J. Appl. Prob. **53,** 633–649 (2016)
*doi:*10.1017/jpr.2016.31
© *Applied Probability Trust* 2016

SEMIPARAMETRIC CROSS ENTROPY FOR RARE-EVENT SIMULATION

Z. I. BOTEV,* *The University of New South Wales*
A. RIDDER,** *Vrije Universiteit*
L. ROJAS-NANDAYAPA,*** *The University of Queensland*

Abstract

The cross entropy is a well-known adaptive importance sampling method which requires estimating an optimal importance sampling distribution within a parametric class. In this paper we analyze an alternative version of the cross entropy, where the importance sampling distribution is selected instead within a general semiparametric class of distributions. We show that the semiparametric cross entropy method delivers efficient estimators in a wide variety of rare-event problems. We illustrate the favourable performance of the method with numerical experiments.

Keywords: Light-tailed; regularly-varying; subexponential; rare-event probability; cross entropy method

2010 Mathematics Subject Classification: Primary 65C05
Secondary 65C60; 65C40

1. Introduction

In this paper we further analyze both numerically and theoretically a semiparametric version of the well-known cross entropy (CE) method for estimating rare-event probabilities of the type $\ell = \mathbb{P}(\boldsymbol{X} \in A_\gamma)$, where A_γ is a family of rare events. Recall that for estimating such probabilities, the CE proposes a methodology that selects an optimal multivariate importance sampling distribution taken from some appropriate predefined parametric family of distributions. In contrast, the semiparametric version deviates from the standard CE approach by considering instead a general class of importance sampling distributions with product-form densities. Such a rich class is simply characterized by having independent and absolutely continuous components; in particular, it can be proved that the optimal set of marginal components coincides with the marginal densities of the zero-variance importance sampling distribution.

An outstanding advantage of the methodology advocated here is that a single and broadly-applicable algorithm provides satisfactory practical performance on a wide range of rare-event estimation problems. We consider several key cases for which we formally prove that the resulting estimator can theoretical achieve either logarithmic or bounded relative error efficiencies for estimating tail probabilities of sums of nonnegative random variables. In particular, we consider the Weibull case where the decay of the tail probability is determined by the tail index α: light-tailed ($\alpha \geq 1$) or heavy-tailed ($\alpha < 1$). Remarkably, the proposed

Received 28 August 2013; revision received 26 August 2015.

* Postal address: School of Mathematics and Statistics, University of New South Wales, Sydney, NSW 2052 Australia. Email address: botev@unsw.edu.au

** Postal address: Department of Econometrics and Operations Research, Vrije Universiteit, 1081 HV Amsterdam, The Netherlands.

*** Postal address: School of Mathematics and Physics, The University of Queensland, Brisbane, QLD 4072, Australia.

estimator delivers an efficient estimator for all values of the tail index $0 < \alpha < 1$, including the elusive case where $\ln 3/2 \ln 2 < \alpha < 1$ (unlike our proposal, other existing procedures feature efficiency only for restricted sets of values of the tail index [2, Remark 3.1]). In addition, we provide proofs of efficiency for a general class of light-tailed distributions known as the *exponential class* and also for the archetypal regularly varying distribution: the Pareto case.

Numerically, we show that the proposed method not only performs satisfactorily for both light- and heavy-tailed problems, but can sometimes deliver numerical accuracy superior to that of estimation schemes specifically designed for heavy-tailed random variables.

Our numerical investigations suggest that the proposed methodology produces efficient estimators in a wide variety of settings which cover cases well beyond the ones presented in this paper. In fact, recently Perrakis *et al.* [14] have shown how the semiparametric idea described here can be used to estimate the marginal likelihood in some Bayesian computational problems.

The construction of the semiparametric CE estimator requires the following ingredients. First, we execute a pilot run (using Markov chain Monte Carlo (MCMC), for example) to generate random variables from a distribution which approximates the zero-variance importance sampling distribution. The MCMC approach of generating from the zero-variance measure is the same used in [12], for example. Second, with the sample at hand, we construct a conditional (Rao–Blackwell) estimator of each of the marginal densities of the zero-variance importance sampling distribution. Finally, we use the product of the (estimated) marginal densities as our importance sampling density in order to estimate ℓ.

The rest of the paper is organized as follows. In Section 2 we provide a brief review of the parametric CE method and then we introduce its semiparametric version. In Section 3 we provide a theoretical analysis of the efficiency of a simple version of the estimator for estimating tail probabilities of sums of light- and heavy- tailed random variables. This is followed by a number of examples in Section 4 with details about the practical implementation. The examples show that, at least in the heavy-tailed case, the proposed algorithm can yield an improvement in the relative error in the orders of magnitude.

2. The CE method

2.1. Parametric CE method

We wish to estimate $\ell = \mathbb{P}(S(\boldsymbol{X}) > \gamma)$, where $\boldsymbol{X} = (X_1, \ldots, X_d)$ is a random vector, $S\colon \mathbb{R}^d \to \mathbb{R}$ is some given function, and γ is large. In order to introduce the semiparametric version of the CE method for estimating the quantity ℓ, we briefly review the CE method itself. Assume that the random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ has a known density $f(\boldsymbol{x})$ which belongs to a parametric family of density functions $\mathcal{F} = \{f(\cdot;\boldsymbol{v})\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0}\colon \int f(\boldsymbol{x};\boldsymbol{v})\,\mathrm{d}\boldsymbol{x} = 1;\ \boldsymbol{v} \in \mathcal{V}\}$, where $\mathcal{V} \subset \mathbb{R}^p$ is a feasible parameter set; hence, $f(\boldsymbol{x}) := f(\boldsymbol{x};\boldsymbol{u}) \in \mathcal{F}$ for some $\boldsymbol{u} \in \mathcal{V}$. The objective is to find a parameter $\boldsymbol{v} \in \mathcal{V}$ that yields an optimal importance sampling estimator of the form: $\widehat{\ell}_{\mathrm{CE}} = m^{-1}\sum_{i=1}^m \mathbf{1}_{\{S(\boldsymbol{Y}_i)>\gamma\}} f(\boldsymbol{Y}_i;\boldsymbol{u})/f(\boldsymbol{Y}_i;\boldsymbol{v})$, where $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \sim f(\boldsymbol{y};\boldsymbol{v})$ independent and identically distributed (i.i.d.) and $\mathbf{1}$ is the indicator function. In the CE method, the optimal parameter $\boldsymbol{v}^* \in \mathcal{V}$ minimizes the cross entropy distance of $f(\cdot;\boldsymbol{v}) \in \mathcal{F}$ with respect to the zero-variance importance sampling density $\pi(\boldsymbol{x}) := \mathbf{1}_{\{S(\boldsymbol{x})>\gamma\}} f(\boldsymbol{x})/\mathbb{P}(S(\boldsymbol{X}) > \gamma)$. In other words,

$$\boldsymbol{v}^* = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{V}} \int \pi(\boldsymbol{x}) \ln\biggl(\frac{\pi(\boldsymbol{x})}{f(\boldsymbol{x};\boldsymbol{v})}\biggr)\,\mathrm{d}\boldsymbol{x} = \operatorname*{arg\,max}_{\boldsymbol{v}\in\mathcal{V}} \int \pi(\boldsymbol{x}) \ln f(\boldsymbol{x};\boldsymbol{v})\,\mathrm{d}\boldsymbol{x}. \tag{1}$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

In practice, the integral $\int \pi(\boldsymbol{x}) \ln(\pi(\boldsymbol{x})/f(\boldsymbol{x};\boldsymbol{v}))\,\mathrm{d}\boldsymbol{x}$ is estimated from a preliminary simulation so that an estimator of the optimal parameter $\boldsymbol{v}^*$ is given as $\widehat{\boldsymbol{v}^*} = \arg\max_{\boldsymbol{v}\in\mathcal{V}} \sum_{i=1}^n \ln f(\boldsymbol{X}_i, \boldsymbol{v})$, where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is a sample from a distribution approximating π. That sample can be obtained by, for example, MCMC sampling over the restricted set $\{\boldsymbol{x} : S(\boldsymbol{x}) > \gamma\}$; see [4], [5], and [12]. In this way we use a preliminary (pilot) run to learn about the optimal (in the cross entropy sense) parameter $\boldsymbol{v}^*$.

