Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 1–18
https://doi.org/10.1214/25-BJPS621

Finite-sample bounds to the normal limit under group sequential sampling

Julian Aronowitz[1,a] **and Jay Bartroff**[2,b]

[1]*Google Inc., New York City, U.S.A.,* [a]*julian.aronowitz@gmail.com*
[2]*Department of Statistics & Data Sciences, University of Texas at Austin, Austin, U.S.A.,* [b]*bartroff@austin.utexas.edu*

Abstract. In group sequential analysis, data is collected and analyzed in batches until pre-defined stopping criteria are met. Inference in the parametric setup typically relies on the limiting asymptotic multivariate normality of the repeatedly computed maximum likelihood estimators (MLEs), a result first rigorously proved by Jennison and Turnbull (1997) under general regularity conditions. In this work, using Stein's method we provide optimal order, non-asymptotic bounds on the distance for smooth test functions between the joint group sequential MLEs and the appropriate normal distribution under the same conditions. Our results assume independent observations but allow heterogeneous (i.e., non-identically distributed) data. We examine how the resulting bounds simplify when the data comes from an exponential family. Finally, we present a general result relating multivariate Kolmogorov distance to smooth function distance which, in addition to extending our results to the former metric, may be of independent interest.

1 Introduction

Sequential analysis is a powerful statistical framework in which sample sizes are not determined before conducting a study, and is the dominant statistical methodology in many types of clinical trials (Bartroff, Lai and Shih, 2013), among other applications. In clinical trials in particular, data is often collected and analyzed in *groups* until the conditions of a pre-defined stopping criteria are met. In their authoritative textbook on this topic, Jennison and Turnbull (2000) note that group sequential analysis can significantly reduce the time required to conduct clinical trials over the traditional fixed sample size designs.

Group sequential analysis (e.g., O'Brien and Fleming, 1979; Pocock, 1977) is typically based on repeatedly computed (after each group) maximum likelihood estimators (MLEs), and therefore all the properties of the statistical test (e.g., expected sample size, type I error probability, power) are functions of the *joint* distribution of these repeatedly computed MLEs. Suppose independent but not necessarily identically distributed observations $Y_1, \dots, Y_n \in \mathbb{R}^t$ are divided into K groups with n_k denoting the number of observations seen up to and including group k for $k = 1, \dots, K$. Let $f_i(y_i; \theta)$ be the probability mass or density function of Y_i with $\theta \in \Theta \subset \mathbb{R}^d$. We let $\hat{\theta}_k = \hat{\theta}_k(Y_1, Y_2, \dots, Y_{n_k}) \in \mathbb{R}^{d\times 1}$ be the MLE based on the observations in the first k groups and $\hat{\theta}^K = [\hat{\theta}_1^\intercal, \hat{\theta}_2^\intercal, \dots, \hat{\theta}_K^\intercal]^\intercal \in \mathbb{R}^q$, where $q = dK$, be the group sequential MLE. Here and throughout, $(\cdot)^\intercal$ denotes transpose. Although the exact identification of the distribution of $\hat{\theta}^K$ is often intractable, Jennison and Turnbull (1997a) showed that it is asymptotically multivariate normal under suitable regularity conditions (see Theorem 2.2). This result is perhaps unsurprising in light of the classical result that MLEs are asymptotically normal under the same regularity conditions; see Theorem 2.1 below.

Key words and phrases. Central limit theorem, group sequential, maximum likelihood estimation, Stein's method.

Received November 2023; accepted January 2025.

Natural questions that then arise from this formulation are how many samples are required, and in which groups should the samples be allocated, for the asymptotics to 'kick in' and a normal approximation to become appropriate without introducing too much error? One form of an answer to this question would be an upper bound on the error of the multivariate normal approximation of the group sequential MLE. In this work we provide an optimal order bound on the distance between the group sequential MLE and the appropriate normal distribution under the same regularity conditions with heterogeneous, independent data (our Theorem 3.1). The proof relies on Stein's method (Stein, 1972) and Taylor series arguments and has similarities to the approach of Anastasiou (2018) in which a bound was derived on the multivariate MLE but without the joint, repeatedly computed group sequential structure. Our result generalizes Anastasiou (2018) since when there is only one group our bound reduces to that found in that work.

In this work, we give an upper bound on

$$\big|\mathbb{E}\big[h(X)\big] - \mathbb{E}\big[h(Z)\big]\big|,$$

where Z is the standard multivariate normal, X is the normalized group sequential MLE, and h is a test function in some function class $\mathcal{H}$. In particular, for any three times differentiable function $h : \mathbb{R}^q \to \mathbb{R}$, we denote $\|h\| := \sup|h|$, $\|h\|_1 := \sup_i |\frac{\partial}{\partial x_i} h|$, $\|h\|_2 := \sup_{i,j} |\frac{\partial^2}{\partial x_i \partial x_j} h|$, and $\|h\|_3 := \sup_{i,j,k} |\frac{\partial^3}{\partial x_i \partial x_j \partial x_k} h|$. Then let

$$\begin{aligned}\mathcal{H} = \big\{&h : \mathbb{R}^q \to \mathbb{R} : h \text{ is three times differentiable with bounded} \\ &\|h\|, \|h\|_1, \|h\|_2, \|h\|_3\big\}. \end{aligned} \tag{1}$$

A non-asymptotic bound for the univariate MLE was first developed by Anastasiou and Reinert (2017) and then expanded to the multivariate case by Anastasiou (2018). In Anastasiou and Ley (2017) and Anastasiou and Gaunt (2020), the bounds for the univariate and multivariate cases, respectively, were sharpened and simplified under the additional assumption that the MLE follows a special form. Let $\boldsymbol{Y} = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n})$ be a random sample of n i.i.d. t-dimensional random vectors, $\Theta \subset \mathbb{R}^d$, and $\hat{\boldsymbol{\theta}}_n(\boldsymbol{Y})$ be the resulting MLE. The special form considered by these authors is

$$p\big(\hat{\boldsymbol{\theta}}_n(\boldsymbol{Y})\big) = \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{Y_i}) = \frac{1}{n}\sum_{i=1}^{n}\big(g_1(\boldsymbol{y_i}), \ldots, g_d(\boldsymbol{y_i})\big)^{\intercal} \tag{2}$$

for some $g : \mathbb{R}^t \to \mathbb{R}^d$ and $p : \Theta \to \mathbb{R}^d$. One simple example of an MLE that follows this form is the case of independent normal data with unknown mean and variance. Anastasiou (2017) considers data that is m-dependent for identically distributed scalar parameter MLEs. The group sequential MLE is necessarily multivariate and so this particular result will not be generalizable to our setting.

Each of the bounds in Anastasiou and Reinert (2017), Anastasiou (2017, 2018), Anastasiou and Ley (2017), and Anastasiou and Gaunt (2020) are of optimal order $\mathcal{O}(n^{-1/2})$ and defined in terms of slightly different $\mathcal{H}$'s where in general the multidimensional bounds require additional bounded derivatives. All of these results use Stein's method techniques and require $h \in \mathcal{H}$ to be at least bounded and absolutely continuous. Pinelis (2017) gives a bound of the optimal order on the rate of convergence to normality for univariate MLEs in terms of the Kolmogorov distance, $\mathcal{H} = \{\mathbb{1}\{x \le a\} : a \in \mathbb{R}\}$. Unlike the above results this bound does not make use of Stein's method. Our Corollary 6.2 can be viewed as a generalization of this to multivariate MLEs and the group sequential setting, and is the first that we are aware of by any method.

Bounds derived from characteristic functions (Ulyanov, 1979, 1986, 1987) could be of use in our setting, if not for two potential hurdles. First, in the multivariate setting these methods require independence. As one goal here is to develop a method that may later be applicable to observations with some form of dependence, characteristic function methods appear to be insufficient for that goal. Second, these methods do not provide explicit constants for the bounds produced.

The rest of this paper is organized as follows. In Section 2, we introduce the remaining needed background on MLEs and Stein's method. In Section 3, we present our main result, which we then specialize to exponential family distributions and the exponential distribution in Section 4. And in Section 6, we present results relating multivariate Kolmogorov distance to smooth function distance, extending our results to the former metric which gives bounds to normality for multivariate MLEs in terms of Kolmogorov distance.

2 Background

2.1 Stein's method

The results detailed below make heavy use of Stein's method bounds in combination with multivariate Taylor series arguments. We do not attempt to give a complete introduction to Stein's method here, but rather focus on the relevant multivariate results. Readers interested in a more broad introduction are referred to Chen, Goldstein and Shao (2010).

Stein's method for multivariate normal approximations hinges on the following multivariate version of the Stein equation. Let $f : \mathbb{R}^q \to \mathbb{R}$ be twice differentiable and D^2 the second derivative, or Hessian matrix, of f. Then for $h : \mathbb{R}^q \to \mathbb{R}$ a multivariate Stein equation is

$$\operatorname{Tr} D^2 f(\boldsymbol{w}) - \boldsymbol{w} \cdot \nabla f(\boldsymbol{w}) = h(\boldsymbol{w}) - Nh.$$

Here Nh denotes the expectation $\mathbb{E}h(Z)$, where Z is a q-dimensional standard normal vector. The following bound of Goldstein and Rinott (1996) relates derivatives of the Stein equation solution f to those of the test function h:

$$\left\| \frac{\partial^k f(\boldsymbol{w})}{\prod_{j=1}^k \partial w_{i_j}} \right\| \le \frac{1}{k} \left\| \frac{\partial^k h(\boldsymbol{w})}{\prod_{j=1}^k \partial w_{i_j}} \right\|, \quad k \ge 1, \tag{3}$$

when the kth partial derivative of h exists.

