HAL open science

Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization

Thomas Guilmeau, Nicola Branchini, Emilie Chouzenoux, Víctor Elvira

To cite this version:

Thomas Guilmeau, Nicola Branchini, Emilie Chouzenoux, Víctor Elvira. Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization. AISTATS 2024 - 27th International Conference on Artificial Intelligence and Statistics, May 2024, Valencia, Spain. hal-04506761

HAL Id: hal-04506761

https://hal.science/hal-04506761v1

Submitted on 15 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization

Thomas Guilmeau[⋆,◊] **Nicola Branchini**[†,◊] **Emilie Chouzenoux**[⋆] **Víctor Elvira**[†]
[⋆] Université Paris-Saclay, CentraleSupélec, INRIA, CVN, France,
[†] School of Mathematics, University of Edinburgh, United Kingdom.
[◊] equal contribution.

Abstract

Adaptive importance sampling (AIS) algorithms are widely used to approximate moments of target probability distributions. When the target has heavy tails, existing AIS algorithms can provide inconsistent estimators or exhibit slow convergence, as they often neglect the target's tail behaviour. To avoid this pitfall, we propose an AIS algorithm that approximates the target by Student-t proposal distributions. We adapt location and scale parameters by matching the *escort* moments (defined even for heavy-tailed distributions) of the target and proposal. The resulting updates minimize the α-divergence between the target and the proposal, thereby connecting with variational inference methods. We then show that the α-divergence can be approximated by a generalized notion of effective sample size. We leverage this new perspective to adapt the proposal tail parameter using Bayesian optimization. We demonstrate the efficacy of our approach through applications to synthetic targets and a Bayesian Student-t regression task on real clinical trial data.

1 INTRODUCTION

Expectations that are challenging to compute arise repeatedly in probabilistic machine learning (Ghahramani, 2015), Bayesian statistics (Robert et al., 2007), statistical signal processing (Särkkä and Svensson, 2023), option pricing in mathematical finance (L'Ecuyer, 2004), and many other fields where Monte Carlo methods are often the de-facto standard. Importance sampling (IS) generalizes the Monte Carlo integration principle to approximate expectations with respect to a target distribution π (Robert and Casella, 1999; Owen, 2013; Kroese et al., 2014). In IS, samples are obtained from a distribution q called proposal that is not necessarily equal to π.

Constructing an adequate proposal q is difficult yet crucial for the performance of IS. Adaptive IS (AIS) algorithms, which iteratively refine the proposal distributions, have become the standard to construct efficient samplers (Bugallo et al., 2017) and have also shown connections with particle filtering (Branchini and Elvira, 2024). AIS proposal adaptation procedures can be based on moment matching (Cornuet et al., 2012), gradient updates (Elvira et al., 2015, 2023; Elvira and Chouzenoux, 2022), or combined with Markov Chain Monte Carlo (Botev et al., 2013; Martino et al., 2017b; Thin et al., 2021).

Several recent works have also highlighted connections between AIS and variational inference (VI) (Mnih and Rezende, 2016; Sakaya and Klami, 2017; Domke and Sheldon, 2018; Finke and Thiery, 2019; Dhaka et al., 2021; Zhang et al., 2022; Mattei and Frellsen, 2022; Kviman et al., 2022; Doucet et al., 2023), a framework popular in Bayesian statistics, machine learning and signal/image processing (Jordan et al., 1999; Blei et al., 2017; Marnissi et al., 2017). Indeed, VI methods aim at approximating a target π with a distribution q, by explicitly minimizing a statistical divergence, typically the Kullback-Leibler (KL) divergence. In IS and AIS, the most widely used criterion to evaluate performance is the effective sample size (ESS), which has some connections with a statistical divergence.

