When Composite Likelihood meets Stochastic Approximation

Giuseppe Alfonzetti and
Ruggero Bellio
Department of Economics and Statistics, University of Udine,
via Tomadini, 33100, Udine, Italy

Yunxiao Chen
and
Irini Moustaki
Department of Statistics, London School of Economics,
Houghton Street, WC2A 2AE, London, UK

November 25, 2024

Abstract

A composite likelihood is an inference function derived by multiplying a set of likelihood components. This approach provides a flexible framework for drawing inferences when the likelihood function of a statistical model is computationally intractable. While composite likelihood has computational advantages, it can still be demanding when dealing with numerous likelihood components and a large sample size. This paper tackles this challenge by employing an approximation of the conventional composite likelihood estimator based on a stochastic optimization procedure. This novel estimator is shown to be asymptotically normally distributed around the true parameter. In particular, based on the relative divergent rate of the sample size and the number of iterations of the optimization, the variance of the limiting distribution is shown to compound for two sources of uncertainty: the sampling variability of the data and the optimization noise, with the latter depending on the sampling distribution used to construct the stochastic gradients. The advantages of the proposed framework are illustrated through simulation studies on two working examples: an Ising model for binary data and a gamma frailty model for count data. Finally, a real-data application is presented, showing its effectiveness in a large-scale mental health survey.

Keywords: Exchangeable variables central limit theorem, Ising model, Gamma frailty model, Pairwise likelihood, Stochastic gradient

1 Introduction

The seminal work of Besag (1974) and the general framework proposed by Lindsay (1988) have paved the way for the wide adoption of composite likelihood methods as a practical approach for modelling multivariate responses with complex dependence structures (e.g., Henderson & Shimakura 2003, Bellio & Varin 2005, Katsikatsou et al. 2012, Lee & Hastie 2015). Such methods replace an intractable likelihood with an inference function constructed by multiplying many lower-dimensional marginal or conditional likelihood components, enabling frequentist estimation when traditional maximum likelihood approaches are infeasible or unattainable; see Varin et al. (2011) for an overview. However, in settings with large sample sizes and moderate response dimensions, the numerical optimization of the composite likelihood function requires evaluating many likelihood components at each iteration. Thus, it becomes, in turn, computationally unattainable.

Natural candidates for such settings are stochastic approximations, computationally convenient alternatives to numerical optimization that replace the score used by gradient-based routines with an adequately defined stochastic substitute (Robbins & Monro 1951). Thanks to their computational convenience, methods based on stochastic gradients (SGs) have quickly gained popularity among practitioners, becoming the standard choice for estimating complex models on large-scale data (Bottou et al. 2018). While their success is due to the capability of providing computationally affordable point estimates of the parameters of interest, the last decade has seen rising attention to conducting statistical inference with such estimates. Most of the recent developments in this regard build on the seminal work of Ruppert (1988) and Polyak & Juditsky (1992), who first established the asymptotic normality and statistical optimality of averaged stochastic estimators. We identify two challenges that have attracted the interest of researchers in recent years. The first one is the theoretical extension of the asymptotic results of Polyak & Juditsky (1992) to more general and flexible settings than the ones outlined in the original paper. In this regard, Toulis & Airoldi (2017) establish the asymptotic optimality of the Polyak-Ruppert averaged version of their proposed estimator based on implicit stochastic updates; Lee et al. (2022) extend Polyak & Juditsky (1992) results to a functional form; Su & Zhu (2023) relax the original assumptions to allow for globally convex and locally strongly convex objective functions; Wei et al. (2023) include the effect of general averaging schemes in the asymptotic distribution of the averaged estimates; Chen et al. (2024) establish the asymptotic properties of the averaged version of the Kiefer-Wolfowitz algorithm. The second challenge is the online estimation of the uncertainty of stochastic estimates. With few exceptions (e.g. Chee et al. 2023), most recent works in this area focus on Polyak-Ruppert type estimators, including the online bootstrap (Fang et al. 2018), mean-batch estimators (Chen et al. 2020, Zhu et al. 2023), the random scaling approach (Lee et al. 2022, Chen et al. 2023, 2024), and HiGrad (Su & Zhu 2023).

While most of the papers mentioned above explicitly refer to settings with online data (i.e., new observations sampled from the true data-generating distribution), stochastic optimization can also be used offline on an observed dataset, which is the classical setting of frequentist estimation. In the online-data setting, the result in Polyak & Juditsky (1992) and its extensions directly guarantee the asymptotic statistical optimality of the averaged stochastic estimator for the true parameter value. However, in the offline-data scenario, at each iteration, new observations are resampled from the empirical distribution of data, and

stochastic estimators converge to the maximum likelihood estimator (MLE) (e.g., Moulines & Bach 2011, Needell et al. 2014). It follows that, for a given dataset, the inferential procedures based on Polyak & Juditsky (1992) only quantify the variability of stochastic estimators around their target, i.e. the MLE, but neglect the sampling variability of the data. In addition, to our knowledge, the combination of stochastic approximations and composite likelihood inference has not been formally investigated.

Thus, our contribution in the following is two-fold. First, in Section 2, we show how different sampling schemes for the margins involved in the composite likelihood affect the statistical efficiency of SGs. Second, in Section 3, we extend Polyak & Juditsky (1992) result by establishing the consistency and asymptotic normality of the stochastic estimator around the true parameter in the offline-data setting. In particular, we show that, according to the relative divergence rate of the sample size and the number of iterations, the variance of the limiting distribution compounds for two sources of uncertainty: the sampling variability of the data and the noise injected by the SGs. While intuitive, combining the two sources of variability is technically non-trivial. Allowing the data to be random implies that all the SGs share a common source of variability, which complicates the applicability of central limit theorems for martingale sequences. Nevertheless, the asymptotic distribution can still be identified with the sum of exchangeable summands, which allows us to use the central limit theorem outlined in Blum et al. (1958) to establish the asymptotic normality of the estimator. Furthermore, by taking advantage of the second Bartlett's identity for the single likelihood components, it is possible to lower the noise injected in the optimization with the SGs at a fixed computational cost. In Section 4, we investigate the established theoretical results with simulation experiments on two working examples, one of which concerns estimating the Ising model based on full-conditional margins and the other concerns estimating a gamma-frailty model based on bivariate margins. Finally, Section 5 provides a real data application to a United States national health survey[1].

2 Composite Likelihood, Stochastic Approximations

2.1 Composite Likelihood

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ be a p-variate random vector that follows a parametric distribution with probability density/mass function $p(\boldsymbol{y}; \boldsymbol{\theta})$ and parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. Let $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ be a set of marginal or conditional events with likelihood functions $\mathcal{L}_k(\boldsymbol{\theta}; \boldsymbol{y}) \propto p(\boldsymbol{y} \in \mathcal{A}_k; \boldsymbol{\theta})$. A composite log-likelihood is obtained by summing the logarithms of the K likelihood objects. Thus, with $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$ being independent and identically distributed (i.i.d.) realizations of $\boldsymbol{Y}$, inference on $\boldsymbol{\theta}$ can be drawn based on the composite log-likelihood function

$$c\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k \ell_k(\boldsymbol{\theta}; \boldsymbol{y}_i), \tag{1}$$

where $\ell_k(\boldsymbol{\theta}; \boldsymbol{y}_i) = \log \mathcal{L}_k(\boldsymbol{\theta}; \boldsymbol{y}_i)$ is usually referred to as the k-th log-likelihood component or sub-log-likelihood and $\{w_1, \ldots, w_K\}$ is a set of weights to be defined depending on the

[1] https://catalog.data.gov/dataset/national-epidemiologic-survey-on-alcohol-and-related-conditions-nesarcwave-1-20012002-and-

model being estimated. The Composite Likelihood Estimator (CLE) is given by

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \{-c\ell_n(\boldsymbol{\theta})\}. \tag{2}$$

Typically, the CLE is used when the sub-log-likelihoods have simple forms while the joint density function $p(\boldsymbol{y};\boldsymbol{\theta})$ is analytically intractable. We give two illustrative examples below.

Example 1 *Let $\boldsymbol{Y}$ be a binary vector with $Y_j \in \{0,1\}$ following an Ising model (Ising 1924). Under this model, $p(\boldsymbol{y};\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{p}\beta_{j0}y_j + \sum_{j<j'}\beta_{jj'}y_jy_{j'}\right\}/Z(\boldsymbol{\theta})$, where $Z(\boldsymbol{\theta}) = \sum_{y\in\{0,1\}^p}\exp\left\{\sum_{j=1}^{p}\beta_{j0}y_j + \sum_{j<j'}\beta_{jj'}y_jy_{j'}\right\}$ is the so-called partition function which is needed to guarantee $p(\boldsymbol{y};\boldsymbol{\theta})$ to be a proper probability mass function. In this model, the parameters of interest are $\boldsymbol{\theta} = (\beta_{i0}, \beta_{jj'} : i = 1,\ldots,p, 1 \le j < j' \le p)^\top$, whose dimension is $d = p + p(p-1)/2$. As $Z(\boldsymbol{\theta})$ involves a summation over all possible binary vectors, the complexity of computing $Z(\boldsymbol{\theta})$ grows exponentially with p. Thus, the likelihood function quickly becomes intractable when p is large. To draw inference under the Ising model, Besag (1974) proposed a composite likelihood estimator that considers $K = p$ component likelihood terms $\mathcal{L}_j(\boldsymbol{\theta};\boldsymbol{y}) = \exp(y_j(\beta_{j0} + \sum_{j'}\beta_{jj'}y_{j'}))/(1+\exp(\beta_{j0}+\sum_{j'}\beta_{jj'}y_{j'})), j = 1,\ldots,p$, which are derived from the conditional distribution of Y_j given the rest of the entries of Y. The composite log-likelihood function for a random sample of size n can then be written as $c\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}\big(\beta_{j0} + \sum_{j'\neq j}\beta_{jj'}y_{ij'}\big) - \log\big\{1 + \exp\big(\beta_{j0} + \sum_{j'\neq j}\beta_{jj'}y_{ij'}\big)\big\}$.*