2.2. Semiparametric importance sampling

In the semiparametric CE method the objective is to find an optimal importance sampling distribution amongst a family of distributions with product-form densities, while the optimality criterion remains to minimize the cross-entropy distance from the zero-variance density. Denote by $\mathcal{G}_1$ the set of all single-variate probability density functions; that is, if $g \in \mathcal{G}_1$ then $g(x) : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is absolutely continuous with $\int g(x)\,\mathrm{d}x = 1$. Let $\mathcal{G}$ be the family of product-form densities on $\mathbb{R}^d$: $\mathcal{G} = \{g(\cdot) : \mathbb{R}^d \to \mathbb{R}_{\geq 0} : g(\boldsymbol{x}) = \prod_{i=1}^d g_i(x_i);\ g_i \in \mathcal{G}_1,\ i = 1, \ldots, d\}$. In this paper we consider $\mathcal{G}$ as the target set of importance sampling densities. Hence, the objective is to solve the functional optimization program $\min_{g\in\mathcal{G}} \int \pi(\boldsymbol{x}) \ln(\pi(\boldsymbol{x})/g(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}$. This is equivalent to solving

$$g(\boldsymbol{x}) = \underset{g_1,\ldots,g_d\in\mathcal{G}_1}{\arg\min} \int \pi(\boldsymbol{x}) \ln\biggl(\frac{\pi(\boldsymbol{x})}{\prod_{i=1}^d g_i(x_i)}\biggr)\,\mathrm{d}\boldsymbol{x} = \underset{g_1,\ldots,g_d\in\mathcal{G}_1}{\arg\max} \int \pi(\boldsymbol{x}) \ln\biggl(\prod_{i=1}^d g_i(x_i)\biggr)\,\mathrm{d}\boldsymbol{x}. \tag{2}$$

Lemma 1. *Let $\pi_i(x_i)$ be the ith marginal of the zero-variance density $\pi(\boldsymbol{x})$. Then the solution to the semiparametric CE program (2) is $g_i = \pi_i$ for all $i = 1, \ldots, d$.*

In other words, the optimal importance sampling density within the space of all product-form densities is given by the product of the marginals of $\pi(\boldsymbol{x})$. The straightforward proof is given in the appendix.

In practice, the marginal densities of π are typically unknown (just like the exact $\boldsymbol{v}^*$ in (1) is not available), but these can be easily estimated using simulation. In this paper we approximate the marginal densities of π as follows. Assume that both the conditional densities $\pi(x_i \mid \boldsymbol{x}_{-i}), i = 1, \ldots, d$, of the zero-variance importance sampling distribution are known, and a sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from a distribution approximating π is available (for example, from running the MCMC algorithm in [12]). Hence, we define the approximating multivariate density $\widehat{g}(\boldsymbol{y})$ to the optimal semiparametric CE solution as $\widehat{g}(\boldsymbol{y}) := \prod_{i=1}^d \widehat{\pi}_i(y_i)$, where $\widehat{\pi}_i(x_i) = (1/n)\sum_{k=1}^n \pi(x_i \mid \boldsymbol{Y}_{k,-i}),\ i = 1, \ldots, d$.

The approximation above is motivated by the fact that the marginal densities of π can be rewritten as $\pi_i(x_i) = \mathbb{E}[\pi_i(x_i \mid \boldsymbol{X}_{-i})]$, where $\boldsymbol{X} \sim \pi$. Further, since the conditional densities of the zero-variance importance distribution are assumed to be known then we naturally employ the Gibbs sampler to generate the sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$; see [12]. Finally, we can estimate ℓ via the importance sampling estimator

$$\widehat{\ell} = \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{S(\boldsymbol{Y}_i)>\gamma\}} \frac{f(\boldsymbol{Y}_i)}{\widehat{g}(\boldsymbol{Y}_i)}, \tag{3}$$

where $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \sim \widehat{g}(\boldsymbol{y})$ i.i.d.

Remark 1. (*Using the exact conditional density.*) Once we have sampled $Y_1, \ldots, Y_{d-1}$ from $\widehat{\pi}_1, \ldots, \widehat{\pi}_{d-1}$, respectively, we have the option of sampling the final Y_d from the exact conditional $\pi(y_d \mid Y_1, \ldots, Y_{d-1})$, instead of from the dth marginal $\widehat{\pi}_d$. This reduces the cross entropy

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

distance to π even further and yields the alternative, and typically more accurate, estimator (3) with $\widehat{g}(\boldsymbol{y})$ redefined as $\widehat{g}(\boldsymbol{y}) \leftarrow \widehat{\pi}_1(y_1) \times \cdots \times \widehat{\pi}_{d-1}(y_{d-1}) \times \pi(y_d \mid y_1, \dots, y_{d-1})$.

3. Robustness properties of semiparametric CE estimator

Although the semiparametric CE method is broadly applicable, in order to achieve theoretical tractability and clarity, we choose to examine its performance on the few frequently occurring prototypical [3], [9] rare-event estimation problem $\ell = \ell(\gamma) = \mathbb{P}(S(\boldsymbol{X}) > \gamma)$, with $S(\boldsymbol{x}) = x_1 + \cdots + x_d$. We assume that the $X_1, \dots, X_d$ are i.i.d. continuous random variables with right unbounded support and common distribution $F(\cdot)$. Let the X_i have density $f(\cdot)$. Let F^{*d} denote the d-fold convolution of F, so the probability of interest is $\ell(\gamma) = \overline{F^{*d}}(\gamma)$. By integrating the zero-variance importance sampling density we find that the ith marginal of the semiparametric CE distribution is $\pi_i(x_i) = f(x_i)\overline{F^{*(d-1)}}(\gamma - x_i)/\ell(\gamma)$, so the single-run estimator Z can be written as

$$Z = \mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}} \prod_{i=1}^{d} \frac{f(X_i)}{\pi_i(X_i)} = \frac{\mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}}\, \ell(\gamma)^d}{\prod_{i=1}^{d} \overline{F^{*(d-1)}}(\gamma - X_i)}, \tag{4}$$

where $\boldsymbol{X} \sim g(\mathbf{x}) = \prod_{i=1}^{d} \pi_i(x_i)$. Clearly, the estimator (4) is unbiased; furthermore, its second moment is given by

$$\mathbb{E}_g Z^2 = \mathbb{E}_g Z \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} = \mathbb{E}_f Z,$$

where $\mathbb{E}_f$ and $\mathbb{E}_g$ are the expectation operators under the densities f and g, respectively.

In the following, we study the robustness properties of the estimator (4) when $\gamma \to \infty$, so that $\ell(\gamma) \to 0$. We are interested in the behavior of the standard error of the estimator in this regime, relative to its mean $\ell(\gamma)$. Since we take a finite constant sample size, it suffices to analyze the robustness of the single-run estimator of $\ell(\gamma)$. We say that an estimator has bounded relative error if $\limsup_{\gamma\to\infty} \sqrt{\operatorname{var}(Z)}/\ell < \infty$, which is equivalent to having a bounded relative second moment: $\limsup_{\gamma\to\infty} \mathbb{E}Z^2/\ell^2 < \infty$, [1].

For our analysis, we assume that the importance sampling density $g = \prod_{i=1}^{d} \pi_i(x_i)$ is available. In practice, we estimate g via $\widehat{g}$ from an MCMC simulation as discussed in Section 2.2. In this respect, our analysis is similar in spirit to that conducted for the parametric cross entropy method [6]. In the following, we will prove the efficiency of the semiparametric importance sampling estimator (4) for a number of light- and heavy-tailed distributions.

3.1. Heavy-tailed case

In this section we assume that the X_i are subexponential [10]. For our proofs, it will be useful to consider the following decomposition of the relative second moment:

$$\frac{\mathbb{E}_g Z^2}{\ell^2(\gamma)} = \frac{\mathbb{E}_g \mathbf{1}_{\{M_d>\gamma\}} Z^2}{\ell^2(\gamma)} + \frac{\mathbb{E}_g \mathbf{1}_{\{M_d\le\gamma\}} Z^2}{\ell^2(\gamma)}, \tag{5}$$

where $M_d := \max_{i\le d} X_i$. The rest of the proof is divided in two lemmas. In Lemma 2 below we show that the first term in (5) is uniformly bounded if the X_i are subexponential. In Lemma 3 we provide an upper bound for the second term in (5), which will be examined later for each particular subexponential distribution considered.

Lemma 2. *We have*

$$\sup_{\gamma} \frac{\mathbb{E}_f \mathbf{1}_{\{M_d>\gamma\}} Z}{\ell^2(\gamma)} < \infty.$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

Proof. Observe that, in $\{M_d > \gamma\}$, there is a j such that $X_j > \gamma$, so $\overline{F^{*(d-1)}}(\gamma - X_j) = 1$, while for all other $i \neq j$ it holds trivially that $F^{*(d-1)}(\gamma - X_i) \geq F^{*(d-1)}(\gamma)$. Hence, using (4), we have

$$\frac{\mathbb{E}_f \, \mathbf{1}_{\{M_d>\gamma\}} \, Z}{\ell^2(\gamma)} \leq \frac{\ell(\gamma)^{d-2}}{(\overline{F^{*(d-1)}}(\gamma))^{d-1}} \mathbb{E}_f \, \mathbf{1}_{\{M_d>\gamma\}} \leq \frac{(\overline{F^{*d}}(\gamma))^{d-1}}{(\overline{F}(\gamma))^{d-1}},$$

the last inequality follows from

$$\mathbb{P}_f(M_d > \gamma) \leq \mathbb{P}_f(S(\boldsymbol{X}) > \gamma) = \overline{F^{*d}}(\gamma)$$

and $\overline{F^{*(d-1)}}(\gamma) \geq \overline{F}(\gamma)$ (recall that $\ell(\gamma) = \overline{F^{*d}}(\gamma)$). The lemma follows by using Kesten's bound [10]. □

Lemma 3. *We have*

$$\frac{\mathbb{E}_f \, \mathbf{1}_{\{M_d\leq\gamma\}} \, Z}{\ell^2(\gamma)} < c\overline{F}(\gamma)^{d-2}\mathbb{E}_f \frac{\mathbf{1}_{\{M_d<\gamma, S_d>\gamma\}}}{\prod_{i=1}^{d} \overline{F}(\gamma - X_i)},$$

where $c > 0$ is fixed.