2.2 Asymptotic normality of maximum likelihood estimators

For $Y_1, Y_2, \dots, Y_n \in \mathbb{R}^t$ independent but not necessarily identically distributed, let $f_i(y_i; \theta)$ be the probability mass or density function of Y_i with the parameter space Θ an open subset of $\mathbb{R}^d$. Suppose the observations $Y_1, \dots, Y_n$ are divided into K groups with n_k denoting the number of observations seen up to and including group k for $k = 1, \dots, K$ with $n = n_K$. For example, $\{Y_1, Y_2, \dots, Y_{n_3}\}$ denotes all observations in the first 3 groups. We let $\hat{\theta}_k = \hat{\theta}_k(Y_1, Y_2, \dots, Y_{n_k}) \in \mathbb{R}^d$ denote the MLE based on the observations in the first k groups, and $G_k = \{n_{k-1} + 1, \dots, n_k\}$ the set of indices of the observations in group k. The likelihood function at analysis k is $L_k(\theta; \boldsymbol{y}) = \prod_{i=1}^{n_k} f_i(y_i|\theta)$ and the log-likelihood function is $\ell_k(\theta, \boldsymbol{y}) = \log(L_k(\theta; \boldsymbol{y}))$.

Below we state well-known regularity conditions that are sufficient for the asymptotic normality of the fixed-sample MLE, which we record as Theorem 2.1. These conditions are also sufficient for asymptotic normality in the group sequential setting, a result due to Jennison and Turnbull (1997a) which we record here as Theorem 2.2.

(R1) $\hat{\theta}_n(Y) \xrightarrow{p} \theta_0$, as $n \to \infty$, where θ_0 is the true parameter vector.

(R2) The score function, $S_i(Y_i;\theta) = \nabla_\theta \log f_i(Y_i;\theta) \in \mathbb{R}^{d\times 1}$ and the information matrix, $I_i(\theta) = \mathrm{Var}_\theta[S_i(Y_i;\theta)] \in \mathbb{R}^{d\times d}, i = 1,\ldots,n$ exist almost surely with respect to the probability measure $\mathbb{P}$.

(R3) $I_i(\theta)$ is a continuous function of $\theta, \forall i = 1,2,\ldots,n$, almost surely with respect to $\mathbb{P}$ and is a measurable function of Y_i.

(R4) $\mathbb{E}_\theta[S_i(Y_i;\theta)] = 0_d \in \mathbb{R}^{d\times 1}$ where 0_d is the d-column vector of all zeros.

(R5) $I_i(\theta) = \mathbb{E}_\theta[[\nabla \log f_i(Y_i;\theta)][\nabla \log f_i(Y_i;\theta)]^\intercal] = \mathbb{E}_\theta[-\nabla_\theta S_i^\intercal(Y_i;\theta)]$.

(R6) For

$$\bar{I}_n(k,\theta) = \frac{1}{n}\sum_{i=1}^{n_k} I_i(\theta), \quad k = 1,\ldots,K,$$

there exists a matrix $\bar{I}(k,\theta) = \lim_{n\to\infty} \bar{I}_n(k,\theta)$. In addition, $\bar{I}_n(k,\theta)$ and $\bar{I}(k,\theta)$ are symmetric and $\bar{I}(k,\theta)$ is positive definite for all θ.

(R7) For some $\delta > 0$,

$$\frac{\sum_i \mathbb{E}_{\theta_0}|\boldsymbol{\lambda}^\intercal S_i(Y_i;\theta)|^{2+\delta}}{n^{\frac{2+\delta}{2}}} \underset{n\to\infty}{\longrightarrow} 0 \quad \text{for all } \boldsymbol{\lambda} \in \mathbb{R}^d.$$

(R8) With $\|\cdot\|$ the ordinary Euclidean norm on $\mathbb{R}^d$, then for $i,j,u \in \{1,2,\ldots,d\}$ there exists $\epsilon > 0, C > 0, \delta > 0$ and random variables $B_{i,j,u}(Y_u)$ such that

(i) $\sup\{|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(f_u(Y_u;t))| \,|\, \|t-\theta_0\| \le \epsilon\} \le B_{i,j,u}(Y_u)$,

(ii) $\mathbb{E}|B_{i,j,u}(Y_u)|^{1+\delta} \le C$.

The classical result that MLEs are asymptotically normal is the following. See, for example, van der Vaart (2007) for a proof, and Hoadley (1971) for a further discussion of the sufficient conditions.

Theorem 2.1. *Let $Y_1, Y_2, \ldots, Y_n$ be independent random vectors with probability density (or mass) function $f_i(y_i;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$. Assume that the MLE exists and is unique and that the regularity conditions (R1)–(R8) hold. Also let $Z \sim \mathcal{N}_d(0_d, 1_{d\times d})$ be the standard multivariate d-dimensional normal. Then*

$$\sqrt{n}[\bar{I}_n(\theta_0)]^{\frac{1}{2}}(\hat{\theta}_n(Y) - \theta_0) \underset{n\to\infty}{\xrightarrow{d}} Z$$

Theorem 2.2 (Jennison and Turnbull (1997a)). *Suppose that observations Y_i are independent with distributions $f_i(y_i;\theta)$, where θ is d-dimensional, and that observations $Y_1,\ldots,Y_{n_k}$ are available at analysis $k, k = 1,\ldots,K$. Let $n_k - n_{k-1} \to \infty$ such that $(n_k - n_{k-1})/n \to \gamma_k \in (0,1)$ for all $k = 1,\ldots,K$. Furthermore let $\hat{\theta}_k$ denote the MLE of θ based on $Y_1,\ldots,Y_{n_k}$. Suppose that the distributions f_i are sufficiently regular so that (R1)–(R8) hold for each k. Also let $Z \sim \mathcal{N}_q(0_q, 1_{q\times q})$ and $\theta_0^K := [\theta_0^\intercal, \theta_0^\intercal, \ldots, \theta_0^\intercal]^\intercal \in \mathbb{R}^q$. Then $\hat{\theta}^K = [\hat{\theta}_1^\intercal, \ldots, \hat{\theta}_K^\intercal]^\intercal$ is asymptotically multivariate normal,*

$$\sqrt{n}[J_n]^{-\frac{1}{2}}(\hat{\theta}^K - \theta_0^K) \underset{n\to\infty}{\xrightarrow{d}} Z,$$

where,

$$J_n = J_n(\theta_0) = \begin{bmatrix} \bar{I}_n^{-1}(1,\theta_0) & \bar{I}_n^{-1}(2,\theta_0) & \bar{I}_n^{-1}(3,\theta_0) & \ldots & \bar{I}_n^{-1}(K,\theta_0) \\ \bar{I}_n^{-1}(2,\theta_0) & \bar{I}_n^{-1}(2,\theta_0) & \bar{I}_n^{-1}(3,\theta_0) & & \\ \bar{I}_n^{-1}(3,\theta_0) & \bar{I}_n^{-1}(3,\theta_0) & \bar{I}_n^{-1}(3,\theta_0) & & \\ & \vdots & & \ddots & \\ \bar{I}_n^{-1}(K,\theta_0) & \bar{I}_n^{-1}(K,\theta_0) & \bar{I}_n^{-1}(K,\theta_0) & \ldots & \bar{I}_n^{-1}(K,\theta_0) \end{bmatrix}.$$

The regularity conditions (R1)–(R8) are assumed for limiting normality of the MLE in the fixed-sample (Theorem 2.1) and group sequential (Theorem 2.2) cases, and in our finite-sample bounds to these limits we assume (R1)–(R6), plus some specialized conditions analogous to (R7)–(R8) stated in those theorems, below. There is no simple description of all distributions which satisfy these sets of regularity conditions, which depend not only on the particular sequence of densities $f_1, f_2, \ldots$ and their parameterizations, but also on design considerations such as group sizes which determine the amount of information accumulated with each observation, or group of observations. For example, condition (R6) constrains the experimental design so that the fraction of information contributed by any individual observation decreases to zero as the sample size increases. However, here we mention a few large classes of distributions for which the needed distributional conditions can be shown to hold. In the normal linear model, if the covariance structure is known then the normal limits hold exactly. If the variance is unknown, then the asymptotic theory can be applied to the mean and variance estimators for the joint distribution of group sequential Student's t-statistics, say (Jennison and Turnbull, 1997b). For non-normal data, the asymptotic theory applies to a large class of parametric regression models, which Jennison and Turnbull (1997a, p. 1335) write includes "the entire family of generalized linear models" (GLMs), again subject to the above regularity and design conditions. There are well-known conditions which cause MLEs to not exist in GLMs including separation, sparsity, parameter space constraints, and ill-conditioned design matrices (see Verbeek, 1992), however the regularity conditions rule these out; for more details on GLMs see McCullagh and Nelder (1989). A special case of GLMs are non-regression exponential family models, to which we apply our results in Section 4. Other authors have shown that the needed conditions can be verified even in some spurious cases in GLMs, and non GLM-models. For example, Hoadley (1971) shows that independent but non-identically distributed observations from censored exponential distributions satisfy the conditions, and Jennison and Turnbull (2000, Chapters 3, 13) show that the log-rank test statistics and semi-parametric models for survival data subject to censoring have the needed regularity for asymptotic normality of the MLE.