Example 2 *Consider the correlated gamma frailty model proposed in Henderson & Shimakura (2003). The authors impose an autoregressive correlation structure to model the underlying gamma process. In such a setting, it is convenient to consider only pairs within a certain time lag, drastically lowering the model's estimation cost. For this reason, we consider an exchangeable correlation structure such that no pairs can be ignored a priori. For illustration purposes, we consider a simplified version without covariates. Let $\boldsymbol{Y}$ be a multivariate count vector of dimension p. Its generic element, $Y_j \in \mathbb{N}$ for $j = 1,\ldots,p$, is distributed as $Y_j|V_j \sim$ Poisson$\{V_j\exp(\lambda_j)\}$, for $j = 1,\ldots p$, where λ_j the j-th baseline rate. The p-dimensional frailty term $\boldsymbol{V}$ has unidimensional margins distributed as $V_j \sim$ Gamma(ξ^{-1},ξ^{-1}), and correlation matrix $\boldsymbol{C}$, with generic element $C_{jj'} = \rho$, for $0 \le \rho \le 1$. The interest is in estimating $\boldsymbol{\theta} = (\boldsymbol{\lambda},\rho,\xi)$, with $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_p)^\top$, which has dimension $d = p+2$. The density function of the model can be written as*

$$\begin{aligned} p(Y_1 = y_1,\ldots,Y_p = y_p;\boldsymbol{\theta}) &= \left(\prod_{j=1}^{p}\frac{u_j^{y_j}}{y_j!}\right)\int_{\mathbb{R}^p}\left(\prod_{j=1}^{p}V_j^{y_j}\right)\exp(-\boldsymbol{u}^\top\boldsymbol{V})d\boldsymbol{V} \\ &= (-1)^{\sum_j y_j}\left(\prod_{j=1}^{p}\frac{u_j^{n_j}}{n_j!}\right)\frac{\partial^{(\sum_j y_j)}L(\boldsymbol{u})}{\partial^{y_1}u_1,\ldots,\partial^{y_p}u_p}, \end{aligned} \tag{3}$$

where $u_j = \exp(\lambda_{j0})$, $\boldsymbol{u} = (u_1,\ldots,u_p)^\top$, $\boldsymbol{V} = (V_1,\ldots,V_p)^\top$ and $L(\boldsymbol{u}) = E_V\left\{\exp(-\boldsymbol{u}^\top\boldsymbol{V})\right\}$. The computational challenge is that the random number of derivatives involved in (3), $\sum_j y_j$, can be too large to handle even in small p settings. Henderson & Shimakura (2003)

substitute (3) *with the composition of bivariate log-margins* $\ell_{jj'}(\boldsymbol{\theta}; y_j, y_{j'})$*, computed via*

$$\begin{aligned}\ell_{jj'}(\boldsymbol{\theta}; y_j, y_{j'}) = {} & y_j \log u_j + y_{j'} \log u_{j'} - \log(y_j!) - \log(y_{j'}!) + \\ & + \sum_{s=0}^{m_2-1} \log(1+s\xi) + y_j \log D_j + y_{j'} \log D_{j'} - (y_j + y_{j'} + \xi^{-1}) \log \Delta + \\ & + \log \sum_{s=0}^{m_1} \left[(-1)^s \binom{m_1}{s} \binom{m_2}{s} s! \left\{ \prod_{s'=m_2}^{m_1+m_2-s-1} (1+s'\xi) \right\} \xi^s f^s \right],\end{aligned}$$

with $m_1 = \min(y_j, y_{j'})$*,* $m_2 = \max(y_j, y_{j'})$*,* $\Delta = 1 + \xi u_j + \xi u_{j'} + \xi^2 u_j u_{j'} (1 - \rho^{|j-j'|})$*,* $D_j = 1 + \xi u_{j'}(1 - \rho^{|j-j'|})$ *and* $f = \frac{\Delta(1-\rho)}{D_j D_{j'}}$*. Therefore, the composite log-likelihood for a random sample of size* n *can then be written as* $c\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j<j'} \ell_{jj'}(\boldsymbol{\theta}; y_{ij}, y_{ij'})$*.*

When the parametric model is correctly specified, the CLE is consistent and asymptotically normal (Lindsay 1988). Let $\boldsymbol{\theta}^*$ denote the true parameter value. Then $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^*$, and further

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathcal{N}_p\left(0, \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}\right), \tag{4}$$

where $\boldsymbol{H} = E_{\boldsymbol{Y}}\{-\nabla^2 c\ell_n(\boldsymbol{\theta}^*)/n\}$ and $\boldsymbol{J} = \mathrm{Var}_{\boldsymbol{Y}}\{\nabla c\ell_n(\boldsymbol{\theta}^*)/n\}$. Note that such asymptotic results are obtained under a classical asymptotic regime where p is fixed while n diverges. In the case of (1), such an assumption implies considering both the number of contributions, K, and the parameter space, d, as fixed.

2.2 Proposed Method

As discussed in Section 2.3, it can be computationally intensive to obtain a numerical solution to (2) that can be used for statistical inference when large sample sizes and numerous sub-likelihood components are involved. To leverage the trade-off between statistical and computational efficiencies, we propose to use an algorithm based on SGs to get an approximation of $\hat{\boldsymbol{\theta}}$. First, consider that the gradient of $-c\ell_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is simply given by $-\sum_{i=1}^n \sum_{k=1}^K \nabla \ell_k(\boldsymbol{\theta}; \boldsymbol{y}_i)$. Considering its double-sum structure and the proper-likelihood nature of each $\ell_k(\boldsymbol{\theta}; \boldsymbol{y}_i)$, a SG of $-c\ell_n(\boldsymbol{\theta})$ can be constructed as

$$S(\boldsymbol{\theta}; \boldsymbol{W}) = -\frac{1}{\gamma_1} \sum_{i=1}^n \sum_{k=1}^K w_{ik} \nabla \ell_k(\boldsymbol{\theta}; \boldsymbol{y}_i). \tag{5}$$

Here, $\boldsymbol{W} = (w_{ik})_{n \times K}$ is a random matrix following a joint distribution $\mathcal{P}$, under which $w_{ik} \in \{0, 1\}$ and $P(w_{ik} = 1) = \gamma_1$. Differently from (1), we let the weights change across observations and iterations. Such broader specification introduces greater flexibility in managing the variability of the SG by allowing for different choices of $\mathcal{P}$.

As summarized in Algorithm 1 below, the proposed method iterates between updating $\boldsymbol{\theta}$ and constructing an SG under the current value of $\boldsymbol{\theta}$. We anchor the number of iterations to the sample size for theoretical reasons that will be shown in detail in Section 3. For the moment, just let the notation T_n refer to the number of iterations performed, where the dependence on n is such that $T_n \to \infty$ as $n \to \infty$.

Algorithm 1 Composite Stochastic Gradient Descent

Input: $\boldsymbol{y}$,$\mathcal{P}$, $\boldsymbol{\theta}_0$, η_0, T_n, B.
for t **in** $1, \ldots, T_n$ **do**
    Sampling Step: Draw $\boldsymbol{W}_t$ from $\mathcal{P}$;
    Approximation Step: Construct a stochastic gradient $\boldsymbol{S}_t = S(\boldsymbol{\theta}_{t-1}; \boldsymbol{W}_t)$;
    Update Step: Update the parameter estimate via $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \frac{\eta_t}{n}\boldsymbol{S}_t$;
end for
Trajectories Averaging: Compute $\bar{\boldsymbol{\theta}}_{\mathcal{P}} = \frac{1}{T_n - B}\sum_{t=B+1}^{T_n} \boldsymbol{\theta}_t$;
Output: Return $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$.

In each iteration, the Sampling Step injects some randomness into the procedure by drawing $\boldsymbol{W}_t$ from $\mathcal{P}$. Note that $\mathcal{P}$ must be defined such that the SG computed during the Approximation Step, $\boldsymbol{S}_t$, is computationally cheaper than the full gradient $-\nabla c\ell_n(\boldsymbol{\theta}_{t-1})$, but still unbiased, i.e. $E_{\boldsymbol{W}}\{\boldsymbol{S}_t\} = -\nabla c\ell_n(\boldsymbol{\theta}_{t-1})$. In other words, although the descent direction identified by $\boldsymbol{S}_t$ is noisy due to the randomness of $\boldsymbol{W}$, it still recovers the exact negative gradient on average. The computational convenience of $\boldsymbol{S}_t$ is the key advantage of Algorithm 1 and, together with $\mathcal{P}$, it controls the trade-off between computational and statistical efficiencies. At the end of each iteration, the estimates are updated via a Robbins-Monro step (Robbins & Monro 1951), where η_t is the stepsize at the t-th iteration computed given an initial value η_0 and a suitable decreasing schedule, formalized in Assumption 5 in Section 3. The rescaling by $1/n$ only affects the initial value η_0 and not the scheduling per se. It serves to standardize the SG, such that $n^{-1}\boldsymbol{S}_t$ is of magnitude $O(1)$, which is used in the proof of Proposition 1. It is worth remarking that the Update Step can be generalized to a second-order update, where the stepsize is substituted by a $d \times d$ matrix approximating $\boldsymbol{H}^{-1}$. For ease of exposition, we limit the discussion to one-dimensional stepsizes. Finally, after completing T_n iterations, the output of the algorithm, $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$, is computed by averaging the stochastic estimates along their trajectories (Ruppert 1988, Polyak & Juditsky 1992). In practice, it is often useful to account for a burn-in period, B, to avoid including estimates too close to the starting point $\boldsymbol{\theta}_0$ in the computation of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$.

It is straightforward to notice that Algorithm 1 closely follows the averaging SG descent outlined in Polyak & Juditsky (1992), and it inherits, in fact, its theoretical properties in approximating $\hat{\boldsymbol{\theta}}$. However, the theoretical framework discussed in Section 3 allows the output of Algorithm 1 to be directly used to draw inference on $\boldsymbol{\theta}^*$. A further advantage of our proposal is that, by specifying $\boldsymbol{S}_t$ as in (5), Algorithm 1 gives the user the freedom to leverage the peculiar structure of the composite likelihood to improve the efficiency of the approximation by adequately choosing $\mathcal{P}$. To this end, let us introduce three possible choices for the distribution of the weights, as described in Definitions 1 through 3.