Proof. The result follows by substituting (4) and then applying $\overline{F^{*(d-1)}}(x) \geq \overline{F}(x)$ for the denominator, and Kesten's bound [10] for the numerator. □

Next we analyze the bound provided by Lemma 3 for particular cases of subexponential distributions. By definition, subexponential distributions are nonnegative but suitable generalizations can be readily made for allowing distributions supported all over real numbers. First, we consider the Weibull case, where the X_i have a density $\alpha x^{\alpha-1}\mathrm{e}^{-x^{\alpha}}$ with $\alpha > 0$. Recall that if $0 < \alpha < 1$ then the X_i are subexponential.

Proposition 1. *The semiparametric CE has bounded relative error in the Weibull case with tail index $0 < \alpha < 1$.*

Proof. It is enough to prove that the right-hand side of the bound in Lemma 3 is finite as $\gamma \to \infty$. Consider the set $\mathcal{C} = \{\boldsymbol{x} \colon 0 < x_i < \gamma, \ \sum_i x_i > \gamma\}$ and write the bound in Lemma 3 as

$$\frac{\mathbb{E}_f \, \mathbf{1}_{\{M_d\leq\gamma\}} \, Z}{\ell^2(\gamma)}$$
$$< c\alpha^d \int \cdots \int_{\mathcal{C}} \bigg(\prod_{i=1}^{d} x_i^{\alpha-1}\bigg) \exp\bigg(-(d-2)\gamma^{\alpha} + \sum_{i=1}^{d}((\gamma - x_i)^{\alpha} - x_i^{\alpha})\bigg) \, \mathrm{d}\boldsymbol{x}.$$

After the change of variable $u_i = x_i/\gamma, i = 1, \ldots, n$, we obtain a Laplace-type integral: $\alpha^d\gamma^{d\alpha}\int\cdots\int_{\mathcal{D}} h(\boldsymbol{u})\mathrm{e}^{-\gamma^{\alpha}\phi(\boldsymbol{u})} \, \mathrm{d}\boldsymbol{u}$, where $\mathcal{D} := \{\boldsymbol{u} \colon 0 < u_i < 1, \sum_i u_i > 1\}, h(\boldsymbol{u}) := \prod_{i=1}^{d} u_i^{\alpha-1}$, and $\phi(\boldsymbol{u}) := d - 2 + \sum_{i=1}^{d}(u_i^{\alpha} - (1 - u_i)^{\alpha})$. Laplace-type integrals have the following properties. First, if $\bar{\mathcal{D}}$ denotes the closure of the open set $\mathcal{D}$, the function $\phi(\boldsymbol{u})$ attains its unique global minimum within the bounded domain $\bar{\mathcal{D}} \subseteq \mathbb{R}^d$ on the boundary at $\boldsymbol{u}^* = (1/d, \ldots, 1/d)$. This can be seen by either applying the Lagrange constraint optimization method or more simply by noting that $u^{\alpha} - (1 - u)^{\alpha}$ is monotonically increasing and $\phi(\boldsymbol{u})$ is a invariant to permutations of the components of $\boldsymbol{u}$. The minimum $\phi(\boldsymbol{u}^*) = d - 2 + d^{1-\alpha} - d^{1-\alpha}(d - 1)^{\alpha}$, as a function of d is such that for $d > 2$ we have the strict inequality

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

$\phi(\boldsymbol{u}) \geq \phi(\boldsymbol{u}^*) > 0$ for all $\boldsymbol{u} \in \bar{\mathscr{D}}$. The point $\boldsymbol{u}^*$ is not a critical point, since $(\partial\phi/\partial u_i)(\boldsymbol{u}) = \alpha(u_i^{\alpha-1} + (1-u_i)^{\alpha-1}) > 0$ for all i and $\boldsymbol{u} \in \mathscr{D}$. Second, the function $h\colon \mathbb{R}^d \to \mathbb{R}$ is continuous and the Hessian of the surface $p(u_1, \dots, u_{d-1}) = \phi(u_1, u_2, \dots, u_{d-1}, 1-u_1-u_2-\cdots-u_{d-1})$ is

$$\frac{\partial^2 p}{\partial u_i \partial u_j} = \alpha(\alpha-1)\begin{cases}\Big(1 - \sum\limits_{k<d} u_k\Big)^{\alpha-2} - \Big(\sum\limits_{k<d} u_k\Big)^{\alpha-2}, & i \neq j,\\ u_i^{\alpha-2} - (1-u_i)^{\alpha-2} + \Big(1 - \sum\limits_{k<d} u_k\Big)^{\alpha-2} - \Big(\sum\limits_{k<d} u_k\Big)^{\alpha-2}, & i = j,\end{cases}$$

which when evaluated at $\boldsymbol{u}^*$ yields a nondegenerate Hessian matrix. As a result of all these conditions, we have the Laplace-type asymptotic expansion at a boundary point (see [15, p. 500]), which is not a critical point $\int \cdots \int_{\mathscr{D}} h(\boldsymbol{u})\mathrm{e}^{-\gamma^\alpha \phi(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u} = \mathcal{O}(\gamma^{-\alpha(d+1)/2}\mathrm{e}^{-\gamma^\alpha\phi(\boldsymbol{u}^*)})$, where the constant $\phi(\boldsymbol{u}^*) > 0$. It follows that

$$\begin{aligned}\frac{\mathbb{E}\,\mathbf{1}_{\{M_d<\gamma\}}\,Z^2}{\ell^2} &\leq c_2\alpha^d\gamma^{d\alpha}\int\cdots\int_{\mathscr{D}} h(\boldsymbol{u})\mathrm{e}^{-\gamma^\alpha\phi(\boldsymbol{u})}\,\mathrm{d}\boldsymbol{u}\\ &= \mathcal{O}(\gamma^{\alpha(d-1)/2}\mathrm{e}^{-\gamma^\alpha\phi(\boldsymbol{u}^*)})\\ &= \mathcal{O}(\mathrm{e}^{\alpha(d-1)/2\ln\gamma - \gamma^\alpha\phi(\boldsymbol{u}^*)})\\ &\to 0 \quad (\gamma \to \infty).\end{aligned}$$

Hence, the second term in (5) vanishes as $\gamma \to \infty$. □

For our second example, we consider the Pareto case. More precisely, we assume that the X_i have common density $x^{-\alpha}$ with $\alpha > 0$ and $x > 1$.

Proposition 2. *The semiparametric CE is logarithmically efficient in the Pareto case.*

Proof. Define

$$H_n(\gamma) := \mathbb{E}_f\bigg[\prod_{k=1}^n \frac{\overline{F}(\gamma)}{\overline{F}(\gamma - X_k)};\, B_n\bigg],$$

with $B_n = \{S_{n-1} \leq \gamma, S_n > \gamma, M_n \leq \gamma\}$ for $n \geq 2$. Observing that $\{M_d \leq \gamma, S_d > \gamma\} \subset \bigcup_{n=2}^d B_n$ and $\overline{F}(\gamma)/\overline{F}(\gamma - x) \leq 1$, we arrive at (see also the proof of Lemma 3)

$$\mathbb{E}_f\,\mathbf{1}_{\{M_d\leq\gamma, S_d>\gamma\}}\,Z \leq c\sum_{n=2}^d \mathbb{E}_f\bigg[\prod_{i=1}^n \frac{\overline{F}(\gamma)}{\overline{F}(\gamma - X_i)};\, B_n\bigg] = c\sum_{n=2}^d H_n(\gamma). \tag{6}$$

In particular, with the aid of an appropriate change of variable it is possible to rewrite the quantities $H_n(\gamma) = \alpha^n\gamma^{-n\alpha}I_n(\gamma, 1)$, where the function $I_n(\gamma, 1)$ is the multiple integral

$$\begin{aligned}I_n(\gamma, \zeta) := \int_{\gamma^{-1}}^{1-(n-2)\gamma^{-1}} &\int_{\gamma^{-1}}^{1-y_1-(n-3)\gamma^{-1}}\\ &\cdots\int_{\gamma^{-1}}^{1-y_1-\cdots-y_{n-2}}\int_{(1-y_1-\cdots-y_{n-1})\vee\gamma^{-1}}^{1}\prod_{k=1}^n L(y_k)\,\mathrm{d}y_n\,\mathrm{d}y_{n-1}\cdots\mathrm{d}y_2\,\mathrm{d}y_1,\end{aligned}$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

with $L(y) := (1-y)^{\alpha} y^{-(\alpha+1)},\ y \in (0,1]$. Moreover, $I_n(\gamma,\zeta)$ can be defined recursively via

$$I_n(\gamma,\zeta) := \begin{cases} \displaystyle\int_{\zeta\vee\gamma^{-1}}^{1} L(y)\,\mathrm{d}y, & n=1, \\ \displaystyle\int_{\gamma^{-1}}^{\zeta-(n-2)\gamma^{-1}} L(y) I_{n-1}(\gamma,\zeta-y)\,\mathrm{d}y, & n \geq 2. \end{cases} \tag{7}$$