3 A bound to the normal for group sequential maximum likelihood estimators

The asymptotic theory of group sequential MLEs in Theorem 2.2 due to Jennison and Turnbull (1997a) guarantee asymptotic normality under the appropriate conditions. In this section, we present results that give an error bound for the normal approximation under the same conditions. In Section 3.1, we present the main result in Theorem 3.1, whose proof is in the Supplementary Material (Aronowitz and Bartroff (2025)). After the theorem, we outline an argument showing that the bound is asymptotically $\mathcal{O}(n^{-1/2})$. In Section 3.2, we discuss a result of Gaunt (2016) allowing a relaxation on the number of needed derivatives of the test function h, and apply this to achieve a similar relaxation to Reinert and Röllin (2009, Theorem 2.1) in Theorem 3.3, which may be of independent interest. Then in Theorem 3.3, we pass this relaxation along to Theorem 3.1.

3.1 Main result

For ease of presentation, we introduce the following additional notation. For

$$\theta_0^K := [\theta_0^\intercal, \theta_0^\intercal, \ldots, \theta_0^\intercal]^\intercal \in \mathbb{R}^q,$$

let $Q_i := [\hat{\theta}^K - \theta_0^K]_i$ and $Q_{(m)} := \max_{i\in\{1,\ldots,q\}} Q_i$. Let Y_i' be an independent copy of Y,

$$\xi_{ij} = \left[n^{-\frac{1}{2}} S_i(Y_i, \theta_0)\right]_j \quad \text{and} \quad \xi_{ij}' = \left[n^{-\frac{1}{2}} S_i(Y_i', \theta_0)\right]_j.$$

Define

$$A=\begin{bmatrix} 1_{d\times d} & 0 & 0 & \dots & 0 \\ 1_{d\times d} & 1_{d\times d} & 0 & & 0 \\ 1_{d\times d} & 1_{d\times d} & 1_{d\times d} & & 0 \\ & \vdots & & \ddots & \\ 1_{d\times d} & 1_{d\times d} & 1_{d\times d} & \dots & 1_{d\times d} \end{bmatrix} \in \mathbb{R}^{q\times q},$$

$$\tilde{J}_n=\tilde{J}_n(\theta)=\begin{bmatrix} \bar{I}_n(1,\theta) & \bar{I}_n(1,\theta) & \bar{I}_n(1,\theta) & & \bar{I}_n(1,\theta) \\ \bar{I}_n(1,\theta) & \bar{I}_n(2,\theta) & \bar{I}_n(2,\theta) & \dots & \bar{I}_n(2,\theta) \\ \bar{I}_n(1,\theta) & \bar{I}_n(2,\theta) & \bar{I}_n(3,\theta) & & \bar{I}_n(3,\theta) \\ & \vdots & & \ddots & \\ \bar{I}_n(1,\theta) & \bar{I}_n(2,\theta) & \bar{I}_n(3,\theta) & \dots & \bar{I}_n(K,\theta) \end{bmatrix} \in \mathbb{R}^{q\times q},$$

where $1_{d\times d}$ is the d-dimensional identity matrix. We denote the nth row and mth column of a matrix W as

$$n\text{th row: } W_{n*}=W_{n,*},$$

$$m\text{th columns: } W_{*m}=W_{*,m}.$$

When W is defined as a block matrix, we let $W_{[i][j]}$ be the sub matrix in 'block row' i and 'block column' j. Similarly when $W\in\mathbb{R}^q$ is defined as a block vector, we let $W_{[i]}\in\mathbb{R}^d$ be the ith sub vector. We denote the nth block row and mth block column as

$$n\text{th block row: } W_{[n][*]}=W_{[n],[*]},$$

$$m\text{th block columns: } W_{[*][m]}=W_{[*],[m]}.$$

We have the notation in place to state the main result of the paper.

Theorem 3.1. *Let $Y_1,Y_2,\dots,Y_n$ be independent but possibly non-identically distributed $\mathbb{R}^t$-valued random vectors with probability density (or mass) functions $f_i(y_i|\theta)$, for which the parameter space Θ is an open subset of $\mathbb{R}^d$. Suppose the observations $Y_1,Y_2,\dots,Y_{n_k}$ are available at analysis $k=1,\dots,K$. Assume that the MLE $\hat{\theta}_k$ exists and is unique and that conditions (R1)–(R6) are satisfied at each analysis k. In addition, assume that for any θ_0 there exists $0<\epsilon=\epsilon(\theta_0)$ and functions $M^k_{iuj}(y)$, $\forall i,u,j=1,2,\dots,d$ such that $|\frac{\partial^3}{\partial\theta_i\partial\theta_u\partial\theta_j}\ell_k(\theta,y)|\le M^k_{iuj}(y)$ for all $\theta\in\Theta$ with $|\theta_j-\theta_{0,j}|<\epsilon$ $\forall j=1,2,\dots,d$. Also, assume that $\mathbb{E}[(\frac{\partial^3}{\partial\theta_i\partial\theta_u\partial\theta_j}M^k_{iuj}(Y))^2||Q_{(m)}|<\epsilon]<\infty$ for all $k=1,\dots,K$. Let $\{Y'_i, i=1,2,\dots,n\}$ be an independent copy of $\{Y_i, i=1,2,\dots,n\}$. For $Z\sim N_q(0_q,1_{q\times q}), h\in\mathcal{H}$, where $\mathcal{H}$ is as in (1), it holds that*

$$\begin{aligned}&\left|\mathbb{E}[h(\sqrt{n}J_n^{-\frac{1}{2}}(\hat{\theta}^K-\theta_0^K))]-\mathbb{E}[h(Z)]\right| \\ &\quad\le \frac{\|h\|_1}{\sqrt{n}}K_1(\theta_0)+\frac{q^2c^2\|h\|_2}{4}K_2(\theta_0)+\frac{q^3c^3\|h\|_3}{12}K_3(\theta_0)+\frac{2\|h\|}{\epsilon^2}\mathbb{E}\left[\sum_{j=1}^{q}Q_j^2\right], \qquad (4)\end{aligned}$$

where

$$\begin{aligned}K_1(\theta_0)=&\sum_{k_1=1}^{K}\sum_{k_2=k_1}^{\min\{k_1+1,K\}}\sum_{l=1}^{d}\sum_{j=1}^{d}|\bar{I}_n^{-\frac{1}{2}}(G_{k_2};\theta_0)_{lj}| \\ &\times\left\{\left(\sum_{i=1}^{d}\sqrt{\mathbb{E}[[\hat{\theta}_{[k_1]}-\theta_0]_i]^2\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_l}\ell_{k_1}(\theta_0,Y)+n\bar{I}_n(k_1,\theta_0)_{li}\right)^2}\right)\right.\end{aligned}$$

$$+\frac{1}{2}\sum_{i=1}^{d}\sum_{u=1}^{d}\left(\mathbb{E}\left\{\left([\hat{\theta}_{[k_1]}-\theta_0]_i\right)^2\left([\hat{\theta}_{[k_1]}-\theta_0]_u\right)^2\right\}\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left(M_{iul}^{k_1}(Y)\right)^2 \mid |Q_{(m)}|<\epsilon\right]\right)^{\frac{1}{2}}\Bigg\},$$

$$K_2(\theta_0)=\sum_{k=1}^{K}\left\{\sum_{j=1}^{d}\left[\sum_{i\in G_k}\mathrm{Var}[\xi_{ij}^2]\right]^{\frac{1}{2}}+2\sum_{i<j}\left[\sum_{v\in G_k}\mathrm{Var}[\xi_{vi}\xi_{vj}]\right]^{\frac{1}{2}}\right\},$$

$$K_3(\theta_0)=\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=1}^{d}\left|(\xi'_{ij}-\xi_{ij})\right|\right]^3,$$

$$c=\max_{k\in\{1,\dots,K\}}\left|\bar{I}_n^{-1/2}(G_k;\theta_0)\right|.$$

The theorem is proved in the Supplementary Material (Aronowitz and Bartroff (2025)).