Definition 1 *Let $\mathcal{P}_1$ be the joint distribution of $\boldsymbol{W}$ such that $\boldsymbol{W} = \boldsymbol{D}\mathbf{1}_{nK}$, where $\mathbf{1}_{nK}$ is a $n \times K$ matrix of ones and $\boldsymbol{D}$ is a $n \times n$ diagonal matrix, with diagonal distributed as Multinomial$\{1, (1/n, \ldots, 1/n)\}$.*

Definition 2 *Let $\mathcal{P}_2$ be the joint distribution of $\boldsymbol{W}$ such that $W_{i,k} \overset{i.i.d.}{\sim}$ Bernoulli$(1/n)$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.*

Definition 3 *Let $\mathcal{P}_3$ be the joint distribution of $\boldsymbol{W}$ such that $W_{i,k} = v_{iK-K+k}$, where $\boldsymbol{v}$ is a nK-dimensional random vector following a multivariate hypergeometric distribution with K draws over nK categories of dimension 1 and v_j is its j-th element.*

All three definitions lead to drawing, on average, K weights equal to 1 and the remaining $nK-K$ equal to zero. The implied per-iteration complexity is $O(K)$, independent of n, and thus particularly suitable for large-scale applications. However, according to the sampling scheme chosen, a different dependence layout is induced on the cells of $\boldsymbol{W}$. In Section 3, we show how this noise structure affects the asymptotic efficiency of the stochastic estimates.

We discuss the three proposed sampling schemes in more detail. A weight matrix $\boldsymbol{W}$ sampled according to Definition 1 is constrained to have all elements equal to zero, apart from a single row filled with ones. With such weights, $\boldsymbol{S}_t$ evaluates the gradient on a single observation selected randomly from $\{1, \dots, n\}$. We refer to this construction as *standard* SG to stress its widespread adoption at the core of many stochastic algorithms. Note that using such a sampling scheme, we are ignoring the double sum structure of (1) since $\mathcal{P}_1$ constraints the K selected sub-likelihood component to belong to the same observation.

Nevertheless, (5) is very flexible in defining the SG and allows for different choices of $\mathcal{P}$. Consider $\mathcal{P}_2$ as described by Definition 2. All the elements of the matrix are now independent and identically distributed as Bernoulli random variables with proportion parameter $1/n$. It means that, at each iteration, K sub-likelihood components are selected on average by the weighting matrix. Therefore, the complexity of the approximation matches the standard SG one. However, the proportion parameter in Definition 2 can be set as low as $(nK)^{-1}$, implying an iteration complexity $O(1)$, which is unattainable using Definition 1. Regardless, for comparison purposes, we stick to the proportion parameter $1/n$. Note that the structure of the noise injected by $\mathcal{P}_2$ is very different from the one implied by $\mathcal{P}_1$. While the K components drawn by $\mathcal{P}_1$ share a very specific covariance stemming from the dependence among the summands of $c\ell_n(\boldsymbol{\theta})$, $\mathcal{P}_2$ completely breaks this structure by independently selecting sub-likelihoods possibly belonging to independent observations. Finally, consider the sampling scheme $\mathcal{P}_3$. It can be seen as a random scramble of the vectorization of $\boldsymbol{W}$, where only the elements in the first K positions are retained. Like $\mathcal{P}_2$, the complexity per iteration can be lowered to $O(1)$ by decreasing the number of components retained per iteration. However, in this case, the weights are not completely independent since, given the fixed number of components drawn, a weak negative correlation is induced among the elements of $\boldsymbol{W}$. In Section 3, we show how $\mathcal{P}_2$ and $\mathcal{P}_3$ improve the efficiency of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ while maintaining the same computational cost as the standard SG.

2.3 Comparison with Gradient Descent

Before investigating the theoretical properties of Algorithm 1 in the next section, let us consider solving (2) by a gradient descent algorithm with fixed stepsize to compute $\tilde{\boldsymbol{\theta}}$, a numerical estimate of $\boldsymbol{\theta}^*$. Given the subtleness of the notation, see Table 1 as a reference for the symbols used. To describe the relative divergence rate of two positive sequences a_n and b_n, we write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, $a_n = \omega(b_n)$ if $\lim_{n\to\infty} a_n/b_n = \infty$, while $a_n = \Theta(b_n)$ if $\lim_{n\to\infty} a_n/b_n = \gamma$ with $0 < \gamma < \infty$. Given a fixed starting point $\boldsymbol{\theta}_0$ and assuming $c\ell_n(\boldsymbol{\theta})$ to be strongly convex and its gradient to be Lipschitz continuous with constant $L > 0$, then, at each iteration t, the numerical procedure updates via $\boldsymbol{\theta}_t =$

Table 1: Notation summary for $\boldsymbol{\theta}$ values.

$\boldsymbol{\theta}^*$	$\hat{\boldsymbol{\theta}}$	$\tilde{\boldsymbol{\theta}}$	$\bar{\boldsymbol{\theta}}_{\mathcal{P}}$
True parameter. Not observed.	CLE. Usually not available analytically.	Numerical approximation of $\hat{\boldsymbol{\theta}}$ used as estimate of $\boldsymbol{\theta}^*$.	Stochastic estimator of $\boldsymbol{\theta}^*$ based on the chosen $\mathcal{P}$.

$\boldsymbol{\theta}_{t-1} + \eta \nabla c\ell_n(\boldsymbol{\theta}_{t-1})$, $t = 1 \ldots, T_n$, where $0 < \eta < 2/L$ is a fixed stepsize. The final parameter estimate is taken as the output of the algorithm, namely $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_{T_n}$. This gradient-based update highlights that the critical quantity to be computed at each iteration is $\nabla c\ell_n(\boldsymbol{\theta})$, which costs $O(nK)$ operations. Furthermore, to draw statistical inference with the numerical solution $\tilde{\boldsymbol{\theta}}$, one requires $\tilde{\boldsymbol{\theta}}$ to have the same limiting distribution as $\hat{\boldsymbol{\theta}}$, which implies that $\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = o(1/\sqrt{n})$. Given the linear convergence rate of gradient descent (see, for example, Theorem 2 in Section 1.4.2 of Polyak (1987)), the total number of iterations needs to satisfy $T_n = \omega(-\frac{1}{2}\log_c n)$ with $c \in [0, 1)$, hence $T_n = \omega(\log n)$. In Table 2, we compare the total computational budget, $\mathcal{B}$, and asymptotic variance of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ and $\tilde{\boldsymbol{\theta}}$ according to the relative divergence rate of T_n and n.

Table 2: Computational and statistical efficiency comparison between gradient descent and Algorithm 1 with $\mathcal{P}$ chosen according to Definitions 1, 2 and 3.

	Per iteration complexity	Number of iterations	Total complexity	Asymptotic variance
GD	$O(nK)$	$T_n = \omega(\log n)$	$\mathcal{B} = \omega(nK \log n)$	$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$
$\mathcal{P}_1$	$O(K)$	$T_n = \omega(n)$	$\mathcal{B} = \omega(nK)$	$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$
		$T_n = o(n)$	$\mathcal{B} = o(nK)$	$T_n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$
		$T_n = \Theta(n)$	$\mathcal{B} = \Theta(nK)$	$(T_n^{-1} + n^{-1})\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$
$\mathcal{P}_2$	$O(K)$	$T_n = \omega(n)$	$\mathcal{B} = \omega(nK)$	$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$
		$T_n = o(n)$	$\mathcal{B} = o(nK)$	$T_n^{-1}\boldsymbol{H}^{-1}$
		$T_n = \Theta(n)$	$\mathcal{B} = \Theta(nK)$	$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1} + T_n^{-1}\boldsymbol{H}^{-1}$
$\mathcal{P}_3$	$O(K)$	$T_n = \omega(n)$	$\mathcal{B} = \omega(nK)$	$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$
		$T_n = o(n)$	$\mathcal{B} = o(nK)$	$T_n^{-1}\boldsymbol{H}^{-1}$
		$T_n = \Theta(n)$	$\mathcal{B} = \Theta(nK)$	$n^{-1}\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1} + T_n^{-1}\boldsymbol{H}^{-1}$

First, to achieve the asymptotic efficiency in (4), Algorithm 1 needs many more iterations compared to gradient descent. However, given the extreme computational affordability of its iterations, the total budget needed to reach such an asymptotic behavior is lower than what is needed by gradient descent, whatever the choice of $\mathcal{P}$. While theoretically appealing, Algorithm 1 might need to tune η_0 adequately in practice, like most stochastic optimization methods, which increases its computational cost. Furthermore, numerical optimization is typically carried out with Newton and quasi-Newton algorithms. They are well-known to be more efficient than gradient descent in practical applications. Still, they are particularly sensitive to the dimension of the parameter space since they require to compute, or approximate, the inverse of the $d \times d$ Hessian matrix. From a practical point of view, on problems of moderate dimensions, it might still be preferable to use numerical optimizers to take advantage of the asymptotic behavior of $\hat{\boldsymbol{\theta}}$.

The real advantage of Algorithm 1 shows up when $O(nK)$ is the maximum compu-

tational budget available, and we want to quantify uncertainty around our estimates. In such a scenario, using the numerical solution $\tilde{\boldsymbol{\theta}}$ can be infeasible since the computational inaccuracy remaining might be too large. Regardless, using $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ is a viable option, and the reason is rather intuitive. Although subtle, when running numerical optimization, one does not use $\tilde{\boldsymbol{\theta}}$ directly to draw inference on $\boldsymbol{\theta}^*$. Instead, the requirement for $\tilde{\boldsymbol{\theta}}$ is to be close enough to $\hat{\boldsymbol{\theta}}$ in order to safely replace $\hat{\boldsymbol{\theta}}$ with $\tilde{\boldsymbol{\theta}}$ in (4). We argue that this is not the case when using stochastic optimizers since $\bar{\boldsymbol{\theta}}_{\mathcal{P}} - \hat{\boldsymbol{\theta}}$ is a random variable itself, with distribution depending on $\mathcal{P}$. Hence, quantifying the noise injected by $\boldsymbol{W}$ in the optimization makes it possible to directly use $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ for inference on $\boldsymbol{\theta}^*$, without strict requirements on its distance from $\hat{\boldsymbol{\theta}}$. Fixing $\mathcal{B} = O(nK)$ implies the running length of the algorithm to be either $T_n = o(n)$ or $T_n = \Theta(n)$. With such divergence rates, the choice of $\mathcal{P}$ plays a central role in the asymptotic variance of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ since the noise in the stochastic approximation is still non-negligible. Table 2 shows that, in those cases, relying on different choices of $\mathcal{P}$ is not equivalent. Estimates based on $\mathcal{P}_2$ and $\mathcal{P}_3$ enjoy a lower asymptotic variance than $\mathcal{P}_1$, as will become more apparent in the next section. Furthermore, Section 3 also establishes the asymptotic distribution of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ around $\boldsymbol{\theta}^*$ under some mild assumptions on the choice of $\mathcal{P}$.