Next, we prove that, for $n=2,3,\ldots$, it holds that

$$\limsup_{\gamma\to\infty} \frac{I_n(\gamma,1)}{\gamma^{\alpha(n-2)}\ln\gamma} = 0. \tag{8}$$

From Lemma 5 in the appendix, we have a recursive expression for the derivative of the functions $I_n(\gamma,\zeta)$, i.e.

$$\frac{\partial}{\partial\gamma} I_n(\gamma,\zeta) = nL(\gamma^{-1}) I_{n-1}(\gamma,\zeta-\gamma^{-1})\gamma^{-2}, \qquad n=2,3,\ldots.$$

Therefore, we obtain, for $n=2,3,\ldots$,

$$\begin{aligned} \limsup_{\gamma\to\infty} \frac{I_n(\gamma,1)}{\gamma^{\alpha(n-2)}\ln\gamma} &= \limsup_{\gamma\to\infty} \frac{(\mathrm{d}/\mathrm{d}\gamma) I_n(\gamma,1)}{(\mathrm{d}/\mathrm{d}\gamma)\gamma^{\alpha(n-2)}\ln\gamma} \\ &= \limsup_{\gamma\to\infty} \frac{nL(\gamma^{-1}) I_{n-1}(\gamma,1-\gamma^{-1})\gamma^{-2}}{(1+\alpha(n-2)\ln\gamma)\gamma^{\alpha(n-2)-1}}. \end{aligned}$$

For $n=2$, the last expression can be written as $2L(\gamma^{-1})I_1(\gamma,1-\gamma^{-1})\gamma^{-1}$. Furthermore, observe that $L(\gamma^{-1}) = (1-\gamma^{-1})^{\alpha}\gamma^{\alpha+1} = \mathcal{O}(\gamma^{\alpha+1})$, while

$$I_1(\gamma,1-\gamma^{-1}) = \int_{1-\gamma^{-1}}^{1} L(y)\,\mathrm{d}y \leq \gamma^{-1}L(1-\gamma^{-1}) = \mathcal{O}(\gamma^{-(\alpha+1)})$$

(the inequality follows because the function $L(y)$ is decreasing on $(0,1]$). Hence,

$$\limsup_{\gamma\to\infty} \frac{2L(\gamma^{-1})I_1(\gamma,1-\gamma^{-1})}{\gamma} = \limsup_{\gamma\to\infty} \frac{c_1 \times \gamma^{\alpha+1}\gamma^{-(\alpha+1)}}{\gamma} = 0.$$

Assume that (8) holds for $n \geq 2$. Then reasoning as above and using Lemma 6 in the appendix for (7), we obtain, for $n+1$,

$$\begin{aligned} \limsup_{\gamma\to\infty} \frac{I_{n+1}(\gamma,1)}{\gamma^{\alpha(n-1)}\ln\gamma} &= \limsup_{\gamma\to\infty} \frac{(\mathrm{d}/\mathrm{d}\gamma) I_{n+1}(\gamma,1)}{(\mathrm{d}/\mathrm{d}\gamma)\gamma^{\alpha(n-1)}\ln\gamma} \\ &= \limsup_{\gamma\to\infty} \frac{(n+1)L(\gamma^{-1}) I_n(\gamma,1-\gamma^{-1})\gamma^{-2}}{(1+\alpha(n-1)\ln\gamma)\gamma^{\alpha(n-1)-1}} \\ &= \limsup_{\gamma\to\infty} \frac{(n+1)L(\gamma^{-1})(I_n(\gamma,1)+o(1))\gamma^{-2}}{(1+\alpha(n-1)\ln\gamma)\gamma^{\alpha(n-1)-1}} \qquad \text{(from (7))} \\ &= \limsup_{\gamma\to\infty} \frac{c_2\gamma^{\alpha+1} I_n(\gamma,1)\gamma^{-2} + o(1)}{c_3\gamma^{\alpha(n-1)-1}\ln\gamma} \\ &= \limsup_{\gamma\to\infty} c_4 \frac{I_n(\gamma,1)+o(1)}{\gamma^{\alpha(n-2)}\ln\gamma} \\ &= 0. \end{aligned}$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

Combining these arguments, we can complete the proof of the proposition, i.e.

$$\begin{aligned}\limsup_{\gamma\to\infty}\frac{\mathbb{E}\,\mathbf{1}_{\{M_d\le\gamma\}}\,Z^2}{\ell^{2-\varepsilon}} &\le \limsup_{\gamma\to\infty}\frac{c\sum_{n=2}^{d}H_n(\gamma)}{\ell^{2-\varepsilon}} \qquad \text{(from (6))}\\ &= c\limsup_{\gamma\to\infty}\sum_{n=2}^{d}\frac{\alpha^n I_n(\gamma)}{\gamma^{\alpha n}\ell^{2-\varepsilon}}.\end{aligned}$$

Note that $\ell=\overline{F^{*d}}(\gamma)\ge\overline{F}(\gamma)=\gamma^{-\alpha}$; thus, for $\varepsilon<1/\alpha$ (i.e. $\varepsilon\alpha<1$), we have $\ell^{2-\varepsilon}\ge\gamma^{-2\alpha}\gamma^{\alpha\varepsilon}\ge\gamma^{-2\alpha}\ln\gamma,\ \gamma\to\infty$. Combining this with the above, we obtain

$$\limsup_{\gamma\to\infty}\sum_{n=2}^{d}\frac{\alpha^n I_n(\gamma)}{\gamma^{\alpha n}\ell^{2-\varepsilon}}\le\sum_{n=2}^{d}\alpha^n\limsup_{\gamma\to\infty}\frac{I_n(\gamma)}{\gamma^{\alpha(n-2)}\ln\gamma}=0. \qquad \square$$

Remark 2. (*Sensitivity to deviations from* g.) Quantifying the error in approximating g via the MCMC approximation $\widehat{g}$ is beyond the scope of this paper. Nevertheless, similar to [6, Proposition 5.2], we consider whether any deviation from the true, but unknown, product of marginals probability density function (PDF) $g(\boldsymbol{x})=\prod_i\pi_i(\boldsymbol{x})$ wrecks the asymptotic efficiency of (4).

The following argument illustrates that, even though g is the best (in the cross entropy sense) importance sampling PDF of product form, it is not the only product-form PDF giving the same asymptotic efficiency. Suppose that instead of the exact $\pi_i(x_i)$ in (4), we use the much simpler $\breve{\pi}_i(x_i)=f(x_i)\overline{G}(\gamma-x_i)/\overline{F*G}(\gamma)$, where G is a weak tail equivalent to F [10, p. 45], and $F*G$ denotes the convolution of F and G. Then, in the subexponential case,

$$\begin{aligned}\prod_i\frac{\pi_i(x_i)}{\breve{\pi}_i(x_i)}&=\prod_i\frac{\overline{F^{*(d-1)}}(\gamma-x_i)\,\overline{F*G}(\gamma)}{\overline{G}(\gamma-x_i)\,\ell(\gamma)}\\ &\le\left(\frac{\overline{F*G}(\gamma)}{\overline{F}(\gamma)}\right)^d\prod_i\frac{\overline{F^{*(d-1)}}(\gamma-x_i)}{\overline{G}(\gamma-x_i)}\\ &\le\left(\frac{\overline{F*G}(\gamma)}{\overline{F}(\gamma)}\right)^d\prod_i\frac{c_1\overline{F}(\gamma-x_i)}{\overline{G}(\gamma-x_i)} \qquad \text{(Kesten's bound)}\\ &\le\left(\frac{\overline{F*G}(\gamma)}{\overline{F}(\gamma)}\right)^d\prod_i c_1c_2\\ &=c_3\left(\frac{\overline{F*G}(\gamma)}{\overline{F}(\gamma)}\right)^d \qquad \text{(tail equivalence)}\end{aligned}$$

for some constants c_1,c_2,c_3. Hence, if $\breve{Z}=Z\prod_i(\pi_i(X_i)/\breve{\pi}_i(X_i))$ is the estimator using the simpler marginals, then

$$\mathbb{E}_f\breve{Z}=\mathbb{E}_fZ\prod_i\frac{\pi_i(X_i)}{\breve{\pi}_i(X_i)}\le c_3\left(\frac{\overline{F*G}(\gamma)}{\overline{F}(\gamma)}\right)^d\mathbb{E}_fZ.$$

In other words, $\limsup_\gamma\mathbb{E}_f\breve{Z}/\mathbb{E}_fZ<\infty$, since for subexponential F, we have

$$\frac{\overline{F*G}(\gamma)}{\overline{F}(\gamma)}=\mathcal{O}(1)\quad\text{as }\gamma\uparrow\infty$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

using [10, Corollary 3.19]. The last inequality suggests that $\breve{Z}$ inherits the asymptotic properties of Z in the subexponential case. As a consequence, if $\widehat{Z} = \mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}} \prod_i f(X_i)/\widehat{\pi}_i(X_i)$ with $\boldsymbol{X} \sim \widehat{g}(\boldsymbol{x})$ is the estimator using the approximate marginals $\{\widehat{\pi}_i\}$, then the combined estimator $\widetilde{Z} = w\widehat{Z} + (1-w)\breve{Z}$, where $w \in (0,1)$ and $w = \mathcal{O}(\overline{F}(\gamma))$ has relative error of the same order as the relative error of (4).