Although the bound in the theorem appears complex, the fact that it is in general of optimal order $\mathcal{O}(n^{-1/2})$ is not hard to see. Assume that $[\bar{I}_n(k,\theta_0)]_j=\mathcal{O}(1)$. By Theorem 2.2,

$$\sqrt{n}\mathbb{E}[\hat{\theta}_k-\theta_0]\to 0_d.$$

Thus, $\mathbb{E}[\hat{\theta}_k-\theta_0]_j=o(1/\sqrt{n})$. Again from Theorem 2.2,

$$n\bar{I}_n^{\frac{1}{2}}(k,\theta_0)\,\mathrm{Cov}[\hat{\theta}_k]\bar{I}_n^{\frac{1}{2}}(k,\theta_0)\to 1_{d\times d}.$$

Using $[\bar{I}_n(k,\theta_0)]_j=\mathcal{O}(1)$, we see that $n\,\mathrm{Var}[[\hat{\theta}_k]_j]\to 1$ and thus $\mathrm{Var}[[\hat{\theta}_k]_j]=\mathcal{O}(1/n)$. It follows that

$$\mathbb{E}([\hat{\theta}_k-\theta_0]_j^2)=\mathrm{Var}\big[[\hat{\theta}_k]_j\big]+\big(\mathbb{E}[\hat{\theta}_k-\theta_0]_j\big)^2=\mathcal{O}\big(n^{-1}\big). \tag{5}$$

From (R2) and (R6) and independence of $Y_1,\dots,Y_{n_k}$ it can be seen that,

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_l}\ell_k(\theta_0,Y)+n\bar{I}_n(k,\theta_0)_{li}\right)^2=\sum_{j=1}^{n_k}\mathrm{Var}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_l}\log\big(f_j(Y_j|\theta_0)\big)\right] \tag{6}$$

showing that the left-hand side of the above equation is $\mathcal{O}(n)$. It follows from (5), (6), and $[\bar{I}_n(k,\theta_0)]_j=\mathcal{O}(1)$ that $K_1(\theta_0)=\mathcal{O}(1)$ and $K_3(\theta_0)=\mathcal{O}(n^{-\frac{1}{2}})$. That $K_2(\theta_0)=\mathcal{O}(n^{-\frac{1}{2}})$ depends on the fourth moment of ξ_{ij} being $\mathcal{O}(\frac{1}{n^2})$. In general, we have reason to believe this is true but have not proved it. Below we show for exponential families that K_2 is indeed $\mathcal{O}(n^{-1/2})$. When all this holds, the right-hand side of (4) is indeed $\mathcal{O}(n^{-1/2})$. For extensive details of the order of the bound one may refer to the analysis in Section 4.2 for the simpler case of exponential families.

3.2 Extensions of Theorem 3.1

Reinert and Röllin (2009, Theorem 2.1), which is the key tool in proving our Theorem 3.1, relies on (3) to bound the derivatives of the multivariate Stein solution. Gaunt (2016, Proposition 2.1) found new bounds on the derivatives of the Stein solution that require one fewer derivative of the function h in (3). In Theorem 3.2, we use this result to relax the conditions of Reinert and Röllin (2009, Theorem 2.1) by requiring that h be two times differentiable instead of three. The price paid is an increase in the order of the bound with respect to d by a factor of $d^{1/2}$.

Theorem 3.2. *Assume that (W,W') is an exchangeable pair of $\mathbb{R}^d$-valued random vectors such that*

$$\mathbb{E}W=0,\mathbb{E}WW^{\intercal}=\Sigma$$

with $\Sigma \in \mathbb{R}^{d\times d}$ symmetric and positive definite. Suppose further that

$$\mathbb{E}[W' - W|W] = -\Lambda W + R$$

is satisfied for an invertible matrix Λ and a $\sigma(W)$-measurable random vector R. Then, if Z has d-dimensional standard normal distribution, we have for every two times differentiable function h,

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \le d^{1/2}|\Sigma^{-\frac{1}{2}}|\Big(\frac{\|h\|_1}{\sqrt{\pi}}A + \frac{\|h\|_2\sqrt{2\pi}}{8}B$$
$$+ \Big(\sqrt{\frac{\pi}{2}}\|h - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\| + \frac{2d}{\sqrt{\pi}}|\Sigma|^{1/2}\|h\|\Big)C\Big),$$

where, with $\lambda^{(i)} := \sum_{m=1}^{d} |(\Lambda^{-1})_{m,i}|$,

$$A = \sum_{i,j=1}^{d} \lambda^{(i)}\sqrt{\mathrm{Var}[\mathbb{E}\{(W_i' - W_i)(W_j' - W_j)|W\}]},$$

$$B = \sum_{i,j,k=1}^{d} \lambda^{(i)}\mathbb{E}|(W_i' - W_i)(W_j' - W_j)(W_k' - W_k)|,$$

$$C = \sum_{i=1}^{d} \lambda^{(i)}\sqrt{\mathrm{Var}[R_i]}.$$

Proof. The proof is similar to the proof of Reinert and Röllin (2009, Theorem 2.1), with Gaunt (2016, Proposition 2.1) used in place of (3). The details are thus omitted. □

Recently, Gaunt and Li (2023, Theorem 3.6) have an improved result which permits a version of this bound that depends on $|h|_1$ and $|h|_2$ but not on $|h|$.

Next, we apply Theorem 3.2 to enhance our main result, Theorem 3.1.

Theorem 3.3. *Let $Y_1, Y_2, \ldots, Y_n$ be independent non-identically distributed $\mathbb{R}^t$-valued random vectors with probability density (or mass) functions $f_i(y_i|\theta)$, for which the parameter space Θ is an open subset of $\mathbb{R}^d$. Suppose the observations $Y_1, Y_2, \ldots, Y_{n_k}$ are available at analysis $k = 1, \ldots, K$. Assume that the MLE $\hat{\theta}_k$ exists and is unique and that conditions (R1)–(R6) are satisfied at each analysis k. In addition, assume that for any θ_0 there exists $0 < \epsilon = \epsilon(\theta_0)$ and functions $M^k_{iuj}(y)$, $\forall i, u, j = 1, 2, \ldots, d$ such that $|\frac{\partial^3}{\partial\theta_i\partial\theta_u\partial\theta_j}\ell_k(\theta, y)| \le M^k_{iuj}(y)$ for all $\theta \in \Theta$ with $|\theta_j - \theta_{0,j}| < \epsilon$ $\forall j = 1, 2, \ldots, d$. Also, assume that $\mathbb{E}[(\frac{\partial^3}{\partial\theta_i\partial\theta_u\partial\theta_j}M^k_{iuj}(Y))^2||Q_{(m)}| < \epsilon] < \infty$ for all $k = 1, \ldots, K$. Let $\{Y_i', i = 1, 2, \ldots, n\}$ be an independent copy of $\{Y_i, i = 1, 2, \ldots, n\}$. For $Z \sim N_q(0_q, 1_{q\times q})$, $h \in \mathcal{H}$, where $\mathcal{H}$ is the class of all bounded functions with bounded first and second order derivatives, it holds that*

$$\left|\mathbb{E}[h(\sqrt{n}J_n^{-\frac{1}{2}}(\hat{\theta}^K - \theta_0^K))] - \mathbb{E}[h(Z)]\right|$$
$$\le \frac{\|h\|_1}{\sqrt{n}}K_1(\theta_0) + \frac{q^{3/2}c^2\|h\|_1}{\sqrt{\pi}}K_2(\theta_0) + \frac{\sqrt{2\pi}q^{5/2}c^3\|h\|_2}{8}K_3(\theta_0) + \frac{2\|h\|}{\epsilon^2}\mathbb{E}\left[\sum_{j=1}^{q} Q_j^2\right],$$

where

$$K_1(\theta_0) = \sum_{k_1=1}^{K} \sum_{k_2=k_1}^{\min\{k_1+1,K\}} \sum_{l=1}^{d} \sum_{j=1}^{d} |\bar{I}_n^{-\frac{1}{2}}(G_{k_2};\theta_0)_{lj}|$$
$$\times \left\{ \left(\sum_{i=1}^{d} \sqrt{\mathbb{E}[[\hat{\theta}_{[k_1]} - \theta_0]_i]^2 \mathbb{E}\left(\frac{\partial^2}{\partial\theta_i \partial\theta_l} \ell_{k_1}(\theta_0, Y) + n\bar{I}_n(k_1, \theta_0)_{li} \right)^2} \right) \right.$$
$$\left. + \frac{1}{2} \sum_{i=1}^{d} \sum_{u=1}^{d} (\mathbb{E}\{([\hat{\theta}_{[k_1]} - \theta_0]_i)^2 ([\hat{\theta}_{[k_1]} - \theta_0]_u)^2\})^{\frac{1}{2}} (\mathbb{E}[(M_{iul}^{k_1}(Y))^2 || Q_{(m)}| < \epsilon])^{\frac{1}{2}} \right\},$$
$$K_2(\theta_0) = \sum_{k=1}^{K} \left\{ \sum_{j=1}^{d} \left[\sum_{i \in G_k} \operatorname{Var}[\xi_{ij}^2] \right]^{\frac{1}{2}} + 2 \sum_{i<j} \left[\sum_{v \in G_k} \operatorname{Var}[\xi_{vi}\xi_{vj}] \right]^{\frac{1}{2}} \right\},$$
$$K_3(\theta_0) = \sum_{i=1}^{n} \mathbb{E}\left[\sum_{j=1}^{d} |(\xi'_{ij} - \xi_{ij})| \right]^3,$$
$$c = \max_{k \in \{1,\ldots,K\}} |\bar{I}_n^{-1/2}(G_k;\theta_0)|.$$

Proof. The proof of Theorem 3.1 (see the Supplementary Material (Aronowitz and Bartroff (2025))) only needs to be augmented by using Theorem 3.2 in place of Reinert and Röllin (2009, Theorem 2.1) and noting that by

$$\tilde{J}_n^{-1/2} A = \Sigma^{-1/2} = \operatorname{diag}(\bar{I}_n^{-1/2}(G_1;\theta_0), \ldots, \bar{I}_n^{-1/2}(G_K;\theta_0)),$$

we have $|\Sigma^{-1/2}| = c$. With this adjustment, the rest of the proof is similar and the details are thus omitted. □

Fang, Shao and Xu (2019) use Malliavin calculus techniques along with an exchangeable pair approach to attain a near-optimal error bound on the Wasserstein distance for multivariate approximations. Using this, a further improvement of Theorem 3.3 is possible which removes the requirement that h have bounded second order derivatives at the expense of sacrificing optimal order convergence. Bonis (2020) obtained the optimal $n^{-1/2}$ rate of convergence in the Wasserstein distance in the multivariate central limit theorem for properly normalized sums of i.i.d. random vectors. Although this does not apply directly to our the smooth test functions used here, it could provide a different avenue to further improvements of Theorem 3.3.