2.4 Implementation Remarks

From an implementation perspective, Algorithm 1 allows for some practical expedients to enhance the computational performance. First, the computation of $\boldsymbol{S}_t$ at each iteration can be easily parallelized by taking advantage of the double-sum structure of $\nabla c\ell_n(\boldsymbol{\theta})$, assigning the gradient computation for a different sub-likelihood component to each available thread.

Second, as typical of stochastic algorithms, not all the data are needed at each iteration, such that memory resources can be saved by carefully passing only the portion of the data needed to compute $\boldsymbol{S}_t$ at a given t. In this regard, $\mathcal{P}_1$ is the cheaper choice memory-wise since it fixes the memory cost to $O(K)$ no matter the structure of $\nabla c\ell_n(\boldsymbol{\theta})$. By choosing $\mathcal{P}_2$ or $\mathcal{P}_3$, one typically needs to store data coming from multiple observations, such that the maximum amount of memory necessary strictly depends on the model. In the case of Example 1, each sub-log-likelihood component has a memory cost $O(K)$. Thus, since $\mathcal{P}_2$ and $\mathcal{P}_3$ can draw components potentially belonging to K different observations, their maximum memory cost per iteration is $O(K^2)$. If we consider Example 2, each component accesses only $O(1)$ data instead. Consequently, by drawing K components on different observations, both $\mathcal{P}_2$ and $\mathcal{P}_3$ have a maximum per-iteration memory cost of $O(K)$.

Third, the Sampling Step can be recycled across iterations to save computational resources. Namely, with $\mathcal{P}_1$, one can scramble the vector $(1, \dots, n)$ once and then use each of the first l elements as indices of the observations drawn in the following l-dimensional window of iterations. Intuitively, as long as l is low enough, the dependence induced by the recycling among iterations within the same window is negligible, such that they can still be considered independent. Thus, recycling trims the cost of the Sampling Step by a factor of l. The same approach can be implemented when using $\mathcal{P}_3$ by scrambling the vector $(1, \dots, nK)$ and allocating the first l K-dimensional sequences of indices to the subsequent l iterations. Unfortunately, recycling is impossible when $\mathcal{P}_2$ is chosen since the number of components drawn per iteration is not deterministic.

3 Theoretical Results

In what follows, we establish the asymptotic properties of the proposed estimator $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$. Proposition 1 states the convergence of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ to $\boldsymbol{\theta}^*$, while Theorem 1 and Corollary 1 provide novel theoretical results describing the asymptotic distribution of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ under different choices of T_n and $\mathcal{P}$. The following assumptions combine classical domination conditions on the log-likelihood components (e.g. White 1982), with the flexible setting outlined in Su & Zhu (2023) for globally convex and locally strongly convex objective functions.

Assumption 1 *For $k = 1, \dots, K$, $\ell_k(\boldsymbol{\theta}; \boldsymbol{y})$ exists for every $\boldsymbol{\theta}$, is continuous in $\boldsymbol{\theta}$ for all $\boldsymbol{y} \in \mathcal{Y}$ and $|\ell_k(\boldsymbol{\theta}; \boldsymbol{y})|$ is dominated by a function integrable with respect to the distribution of $\boldsymbol{Y}$. The vector $\boldsymbol{\theta}^*$ is the unique solution to $E_{\boldsymbol{Y}}\{\sum_{k=1}^{K} \ell_k(\boldsymbol{\theta}; \boldsymbol{Y})\} = 0$.*

Assumption 2 *Each sub log-likelihood $\ell_k(\boldsymbol{\theta}; \boldsymbol{y})$ is differentiable in $\boldsymbol{\theta}$, and its gradient continuous, for all $\boldsymbol{y} \in \mathcal{Y}$. Furthermore, with θ_r being the generic r-th element of $\boldsymbol{\theta}$, the quantity $|\partial \ell_k(\boldsymbol{\theta}; y)/\partial \theta_r|$ is dominated by a function integrable with respect to $p(\boldsymbol{Y}; \boldsymbol{\theta}^*)$ up to the fourth power. In addition, $|\partial^2 \ell_k(\boldsymbol{\theta}; y)/\partial \theta_r \partial \theta_{r'}|$ exists in a δ-neighbourhood of $\boldsymbol{\theta}^*$, i.e. for some $\delta > 0$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta$, where it is continuous in $\boldsymbol{\theta}$ and dominated by a function integrable with respect to $p(\boldsymbol{Y}; \boldsymbol{\theta}^*)$.*

Assumption 3 *The expected sub log-likelihoods are concave in $\boldsymbol{\theta}$ and their gradients satisfy $\|\nabla E_{\boldsymbol{Y}}\{\ell_k(\boldsymbol{\theta}', \boldsymbol{Y})\} - \nabla E_{\boldsymbol{Y}}\{\ell_k(\boldsymbol{\theta}, \boldsymbol{Y})\}\| \leq L_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$, for $L_1 > 0$, $k = 1, \dots, K$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$. Furthermore, for some $L_2, \delta > 0$, the expected Hessian of each sub log-likelihood verifies $\|\nabla^2 E_{\boldsymbol{Y}}\{\ell_k(\boldsymbol{\theta}, \boldsymbol{Y})\} - \nabla^2 E_{\boldsymbol{Y}}\{\ell_k(\boldsymbol{\theta}^*, \boldsymbol{Y})\}\| \leq L_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$ with $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta$. Additionally $\sum_{k=1}^{K} \nabla^2 E_{\boldsymbol{Y}}\{\ell_k(\boldsymbol{\theta}^*, \boldsymbol{Y})\}$ is negative-definite.*

Assumption 4 *The sampling scheme $\mathcal{P}$ is such that $E\{W_{i,k}\} = \gamma_1$, with $0 < \gamma_1 < 1$, for $i = 1, \dots, n$ and $k = 1, \dots, K$. Additionally, $\lim_{n \to \infty} n\gamma_1 > 0$, $E\{W_{i,k} W_{i,k'}\}, E\{W_{i,k} W_{i,k'} W_{i,k''}\}$, and $E\{W_{i,k} W_{i,k'} W_{i,k''} W_{i,k'''}\}$ are of order $O(\gamma_1)$, and $E\{W_{i,k} W_{j,k'}\}, E\{W_{i,k} W_{i,k'} W_{j,k''}\}$, and $E\{W_{i,k} W_{i,k'} W_{j,k''} W_{j,k'''}\}$ are of order $O(\gamma_1^2)$, with $i \neq j$ and $k, k', k'', k''' = 1, \dots, K$. For notation aims, let $E\{W_{i,k} W_{i,k'}\} = \gamma_2$ if $k \neq k'$.*

Assumption 5 *Given $\eta_0 > 0$, the stepsize scheduling is chosen as $\eta_t = \eta_0 t^{-c}$ with $1/2 < c < 1$, implying $\lim_{n \to \infty} \sum_t^{T_n} \eta_t = \infty$, $\lim_{n \to \infty} \sum_t^{T_n} \eta_t / \sqrt{t} < \infty$, $\lim_{n \to \infty} \sum_t^{T_n} \eta_t^2 < \infty$.*

Assumption 1 collects the regularity conditions on the behavior of the likelihood function that guarantee the existence and uniqueness of $\boldsymbol{\theta}^*$ (see Varin & Vidoni 2005). Assumption 2 allows exchanging the order of integration and differentiation when working with sub-log-likelihood objects. In addition, it guarantees the existence of $E_{\boldsymbol{Y}}\{\nabla^2 \ell_k(\boldsymbol{\theta}; \boldsymbol{Y})\}$ on $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta$, with δ small enough. Assumption 3, imposes Lipschitz regularity for $E_{\boldsymbol{Y}}\{\nabla \ell_k(\boldsymbol{\theta})\}$ on all $\mathbb{R}^d$, and for $E_{\boldsymbol{Y}}\{\nabla^2 \ell_k(\boldsymbol{\theta})\}$ on the ball $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta$. Furthermore, it outlines the local strong concavity of the composition of the expected composite log-likelihood in a neighborhood of $\boldsymbol{\theta}^*$, thus guaranteeing its local identifiability. Together with Assumption 2, it allows recovering the conditions on the objective function required by Su & Zhu (2023). Assumption 4 requires all the weights to share the same expected value, γ_1, which does not decay faster than $1/n$. Furthermore, it assumes the cross-products across the weights of different sub log-likelihoods to be bounded up to a constant by γ_1, when referring to the same observations, and by γ_1^2, when referring to two different ones. Together with

Assumption 2, such bounds control the stochastic behavior of the SGs constructed via (5). Assumption 4 is more general than what is needed for $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ and potentially allows for different sampling schemes. Note that γ_1 can be fixed across different choices of $\mathcal{P}$ to match the computational cost of constructing (5). However, for a given γ_1, different sampling schemes lead to different γ_2, which affects the correlation across weights belonging to the same observation. Theorem 1 shows that the pair γ_1,γ_2 is sufficient to explain the statistical efficiency implied by different choices of $\mathcal{P}$. Assumption 5 is a standard requirement on the decreasing scheduling of the stepsize (e.g. Polyak & Juditsky 1992).

With Assumptions 1-5, it is straightforward to adapt Su & Zhu (2023) results to prove $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ being a stochastic approximation of $\hat{\boldsymbol{\theta}}$ as the number of iterations T_n diverge. However, $\hat{\boldsymbol{\theta}}$ gets closer to $\boldsymbol{\theta}^*$ as the sample size grows. It follows that $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ can also be used as a consistent estimator of $\boldsymbol{\theta}^*$ as long as both T_n and n diverge simultaneously.

Proposition 1 *With Assumptions 1-5, the output of Algorithm 1 provides a consistent estimator of* $\boldsymbol{\theta}^*$ *when* $T_n \to \infty$ *as* $n \to \infty$, *i.e.* $\bar{\boldsymbol{\theta}}_{\mathcal{P}} \xrightarrow[n]{a.s.} \boldsymbol{\theta}^*$.