3.2. Light-tailed case

We consider the case when F belongs to a subfamily of light-tailed distributions as defined by Embrechts and Goldie [8]. We say that a distribution F belongs to the *Embrechts–Goldie* family of distributions indexed by $\theta \geq 0$ and denoted by $\mathcal{L}(\theta)$ if $\lim_{\gamma\to\infty} \overline{F}(\gamma+x)/\overline{F}(\gamma) = \mathrm{e}^{-\theta x}$. If θ is strictly larger than 0 then $\mathcal{L}(\theta)$ contains light-tailed distributions exclusively and the whole class is often referred to as the *exponential class*. The exponential class is very rich as it contains distributions with an exponential-type tail decay. For instance, the class of nonnegative *matrix exponential* distributions (a dense class within the continuous nonnegative distributions) is a particular subset of the exponential class. In contrast, if $\theta = 0$ then $\mathcal{L}(0)$ corresponds to the class of *long-tailed* (and thus heavy-tailed) distributions.

We concentrate on the efficiency of the semiparametric CE method for estimating tail probabilities of sums of random variables in the light-tailed exponential class (i.e. $\theta > 0$). In doing so, we require some results of the so-called *long-tailed functions* (cf. [10, Definition 2.14]). Recall that h is long-tailed if it is ultimately positive and $\lim_{\gamma\to\infty} h(\gamma+x)/h(\gamma) = 1$ for all x. Thus, if $F \in \mathcal{L}(0)$ then the tail probability $\overline{F}$ is long-tailed. We summarize important properties for both the exponential class and long-tailed functions as follows:

(i) $\mathcal{L}(\theta)$ is closed under convolutions [8, Theorem 3];

(ii) If $F \in \mathcal{L}$ and we let $G(x) := 1 - (\overline{F}(x))^{\alpha}$, then $G \in \mathcal{L}(\alpha\theta)$ for all $\alpha > 0$;

(iii) If $F \in \mathcal{L}(\theta)$ then $\overline{F}(\gamma) = \mathrm{e}^{-\theta\gamma}h(\gamma)$, with h long-tailed.

(iv) If h is long-tailed then we have the *long-decay condition* $\lim_{\gamma\to\infty} h(\gamma)/\mathrm{e}^{-\varepsilon\gamma} = \infty$ for all $\varepsilon > 0$; hence, a long-tailed function decays at a slower asymptotic rate than any exponential function [10, Lemma 2.17].

These properties will be employed in order to construct an asymptotic upper bound for the semiparametric estimator. In particular, in the following lemma we show that the ratio of two tail convolutions of a distribution in $\mathcal{L}(\theta)$ cannot increase/decrease at a faster rate than exponential.

Lemma 4. *Let $F \in \mathcal{L}(\theta), \theta > 0$, and $d_1, d_2 \in \mathbb{N}$. Then $\overline{F^{*d_1}}(\gamma)/\overline{F^{*d_2}}(\gamma) = o(\mathrm{e}^{\varepsilon\gamma})$ for all $\varepsilon > 0$.*

Proof. By property (i), $\mathcal{L}(\theta)$ is closed by convolution. Hence, by (iii) both $F^{*d_1}, F^{*d_2} \in \mathcal{L}(\theta)$, and their tail distributions have decompositions as in $\overline{F}(\gamma) = \mathrm{e}^{-\theta\gamma}h(\gamma)$ for some long-tailed functions h_1 and h_2. Therefore,

$$\frac{\overline{F^{*d_1}}(\gamma)}{\overline{F^{*d_2}}(\gamma)} = \frac{h_1(\gamma)\mathrm{e}^{-\theta\gamma}}{h_2(\gamma)\mathrm{e}^{-\theta\gamma}} = \frac{h_1(\gamma)}{h_2(\gamma)}.$$

First, we argue that both $h_1(\cdot)/h_2(\cdot)$ and its reciprocal function are long-tailed. This holds, since they are ultimately positive, and

$$\frac{h_1(\gamma+x)h_1(\gamma)}{h_2(\gamma)h_2(\gamma+x)} = \frac{h_1(\gamma+x)h_2(\gamma)}{h_1(\gamma)h_2(\gamma+x)} \to 1.$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

The reciprocal function behaves similarly. Thus, $h_2(\cdot)/h_1(\cdot)$ satisfies the long-decay condition (property (iv)), which says $\lim_{\gamma\to\infty} h_2(\gamma)/h_1(\gamma)\mathrm{e}^{-\varepsilon\gamma} = \infty$. Clearly, this is equivalent to $\lim_{\gamma\to\infty} h_1(\gamma)/h_2(\gamma)\mathrm{e}^{\varepsilon\gamma} = 0$. □

We also have the following mild assumption. For instance, it is trivially satisfied by the exponential and gamma distributions.

Assumption 1. *Let $F(x) = h(x)\mathrm{e}^{-\theta x}$, where h is a long-tailed function. We assume that*

$$G(\gamma) := \sup\left\{\frac{h(\gamma)}{h(x)} : 0 \le x \le \gamma\right\} = o(\mathrm{e}^{\varepsilon\gamma}) \quad \textit{for all } \varepsilon > 0.$$

Proposition 3. *Let $F \in \mathcal{L}(\theta)$. If Assumption 1 holds then the semiparametric CE estimator is logarithmically efficient.*

Proof. Recall that $\mathbb{E}Z^2 = \mathbb{E}_f \mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}} \prod_{i=1}^d \overline{F^{*d}}(\gamma)/\overline{F^{*(d-1)}}(\gamma - X_i)$. Since $F^{*(d-1)} \in \mathcal{L}(\theta)$, we can use the decomposition $\overline{F}(\gamma) = \mathrm{e}^{-\theta\gamma}h(\gamma)$ to write $\overline{F^{*(d-1)}}(\gamma) = h(\gamma)\mathrm{e}^{-\theta\gamma}$ for some $h(\cdot)$ long-tailed function, and $\theta > 0$. Then, we obtain the bound

$$\begin{aligned}
\prod_{i=1}^d \frac{\overline{F^{*(d-1)}}(\gamma)}{\overline{F^{*(d-1)}}(\gamma - X_i)} &= \prod_{i=1}^d \frac{h(\gamma)}{h(\gamma - X_i)} \frac{\mathrm{e}^{-\theta\gamma}}{\mathrm{e}^{-\theta(\gamma - X_i)}} \\
&\le \left(\sup_{0\le x\le\gamma} \frac{h(\gamma)}{h(\gamma - x)}\right)^d \prod_{i=1}^d \mathrm{e}^{-\theta X_i} \\
&= (G(\gamma))^d \mathrm{e}^{-\theta S(\boldsymbol{X})}.
\end{aligned}$$

Define $H(\gamma) = [\overline{F^{*d}}(\gamma)\big/\overline{F^{*(d-1)}}(\gamma)]^d$, then, using the bound and $\theta > 0$,

$$\begin{aligned}
\frac{\mathbb{E}Z^2}{\ell^{2-\varepsilon}(\gamma)} &= \frac{H(\gamma)}{\ell^{2-\varepsilon}(\gamma)} \mathbb{E}_f \mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}} \prod_{i=1}^d \frac{\overline{F^{*(d-1)}}(\gamma)}{\overline{F^{*(d-1)}}(\gamma - X_i)} \\
&\le \frac{H(\gamma)G^d(\gamma)}{\ell^{2-\varepsilon}(\gamma)} \mathbb{E}_f \mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}} \mathrm{e}^{-\theta S(\boldsymbol{X})} \\
&\le \frac{H(\gamma)G^d(\gamma)}{\ell^{2-\varepsilon}(\gamma)} \mathrm{e}^{-\theta\gamma} \mathbb{P}_f(S(\boldsymbol{X}) > \gamma) \\
&= \frac{H(\gamma)G^d(\gamma)\mathrm{e}^{-\theta\gamma}}{\ell^{1-\varepsilon}(\gamma)}.
\end{aligned}$$

Applying the properties of the exponential class, we obtain $\ell^{1-\varepsilon}(\gamma) = (\overline{F^{*d}}(\gamma))^{1-\varepsilon} = \mathrm{e}^{-\theta(1-\varepsilon)\gamma} h_d^{1-\varepsilon}(\gamma)$ for some long-tailed function h_d. In consequence,

$$\limsup_{\gamma\to\infty} \frac{\mathbb{E}Z^2}{\ell^{2-\varepsilon}(\gamma)} \le \limsup_{\gamma\to\infty} \frac{H(\gamma)G^d(\gamma)\mathrm{e}^{-\theta\gamma}}{\ell^{1-\varepsilon}(\gamma)} = \limsup_{\gamma\to\infty} \frac{H(\gamma)G^d(\gamma)}{h_d^{1-\varepsilon}(\gamma)} \mathrm{e}^{-\varepsilon\theta\gamma}.$$