4 Application to observations from an exponential family

In this section, we specialize Theorem 3.1 to the case where observations are i.i.d. from an exponential family. We then show that this bound is of optimal order $\mathcal{O}(n^{-1/2})$ and calculate the bound explicitly in the case of lifetime data from an exponential distribution.

4.1 Notation and setup

We slightly modify some of our notation to be more in line with the exponential family literature. We say that the distribution of Y is a canonical multi-parameter exponential family distribution if for $\eta \in \mathbb{R}^d$ the density of Y is

$$f(y;\eta) = \exp\left\{ \sum_{j=1}^{d} \eta_j T_j(y) - A(\eta) + S(y) \right\} \mathbb{1}_{\{y \in B\}}, \tag{7}$$

where B is the support of y that does not depend on η, and $T(y) = [T_1(y), \ldots, T_d(y)]^\intercal$ is the natural sufficient statistic. Here, the natural parameter η plays the role of the parameter of interest θ above. The cumulant function $A(\eta)$ satisfies

$$\frac{\partial}{\partial \eta_i} A(\eta) = \mathbb{E}_\eta\big[T_i(Y)\big], \mathrm{Var}_\eta\big[T(Y)\big] = H\big(A(\eta)\big) = \nabla_\eta^2 A(\eta). \tag{8}$$

To ease the notational burden, we use the following simplified notation for the mixed moments of the sufficient statistics:

$$\begin{aligned} \mu_i &= \mu_i(\eta) = \mathbb{E}_\eta\big[T_i(Y)\big], \\ \mu_{ij} &= \mu_{ij}(\eta) = \mathbb{E}_\eta\big[T_i(Y)T_j(Y)\big], \\ \mu_{ijk} &= \mu_{ijk}(\eta) = \mathbb{E}_\eta\big[T_i(Y)T_j(Y)T_k(Y)\big]. \end{aligned}$$

In Section 4.2, we will also make use of the third partial derivative of $A(\eta)$ which appear in Theorem 3.1. We have

$$\begin{aligned} &\frac{\partial^3}{\partial\eta_i\partial\eta_j\partial\eta_k} A(\eta) \\ &= \int \exp\{S(y)\} T_i(y) \bigg[\bigg(T_j(y) - \frac{\partial}{\partial \eta_j} A(\eta)\bigg)\bigg(T_k(y) - \frac{\partial}{\partial \eta_k} A(\eta)\bigg) a(y,\eta) \\ &\quad - a(y,\eta) \frac{\partial^2}{\partial\eta_j\partial\eta_k} A(\eta)\bigg] dy \\ &= \mu_{ijk} - \mu_{ij}\mu_k - \mu_{ik}\mu_j + \mu_i\mu_j\mu_k - \mathrm{Cov}\big[T_j(Y), T_k(Y)\big]\mu_i \\ &= \mu_{ijk} - \mu_{ij}\mu_k - \mu_{ik}\mu_j - \mu_{jk}\mu_i + 2\mu_i\mu_j\mu_k. \end{aligned} \tag{9}$$

Define

$$\begin{aligned} \mu_{iul}^\epsilon &= \mu_{iul}^\epsilon(\eta_0) \\ &= \max_{\{\eta : |\eta_j - \eta_{0,j}| < \epsilon, \forall j \in \{1,\ldots,d\}\}} |\mu_{ilk} - \mu_{il}\mu_k - \mu_{ik}\mu_l - \mu_{lk}\mu_i + 2\mu_i\mu_l\mu_k| \end{aligned} \tag{10}$$

The joint cumulant generating function of $W = (T_1(Y), \ldots, T_d(Y))^\intercal$ is then

$$K(s) = \log \mathbb{E}\big[e^{s^\intercal W}\big] = A(s+\eta) - A(\eta).$$

Via this expression, the cumulants can be found by taking derivatives of $A(\eta)$, as follows:

$$\nabla_{s^n} K(s)|_{s=0} = \nabla_{s^n} A(s+\eta)|_{s=0} = \nabla_{\eta^n} A(\eta).$$

Thus partial derivatives of $A(\eta)$ yield joint cummulants of $[T_1(Y), \ldots, T_d(Y)]$ which in turn are polynomial functions of the moments of $[T_1(Y), \ldots, T_d(Y)]$.

We now turn to MLEs of exponential families, which we denote by $\hat{\eta}$. The score function is

$$\nabla_\eta \ell(y^n; \eta) = \nabla_\eta \sum_{i=1}^n \bigg(\sum_{j=1}^d \eta_j T_j(y_i) - A(\eta) + S(y_i)\bigg) = \sum_{i=1}^n T(y_i) - n\nabla_\eta A(\eta)$$

and setting this equal to zero and using (8) yields

$$\frac{1}{n}\sum_{i=1}^n T(y_i) = \nabla_\eta A(\eta)|_{\eta=\hat{\eta}} = \mathbb{E}_{\hat{\eta}}\big[T(Y)\big].$$

Since the log-likelihood of an exponential family is strictly concave, if the MLE exists it is a global maximum. The mean function is

$$\tau(\eta) = [\tau_1(\eta), \ldots, \tau_d(\eta)]^\intercal = [\mathbb{E}_\eta[T_1(Y)], \ldots, \mathbb{E}_\eta[T_d(Y)]]^\intercal,$$

which we write compactly as

$$\tau(\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} T(y_i) \quad \text{or, equivalently,} \quad \hat{\eta} = \tau^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} T(y_i)\right). \tag{11}$$

We note that (11) matches the form (2) and thus the bounds found in Anastasiou and Ley (2017) and Anastasiou and Gaunt (2020) can be used on the distance to normality for MLEs of exponential families. This fact was noted in Anastasiou and Ley (2017) for single parameter exponential families but the generalization to multi-parameter exponential families was not explicitly addressed in Anastasiou and Gaunt (2020).

4.2 Bound to normality for MLEs of exponential families

With the notation in place we are ready to state the specialization of Theorem 3.1 to exponential families in Corollary 4.1. The bound (12) that results can be written completely in terms of mixed moments up to the third order of the sufficient statistics $T_1(Y), \ldots, T_d(Y)$, the true value of the natural parameter η_0, and the sample. Additional knowledge of the distribution, in particular the cumulant function $A(\eta)$ and $S(Y)$, is not required. Although the bound in Corollary 4.1 inherits the optimal order of $n^{-1/2}$ from Theorem 3.1, a direct analysis of the bound's order is more straightforward due to the observations being i.i.d., and we carry this out following the corollary in Section 4.3.

Corollary 4.1. *Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. with probability density (or mass) function $f(y|\eta)$, for which the parameter space is an open subset of $\mathbb{R}^d$. Suppose f is an exponential family defined by (7) and the observations $Y_1, Y_2, \ldots, Y_{n_k}$ are available at analysis $k = 1, \ldots, K$. Assume that the MLE $\hat{\eta}_k$ exists and that conditions (R1)–(R6) are satisfied at each analysis k. Let $\{Y_i', i = 1, 2, \ldots, n\}$ be an independent copy of $\{Y_i, i = 1, 2, \ldots, n\}$. For $Z \sim N_q(0_q, 1_{q\times q})$, $h \in \mathcal{H}$, where $\mathcal{H}$ is as in (1) and μ^{ϵ}_{iul} defined by (10), it holds that*

$$\begin{aligned}&\left|\mathbb{E}[h(\sqrt{n}J_n^{-\frac{1}{2}}(\hat{\eta}^K - \eta_0^K))] - \mathbb{E}[h(Z)]\right| \\ &\quad\leq \frac{\|h\|_1}{\sqrt{n}}K_1(\eta_0) + \frac{q^2c^2\|h\|_2}{4}K_2(\eta_0) + \frac{q^3c^3\|h\|_3}{12}K_3(\eta_0) + \frac{2\|h\|}{\epsilon^2}\mathbb{E}\left[\sum_{j=1}^{q} Q_j^2\right],\end{aligned} \tag{12}$$

where,

$$\begin{aligned}K_1(\eta_0) = {}& \frac{1}{2}\sum_{k_1=1}^{K}\sum_{k_2=k_1}^{\min\{k_1+1,K\}} n_{k_1}\left(\frac{|G_{k_2}|}{n}\right)^{-\frac{1}{2}}\sum_{l=1}^{d}\sum_{j=1}^{d}\mathrm{Var}^{-\frac{1}{2}}[T(Y)]_{lj} \\ &\times\left\{\sum_{i=1}^{d}\sum_{u=1}^{d}\mu^{\epsilon}_{iul}\left(\mathbb{E}\left[\left(\tau^{-1}\left(\frac{1}{n_{k_1}}\sum_{s=1}^{n_{k_1}}T(y_s)\right)_i - \eta_{0,i}\right)^2\right.\right.\right. \\ &\left.\left.\left.\times\left(\tau^{-1}\left(\frac{1}{n_{k_1}}\sum_{s=1}^{n_{k_1}}T(y_s)\right)_u - \eta_{0,u}\right)^2\right]\right)^{\frac{1}{2}}\right\},\end{aligned}$$

$$K_2(\eta) = \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\left(\frac{|G_k|}{n}\right)^{\frac{1}{2}}\left\{\sum_{j=1}^{d}[\mathrm{Var}[(T_j(Y) - \mu_j(\eta_0))^2]]^{\frac{1}{2}}\right.$$

$$+ 2\sum_{i<j}\big[\mathrm{Var}\big[(T_i(Y)-\mu_i(\eta_0))(T_j(Y)-\mu_j(\eta_0))\big]\big]^{\frac{1}{2}}\Bigg\},$$

$$K_3(\eta_0) = \frac{1}{\sqrt{n}}\mathbb{E}\Bigg[\sum_{j=1}^{d}\big|T_j(Y')-T_j(Y)\big|\Bigg]^3,$$

$$c = \big|\mathrm{Var}^{-\frac{1}{2}}[T(Y)]\big| \max_{k\in\{1,\dots,K\}}\left(\frac{|G_k|}{n}\right)^{-\frac{1}{2}}.$$