The proof of Proposition 1 in the online Appendix is an adaptation of Lemma 17 in Su & Zhu (2023). However, it is worth emphasizing that it takes advantage of the expected behavior of $\boldsymbol{S}_t$ when evaluated at $\boldsymbol{\theta}^*$ rather than $\hat{\boldsymbol{\theta}}$. In other words, it acknowledges the data and $\boldsymbol{W}_t$ as random variables when constructing $\boldsymbol{S}_t$. This is critical to stress since it allows considering $\boldsymbol{\theta}^*$ as the target of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ given the double asymptotics in n and T_n.

While Proposition 1 formalizes the consistency of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$, nothing has been said so far about its distributional properties. Theorem 1 tackles this aspect by outlining the asymptotic distribution of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ according to the relative divergence rate of n and T_n. Let us anticipate how Theorem 1 differs from the original inference framework for averaged stochastic optimization. When used to find the root of $\nabla c\ell_n(\boldsymbol{\theta}) = 0$, the original result in Theorem 2 in Polyak & Juditsky (1992) describes the asymptotic behavior of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ around $\hat{\boldsymbol{\theta}}$. It assumes the data is fixed and does not quantify the uncertainty of the stochastic estimates around $\boldsymbol{\theta}^*$. From a technical point of view, directly combining such a result with (4) is not straightforward since the two asymptotic statements are defined on different probability spaces, namely with and without the conditioning on the observed values of $\boldsymbol{Y}$. Furthermore, as soon as we allow the data to be random, we are not able anymore to take advantage of the independence of the stochastic quantities $S_t(\boldsymbol{\theta}^*; \boldsymbol{W}_t)$, with $t = 1, \ldots, T_n$, which is a critical step in the proofs presented in Polyak & Juditsky (1992). In other words, while all the iterations still share the same dataset, its random nature induces dependence among them.

It follows that, when stochastically optimizing $c\ell_n(\boldsymbol{\theta})$, if the interest is drawing inference about $\boldsymbol{\theta}^*$, which is typically the case of composite likelihood methods and maximum likelihood estimation in general, the available results building on the asymptotic covariance matrix outlined in Polyak & Juditsky (1992) only provide a partial answer to the research question. To fill this gap, we provide Theorem 1, which shows that $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ is asymptotically normally distributed around $\boldsymbol{\theta}^*$ and its covariance matrix changes according to both $\mathcal{P}$ and the relative divergent behavior of T_n and n. The choice of $\mathcal{P}$ affects the shape of the noise coming from the optimization, while the divergent behavior of T_n and n quantifies its relative magnitude compared to the sampling variability of the data.

Theorem 1 *Let* $n/(T_n + n) \underset{n}{\to} \alpha$, *with* $0 \leq \alpha \leq 1$. *Under Assumptions 1-5, for the sampling schemes* $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$ *in Definitions 1-3, it holds that:*

Regime 1: If $\alpha = 0$*, then* $\sqrt{n}(\bar{\boldsymbol{\theta}}_{\mathcal{P}} - \boldsymbol{\theta}^*) \xrightarrow[n]{d} \mathcal{N}(0, \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1})$;

Regime 2: If $\alpha = 1$ *and* $n^{7/9} = o(T_n)$*, then* $\sqrt{T_n}(\bar{\boldsymbol{\theta}}_{\mathcal{P}} - \boldsymbol{\theta}^*) \xrightarrow[n]{d} \mathcal{N}(0, \boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}}\boldsymbol{H}^{-1})$;

Regime 3: If $0 < \alpha < 1$*, then* $\sqrt{T_n + n}(\bar{\boldsymbol{\theta}}_{\mathcal{P}} - \boldsymbol{\theta}^*) \xrightarrow[n]{d} \mathcal{N}(0, \boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}}\boldsymbol{H}^{-1}/(1-\alpha) + \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}/\alpha)$,

where $\boldsymbol{V}_{\mathcal{P}} = \lim_{n\to\infty} \gamma_1^{-2} n^{-1}(\gamma_1 - \gamma_2)\boldsymbol{H} + n^{-1}\left(\gamma_1^{-2}\gamma_2 - 1\right)\boldsymbol{J} = O(1)$.

The asymptotic covariance matrices in Theorem 1 can be described as a weighted average between $\boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}}\boldsymbol{H}^{-1}$ and $\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$, with weights depending on the divergence rate of T_n and n. Note that $\boldsymbol{H}$ and $\boldsymbol{J}$ are the usual matrices entering the asymptotic efficiency of the CLE, as discussed in Section 2. While $\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$ is already well known from (4) and quantifies the variability due to $\boldsymbol{Y}$, the matrix $\boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}}\boldsymbol{H}^{-1}$ can be shown to describe the noise coming from the optimization. As the notation stresses, the value of $\boldsymbol{V}_{\mathcal{P}}$ depends on the choice of $\mathcal{P}$. In particular, it results in a linear combination of the matrices $\boldsymbol{H}$ and $\boldsymbol{J}$, with coefficients based on the quantities γ_1 and γ_2. For a detailed derivation of $\boldsymbol{V}_{\mathcal{P}}$, see the proof of the theorem in the online Appendix B. Before describing Corollary 1, which outlines the different shapes of $\boldsymbol{V}_{\mathcal{P}}$ according to the choices $\mathcal{P}$, we briefly summarize the three asymptotic regimes described in Theorem 1.

When the algorithm runs for $T_n = \omega(n)$ iterations, it holds that $n/(T_n + n) \to 0$, and the estimates obtained fall under Regime 1. With such a setting, the noise component generated by the optimization is negligible compared to the sampling variability of the data. In this scenario, the algorithm runs until closely approximating the CLE, i.e., there is no difference between the asymptotic behaviors of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ and $\hat{\boldsymbol{\theta}}$. Inference can be carried out based on the familiar $\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$, as described in (4).

In the opposite setting, where the algorithm is stopped at $T_n = o(n)$, we get $n/(T_n + n) \to 1$ and Regime 2 holds. Note that setting $\gamma_1 = \Theta(1/n)$, as in the three sampling schemes considered, requires a minimum growing rate on the number of iterations to establish asymptotic normality, i.e. $n^{7/9} = o(T_n)$. While such a condition is not particularly restrictive in practice, its technical derivation can be found in Lemma 4 in the online Appendix. In such an asymptotic regime, the dominant variance component is the one induced by $\mathcal{P}$, such that inference can potentially ignore the variability of the data. In this case, the asymptotic distribution resembles the one in Polyak & Juditsky (1992) and related works, apart from having the parameter-dependent quantities evaluated at $\boldsymbol{\theta}^*$ rather than $\hat{\boldsymbol{\theta}}$. To glimpse the connection between Regime 2 and the conditional inference setting traditionally adopted in stochastic optimization, imagine n being so large that the distance between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ is negligible. Then, there is not much difference in practice between using $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ to infer either on $\boldsymbol{\theta}^*$ or $\hat{\boldsymbol{\theta}}$. In other words, the results in Polyak & Juditsky (1992) are equivalent, in a suitable sense, to inference under Regime 2 of Theorem 1.

Regime 3 describes an intermediate setting between the previous two. Since T_n and n grow at the same rate, i.e., $T_n = \Theta(n)$, it holds that $0 < \alpha < 1$ and the asymptotic covariance matrix around $\boldsymbol{\theta}^*$ compounds for both the optimization error and the sampling variability of the data. As it is difficult to assess whether $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ is obtained strictly under Regime 1 or Regime 2, it is always recommended to use Regime 3.

Hence, according to the divergent behavior of $n/(T_n + n)$, Theorem 1 highlights which variability component can be neglected and which can not when quantifying the uncertainty around stochastic estimates. Furthermore, note that for ease of exposition, Theorem 1 assumes $\gamma_1 = \Theta(1/n)$ as in the three sampling schemes $\mathcal{P}_1$-$\mathcal{P}_3$. However, such results can be

extended for whatever choice of $\mathcal{P}$ satisfying Assumption 4. The three asymptotic regimes and the variance decomposition are still valid, but different sampling rates γ_1 require a careful analysis of the magnitude of $V_{\mathcal{P}}$. Nevertheless, it is interesting to comprehend how the distribution of $\boldsymbol{W}$ affects the optimization noise. In this regard, Corollary 1 outlines the effects of choosing $\mathcal{P}$ according to Definitions 1, 2 and 3. In particular, the choice of $\mathcal{P}$ affects the values of γ_1 and γ_2 which control the shape of $\boldsymbol{V}_{\mathcal{P}}$.

Corollary 1 *Let Theorem 1 hold. Then, Definition 1, implies* $\boldsymbol{V}_{\mathcal{P}_1} = \boldsymbol{J}$ *and* $\boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}_1}\boldsymbol{H}^{-1} = \boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$*; Definition 2 implies* $\boldsymbol{V}_{\mathcal{P}_2} = \boldsymbol{H}$ *and* $\boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}_2}\boldsymbol{H}^{-1} = \boldsymbol{H}^{-1}$*; Definition 3 implies* $\boldsymbol{V}_{\mathcal{P}_3} = \boldsymbol{H}$ *and* $\boldsymbol{H}^{-1}\boldsymbol{V}_{\mathcal{P}_3}\boldsymbol{H}^{-1} = \boldsymbol{H}^{-1}$.

While it is apparent that $\gamma_1 = 1/n$ for all three sampling schemes, we leave the details about the implied values of γ_2 in the proof of Corollary 1 provided in the online Appendix B. When $\mathcal{P}$ is chosen according to Definition 1, the Sampling Step of Algorithm 1 keeps untouched the correlation structure among the sub-likelihood components of the objective function. In other words, it samples from the empirical distribution of the data. Therefore, the variability due to $\boldsymbol{W}$ takes the same shape as the one stemming from $\boldsymbol{Y}$, represented by the matrix $\boldsymbol{J}$. Note, in fact, that when $\mathcal{P}_1$ is chosen, $\gamma_1 = \gamma_2$ and therefore $\boldsymbol{H}$ asymptotically disappears when computing $\boldsymbol{V}_{\mathcal{P}}$ following Theorem 1. Instead, if $\mathcal{P}$ is chosen according to Definition 2, the sub log-likelihood components are drawn independently, even when belonging to the same observation. This step breaks the original correlation structure among the summands in $\nabla c\ell_n(\boldsymbol{\theta})$, such that $\boldsymbol{V}_{\mathcal{P}}$ collapses onto the expected second derivative of $\boldsymbol{S}_t$, $\boldsymbol{H}$. The weights independence, in fact, implies $\gamma_1^2 = \gamma_2$ and, therefore, a zero weight for $\boldsymbol{J}$ when computing $\boldsymbol{V}_{\mathcal{P}}$. Finally, if $\mathcal{P}$ follows Definition 3, the correlation among the elements keeps the weight for $\boldsymbol{J}$ different from zero but asymptotically negligible because of being $O(\frac{1}{nK})$. Thus, asymptotically, $\mathcal{P}_3$ shares the same asymptotic efficiency of $\mathcal{P}_2$.