Now, property $\overline{F}(\gamma) = \mathrm{e}^{-\theta\gamma}h(\gamma)$, Lemma 4, and Assumption 1 imply that all of the functions H, G, $h_d^{\varepsilon-1}$, and their products increase at less than exponential rate; namely,

$$\frac{H(\gamma)G^d(\gamma)}{h_d^{1-\varepsilon}(\gamma)} = o(\mathrm{e}^{\theta\varepsilon\gamma}).$$

Hence, the last limit is 0. □

Remark 3. (*Sensitivity to deviations from g.*) Similar to Remark 2, we comment on how robust the estimator Z is to deviations from the true, but unknown, product importance sampling PDF g. As in [6, Proposition 4.2], we consider the case in which $f_i(x) = \mathrm{e}^{-x}$, $x \geq 0$. Recall that X_i in (4) is generated from the exact marginal $\pi_i(\cdot)$. Assume that, instead of the true X_i, we can only simulate the noise-polluted $\widetilde{X}_i = X_i + N_i$, where N_i is normally distributed noise with mean 0 and variance σ_i^2 (small relative to the mean of X_i), independent from X_i, and with PDF ϕ_i. The PDF of $\widetilde{X}_i$ is then given by the convolution $\widetilde{\pi}_i(x_i) = (\pi_i * \phi_i)(x_i)$. After some manipulation, the ith likelihood ratio can be expressed as

$$\frac{\widetilde{\pi}_i(x_i)}{\pi_i(x_i)} = \int_0^\infty f_i(y)\phi_i(x_i - y)\,\frac{\overline{F^{*(d-1)}}(\gamma - y)}{\overline{F^{*(d-1)}}(\gamma - x_i)}\,\mathrm{d}y \bigg/ f_i(x_i).$$

Denote the numerator of this expression by $D_{\gamma,i}(x_i)$, and define

$$L_{\gamma,i}(x_i) = \frac{\pi_i(x_i)}{\widetilde{\pi}(x_i)}, \qquad L_i(x_i) = \frac{f_i(x_i)}{(f * \phi_i)(x_i)},$$
$$L_\gamma(\boldsymbol{x}) = \prod_i L_{\gamma,i}(x_i) = \frac{g(\boldsymbol{x})}{\widetilde{g}(\boldsymbol{x})}, \qquad L(\boldsymbol{x}) = \prod_i L_i(x_i).$$

It is not difficult to show that

(i) for each x_i,

$$\lim_{\gamma\to\infty} D_{\gamma,i}(x_i) = \int_0^\infty f_i(y)\phi_i(x_i - y)\,\mathrm{d}y = (f_i * \phi_i)(x_i);$$

(ii) the convergence in mean $\lim_{\gamma\to\infty} \mathbb{E}_f[L_\gamma(\boldsymbol{X})] = \mathbb{E}_f[L(\boldsymbol{X})]$;

(iii) the exact estimator (4) and the likelihood ratio L_γ are negatively correlated.

As a consequence, for the perturbed (noise-polluted) estimator $\widetilde{Z} = \mathbf{1}_{\{S(\boldsymbol{X})>\gamma\}} f(\boldsymbol{X})/\widetilde{g}(\boldsymbol{X})$, we obtain

$$\mathbb{E}_{\widetilde{g}}[\widetilde{Z}^2] = \mathbb{E}_f[\widetilde{Z}] = \mathbb{E}_f[ZL_\gamma(\boldsymbol{X})] \leq \mathbb{E}_f[Z]\mathbb{E}_f[L_\gamma(\boldsymbol{X})] = \mathbb{E}_f[Z] \times (\mathbb{E}_f[L(\boldsymbol{X})] + o(1)),$$

implying that the estimator Z is robust to Gaussian perturbations in the marginal densities $\{\pi_i\}$.

4. Examples and practical implementation

The proposed semiparametric CE estimator can yield practical performance which compares very favorably with respect to alternative procedures such as the Asmussen–Kroese (AK) estimator [2]. To improve the relative time variance of our estimator of $\mathbb{P}(X_1 + \cdots + X_d > \gamma) = \mathbb{P}(S > \gamma)$, we exploit a decomposition proposed in [13] and written as

$$\begin{aligned}\ell &= 1 - \mathbb{P}(M_d < \gamma) + d\mathbb{P}(X_d = M_d < \gamma)\mathbb{P}(S > \gamma \mid X_d = M_d < \gamma)\\ &= \underbrace{1 - [F(\gamma)]^d}_{\text{Dominant term}} + \mathbb{P}(M_d < \gamma)\underbrace{\widetilde{\mathbb{P}}(S > \gamma)}_{\text{Residual probability}},\end{aligned}$$

where the probability measure $\widetilde{\mathbb{P}}(\cdot) := \mathbb{P}(\cdot \mid X_d = M_d < \gamma)$ with corresponding density

$$\widetilde{f}(\boldsymbol{x}) = f(\boldsymbol{x} \mid X_d = M_d < \gamma) = \frac{\mathrm{d}f(\boldsymbol{x})}{[F(\gamma)]^d}\,\mathbf{1}_{\{M_d<\gamma, X_d=M_d\}}.$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

TABLE 1: Comparison of importance sampling method with the AK estimator. Algorithmic parameters were chosen to be $n = 10^3$, $m = 10^6$, and $d = 10$. The AK estimator is based on $m = 10^6$ replications.

γ	$\widehat{\ell}$	Relative Error	Ratio	τ_{AK}/τ	RTVP
		$\alpha = 0.1$			
10^{10}	4.54×10^4	1.7×10^6	13^2	0.4	71
10^{11}	3.40×10^5	4.1×10^7	22^2	0.4	197
10^{12}	1.30×10^6	6.4×10^8	72^2	0.4	2071
10^{13}	2.16×10^8	8.0×10^9	59^2	0.4	1429
10^{15}	1.84×10^{13}	1.3×10^{10}	125^2	0.4	5944
		$\alpha = 0.2$			
10^4	1.97×10^2	6.5×10^5	3.0^2	0.4	3.7
10^5	4.64×10^4	1.8×10^5	5.6^2	0.4	12
10^6	1.31×10^6	3.0×10^6	9.2^2	0.4	33
10^7	1.23×10^{10}	4.3×10^7	10.0^2	0.4	42
10^8	5.13×10^{17}	6.5×10^8	7.0^2	0.4	20
		$\alpha = 0.6$			
10^2	9.47×10^6	2.6×10^4	19^2	0.4	130
150	7.83×10^8	1.5×10^4	41^2	0.3	550
200	1.34×10^9	1.5×10^4	63^2	0.3	1376
500	1.83×10^{17}	1.7×10^4	5.5^2	0.4	11
10^3	7.00×10^{27}	9.5×10^5	6^2	0.4	13
		$\alpha = 0.9$			
30	1.33×10^4	9×10^4	13^2	0.3	50
40	6.27×10^7	9×10^4	78^2	0.3	1758.7
50	2.25×10^9	1×10^3	254^2	0.3	17746
60	7.01×10^{12}	1×10^3	556^2	0.3	87103
100	4.34×10^{22}	1×10^3	300^2	0.3	23768

Estimating the residual probability, we obtain the replication estimator for ℓ as

$$\widehat{\ell} = 1 - [F(\gamma)]^d + \frac{\widetilde{f}(\boldsymbol{Y})}{\widehat{g}(\boldsymbol{Y})} \mathbf{1}_{\{S(\boldsymbol{Y})>\gamma\}}, \qquad \boldsymbol{Y} \sim \widehat{g}(\boldsymbol{y}), \tag{9}$$

where $\widehat{g}(\boldsymbol{y}) := \widehat{\pi}_1(y_1)\cdots\widehat{\pi}_{d-1}(y_{d-1})\pi(y_d \mid y_1, \dots, y_{d-1})$ is the estimated importance sampling PDF described in Remark 1. In the next three examples, we employ the following performance measures: the relative time variance product (RTVP) and the ratio (R) of relative errors as a measure of efficiency, i.e. $\mathrm{R} := \widehat{\sigma}_{\mathrm{AK}}/\widehat{\sigma}$ and $\mathrm{RTVP} := \mathrm{R}^2(\tau_{\mathrm{AK}}/\tau)$, where $\widehat{\sigma}_{\mathrm{AK}}$ and $\widehat{\sigma}$ are the sample standard deviations of the AK estimator and $\widehat{\ell}$ (all based on m replications), respectively; and τ_{AK} and τ are the processor times taken to compute the respective estimators. The quantity τ includes the processor time needed for the preliminary MCMC simulations.

Example 1. (*Weibull case.*) In this example, we assume that the X_i are i.i.d. Weibull distributed. The results of our numerical experimentations are presented in Table 1. The proposed semiparametric CE estimator provides an improvement in RTVP with respect to all other estimators and for all values of the parameters α and γ considered. The improvement, however, is not uniform.