4.3 The bound in Corollary 4.1 is of order $\mathcal{O}(n^{-1/2})$

In Section 3, we analyzed the order of the bound in Theorem 3.1 and found it is of order $n^{-1/2}$ given some mild assumptions. It is still $\mathcal{O}(n^{-1/2})$ when specialized to the exponential family case is no surprise, and can be seen more directly here. Letting $|G_k|/n \to \gamma_k \in (0,1)$ as in Theorem 2.2, the term c in the bound satisfies

$$c \to \big|\mathrm{Var}^{-\frac{1}{2}}[T(Y)]\big| \max_{k\in\{1,\dots,K\}} \gamma_k^{-\frac{1}{2}}.$$

and thus $c = \mathcal{O}(1)$. For the same reason, $K_2(\eta) = \mathcal{O}(n^{-1/2})$. The function $K_3(\eta)$ is exactly equal to a constant times $n^{-1/2}$ and so is $\mathcal{O}(n^{-1/2})$. The last term in the bound,

$$\frac{2\|h\|}{\epsilon^2}\mathbb{E}\Bigg[\sum_{j=1}^{q} Q_j^2\Bigg],$$

depends on the order of

$$\mathbb{E}[Q_j^2] = \mathbb{E}[(\hat{\eta}_j - \eta_{0,j})^2] = \mathrm{Var}[\hat{\eta}_j] + \big(\mathbb{E}[Q_j]\big)^2 = \mathcal{O}(n^{-1}).$$

This leaves only $K_1(\eta_0)$ to be considered. For the entire bound to be $\mathcal{O}(n^{-1/2})$, $K_1(\eta)$ should be $\mathcal{O}(1)$ since $K_1(\eta)$ is multiplied by $n^{-1/2}$. This is the case if and only if

$$\left(\mathbb{E}\left[\left(\tau^{-1}\left(\frac{1}{n_{k_1}}\sum_{s=1}^{n_{k_1}} T(y_s)\right)_i - \eta_{0,i}\right)^2\left(\tau^{-1}\left(\frac{1}{n_{k_1}}\sum_{s=1}^{n_{k_1}} T(y_s)\right)_u - \eta_{0,u}\right)^2\right]\right)^{\frac{1}{2}}$$

is $\mathcal{O}(n^{-1})$. By the Cauchy–Schwarz inequality, the above is bounded by

$$\big(\mathbb{E}[(\hat{\eta}_i - \eta_{0,i})^4]\mathbb{E}[(\hat{\eta}_u - \eta_{0,u})^4]\big)^{\frac{1}{4}}.$$

Thus the desired order follows from the fact that the fourth central moment of a component of the MLE is $\mathcal{O}(n^{-2})$; see de A. Cysneiros et al. (2001, Equation 2) for the univariate exponential case and Peers and Iqbal (1985) for the general multivariate case.

The choice of ϵ in the bound should be handled on a case-by-case basis and can be chosen to minimize the sum of the K_1 term and the final $\frac{2\|h\|}{\epsilon^2}$ term. Notice that smaller ϵ increases the final term while decreases μ_{iul}^{ϵ} in K_2.

4.4 Example: The exponential distribution

In this section, we apply Corollary 4.1 to the case where observations are i.i.d. from an exponential distribution. We use the natural parameterization

$$T(y) = -y, \qquad A(\eta) = -\log(\eta), \qquad S(y) = 0, \qquad B = [0,\infty).$$

We record the application of Corollary 4.1 to this distribution in the following corollary, in which we utilize that $\hat{\eta}_k \sim \text{Inv-Gamma}(n_k, n_k\lambda)$ and $\mu_{iul}^{\epsilon} = 2/(\eta_0 - \epsilon)^3$. The remaining calculations are similar to, and simpler than, those for Corollary 4.1 and so we omit them.

Corollary 4.2. *Let $Y_1, Y_2, \dots, Y_n$ be independent such that $Y_i \sim \mathrm{Exp}(\eta_0)$ with $\mathbb{E}[Y_i] = 1/\eta_0$. Suppose the observations $Y_1, Y_2, \dots, Y_{n_k}$ are available at analysis $k = 1, \dots, K$. For $Z \sim N_q(0_q, 1_{q\times q})$, $h \in \mathcal{H}$, where $\mathcal{H}$ is as in (1), it holds that*

$$\begin{aligned}&\left|\mathbb{E}[h(\sqrt{n}J_n^{-\frac{1}{2}}(\hat{\eta}^K - \eta_0^K))] - \mathbb{E}[h(Z)]\right| \\ &\quad \le \frac{\|h\|_1}{\sqrt{n}} K_1(\eta_0) + \frac{2K^2c^2\|h\|_2}{\eta_0^2\sqrt{n}} K_2 + \frac{2\eta_0^2\|h\|}{\epsilon^2}\sum_{k=1}^{K}\frac{n_k+2}{(n_k-1)(n_k-2)},\end{aligned} \tag{13}$$

where

$$K_1(\eta_0) = \frac{\eta_0^3\sqrt{3}}{(\eta_0-\epsilon)^3}\sum_{k_1=1}^{K}\sum_{k_2=k_1}^{\min\{k_1+1,K\}}\left(\frac{|G_{k_2}|}{n}\right)^{-\frac{1}{2}}\sqrt{\frac{n_{k_1}^4 + (\frac{46}{3})n_{k_1}^3 + 8n_{k_1}^2}{(n_{k_1}-1)\cdots(n_{k_1}-4)}},$$

$$K_2 = \sum_{k=1}^{K}\left(\frac{|G_k|}{n}\right)^{\frac{1}{2}} \quad \text{and} \quad c = \max_{k\in\{1,\dots,K\}}\left(\frac{|G_k|}{n}\right)^{-\frac{1}{2}}.$$

5 Data and simulation example

In this section, we exhibit a data set of time intervals between cancer cell detections that fits the set up of this paper, and compute our bound for the exponential distribution in Corollary 4.2 in the setting of the data set. We continue to use the notation from Section 4.4 for the exponential distribution.

The exponential distribution has been used (Coumans et al., 2012; Zhu et al., 2017, 2021) to model the time intervals between detections of cancer cells circulating in blood, known as circulating tumor cells (CTCs), which are established biomarkers of cancer metastasis, and therefore their temporal dynamics are important for understanding the mechanisms underlying tumor cell dissemination (Zhu et al., 2021). Williams et al. (2020) analyze Lewis lung carcinoma CTC counts and time intervals drawn from tumor bearing mice at various times up to 34 days after inoculation.[1] Williams et al. (2020, Figure 1H) consider measurements at 5, 13, 18, 25, and 28 days post-inoculation which, along with the final analysis at 34 days, we take as the $K = 6$ time points in a group sequential analysis which, in their data, has cumulative sample sizes

$$\boldsymbol{n}_{\mathrm{CTC}} := (n_1, \dots, n_6) = (19, 119, 261, 365, 483, 589). \tag{14}$$

Figure 1 shows histograms (on the density scale) of the time interval data $Y_1, \dots, Y_{n_k}$ at analyses $k = 1, \dots, 6$, along with the estimated exponential rates $\widehat{\eta}_k$ and the fitted exponential density for each $\widehat{\eta}_k$. The rate estimates are somewhat stable, but tend to increase slightly from $\widehat{\eta}_1 = 0.00217$ (to 3 significant figures) after the first analysis on day 5, to the final estimate $\widehat{\eta}_6 = 0.00309$ at day 34 on the complete data.

The CTC interval data in Figure 1 is a single realization of the data $Y_1, \dots, Y_{n_K}$ and vector of estimates $\widehat{\eta}^K$ described in Section 4.4. In order to investigate the behavior of the bound in Corollary 4.2, we simulated data in the setting of the CTC interval data in order to estimate the left hand side of (13), and computed the corresponding right hand side, as follows. Taking $\eta_0 = 0.00309$, the estimated rate from the complete CTC interval data, and the same number $K = 6$ of analyses, we simulated exponential data with sample sizes $M \cdot \boldsymbol{n}_{\mathrm{CTC}}$

[1] More details on the data acquisition can be found in Williams et al. (2020), and the data is available from https://github.com/mark-niedre/Williams-CTC-Dynamics-2020.