It is well known that $\boldsymbol{H} \neq \boldsymbol{J}$ when using composite likelihood methods. Hence, the inference with $\hat{\boldsymbol{\theta}}$ must be based on $\boldsymbol{H}^{-1}\boldsymbol{J}\boldsymbol{H}^{-1}$ rather than $\boldsymbol{H}^{-1}$, which typically results in inflated variances for each parameter. Corollary 1 shows that $\mathcal{P}_1$ injects this same variability as noise in the optimization, while $\mathcal{P}_2$ and $\mathcal{P}_3$ constrain it to $\boldsymbol{H}^{-1}$. Such a difference is evident, as the simulations in Section 4 highlight, with the estimates based on $\mathcal{P}_2$ and $\mathcal{P}_3$ exhibiting lower variability than those obtained with $\mathcal{P}_1$.

4 Simulation Studies

We investigate the results from some simulation experiments with $R = 500$ replications for the models presented in Examples 1 and 2. In particular, our goal is two-fold. First, we provide evidence to support Proposition 1, namely to show that $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ converges to $\boldsymbol{\theta}^*$ when both T_n and n diverge by tracking the average mean square error of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$. i.e. $MSE = \frac{1}{dR}\sum_{r=1}^{R}\|\bar{\boldsymbol{\theta}}_{\mathcal{P}}^{(r,t)} - \boldsymbol{\theta}^*\|^2$, where $\bar{\boldsymbol{\theta}}_{\mathcal{P}}^{(r,t)}$ is the d-dimensional output of Algorithm 1 on the r-th replication when stopped at the t-th iteration. Furthermore, we highlight that different choices of $\mathcal{P}$ characterize different behavior of the MSE trajectories because of the implied asymptotic variabilities outlined in Corollary 1.

Second, we assess the empirical coverage performance of confidence intervals built from the asymptotic covariance matrices outlined in Theorem 1. We aim to highlight the strength of asymptotic Regime 3, which compounds both the sampling variability

of the data and the optimization noise. To construct the confidence intervals, an estimate for both $\boldsymbol{H}$ and $\boldsymbol{J}$ is needed. Here, we use the usual sample estimators (see e.g. Varin et al. 2011, Section 5) $\hat{\boldsymbol{J}}^{(r,t)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\nabla c\ell(\bar{\boldsymbol{\theta}}_{\mathcal{P}}^{(r,t)};\boldsymbol{y}_i)\right\}\left\{\nabla c\ell(\bar{\boldsymbol{\theta}}_{\mathcal{P}}^{(r,t)};\boldsymbol{y}_i)\right\}^{\top}$ and $\hat{\boldsymbol{H}}^{(r,t)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\nabla \ell_k(\bar{\boldsymbol{\theta}}_{\mathcal{P}}^{(r,t)};\boldsymbol{y}_i)\right\}\left\{\nabla \ell_k(\bar{\boldsymbol{\theta}}_{\mathcal{P}}^{(r,t)};\boldsymbol{y}_i)\right\}^{\top}$. In all experiments, we burn the first $0.25n$ iterations and start drawing inference and tracking estimates variability from $T_n = .5n$. To correctly assess the empirical coverage of confidence intervals iteration-wise, all simulations track results for $T_n \in \{0.5n, \ldots, 3n\}$. Hence, for each stopping time, we observe all the R runs of Algorithm 1 for a given $\mathcal{P}$. Results are shown for the decreasing stepsize scheduling outlined in Assumption 5 with c set arbitrarily small at $c = 0.501$. The initial step size instead, η_0, is chosen differently for the two examples.

In the experiments we compare the performances of $\bar{\boldsymbol{\theta}}_{\mathcal{P}_1}$ (`standard`), $\bar{\boldsymbol{\theta}}_{\mathcal{P}_2}$ (`bernoulli`), $\bar{\boldsymbol{\theta}}_{\mathcal{P}_3}$ (`hyper`) together with the implementations of $\bar{\boldsymbol{\theta}}_{\mathcal{P}_1}$ and $\bar{\boldsymbol{\theta}}_{\mathcal{P}_3}$ taking advantage of a recycled Sampling Step (`recycle_standard` and `recycle_hyper` respectively) as described in Section 2.4. We also compute $\hat{\boldsymbol{\theta}}$ numerically as a benchmark.

4.1 Experiments for Example 1

Data are generated by using the exact probabilities of observing each of the possible p-variate realizations of the graph. We assume the true graph follows a two-row grid structure, similar to the simulation setting of Lee & Hastie (2015). In particular, horizontal edges are set at 0.5, vertical ones at -0.5, intercepts at -0.5 for odd nodes and 0.5 for even ones. The optimization always starts at the null vector.

We investigate the performances of Algorithm 1 with $n \in \{2,500, 5,000, 10,000\}$ and $p \in \{10, 20\}$, implying $d \in \{55, 210\}$. The value of η_0, is picked by minimising over a grid of possible candidates the mean square error of `standard` at $T_n = 3n$ in the most challenging simulation setting, i.e. $n = 2,500, p = 20$. However, additional simulation results for different choices of η_0 are available in the online Appendix C. Figure 1 shows the convergence of all instances of the proposed estimator in terms of average mean square distance from $\boldsymbol{\theta}^*$. Interestingly, the different noise levels introduced by different sampling schemes characterize the convergence behavior as the estimation proceeds. That is, $\bar{\boldsymbol{\theta}}_{\mathcal{P}_2}$ and $\bar{\boldsymbol{\theta}}_{\mathcal{P}_3}$ share the same asymptotic performances and are preferable to $\bar{\boldsymbol{\theta}}_{\mathcal{P}_1}$. As expected, the recycled implementations of $\mathcal{P}_1$ and $\mathcal{P}_3$ do not show any relevant discrepancy from their non-recycled versions. Furthermore, after reaching $T_n = 3n$, none of the stochastic estimators has reached the MSE of the numerical optimizer. This phenomenon happens because when the stochastic algorithm is stopped, the optimization noise is still non-negligible, which affects the variance considered in the MSE computation. It follows that, since this noise can not be neglected, it must be appropriately considered when quantifying the uncertainty around $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$.

In this regard, Figure 2 shows that, by appropriately accounting for both sources of variability when the algorithm is stopped, it is possible to draw inference about $\boldsymbol{\theta}^*$ using $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$, whatever the choice of $\mathcal{P}$. It presents the empirical coverage levels obtained by constructing confidence intervals following the covariance matrices outlined in Theorem 1 under the three asymptotic regimes. As predicted by the theory, one should use Regime 1 when $T_n = \omega(n)$, and Regime 2 in the opposite scenario, $T_n = o(n)$. However, Regime 3 is the recommendable choice in practice because it directly compounds both the optimization uncertainty and the

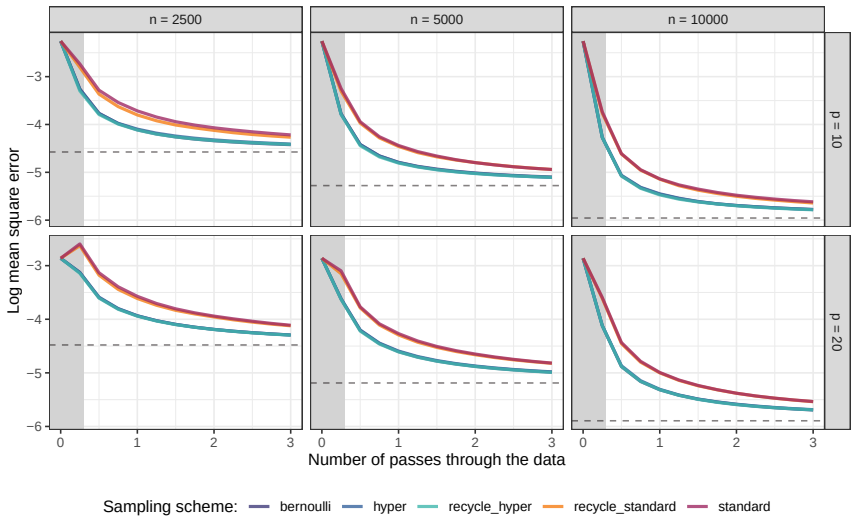

Figure 1: Ising model. Log mean square error trajectories along the optimization, grouped by n and p. Solid lines refer to $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ under different sampling schemes. Dashed lines denote the performance of the numerical approximation of $\hat{\boldsymbol{\theta}}$. Grey areas highlight the burn-in.

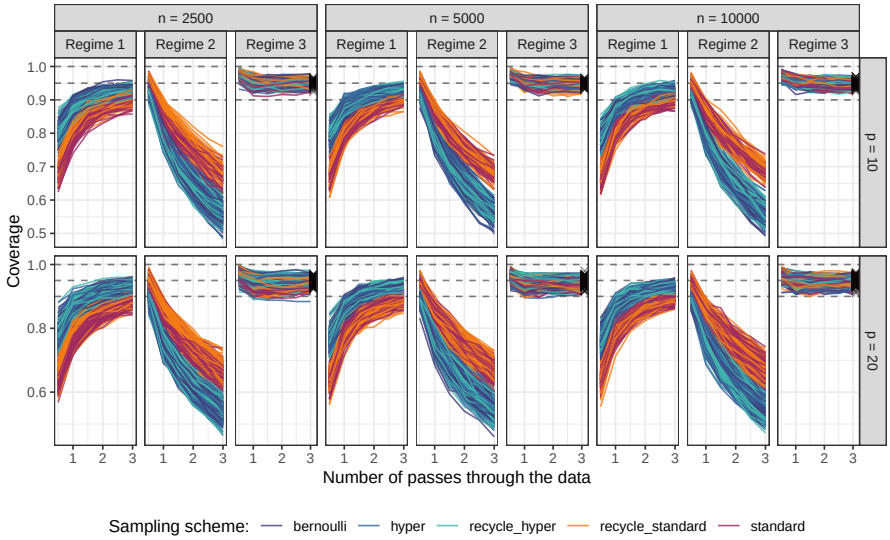

Figure 2: Ising model. Empirical coverage of confidence intervals for $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ constructed according to Theorem 1. Results are grouped by n and p. Dashed lines highlight the nominal coverage level 95%. Solid lines refer to scalar elements of $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ under different sampling schemes. Dark crosses refer to scalar elements of the numerical approximation of $\hat{\boldsymbol{\theta}}$ (placed after the third pass for visualization purposes).

data sampling variability. As a reference, under Regime 3, Figure 2 reports the empirical coverage levels obtained by constructing confidence intervals for the numerical optimizer estimating the asymptotic covariance matrix in (4).