For instance, if $\alpha = 0.1$ then the RTVP varies in the range 71 to 5944. The general trend is that large gains in efficiency occur for smaller values of γ and $\alpha > 0.6$ or $\alpha < 0.3$. In comparison to (9), the AK estimator delivers less efficient estimators for $\alpha \notin [0.3, 0.6]$ and more efficient in the range $\alpha \in [0.3, 0.7]$. Note that the AK estimator is much faster to evaluate than (9), but this speed is insufficient to offset the substantial gains in squared relative error.

Example 2. (*Compound sum.*) We are interested in estimating the tail probability of a compound sum of the form $\mathbb{P}(X_1 + \cdots + X_R > \gamma)$, where the jumps X_i are i.i.d. with Weibull distribution with parameter $0 < \alpha < 1$, and (without loss of generality) $R \sim \text{geom}(\varrho)$ is a geometric random variable with PDF $\varrho(1-\varrho)^{r-1},\ r = 1, 2, \ldots$. We have

$$\begin{aligned}\mathbb{P}(S_R > \gamma) &= \varrho \sum_{r=1}^{\infty} (1-\varrho)^{r-1} \mathbb{P}(S_r > \gamma) \\ &= \underbrace{\frac{\overline{F}(\gamma)}{\overline{F}(\gamma) + \varrho F(\gamma)}}_{\text{Dominant term}} + \frac{\varrho(1-\varrho)(F(\gamma))^2}{\overline{F}(\gamma) + \varrho F(\gamma)} \underbrace{\widetilde{\mathbb{P}}(S_R > \gamma)}_{\text{Residual probability}},\end{aligned}$$

where under the new probability measure $\widetilde{\mathbb{P}}$, we have $(R-1) \sim \text{geom}(\overline{F}(\gamma) + \varrho F(\gamma))$ with PDF $\widetilde{\mathbb{P}}(R = r) = f_R(r),\ r = 2, 3, \ldots$, and $X_1, X_2, \ldots \sim f(x)$ i.i.d. with PDF given by the truncated Weibull density $f(x) = \alpha x^{\alpha-1} \mathrm{e}^{-x^{\alpha}} / (1 - \mathrm{e}^{-\gamma^{\alpha}}),\ 0 < x < \gamma$. Hence, we can again apply our importance sampling estimator to estimate the residual probability $\widetilde{\mathbb{P}}(S_R > \gamma)$. The zero-variance PDF for the estimation of the residual is $\pi(\boldsymbol{y}, r) \propto f_R(r) \prod_{j=1}^{r} f(y_j)\ \mathbf{1}_{\{S_r > \gamma\}}$, which can easily be sampled from using the Gibbs sampler, noting that

$$\pi(r \mid \boldsymbol{Y}) \propto f_R(r) \mathbf{1}_{\{r \geq r^*(\boldsymbol{Y})\}}, \qquad r^*(\boldsymbol{Y}) := \min\{r \colon Y_1 + \cdots + Y_r > \gamma\}.$$

In Table 2 we present the results of a number of numerical experiments. The results of our proposed method are significantly better in all cases, except $\alpha = 0.2$ with $1/\varrho \in \{50, 100\}$. In the latter case, the variance reduction achieved by the proposed method is not sufficient to offset the computational cost of simulating compound sums of expected length of $1/\varrho$. Note that, for $\alpha \geq 0.5$, the proposed method can be thousands of times more efficient. Our proposed method also compares favorably to other methods [7], [12], and in several examples exhibits better empirical efficiency. For example, based on the reported variances and computing time for the improved AK estimator [11, Table 2], in terms of the RTVP our estimator is from 8.5 to 45 times more efficient. We must note, however, that the results given in [11, Table 2] appear to be incorrect. For example, for $\varrho = 0.15, \alpha = 0.75$, and $\gamma = 63.361$, Table 2 reports the estimate 5.23×10^{-4} with relative error of 0.4%. In contrast, we obtain the estimate 5.38×10^{-4} with relative error 0.03%, which we verify with a crude Monte Carlo simulation using 10^9 repetitions.

Example 3. (*M/M/1 queue.*) Consider the classical example corresponding to the M/M/1 queue; see, for example, [1, Example VI.2.3]. Let $X_1, X_2, \ldots \sim f(x)$ i.i.d., where $f(x)$ is the density of $E_\mu - E_\lambda$, with E_μ and E_λ being independent exponential waiting times with means $1/\mu$ and $1/\lambda$, respectively, and $\mu > \lambda$. We are interested in estimating $\ell = \mathbb{P}(\max\{S_1, S_2, \ldots\} > \gamma) = \mathbb{P}(\tau < \infty)$, where $\tau = \min\{n \colon S_n > \gamma\}$. We compare the semiparametric method with the classical Siegmund's estimator, which suggests the importance sampling scheme, in which the rates of E_μ and E_λ are switched from (μ, λ) to (λ, μ). Given n trajectories $\{X_{k,1}, X_{k,2}, \ldots, X_{k,\tau_k}\}$ approximately drawn from the zero-variance measure via

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

TABLE 2: Compound Weibull sum with expected number of jumps $1/\varrho$. Here $n = 10^4$ and $m = 10^6$.

$1/\varrho$	$\widehat{\ell}$	Relative Error	Ratio	τ_{AK}/τ	RTVP
$\alpha = 0.2$ with $\gamma = 10^6$ fixed					
5	6.56×10^7	1.4×10^5	3.6^2	0.70	9.60
10	1.31×10^6	3.1×10^5	2.8^2	0.40	3.50
20	2.65×10^6	5.1×10^5	2.2^2	0.20	1.20
50	6.81×10^6	1.7×10^4	1.4^2	0.02	0.03
100	1.42×10^5	1.7×10^4	2.0^2	0.01	0.04
$\alpha = 0.5$ with $\gamma = 500$ fixed					
3	7.34×10^{10}	7.3×10^4	4.0^2	1.00	16
5	1.60×10^9	1.0×10^3	4.1^2	0.70	12
10	1.17×10^8	1.7×10^3	47^2	0.20	445
20	1.24×10^5	7.2×10^4	246^2	0.10	7300
50	7.90×10^3	2.1×10^4	58^2	0.03	110
$\alpha = 0.8$ with $\gamma = 30/\varrho$ depending on ϱ					
3	6.29×10^{11}	1.2×10^3	330^2	0.400	46000
5	1.65×10^{11}	6.4×10^4	930^2	0.200	200 000
10	6.94×10^{12}	3.8×10^4	2561^2	0.100	780 000
20	4.64×10^{12}	2.7×10^4	3636^2	0.003	34000
50	3.68×10^{12}	2.1×10^4	1485^2	0.010	27000
$\alpha = 0.95$ with $\gamma = 30/\varrho$ depending on ϱ					
5	2.61×10^{13}	4.8×10^4	10^6	0.1	$> 10^5$
10	2.18×10^{13}	3.0×10^4	$> 10^6$	0.1	$> 10^5$
20	2.00×10^{13}	2.2×10^4	$> 10^6$	0.1	$> 10^5$
50	1.91×10^{13}	1.9×10^4	$> 10^6$	0.1	$> 10^5$
100	1.88×10^{13}	1.7×10^4	$> 10^6$	0.1	$> 10^5$

MCMC, one can estimate the marginal density $\pi_1(x_1)$ via

$$\widehat{\pi}_1(x_1) = \frac{1}{n}\sum_{k=1}^{n} f(x_1 \mid x_1 > \gamma - \max\{0, X_{k,2}, X_{k,2} + X_{k,3}, \ldots\}).$$

Similarly, for all $\widehat{\pi}_j(x_j),\ j \le \max \tau_k$.

With this setup, we obtain Table 3, which shows the performance of the semiparametric procedure relative to Sigmund's algorithm with $m = 10^5$ independent replications. From the results we can draw the following conclusions. The semiparametric approach can yield lower relative errors than Siegmund's exponential change of measure (as seen from the 'ratio' column). However, this relative error advantage quickly disappears as γ becomes larger and larger. The results confirm that asymptotically Siegmund's algorithm cannot be improved (within a state-independent importance sampling framework) and that the semiparametric approach can successfully estimate this optimal change of measure. This example also points to the limitations of our method—the reduction in the relative error here is not enough to offset the computational overhead of estimating the optimal change of measure (as seen from the

TABLE 3: The M/M/1 queue example for two values of (μ, λ). Here $n = 10^4$ and $m = 10^5$.

γ	$\widehat{\ell}$	Relative Error	Ratio	τ_{Sig}/τ	RTVP
		$(\mu, \lambda) = (2, \frac{1}{2})$			
1	5.570×10^{-2}	0.20%	1.80^2	6×10^{-3}	0.02
2	1.240×10^{-2}	0.24%	1.30^2	6×10^{-3}	0.01
5	1.380×10^{-4}	0.31%	1.10^2	2×10^{-2}	0.02
7	6.832×10^{-6}	0.33%	1.08^2	2×10^{-2}	0.02
10	7.653×10^{-8}	0.35%	1.03^2	2×10^{-2}	0.02
		$(\mu, \lambda) = (8, \frac{1}{2})$			
0.1	2.943×10^{-2}	0.14%	6.0^2	6×10^{-3}	0.20
0.5	1.474×10^{-3}	0.25%	3.3^2	6×10^{-3}	0.06
1.0	3.445×10^{-5}	0.39%	2.1^2	1×10^{-2}	0.05
2.0	1.950×10^{-8}	0.51%	1.6^2	1×10^{-2}	0.03
3.0	1.063×10^{-11}	0.85%	1.4^2	1×10^{-2}	0.02

RTVP column). Nevertheless, our main point stands—a single broadly applicable importance sampling scheme can estimate the optimal change of measure in both light- and heavy-tailed settings, and in the heavy-tailed setting can be more efficient than current tailor-made schemes.