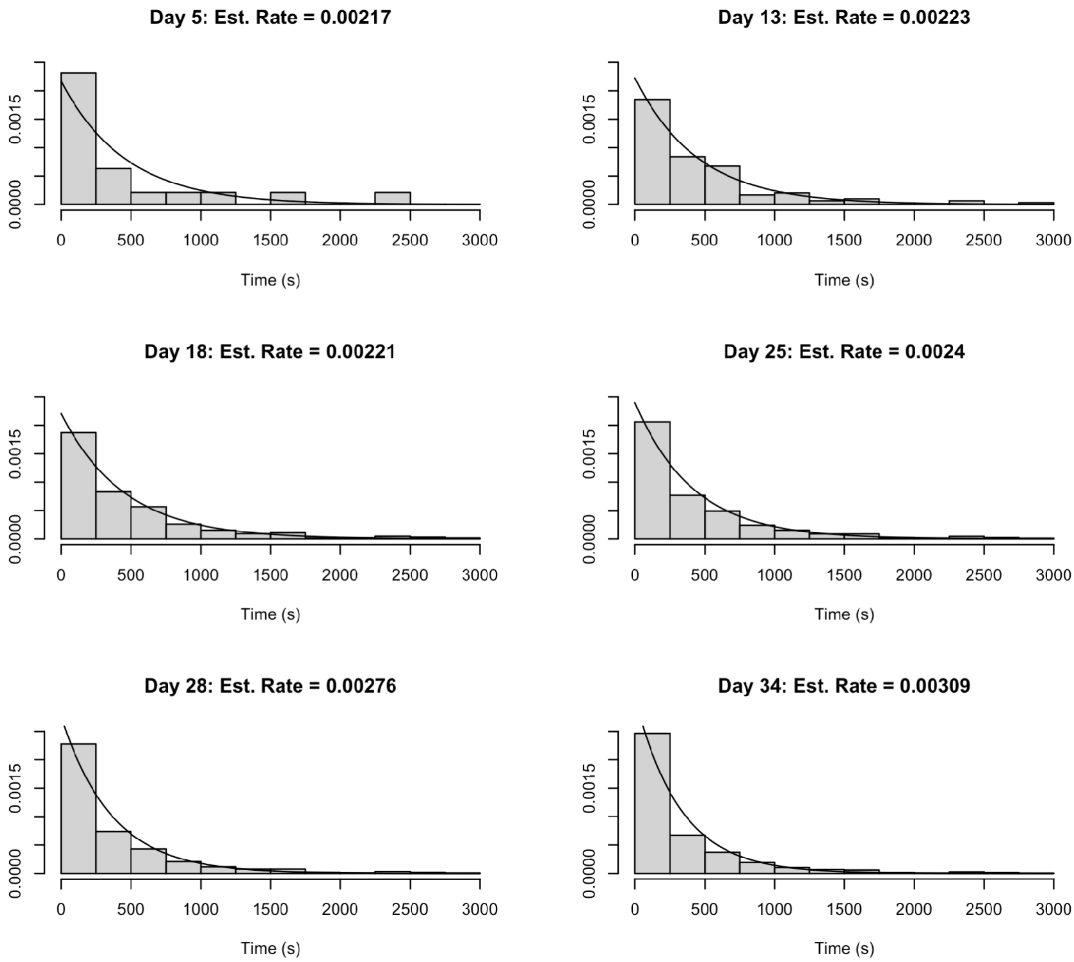

Figure 1 *Histograms of CTC interval data from Williams et al. (2020) at $K = 6$ analyses, with the estimated exponential rates $\widehat{\eta}_k$ and the corresponding fitted exponential densities.*

proportional to the actual sample sizes (14) from the CTC data. For the values of the multiplier M in Table 1, the distributional distance (denoted by $\widehat{D}$ in the table) on the left hand side of (13) was estimated by simulating 5000 exponential data sets and standard normal vectors $Z \in \mathbb{R}^6$. The standardizing matrix $\sqrt{n} J_n^{-1/2}$ in (13) was calculated as $V^{-1/2}$ where V is the data-dependent $K \times K$ covariance matrix with (k, ℓ) entry equal to $\widehat{\eta}^2_{k\vee\ell}/n_{k\vee\ell}$. The bound on the right hand side of (13), denoted by B in the table, was calculated by numerically optimizing over the free parameter ϵ. This was done using the Gaussian test function $h(z_1, \ldots, z_K) = (2\pi)^{-K/2} \exp(-\sum_{k=1}^K z_k^2/2)$, a common choice in studies of distributional distances (e.g., Barron, 1993; Cucker and Zhou, 2007; Smola and Schölkopf, 2004), for which basic calculations give $|h| = (2\pi)^{-K/2}$, $|h|_1 = (2\pi)^{-K/2} e^{-1/2}$, and $|h|_2 = (2\pi)^{-K/2}$. The table also includes the relative error $(B - \widehat{D})/\widehat{D} = B/\widehat{D} - 1$ of B as an estimator of $\widehat{D}$, to 3 significant digits.

In each successive row of the table, M and hence the sample size increases by a factor of 10, and the bound B consistently decreases by a factor close to $1/\sqrt{10} \approx 0.32$, reflecting the fact that the bound is asymptotically $1/\sqrt{n}$, as discussed in Section 4.3. The estimated distance $\widehat{D}$ remains relatively stable for the sample sizes in the table. Although the values of the multiplier M required to make the relative error small are sizeable, they are comparable those found for the state-of-the-art smooth function distance bounds for MLEs and related

Table 1 *Simulation study of the bound B on the right-hand side of (13), the estimated distributional difference $\widehat{D}$ on the left-hand side of (13), and their relative error for 5000 exponential data sets per row simlulated with rate $\eta_0 = 0.00309$ and sample sizes $M \cdot \boldsymbol{n}_{\mathrm{CTC}}$ with $\boldsymbol{n}_{\mathrm{CTC}}$ given by (14) from the CTC interval data, multiplier M, and Gaussian test function h*

Multiplier M	Bound B	Distance $\widehat{D}$	Rel. Error $B/\widehat{D} - 1$
1	0.906243	0.00048	1880
10^1	0.283007	0.00046	612
10^2	0.089108	0.00050	178
10^3	0.028112	0.00049	56.4
10^4	0.008877	0.00051	16.5
10^5	0.002805	0.00050	4.58
10^6	0.000886	0.00049	0.823

estimators in non-group-sequential settings by Anastasiou (2018) and Anastasiou and Gaunt (2020), which do not have the additional variation due to mulitple group sequential analyses considered here.

6 A multivariate Kolmogorov/smooth function bound

In this section, we give a general upper bound on the multivariate Kolmogorov distance

$$d_{kol}(W, Z) = \sup_{\boldsymbol{x} \in \mathbb{R}^p} \left| P(W \leq \boldsymbol{x}) - P(Z \leq \boldsymbol{x}) \right| \tag{15}$$

of two p-dimensional random vectors W, Z, in terms of the smooth function class distances, such as those in our results in Sections 3 and 4. In (15), $P(W \leq \boldsymbol{x})$ is the probability of W being in the p-dimensional "lower quadrant" $\{W \leq \boldsymbol{x}\} = \{W_1 \leq x_1, \ldots, W_p \leq x_p\}$. These bounds may be of independent interest, but another reason we include them here is because they open the door to applying our smooth function bounds for the group sequential MLEs to the normal, to Kolmogorov distance, which may be desirable in statistical applications.

First, we modify our notation slightly to show more explicitly the dependence on dimension and derivative order. For $\boldsymbol{k} = (k_1, \ldots, k_p) \in \mathbb{N}_0^p$, let $|\boldsymbol{k}| = \sum_{i=1}^p k_i$, and for functions $h : \mathbb{R}^p \to \mathbb{R}$ whose partial derivatives

$$h^{\boldsymbol{k}}(\boldsymbol{x}) = \frac{\partial^{k_1 + \cdots + k_p} h}{\partial^{k_1} x_1 \ldots \partial^{k_p} x_p} \quad \text{exist for all } 0 \leq |\boldsymbol{k}| \leq m,$$

and $\|\cdot\|$ the supremum norm, let $L_m^\infty(\mathbb{R}^p)$ be the collection of all functions $h : \mathbb{R}^p \to \mathbb{R}$ with

$$\|h\|_{L_m^\infty(\mathbb{R}^p)} = \max_{0 \leq |\boldsymbol{k}| \leq m} \|h^{(\boldsymbol{k})}\|$$

finite. For random vectors X and Y in $\mathbb{R}^p$, letting

$$\mathcal{H}_{m,\infty,p} = \left\{ h \in L_m^\infty(\mathbb{R}^p) : \|h\|_{L_m^\infty(\mathbb{R}^p)} \leq 1 \right\},$$

define

$$d_{m,\infty,p}(X, Y) = \|\mathcal{L}(X) - \mathcal{L}(Y)\|_{\mathcal{H}_{m,\infty,p}} = \sup_{h \in \mathcal{H}_{m,\infty,p}} \left| \mathbb{E}[h(X)] - \mathbb{E}[h(Y)] \right|.$$

Connecting to the previous sections, note that, for example, Theorem 3.1 gives a bound on $d_{3,\infty,q}$ between the standardized group sequential MLE and the q-dimensional standard normal, and Theorem 3.3 a bound on $d_{2,\infty,q}$.

We first state our general result in Theorem 6.1, and then specialize to the $m = 3$ case in Corollary 6.2. Both are proved in the Supplementary Material (Aronowitz and Bartroff (2025)).

