilde{\boldsymbol{v}})\right].$$

Hence, the first order condition gives us the following relationship:

$$H^2(\widehat{\boldsymbol{\eta}}_0-\widetilde{\boldsymbol{\eta}})=A^T(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^\circ)+\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{0,0},\widetilde{\boldsymbol{v}}).$$

Analogically, the solution for $\widehat{\boldsymbol{\theta}}_1\overset{\text{def}}{=}\widetilde{\boldsymbol{\theta}}(\widehat{\boldsymbol{\eta}}_0)$ has the following form

$$D^2(\widehat{\boldsymbol{\theta}}_1-\widetilde{\boldsymbol{\theta}})=A(\widetilde{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_0)+\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{1,0},\widetilde{\boldsymbol{v}}).$$

Then the iterative process of alternating maximization gives us the following recursive system of equations dependent on the initial guess:

$$\begin{cases}H^2(\widehat{\boldsymbol{\eta}}_k-\widetilde{\boldsymbol{\eta}})=A^T(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_k)+\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k},\widetilde{\boldsymbol{v}})\\ D^2(\widehat{\boldsymbol{\theta}}_{k+1}-\widetilde{\boldsymbol{\theta}})=A(\widetilde{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_k)+\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k+1,k},\widetilde{\boldsymbol{v}}).\end{cases}$$

or, equivalently

$$\begin{cases}H^2(\widehat{\boldsymbol{\eta}}_k-\widetilde{\boldsymbol{\eta}})=A^T(\widetilde{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}_k)+\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k},\widetilde{\boldsymbol{v}})\\ D(\widehat{\boldsymbol{\theta}}_{k+1}-\widetilde{\boldsymbol{\theta}})=D^{-1}A(\widetilde{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_k)+D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k+1,k},\widetilde{\boldsymbol{v}}).\end{cases} \tag{3.3}$$

Then we express $(\widetilde{\boldsymbol{\eta}}-\widehat{\boldsymbol{\eta}}_k)$ from the second equation of (3.3) and plug into the first one, which yields

$$D(\widehat{\boldsymbol{\theta}}_{k+1}-\widetilde{\boldsymbol{\theta}})=D^{-1}AH^{-2}A^T(D^{-1}D)(\widehat{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}})+D^{-1}\left[\nabla_{\boldsymbol{\theta}}\alpha(\boldsymbol{v},\widetilde{\boldsymbol{v}})-AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\boldsymbol{v},\widetilde{\boldsymbol{v}})\right],$$

then defining $M_\circ\overset{\text{def}}{=}D^{-1}AH^{-2}A^TD^{-1}$ as follows and

$$\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})\overset{\text{def}}{=}D^{-1}\left[\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k+1,k},\widetilde{\boldsymbol{v}})-AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k},\widetilde{\boldsymbol{v}})\right]$$

gives us the following recursive formula

$$D(\widehat{\boldsymbol{\theta}}_{k+1}-\widetilde{\boldsymbol{\theta}})=M_\circ\cdot D(\widehat{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}})+\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k}). \tag{3.4}$$

Hence, summing up for all k starting from initial guess, taking the norm and using triangle inequality leads us to the following result:

$$\|D(\widehat{\boldsymbol{\theta}}_{k+1}-\widetilde{\boldsymbol{\theta}})\|\le\|M_\circ\|\cdot\|D(\widehat{\boldsymbol{\theta}}_k-\widetilde{\boldsymbol{\theta}})\|+\|\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})\|\le\|M_\circ\|^k\cdot\|D(\boldsymbol{\theta}^\circ-\widetilde{\boldsymbol{\theta}})\|+$$
$$\sum_{\ell=0}^{k-1}\|M_\circ\|^\ell\cdot\|\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})\|=\|M_\circ\|^k\cdot\|D(\boldsymbol{\theta}^\circ-\widetilde{\boldsymbol{\theta}})\|+\|\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})\|\cdot\frac{1-\|M_\circ\|^k}{1-\|M_\circ\|}.$$

Now using the assumption $\rho\overset{\text{def}}{=}\|D^{-1}AH^{-1}\|<1$ it is easy to see that the first term vanishes to zero when k goes to infinity.