5. Conclusions

In this paper we have studied the theoretical and empirical performance of the semiparametric cross entropy method for estimating a rare-event probability. We show that the same procedure is efficient in both light- and heavy-tailed cases. The numerical examples confirm that the same scheme works in both light- and heavy-tailed settings. Compared to current state-of-the-art estimators, our estimator gives significantly better efficiency in the heavy-tailed case. This is especially relevant for probabilities involving the Weibull distribution with tail index $\alpha < 1$, but close to unity. This setting yields behavior intermediate between the typical heavy- and light-tailed behavior expected of rare events. As a result, while existing procedures are inefficient or fail completely, our method estimates reliably Weibull probabilities for any values of α, including $\alpha > 1$.

Appendix

Proof of Lemma 1. First note that for any single-variate function h,

$$\begin{aligned}\int_{\mathbb{R}^d} h(x_1)\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_{\mathbb{R}} h(x_1)\biggl(\int_{\mathbb{R}^{d-1}} \pi(x_1, x_2, \dots, x_d)\,\mathrm{d}x_2 \cdots \mathrm{d}x_d\biggr)\,\mathrm{d}x_1 \\ &= \int h(x_1)\pi_1(x_1)\,\mathrm{d}x_1.\end{aligned}$$

Next, using the properties of the cross-entropy distance, we have

$$\pi_1 = \operatorname*{arg\,min}_{g_1 \in \mathscr{G}_1} \int \pi_1(x_1) \ln\biggl(\frac{\pi_1(x_1)}{g_1(x_1)}\biggr)\,\mathrm{d}x_1 = \operatorname*{arg\,max}_{g_1 \in \mathscr{G}_1} \int \pi_1(x_1) \ln g_1(x_1)\,\mathrm{d}x_1.$$

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

Applying these two observations for any $i = 1, \dots, d$, we obtain

$$\begin{aligned}
\operatorname*{arg\,max}_{g_1,\dots,g_d \in \mathcal{G}_1} \int \pi(\boldsymbol{x}) \ln\Big(\prod_{i=1}^{d} g_i(x_i)\Big)\,\mathrm{d}\boldsymbol{x} &= \operatorname*{arg\,max}_{g_1,\dots,g_d \in \mathcal{G}_1} \sum_{i=1}^{d} \int \pi(\boldsymbol{x}) \ln g_i(x_i)\,\mathrm{d}\boldsymbol{x} \\
&= \operatorname*{arg\,max}_{g_1,\dots,g_d \in \mathcal{G}_1} \sum_{i=1}^{d} \int \pi_i(x_i) \ln g_i(x_i)\,\mathrm{d}x_i \\
&= \sum_{i=1}^{d} \operatorname*{arg\,max}_{g_i \in \mathcal{G}_1} \int \pi_i(x_i) \ln g_i(x_i)\,\mathrm{d}x_i,
\end{aligned}$$

from where we obtain the solution $g_i = \pi_i$ for all $i = 1, \dots, d$. □

Lemma 5. *Assume that* $\zeta \geq n\gamma^{-1}$. *Then*

$$\frac{\partial}{\partial \gamma} I_n(\gamma, \zeta) = nL(\gamma^{-1}) I_{n-1}(\gamma, \zeta - \gamma^{-1})\gamma^{-2}, \qquad n = 2, 3, \dots.$$

Proof. Recall the recursive property of the I_n functions, i.e.

$$\begin{aligned}
I_1(\gamma, \zeta) &= \int_{\zeta \vee \gamma^{-1}}^{1} L(y)\,\mathrm{d}y, \\
I_n(\gamma, \zeta) &= \int_{\gamma^{-1}}^{\zeta-(n-2)\gamma^{-1}} L(y) I_{n-1}(\gamma, \zeta - y)\,\mathrm{d}y, \quad n = 2, 3, \dots
\end{aligned}$$

The proof is by induction with respect to n, working out carefully the differentation. These are standard algebraic manipulations. □

Lemma 6. *For* $n = 1, 2, \dots$, *we have* $I_n(\gamma, \zeta - \gamma^{-1}) = I_n(\gamma, \zeta) + o(1)$ *as* $\gamma \to \infty$.

Proof. Apply induction and the recursive definition of I_n functions. For $n = 1$, we have

$$I_1(\gamma, \zeta - \gamma^{-1}) = \int_{\zeta-\gamma^{-1}}^{1} L(y)\,\mathrm{d}y = I_1(\gamma, \zeta) + \int_{\zeta-\gamma^{-1}}^{\zeta} L(y)\,\mathrm{d}y = I_1(\gamma, \zeta) + \gamma^{-1} L(\eta)$$

for some $\eta \in (\zeta - \gamma^{-1}, \zeta)$ (mean value theorem). Clearly, the second term is $o(1)$ for $\gamma \to \infty$. Now assume that the statement of Lemma 6 holds for $n \geq 1$. Then

$$\begin{aligned}
I_{n+1}(\gamma, \zeta - \gamma^{-1}) &= \int_{\gamma^{-1}}^{\zeta - n\gamma^{-1}} L(y) I_n(\gamma, \zeta - \gamma^{-1} - y)\,\mathrm{d}y \\
&= \int_{\gamma^{-1}}^{\zeta-(n-1)\gamma^{-1}} L(y)(I_n(\gamma, \zeta - y) + o(1))\,\mathrm{d}y \\
&\quad - \int_{\zeta - n\gamma^{-1}}^{\zeta-(n-1)\gamma^{-1}} L(y) I_n(\gamma, \zeta - \gamma^{-1} - y)\,\mathrm{d}y \\
&= I_{n+1}(\gamma, \zeta) + o(1) \int_{\gamma^{-1}}^{\zeta-(n-1)\gamma^{-1}} L(y)\,\mathrm{d}y - \gamma^{-1} L(\eta) I_n(\gamma, \zeta - \gamma^{-1} - \eta) \\
&= I_{n+1}(\gamma, \zeta) + o(1), \quad \gamma \to \infty.
\end{aligned}$$

□

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press

References

[1] Asmussen, S and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

[2] Asmussen, S. and Kroese, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Adv. Appl. Prob.* **38,** 545–558.

[3] Asmussen, S., Kroese, D. P. and Rubinstein, R. Y. (2005). Heavy tails, importance sampling and cross-entropy. *Stoch. Models* **21,** 57–76.

[4] Botev, Z. I. and Kroese, D. P. (2012). Efficient Monte Carlo simulation via the generalized splitting method. *Statist. Comput.* **22,** 1–16.

[5] Botev, Z. I., L'Ecuyer, P. and Tuffin, B. (2013). Markov chain importance sampling with applications to rare event probability estimation. *Statist. Comput.* **23,** 271–285.

[6] Chan, J. C. C., Glynn, P. W. and Kroese, D. P. (2011). A comparison of cross-entropy and variance minimization strategies. In *New Frontiers in Applied Probability* (J. Appl. Prob. Spec. Vol. **48A**), Applied Probability Trust, Sheffield, pp. 183–194.

[7] Dupuis, P., Leder, K. and Wang, H. (2007). Importance sampling for sums of random variables with regularly varying tails. *ACM TOMACS* **17,** 14.

[8] Embrechts, P. and Goldie, C. M. (1980). On closure and factorization properties of subexponential and related distributions. *J. Austral. Math. Soc. A* **29,** 243–256.

[9] Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events*. Springer, Berlin.

[10] Foss, S., Korshunov, D. and Zachary, S. (2011). *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer, New York.

[11] Ghamami, S. and Ross, S. M. (2012). Improving the Asmussen–Kroese-type simulation estimators. *J. Appl. Prob.* **49,** 1188–1193.

[12] Gudmundsson, T. and Hult, H. (2014). Markov chain Monte Carlo for computing rare-event probabilities for a heavy-tailed random walk. *J. Appl. Prob.* **51,** 359–376.

[13] Juneja, S. (2007). Estimating tail probabilities of heavy tailed distributions with asymptotically zero relative error. *Queueing Systems* **57,** 115–127.

[14] Perrakis, K., Ntzoufras, I. and Tsionas, E. G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance sampling. *Comput. Statist. Data Anal.* **77,** 54–69.

[15] Wong, R. (2001). *Asymptotic Approximation of Integrals* (Classics Appl. Math. **34**). Society for Industrial and Applied Mathematics, Philidelphia, PA.

https://doi.org/10.1017/jpr.2016.31 Published online by Cambridge University Press