In this paper, we focus on a class of AIS procedures based on moment matching, in particular on the AMIS framework of Cornuet et al. (2012) that is behind recent state-of-the-art AIS algorithms (Paananen et al., 2021). Although popular, moment-matching updates can be ill-defined when the target or the proposal is heavy-tailed with undefined moments. Notable ap-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

plications with heavy-tailed π include: Student-t error models in Bayesian regression, realistic posterior distributions that are robust to outliers or promoting sparse solutions (Fernández and Steel, 1998; Tipping and Lawrence, 2005; Amrouche et al., 2022); applied econometrics, where parameter estimation for stochastic volatility models of option pricing involves complicated heavy-tailed distributions (Chib et al., 2002); analysing financial returns datasets (Roy and Hobert, 2010). Similarly, heavy-tailed proposals q can be beneficial in AIS (Owen, 2013, Chapter 9), although they may not have finite moments thus preventing the application of existing moment-matching methods.

Contributions. **(1)** We propose an AIS framework, hereby named `AHTIS` (adaptive heavy-tailed importance sampling), allowing heavy-tailed target and proposal distributions. Its proposal adaptation mechanism is based on matching the moments of *escort* densities associated to the target and proposal, i.e., versions of the density with lighter tails. **(2)** We show that our proposed moment matching corresponds to the minimization of an α-divergence. Our approach generalizes previous AIS moment-matching procedures restricted to the KL divergence. **(3)** We show that a generalized notion of effective sample size, the α-ESS, is an IS approximation of the α-divergence, providing new connections between VI and AIS. **(4)** Finally, we exploit this insight to design a joint adaptation strategy for the tail and the location/scale parameters of the proposals using Bayesian optimization, outperforming existing moment-matching AIS both when a good tail parameter is known in advance and when it needs to be adapted. This advantage of `AHTIS` is illustrated in Fig. 1.

Outline. In Section 2, we introduce the notion of escort probability and our approximating family. In Section 3, we introduce our AIS algorithm, `AHTIS`, with adaptation of the location, scale, and tail parameters of its proposal. Finally, we show the performance of `AHTIS` on heavy-tailed target distributions in Section 5, before concluding in Section 6.

2 BACKGROUND

2.1 Importance Sampling

Importance sampling allows for the Monte Carlo integration of integrals of the form $I = \int h(x)\pi(x)dx$ when samples from the target density π are either unavailable, or even inefficient (such as in rare events). Instead, one samples from a proposal distribution q and uses so-called importance weights to correct the estimation. The simplest IS estimator of I is the un-

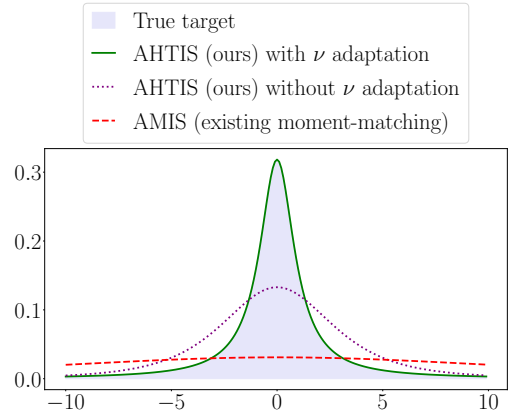


Figure 1: In this illustrative example, the target π is a Student-t distribution with $\nu_\pi = 1$ degrees of freedom (a Cauchy distribution), which is very heavy-tailed and has undefined mean and variance. We show three Student-t approximations using **(i)** existing moment matching with $\nu = 3$ degrees of freedom ($\nu > 2$ is required for the proposals to have moments), **(ii)** `AHTIS` with $\nu = 3$ degrees of freedom, and **(iii)** `AHTIS` with degrees of freedom adaptation.