Theorem 6.1. *Suppose W and Z are p-dimensional, $p \geq 2$, random vectors such that Z has a density bounded by a constant C_1. Then there exists a constant $C_2 \geq 1$ that depends only on m, $m \geq 1$, such that*

$$d_{kol}(W, Z) \leq d_{m,\infty,p}(W, Z)^{\frac{p-1}{m+p-1}} \left(C_2^p + p + C_1 d_{m,\infty,p}(W, Z)^{\frac{1}{m+p-1}}\right).$$

Corollary 6.2. *The Kolmogorov distance is bounded by $d_{3,\infty,p}(\cdot,\cdot)$ as follows,*

$$d_{kol}(W, Z) \leq d_{3,\infty,p}(W, Z)^{\frac{p-1}{2+p}} \left(52.5^p + p + \frac{d_{3,\infty,p}(W, Z)^{\frac{1}{2+p}}}{(2\pi)^{p/2}}\right).$$

Gaunt and Li (2023, Proposition 2.4) also obtained bounds on $d_{\mathrm{kol}}(W, Z)$ in terms of $d_{m,\infty,p}$, but those bounds do not explicitly show the dependence on the dimension p which is important in statistical applications. This is especially true in the group sequential testing considered here, where it represents the maximum number of groups and thus is an important design consideration.

7 Conclusion and future directions

We have generalized the results of Anastasiou and Reinert (2017) and Anastasiou (2018) to find optimal order bounds for the joint distribution of a sequence of maximum likelihood estimates based on accumulating data. The approximate normality of this joint distribution is an essential assumption underlying the statistics of group sequential hypothesis testing, the dominant paradigm in clinical trials. The specialization of this bound to exponential families in Corollary 4.1 is the simplest such bound covering this case even in the non-sequential setting, that we are aware of.

A direction in which these results may be generalized is to examine log-likelihood functions of the form

$$\sum_{i \in G_k} \log f_i(Y_i, \theta) + g_k(\mathcal{Y}_k, \theta),$$

with $\mathcal{Y}_k = \{Y_i : i \in G_k\}$. This form may provide a way of relaxing the independence assumption between samples that could be amenable to Stein's method. It is also the form of some log-likelihood functions of generalized linear mixed models (GLMMs) with random stage effects, for example, the GLMM with Poisson response variable and canonical log link function.

Acknowledgments

The authors thank the Associate Editor, referee, and Prof. Robert Gaunt for helpful comments on the paper. Prof. Mark Niedre and Amber Williams are thanked for assistance with the data in their paper, Williams et al. (2020).

Supplementary Material

Supplement to "Finite-sample bounds to the normal limit under group sequential sampling" (DOI: 10.1214/25-BJPS621SUPP; .pdf). The supplementary material provides background on exchangeable pairs, and proofs of Theorems 3.1 and 6.1 and Corollary 6.2.

References

Anastasiou, A. (2017). Bounds for the normal approximation of the maximum likelihood estimator from m-dependent random variables. *Statistics & Probability Letters* **129**, 171–181. MR3688530 https://doi.org/10.1016/j.spl.2017.04.022

Anastasiou, A. (2018). Assessing the multivariate normal approximation of the maximum likelihood estimator from high-dimensional, heterogeneous data. *Electronic Journal of Statistics* **12**, 3794–3828. MR3881763 https://doi.org/10.1214/18-EJS1492

Anastasiou, A. and Gaunt, R. E. (2020). Multivariate normal approximation of the maximum likelihood estimator via the delta method. *Brazilian Journal of Probability and Statistics* **34**, 136–149. MR4058975 https://doi.org/10.1214/18-BJPS411

Anastasiou, A. and Ley, C. (2017). Bounds for the asymptotic normality of the maximum likelihood estimator using the delta method. *Latin American Journal of Probability and Statistics* **14**, 153–171. MR3622464 https://doi.org/10.30757/alea.v14-09

Anastasiou, A. and Reinert, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli* **23**, 191–218. MR3556771 https://doi.org/10.3150/15-BEJ741

Aronowitz, J. and Bartroff, J. (2025). Supplement to "Finite-sample bounds to the normal limit under group sequential sampling." https://doi.org/10.1214/25-BJPS621SUPP

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930–945. MR1237720 https://doi.org/10.1109/18.256500

Bartroff, J., Lai, T. L. and Shih, M. (2013). *Sequential Experimentation in Clinical Trials: Design and Analysis*. New York: Springer. MR2987767 https://doi.org/10.1007/978-1-4614-6114-2

Bonis, T. (2020). Stein's method for normal approximation in Wasserstein distances with application to the multivariate central limit theorem. *Probability Theory and Related Fields* **178**, 827–860. MR4168389 https://doi.org/10.1007/s00440-020-00989-4

Chen, L. H., Goldstein, L. and Shao, Q.-M. (2010). *Normal Approximation by Stein's Method*. Berlin Heidelberg: Springer. MR2732624 https://doi.org/10.1007/978-3-642-15007-4

Coumans, F. A., Ligthart, S. T., Uhr, J. W. and Terstappen, L. W. (2012). Challenges in the enumeration and phenotyping of CTC. *Clinical Cancer Research* **18**, 5711–5718.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint, Vol. 24*. Cambridge: Cambridge University Press. MR2354721 https://doi.org/10.1017/CBO9780511618796

de A. Cysneiros, F. J., dos Santos, S. J. P. and Cordeiro, G. M. (2001). Skewness and kurtosis for maximum likelihood estimator in one-parameter exponential family models. *Brazilian Journal of Probability and Statistics* **15**, 85–105. MR1907862

Fang, X., Shao, Q.-M. and Xu, L. (2019). Multivariate approximations in Wasserstein distance by Stein's method and Bismut's formula. *Probability Theory and Related Fields* **174**, 945–979. MR3980309 https://doi.org/10.1007/s00440-018-0874-5

Gaunt, R. E. (2016). Rates of convergence in normal approximation under moment conditions via new bounds on solutions of the Stein equation. *Journal of Theoretical Probability* **26**, 231–247. MR3463084 https://doi.org/10.1007/s10959-014-0562-z

Gaunt, R. E. and Li, S. (2023). Bounding Kolmogorov distances through Wasserstein and related integral probability metrics. *Journal of Mathematical Analysis and Applications* **522**, 126985. MR4533915 https://doi.org/10.1016/j.jmaa.2022.126985

Goldstein, L. and Rinott, Y. (1996). Multivariate normal approximations by Stein's method and size bias couplings. *Journal of Applied Probability* **33**, 1–17. MR1371949 https://doi.org/10.1017/s0021900200103675

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics* **42**, 1977–1991. MR0297051 https://doi.org/10.1214/aoms/1177693066

Jennison, C. and Turnbull, B. W. (1997a). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341. MR1615245 https://doi.org/10.2307/2965403

Jennison, C. and Turnbull, B. W. (1997b). Distribution theory of group sequential t, χ^2 and F-tests for general linear models. *Sequential Analysis* **16**, 295–317. MR1491639 https://doi.org/10.1080/07474949708836390

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: Chapman & Hall. MR1710781

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hill. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

O'Brien, P. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Peers, H. W. and Iqbal, M. (1985). Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications. *Journal of the Royal Statistical Society, Series B, Methodological* **47**, 547–554. MR0844486

Pinelis, I. (2017). Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics* **11**, 1160–1179. MR3634332 https://doi.org/10.1214/17-EJS1264

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangable pairs under a general linearity condition. *Annals of Probability* **36**, 2150–2173. MR2573554 https://doi.org/10.1214/09-AOP467

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222. MR2086398 https://doi.org/10.1023/B:STCO.0000035301.49549.88

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 583–602. Berkeley, Calif: University of California Press. MR0402873

Ulyanov, V. (1979). On more precise convergence rate estimates in the central limit theorem. *Theory of Probability and Its Applications* **23**, 660–663.

Ulyanov, V. (1986). Normal approximation for sums of non-identically distributed random variables in Hilbert spaces. *Acta Scientiarum Mathematicarum* **50**, 411–419. MR0882054

Ulyanov, V. (1987). Asymptotic expansions for distributions of sums of independent random variables in H. *Theory of Probability and Its Applications* **31**, 25–39. MR0836950

van der Vaart, A. W. (2007). *Asymptotic Statistics*. New York, NY: Cambrige University Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

Verbeek, A. (1992). The compactification of generalized linear models. *Statistica Neerlandica* **46**, 107–142. MR1178475 https://doi.org/10.1111/j.1467-9574.1992.tb01332.x

Williams, A. L., Fitzgerald, J. E., Ivich, F., Sontag, E. D. and Niedre, M. (2020). Short-term circulating tumor cell dynamics in mouse xenograft models and implications for liquid biopsy. *Frontiers in Oncology* **10**, 601085.

Zhu, X., Suo, Y., Ding, N., He, H. and Wei, X. (2017). Circulating tumor cells occur nonuniformly monitored by in vivo flow cytometry. In *International Conference on Photonics and Imaging in Biology and Medicine*, W3A–104.

Zhu, X., Suo, Y., Fu, Y., Zhang, F., Ding, N., Pang, K., Xie, C., Weng, X., Tian, M., He, H., et al (2021). In vivo flow cytometry reveals a circadian rhythm of circulating tumor cells. *Light: Science & Applications* **10**, 110.