We aim to show that the euclidean norm of $\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})$ adequately decreases with the rise of the number of iterations k.

Considering the term $D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k+1,k},\widetilde{\boldsymbol{v}})$ we refer to the Theorem C.1 in Andresen and Spokoiny (2013) which gives a vanishing in k bound for it. Note that using that theorem we can bound two remainder terms in $\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})$, which will give the bound for $\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})$ as a whole.

Hence, using the Theorem from Andresen and Spokoiny (2013) and triangle inequality gives the claim about $\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})$. More formally,

$$\|\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})\| \le \|D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k+1,k},\widetilde{\boldsymbol{v}})\| + \|D^{-1}AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k},\widetilde{\boldsymbol{v}})\| \to 0 \text{ as } k\to\infty.$$

Combining all provided properties we derive that

$$\|D(\widehat{\boldsymbol{\theta}}_{k+1}-\widetilde{\boldsymbol{\theta}})\| \le \boldsymbol{s}_k \to 0 \text{ as } k\to\infty. \tag{3.5}$$

With the same argument it is clear how to derive the convergence property for *nuisance* parameter $\boldsymbol{\eta}$, which gives the claim of the theorem. □

Remark 3.2. The sequence $\boldsymbol{s}_k$ from (3.5) can be interpreted as the radius of the elliptic set around $\widetilde{\boldsymbol{\theta}}$ in which the estimator $\widehat{\boldsymbol{\theta}}_{k+1}$ derived by alternating least squares lies with high probability.

Remark 3.3. The above result presented in Theorem (3.1) tells us that we only need conditions $\rho \stackrel{\text{def}}{=} \|M_\circ\| < 1$ and $\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(\boldsymbol{v})\mathcal{D}^{-1} - I_{p+q}\| \le \delta$ to claim that the sequence $\left(\widehat{\boldsymbol{\theta}}_k\right)$ derived by alternating maximization method will linearly converge to ML estimator $\widetilde{\boldsymbol{\theta}}$, which is very often too difficult to compute directly.

3.2 Alternating Estimator

Above this section we have showed that the sequence of estimators obtained by using alternating least squares technique converges to the corresponding maximum likelihood estimator. It is a very strong and practically important result. As it has been discussed above, there are two main issues that make the problem non-trivial. The first one is impossibility of deriving closed form solution in most of cases and the second one is the high dimension of nuisance parameter which makes direct implementation of well-known Newton-Raphson method impossible. Alternating maximization technique overcame these issues and delivered estimators "close" to the maximum likelihood estimator.

This part is a bit technical and the only practical value is the statistical property of $D(\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)$.

Further in this section we will show that alternating estimator is also close to the true (unknown) parameter $(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$. Recall

$$\boldsymbol{v}^* = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{v}} \mathbb{E}L(\boldsymbol{v}).$$

The theorem below is known as a Fisher expansion and we formulate it in the framework of GLMs. Recalling the definitions given in the previous section we are now ready to formulate and prove Fisher expansion.

Theorem 3.2. *Consider the conditions of Theorem (3.1) and define* $\breve{\xi} \stackrel{\text{def}}{=} D^{-1}\breve{\nabla}$, *where* $\breve{\nabla} = \nabla_{\boldsymbol{\theta}} - AH^{-2}\nabla_{\boldsymbol{\eta}}$.

Then we have

$$\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{\xi}\| \to 0, \ \text{as } k \to \infty.$$

Proof. The proof is based on the Theorem 1 proof's ideas with the only difference that the Taylor expansion is taken around point $\boldsymbol{v}^*$.

First, we use the first order conditions to obtain the following relations

$$\begin{cases} D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) = D^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{v}^*) - D^{-1}A(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*) + D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*) \\ H(\widetilde{\boldsymbol{\eta}}_k - \boldsymbol{\eta}^*) = H^{-1}\nabla_{\boldsymbol{\eta}} L(\boldsymbol{v}^*) - H^{-1}A^T(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) + H^{-1}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k-1,k}, \boldsymbol{v}^*) \end{cases} \tag{3.6}$$