For space reasons, computational times are reported in the online Appendix C. Briefly, taking advantage of a recycling window of iterations is highly beneficial implementation-wise, especially with diverging n. In particular, it allows `recycle_hyper` to be compu-

tationally competitive with `standard` and `recycle_standard` while being systematically more efficient in statistical terms, whatever the choice of T_n (and of η_0, as remarked in the additional experiments reported in the online Appendix).

4.2 Experiments for Example 2

While the previous example clearly shows the statistical convenience of relying on $\mathcal{P}_2$ or $\mathcal{P}_3$ rather than $\mathcal{P}_1$, the experiments in this second example illustrate how these differences vary based on the model considered. Since the discrepancy in the asymptotic covariance of the considered estimators depends on the matrices $\boldsymbol{H}$ and $\boldsymbol{J}$, such a gap can be more or less evident according to the model analyzed. Compared to the previous example, this difference is much more apparent in the gamma frailty model, as illustrated below. Similarly to the previous experiments, we assess the performances of Algorithm 1 for

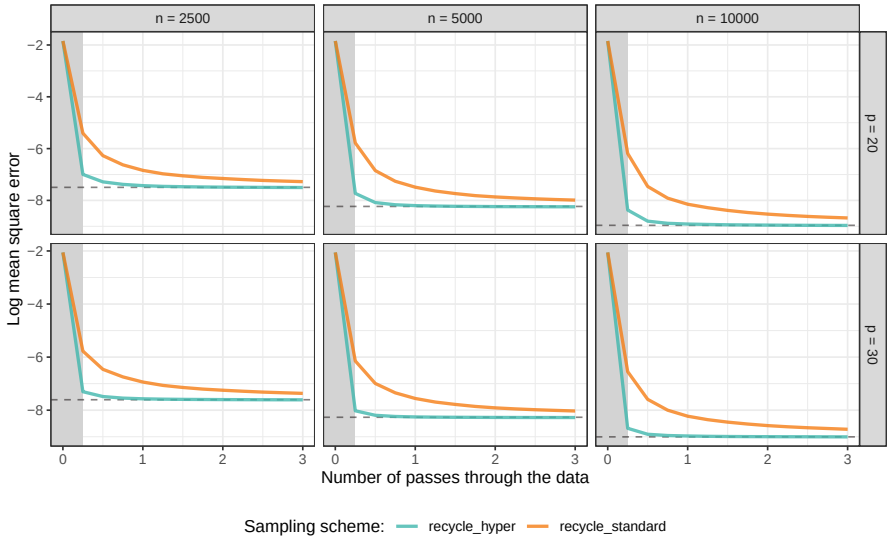

Figure 3: Gamma frailty model. Log mean square error trajectories grouped by n and p. Solid lines refer to $\bar{\boldsymbol{\theta}}_{\mathcal{P}}$ under different sampling schemes. Dashed lines denote the performance of the numerical approximation of $\hat{\boldsymbol{\theta}}$. Grey areas highlight the burn-in.

$n \in \{2,500, 5,000, 10,000\}$ and $p \in \{20, 30\}$. Note that, from $p = 20$ to $p = 30$, the dimension of the parameter space goes from $d = 22$ to $d = 32$ while the computational burden per iteration more than doubles since K increases from $K = 190$ to $K = 435$. Given the large number of likelihood contributions considered compared to the Ising model, we only assess the performances of `recycle_standard` and `recycle_hyper`, not their exact versions. In particular, $\mathcal{P}_3$ is the sampling scheme suffering the most from the hefty K. Accordingly, we increase the recycling window length to $l = 500$ for both estimators to have competitive computational times. For completeness, the online Appendix C provides further experiments showing how different values of l lead to comparable estimates for `recycle_hyper` while having massive impacts on computational times. Similarly to previous experiments, the value of η_0 shown is the stepsize minimizing the mean square error performance at $T_n = 3n$ of `recycle_standard` in the most challenging setting, i.e., $n = 2,500, p = 30$.

Figure 3 shows the trajectories along the optimization of the log mean square error for

the proposed estimators. Similarly to the Ising model, the MSE differences are due to the asymptotic variabilities implied by $\mathcal{P}_1$ and $\mathcal{P}_3$, but they are more pronounced in this case. The estimates based on $\mathcal{P}_3$ exhibit a sharp drop at the beginning of the optimization and reach the performance of the numerical estimator almost always after one pass through the data. For the estimates based on $\mathcal{P}_1$, the convergence is much slower and does not match the numerical approximation even after the maximum length tested of three passes. The optimization noise of $\mathcal{P}_3$ drops almost immediately to negligible levels, leading the variance of $\bar{\boldsymbol{\theta}}_{\mathcal{P}_3}$ to overlap with the one from numerical estimation closely. The noise generated by $\mathcal{P}_1$ persists much longer, translating into higher variances for the stochastic estimates throughout the optimization and, hence, higher MSE. Note that `recycle_standard` is faster than `recycle_hyper` when both are stopped at the same T_n. However, the simulations show that even after $T_n = n$, `recycle_hyper` is already closer to the numerical estimates than `recycle_standard` at $T_n = 3n$. Thus, it represents a more efficient alternative to `recycle_standard` both computationally and statistically. In addition, the low variability of the `recycle_hyper` also allows for larger steps than what shown here, which permits stopping the optimization even earlier than $T_n = n$. For space reasons, we refer to the online Appendix C for additional details and results about the experiments of this section.

5 A Network Analysis of Mental Health Data

To illustrate the power of the proposed methodology, we consider an application of the Ising model to the mental health data from the Epidemiologic Survey on Alcohol and Related Conditions (NESARC) - Wave 1. The NESARC is a nationally representative survey of the United States adult population, which gathered data on alcohol behavior and mental health disorders from April 2001 to June 2002 (Grant et al. 2003). We take the network psychometrics approach (e.g., Epskamp et al. 2018, Borsboom 2022), viewing symptoms as nodes of an unknown graph and direct symptom-to-symptom interactions as edges whose parameters are to be estimated. We select $p = 32$ items related to antisocial disorders, high mood, low mood, panic and personality disorders, and social and other specific forms of phobia. Therefore, the dimension of the parameter space is $d = 528$. The items are selected among the ones with the lowest missing response rates, avoiding screening items and related ones, and the remaining observations with missing values were discarded, leaving the dataset with a total of $31,826$ respondents. See the online Appendix D for the description of the 32 items considered. We hold out 10% of the available observations as a validation set to monitor the out-of-sample behavior of the negative composite log-likelihood during the iterations. The training partition retains $n = 28,643$ observations.

The model is estimated using the hypergeometric sampling of Definition 3. Given the large sample size, we set the recycling window at $l = 1,000$ and burn-in period $B = 0.25n$. The stepsize scheduling is defined by $c = .501$ and η_0 chosen by halving an initial proposal until the holdout negative composite log-likelihood performance ceases improving when evaluated at $T_n = n$. The selected value is $\eta_0 = 5$. After every $0.25n$ iterations, the algorithm performs a new evaluation of the holdout negative log-likelihood. When the improvement falls under 0.1%, the algorithm stops. In our case, it stops at $T_n = 1.75n$. The full estimation procedure, including the initial stepsize selection, took almost 15 seconds

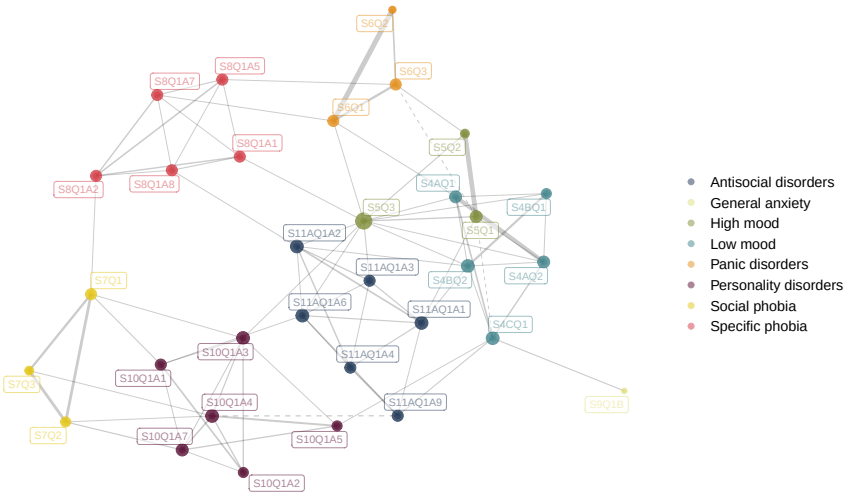

Figure 4: Graphical structure of mental health disorders. Node colors refer to specific survey areas. Solid and dashed lines stand for positive and negative estimated edges. Edge width is proportional to the absolute estimated coefficient.

when run on a single core of a personal laptop[2]. As a benchmark, the numerical estimator took more than half an hour to converge on the same hardware, providing similar results.

At the end of the stochastic estimation, the asymptotic standard errors are computed by estimating the covariance matrix of $\bar{\boldsymbol{\theta}}_{\mathcal{P}_3}$ under Regime 3 of Theorem 1 using the usual sample estimators of $\boldsymbol{H}$ and $\boldsymbol{J}$. To investigate the structure of the estimated graphical model, all the d parameters are tested against the null hypothesis of being zero. The resulting p-values are then adjusted via the Holm correction (Holm 1979) to control for the family-wise error rate across the d hypothesis at level 0.01. The procedure identifies 17.1% of the possible edges as statistically significant, as visualized in Figure 4.