Again referring to Andreas and Spokoiny (2013) result we can bound the norm of $D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*)$ and $H^{-1}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k-1,k}, \boldsymbol{v}^*)$. From the system (3.6) one can easily derive

$$\begin{aligned} D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) = D^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{v}^*) - D^{-1}[AH^{-2}\nabla_{\boldsymbol{\eta}} L(\boldsymbol{v}^*) - AH^{-2}A^T(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) + \\ AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k-1,k}, \boldsymbol{v}^*)] + D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*) \implies \end{aligned}$$

$$\begin{aligned} D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) = M_\circ D(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) + D^{-1}\left[\nabla_{\boldsymbol{\theta}} L(\boldsymbol{v}^*) - AH^{-2}\nabla_{\boldsymbol{\eta}} L(\boldsymbol{v}^*)\right] + \\ D^{-1}\left\{\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*) - AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k-1,k}, \boldsymbol{v}^*)\right\}. \end{aligned} \tag{3.7}$$

Recalling the definitions from Section 3 we notice the term $\breve{\xi}$ as the second term of l.h.s. in (3.7), i.e. it can be rewritten in this way

$$D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{\xi} = M_\circ D(\widetilde{\boldsymbol{\theta}}_{k-1} - \boldsymbol{\theta}^*) + D^{-1}\left\{\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*) - AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k-1,k}, \boldsymbol{v}^*)\right\}$$

Summing up for all k starting from the initial guess after taking the norm of both sides and using the assumption that $\rho = \|M_\circ\| < 1$ leads us to the following result

$$\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{\xi}\| \le \boldsymbol{\mathcal{C}}(\rho) \cdot D^{-1}\left\{\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*) - AH^{-2}\nabla_{\boldsymbol{\eta}}\alpha(\widetilde{\boldsymbol{v}}_{k-1,k}, \boldsymbol{v}^*)\right\}, \tag{3.8}$$

where $\boldsymbol{\mathcal{C}}(\rho)$ is some constant depending on ρ and now, once again, we refer to the same Theorem C.1 from Andersen and Spokoiny (2013) to bound the term on the l.h.s. with the sequence which vanishes to zero when k goes to infinity.

Remember that, in order to obtain such bound for $D^{-1}\nabla_{\boldsymbol{\theta}}\alpha(\widetilde{\boldsymbol{v}}_{k,k}, \boldsymbol{v}^*)$ we need the condition $\|\mathcal{D}^{-1}\nabla^2\mathbb{E}L(\boldsymbol{v})\mathcal{D}^{-1} - I_{p+q}\| \le \delta$ for some constant $\delta > 0$.

Finally,

$$\|D(\widetilde{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*) - \breve{\xi}\| \le \breve{\boldsymbol{s}}_k \to 0, \ \text{as } k \to \infty. \tag{3.9}$$

□

Note that the expectation of random variable $\breve{\xi}$ is equal to zero and moreover $\mathbb{V}ar(\breve{\xi}) = D^{-1}V^2D^{-1}$.

Remark 3.4. It is worth mentioning that in case of finite k Fisher expansion is still working, but the corresponding norm is bounded with the zero-vanishing sequence $\breve{\boldsymbol{s}}_k$, instead of exact zero.

Remark (3.5) comments about the distribution of $\breve{\xi}$ both asymptotically and in finite sample size.

Remark 3.5. Random variable $\breve{\xi} \in \mathbb{R}^p$ under the correct model specification has normal distribution, hence $\|\breve{\xi}\|^2$ has the distribution of χ^2_p with p degrees of freedom. In case of model misspecification we can say nothing about the distribution of $\|\breve{\xi}\|^2$ in non-asymptotic framework, but asymptotically it still has χ^2_p distribution with p degrees of freedom.

3.3 Importance of Initial Guess

The initial guess can play a crucial role in the convergence of alternating method and if we "succeed" with it then the gain is twofold. Firstly, we can come up with weaker conditions than in Theorem (3.1) and secondly, the number of iterations will be significantly less than in case of "bad" initial point. Good initial values of $\boldsymbol{\theta}^\circ$ mostly appeal to the first condition of Theorem (3.1), i.e. $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$.

Denote the vector space of eigenvectors corresponding to the eigenvalues greater than 1 of matrix $M_\circ$ as $\mathcal{V}$, i.e. $\mathcal{V} \stackrel{\text{def}}{=} \operatorname{span}(v_1, v_2, \ldots, v_l)$, where

$$M_\circ v_i = \lambda_i v_i, \forall i \in \{1, \ldots, l\} \quad \text{and} \quad |\lambda_i| \geq 1.$$

Lemma 3.3. *Then if the initial value of $\boldsymbol{\theta}^\circ$ is chosen such that it holds*

$$\boldsymbol{u} \perp \mathcal{V}, \tag{3.10}$$

where $\boldsymbol{u} \stackrel{\text{def}}{=} D(\boldsymbol{\theta}^\circ - \widetilde{\boldsymbol{\theta}})$ then we will have the convergence mentioned above.

The proof of this lemma is straightforward based on simple linear algebra. Nevertheless, we will briefly explain the idea. Note that the assumption $\rho \stackrel{\text{def}}{=} \|D^{-1}AH^{-1}\|^2 < 1$ means that all eigenvalues are inside the unit circle so that the process will converge. It is true independent of the initial guess. Nevertheless, we can weakened this result by the "good" choice of initial values. If the matrix $M_\circ$ has eigenvalues which are outside the unit circle then the initial guess could help to vanish them to zero being just orthogonal to the space of the corresponding eigenvectors.

It is also worth mentioning that the condition of Theorem (3.1)

$$\|\mathcal{D}^{-1}\nabla^2 \mathbb{E}L(\boldsymbol{v})\mathcal{D}^{-1} - I_{p+q}\| \leq \delta(r) \leq \boldsymbol{\delta}$$

has nothing to do with initial values and only needs to bound the remainder term $\Xi(\widetilde{\boldsymbol{v}}_{k(+1),k})$. Thus, this condition can not be weakened depending on the initial point $\boldsymbol{\theta}^\circ$.

4 Numerical Example

In this section, we introduce and implement an algorithm in case of a toy example. Suppose, there are n coins on the table and someone randomly chooses a coin and start flipping k times. For the ease of illustration and implementation we further assume that there are only two coins. Suppose the first coin has a probability of "head" outcome of p_1 and the second coin — p_2. We introduce one *nuisance* parameter π which is the probability of choosing the first coin.

Assuming there are two coins on the table and one randomly chooses one of them 5 times and after each choice makes 10 flips. So, in total, we have 50 outcomes, 2 *target* parameters p_1 and p_2 and one nuisance parameter π.

We describe the algorithm on this simple example and provide the results of the code which computes the estimates of parameters in each step.

We have the following information:

$$\begin{aligned}
\text{Coin 2:}&H,T,H,T,T,H,T,H,H,T.\\
\text{Coin 1:}&H,H,H,H,H,H,T,H,H,H.\\
\text{Coin 1:}&H,T,H,T,H,H,H,H,H,H.\\
\text{Coin 2:}&H,T,H,T,T,T,T,H,H,T.\\
\text{Coin 1:}&H,H,H,T,H,H,T,H,H,T.
\end{aligned}$$

Clearly,

$$\widetilde{p}_1 = \frac{24}{24+6} = 0.8 \qquad \widetilde{p}_2 = \frac{9}{9+11} = 0.45.$$

However, these information is not available for the algorithm. The algorithm starts with some initial values of probabilities, for example, $p_1^\circ = 0.6$ and $p_2^\circ = 0.5$. Recall that in this example $\boldsymbol{\theta}$ from the previous section is the vector (p_1, p_2) and $\boldsymbol{\eta}$ is for π. The estimated probability of k out of 10 tosses of coin $c \in \{1, 2\}$ yielding heads is

$$p_c(k) = C_k^{10} p_c^k (1-p_c)^{10-k}. \tag{4.1}$$

Note that the binomial coefficient is the same for both coins, so it cancels out and only the ratio of the remaining factors determines the result. Using the initial guesses and formula (4.1) we come up with the following table