The identified graphical structure highlights how symptoms of the same disorder tend to cluster together with dense positive connections. Instead, relationships among different disorders are less strict, with some of them being slightly negative. Furthermore, the sparsity induced by the non-significance of many edges allows for providing conditional independence statements among symptom areas. For example, panic disorders are conditionally independent of the rest of the graph, given the specific phobias and mood disorders areas. The same can be said for the social phobia area, which is independent of mood disorders, for example, when conditioned on personality disorders and other specific types of phobias. Similar reasoning can be used to investigate symptoms belonging to the same area. For example, item `S6Q2` concerns the experience of feeling erroneously in danger after a panic attack. This symptom is isolated from the others when the remaining two items related to panic attacks, `S6Q1` and `S643`, are considered. In particular, such items refer to experiencing panic episodes for no real reason and misinterpreting nerves as a heart attack. Finally, it can also happen that single items separate two specific portions of the graph. That is the case of the node `S4CQ1`, which concerns having experienced two or more years of depression and separates from the rest of the symptoms item `S9Q1B`, which is related to experiencing six months or longer of nervousness about everyday problems.

[2]Intel i5-2520M; RAM 8 GB; R version 4.3.0; gcc version 13.1.1; 4x 3.2GHz, OS Manjaro Linux 23.0.0

6 Discussion

When the optimization noise is non-negligible, it is crucial to properly quantify the uncertainty around stochastic approximations if the goal is to run valid frequentist inferences about true parameters. We show how the asymptotic variance of such estimators compounds two sources of uncertainty: the sampling variability of the data and the noise injected in the procedure by the SGs. We optimize composite likelihoods by constructing the SGs using a hypergeometric sampling of sub-likelihood contributions, which enhances their statistical and computational efficiencies. The resulting estimator is a flexible inferential tool for applied research. In contrast to existing methods that discard part of the data to reduce computational times (Dillon & Lebanon 2010, Mazo et al. 2023), our proposal utilizes available information more parsimoniously, leading to improved statistical efficiency. Additionally, a small experiment in online Appendix E compares our method with the randomized pairwise likelihood estimator of Mazo et al. (2023). The results confirm that spreading the usage of likelihood components across iterations via stochastic optimization rather than discarding them leads to improved estimation performances. Various extensions of the proposed method are possible by expanding the scope of the parameter update in Algorithm 1. A first straightforward extension relates to quasi-Newton alternatives of standard SG descent (Byrd et al. 2011). Such extensions can be quite effective in practice because they adapt the steps to the different scales of each parameter, which typically improves the convergence of the estimator. A second forthright extension enriches the update step to account for proximal operators, allowing for non-differentiable terms like projections and lasso penalties, as investigated in Atchadé et al. (2017) and Zhang & Chen (2022).

Nevertheless, the current proposal still has limitations, particularly when making inferences with an increasing parameter space. From a computational perspective, it does not address the challenge of computing $\boldsymbol{H}^{-1}$ and $\boldsymbol{J}$. It is important to have accurate estimates of both $\boldsymbol{H}^{-1}$ and $\boldsymbol{J}$ to construct reliable confidence intervals. However, estimating these quantities can be computationally challenging, especially with large numbers of parameters due to the matrix inversion required. Furthermore, this work focuses on settings where traditional frequentist estimation is theoretically adequate but computationally inconvenient, such as moderate parameter spaces with much larger sample sizes. Further research is needed to expand the current theoretical framework to settings where a regularization term is necessary to identify the parameters of interest. Conducting inference in such settings is complicated due to the bias introduced by regularization. We are exploring potential solutions based on recent advances in debiasing techniques for lasso-based estimators. These methods have gained popularity in both offline settings (Janková & van de Geer 2018) and with streaming data (Han et al. 2023). Addressing these theoretical challenges could enable composite likelihood inference for large-scale data.

References

Atchadé, Y. F., Fort, G. & Moulines, E. (2017), 'On perturbed proximal gradient algorithms', *The Journal of Machine Learning Research* **18**(1), 310–342.

Bellio, R. & Varin, C. (2005), 'A pairwise likelihood approach to generalized linear models with crossed random effects', *Statistical Modelling* **5**(3), 217–227.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2), 192–225.

Blum, J. R., Chernoff, H., Rosenblatt, M. & Teicher, H. (1958), 'Central limit theorems for interchangeable processes', *Canadian Journal of Mathematics* **10**, 222–229.

Borsboom, D. (2022), 'Possible futures for network psychometrics', *Psychometrika* **87**(1), 253–265.

Bottou, L., Curtis, F. E. & Nocedal, J. (2018), 'Optimization methods for large-scale machine learning', *SIAM Review* **60**(2), 223–311.

Byrd, R. H., Chin, G. M., Neveitt, W. & Nocedal, J. (2011), 'On the use of stochastic hessian information in optimization methods for machine learning', *SIAM Journal on Optimization* **21**(3), 977–995.

Chee, J., Kim, H. & Toulis, P. (2023), "plus/minus the learning rate": Easy and scalable statistical inference with SGD, *in* 'Proceedings of The 26th International Conference on Artificial Intelligence and Statistics', PMLR, pp. 2285–2309.

Chen, X., Lai, Z., Li, H. & Zhang, Y. (2024), 'Online statistical inference for stochastic optimization via Kiefer-Wolfowitz Methods', *Journal of the American Statistical Association* pp. 1–24.

Chen, X., Lee, J. D., Tong, X. T. & Zhang, Y. (2020), 'Statistical inference for model parameters in stochastic gradient descent', *The Annals of Statistics* **48**(1), 251–273.

Chen, X., Lee, S., Liao, Y., Seo, M. H., Shin, Y. & Song, M. (2023), 'SGMM: Stochastic approximation to generalized method of moments', *Journal of Financial Econometrics* .

Dillon, J. V. & Lebanon, G. (2010), 'Stochastic composite likelihood', *The Journal of Machine Learning Research* **11**, 2597–2633.

Epskamp, S., Borsboom, D. & Fried, E. I. (2018), 'Estimating psychological networks and their accuracy: A tutorial paper', *Behavior Research Methods* **50**(1), 195–212.

Fang, Y., Xu, J. & Yang, L. (2018), 'Online bootstrap confidence intervals for the stochastic gradient descent estimator', *Journal of Machine Learning Research* **19**(78), 1–21.

Grant, B. F., Moore, T., Shepard, J. & Kaplan, K. (2003), 'Source and accuracy statement: Wave 1 national epidemiologic survey on alcohol and related conditions (nesarc)', *Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism* **52**.

Han, R., Luo, L., Lin, Y. & Huang, J. (2023), 'Online inference with debiased stochastic gradient descent', *Biometrika* **111**(1), 93–108.

Henderson, R. & Shimakura, S. (2003), 'A serially correlated gamma frailty model for longitudinal count data', *Biometrika* **90**(2), 355–366.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**(2), 65–70.

Ising, E. (1924), Beitrag zur theorie des ferro-und paramagnetismus, PhD thesis, Grefe & Tiedemann Hamburg.

Janková, J. & van de Geer, S. (2018), Inference in high-dimensional graphical models, *in* 'Handbook of Graphical Models', CRC Press, pp. 325–350.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F. & Jöreskog, K. G. (2012), 'Pairwise likelihood estimation for factor analysis models with ordinal data', *Computational Statistics & Data Analysis* **56**(12), 4243–4258.

Lee, J. D. & Hastie, T. J. (2015), 'Learning the structure of mixed graphical models', *Journal of Computational and Graphical Statistics* **24**(1), 230–253.

Lee, S., Liao, Y., Seo, M. H. & Shin, Y. (2022), 'Fast and robust online inference with stochastic gradient descent via random scaling', *Proceedings of the AAAI Conference on Artificial Intelligence* **36**(7), 7381–7389.

Lindsay, B. G. (1988), 'Composite likelihood methods', *Contemporary Mathematics* **80**(1), 221–239.

Mazo, G., Karlis, D. & Rau, A. (2023), 'A randomized pairwise likelihood method for complex statistical inferences', *Journal of the American Statistical Association* pp. 1–11.

Moulines, E. & Bach, F. (2011), 'Non-asymptotic analysis of stochastic approximation algorithms for machine learning', *Advances in Neural Information Processing Systems* **24**.

Needell, D., Ward, R. & Srebro, N. (2014), 'Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm', *Advances in Neural Information Processing Systems* **27**.

Polyak, B. T. (1987), *Introduction to Optimization*, Optimization Software, Inc., Publications Division.

Polyak, B. T. & Juditsky, A. B. (1992), 'Acceleration of stochastic approximation by averaging', *SIAM Journal on Control and Optimization* **30**(4), 838–855.

Robbins, H. & Monro, S. (1951), 'A stochastic approximation method', *The Annals of Mathematical Statistics* **22**(3), 400–407.

Ruppert, D. (1988), Efficient estimations from a slowly convergent Robbins-Monro process, Technical report, Cornell University Operations Research and Industrial Engineering.

Su, W. J. & Zhu, Y. (2023), 'Higrad: Uncertainty quantification for online learning and stochastic approximation', *Journal of Machine Learning Research* **24**(124), 1–53.

Toulis, P. & Airoldi, E. M. (2017), 'Asymptotic and finite-sample properties of estimators based on stochastic gradients', *The Annals of Statistics* **45**(4), 1694–1727.

Varin, C., Reid, N. & Firth, D. (2011), 'An overview of composite likelihood methods', *Statistica Sinica* **21**(1), 5–42.

Varin, C. & Vidoni, P. (2005), 'A note on composite likelihood inference and model selection', *Biometrika* **92**(3), 519–528.

Wei, Z., Zhu, W. & Wu, W. B. (2023), 'Weighted averaged stochastic gradient descent: Asymptotic normality and optimality', *arXiv:2307.06915* .

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50**(1), 1–25.

Zhang, S. & Chen, Y. (2022), 'Computation for latent variable model estimation: A unified stochastic proximal framework', *Psychometrika* **87**(4), 1473–1502.

Zhu, W., Chen, X. & Wu, W. B. (2023), 'Online covariance matrix estimation in stochastic gradient descent', *Journal of the American Statistical Association* **118**(541), 393–404.