First iteration			
π	$1-\pi$	Coin 1	Coin 2
0.45	0.55	$\approx 2.2H, 2.2T$	$\approx 2.8H, 2.8T$
0.80	0.20	$\approx 7.2H, 0.8T$	$\approx 1.8H, 0.2T$
0.73	0.27	$\approx 5.9H, 1.5T$	$\approx 2.1H, 0.5T$
0.35	0.65	$\approx 1.4H, 2.1T$	$\approx 2.6H, 3.9T$
0.65	0.35	$\approx 4.5H, 1.9T$	$\approx 2.5H, 1.1T$
		$\approx 21.3H, 8.6T$	$\approx 11.7H, 8.4T$

So, hence, in the next stage we get the following estimators of p_1 and p_2:

$$\widehat{p}_1^{(1)} = \frac{21.3}{21.3+8.6} = 0.71 \qquad \widehat{p}_2^{(1)} = \frac{11.7}{8.4+11.7} = 0.58. \tag{4.2}$$

The number of iterations until convergence depends on the tolerance level, which is the acceptable bias from the ML estimators. In other words, it defines the stopping criteria of the algorithm. At given initial values $p_1^\circ = 0.6$ and $p_2^\circ = 0.5$ of parameters p_1 and p_2 we get the following (see Table

Table 1: Convergence with random and non-random initial values

Iteration	Non-random		Random	
	$\widehat{p}_1^{(i)}$	$\widehat{p}_2^{(i)}$	$\widehat{p}_1^{(i)}$	$\widehat{p}_2^{(i)}$
1	0.600000000	0.500000000	0.935954410	0.0659086863
2	0.713012235	0.581339308	0.759177699	0.434881703
3	0.745292036	0.569255750	0.78052855	0.485752725
4	0.768098834	0.549535914	0.79025212	0.505705724
5	0.7831645	0.534617454	0.794205962	0.513867055
6	0.791055245	0.52628116	0.795774011	0.517227986
7	0.794532537	0.522390437	0.79639082	0.518613125
8	0.79592866	0.520729878	0.796632802	0.519183793
9	0.796465637	0.520047189	0.796727694	0.519418794
10	0.796668307	0.519770389	0.796764932	0.519515523
11	0.796744149	0.519658662	0.796779561	0.51955532
12	0.796772404	0.519613607	0.796785317	0.519571692
13	0.796782900	0.519595434		

1) sequence of estimated parameters. The table also contains the iterations of the algorithm with randomly chosen initial values from uniform distribution at certain seed.

In both cases we see that it converges to the ML estimator with high accuracy.

Remark 4.1. Note that the coins can be easily extended to any distribution. For example, the first part of observed data can be from one Gaussian distribution and the second part is from another Gaussian distribution. Formally, $X_1, \dots, X_\ell \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_{\ell+1}, \dots, X_n \sim \mathcal{N}(\mu_2, \sigma_2^2)$. The idea is general and the same as in the case of coins (Bernoulli distribution) when we know that our data comes from n different normal distributions $\mathcal{N}(\mu_i, \sigma_i^2)$,$\forall i \in \{1, \dots, n\}$ and we introduce new parameters p_i, which are the probabilities of the corresponding distributions.

References

[1] Andresen A., (2015). Finite sample analysis of profile m-estimation in the single index model. *Electronic Journal of Statistics*, 9(2):2528–2641.

[2] Andresen, A., Spokoiny, V., (2015) Two convergence results for an alternation maximization. arXiv: 1501.01525.

[3] Bezdek, J., Hathaway, R. (2003). Convergence of Alternating Optimization. *Neural, Parallel & Pacific Computations 11*, 351 — 368.

[4] Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequialities for inverse problems. *The Annals of Statistics*, 30(3): 843 — 874.

[5] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:55495632.

[6] Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 34(4):1653 — 1677.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39:1 — 38.*

[8] Ibragimov, I.A. and R.Z. Khas'minskij (1981). *"Statistical estimation. Asymptotic theory."* "New York - Heidelberg - Berlin: Springer-Verlag".

[9] Kneip, A., (1994). Ordered linear smoothers. *The Annals of Statistics*, 22(2):835—866.

[10] M. R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference* (2005). *Springer in Statistics.*

[11] G. J. McLachlan and T. Krishman (1997). *The EM Algorithm and Extensions.* Wiley, New York.

[12] Spokoiny, V., Dickhaus, T., (2015). Basics of Modern Mathematical Statistics . *Springer Texts in Statistics.*

[13] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95 — 103.