normalized IS estimator (UIS), given by

$$\widehat{I}_{\text{UIS}} = \frac{1}{M}\sum_{m=1}^{M} w^{(m)} h(x^{(m)}), \quad \{x^{(m)}\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} q, \quad (1)$$

where $w^{(m)} = \pi(x^{(m)})/q(x^{(m)})$ are the (unnormalized) importance weights using the target probability density function (pdf), $\pi(x)$. When $q = \pi$, $\widehat{I}_{\text{UIS}}$ recovers the plain Monte Carlo estimator, $\widehat{I}_{\text{MC}}$. In many cases, we only have access to the unnormalized target density $\widetilde{\pi}(x) = \pi(x)Z_\pi$, i.e., the normalizing constant Z_π is unknown. The standard estimator for Z_π is

$$Z_\pi \approx \widehat{Z}_\pi = \frac{1}{M}\sum_{m=1}^{M} \frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})} = \frac{1}{M}\sum_{m=1}^{M} \widetilde{w}^{(m)}. \quad (2)$$

Eq. (2) allows one to estimate I when Z_π is unknown, leading to the self-normalized IS (SNIS) estimator

$$I \approx \widehat{I}_{\text{SNIS}} = \sum_{m=1}^{M} \overline{w}^{(m)} h(x^{(m)}), \quad (3)$$

where $\overline{w}^{(m)} = \widetilde{w}^{(m)} / \sum_{\ell=1}^{M} \widetilde{w}^{(\ell)}$. The almost sure convergence $\widehat{I}_{\text{SNIS}} \xrightarrow[M\to+\infty]{a.s.} I$ is guaranteed as soon as $\pi(x) > 0 \Rightarrow q(x) > 0$ (Owen, 2013, Chapter 4).

Assessing IS performance. The mean-squared error (MSE) is a common way to evaluate the performance of estimators (Owen, 2013) and, for both $\widehat{I}_{\text{UIS}}$ and $\widehat{I}_{\text{SNIS}}$, the MSE decays at the standard

Monte Carlo rate $\Theta(1/M)$. See, e.g., (Chopin and Papaspiliopoulos, 2020, Chapter 8) for more theoretical properties of IS estimators. However, it is difficult to devise good estimators of the MSE. The effective sample size (ESS) is a more practical, and widely used, metric to assess the quality of IS estimators. It is a sample approximation of the ratio of variances between the SNIS estimator and a Monte Carlo estimator with π (Kong, 1992; Elvira et al., 2022), computed as

$$\widehat{\mathrm{ESS}} = \frac{1}{\sum_{m=1}^{M} \left(\overline{w}^{(m)}\right)^2} \approx \mathrm{ESS} = \frac{\mathbb{V}_q[\widehat{I}_{\mathrm{SNIS}}]}{\mathbb{V}_\pi[\widehat{I}_{\mathrm{MC}}]}. \quad (4)$$

While the original motivation is the above approximation of a ratio of variances, $\widehat{\mathrm{ESS}}$ has been shown to be connected with the Pearson chi-squared divergence $\chi^2(\pi, q)$ (Agapiou et al., 2017; Sanz-Alonso, 2018; Sanz-Alonso and Wang, 2020; Agarwal et al., 2022; Elvira et al., 2022), an information-theoretic measure that plays a key role in the theory of IS (Orsak and Aazhang, 1991; Akyildiz and Míguez, 2021). The choice of proposal q is crucial to achieve good performance in the above metrics, which led to the development of adaptive IS algorithms (AIS), where proposals are iteratively adapted (Bugallo et al., 2017).

Adaptive multiple IS (AMIS). AIS algorithms recycle samples to improve the quality of $\widehat{I}_{\mathrm{SNIS}}$. Suppose we have T proposals $\{q_t\}_{t=1}^T$ and for every $t \in \{1, \dots, T\}$ the samples are $\{x_t^{(m)}\}_{m=1}^M$. One way to re-use all the TM samples is to assign to each of them an unnormalized weight $\widetilde{w}_t^{(m)} = \widetilde{\pi}(x_t^{(m)})/q_t(x_t^{(m)})$, and possibly perform a resampling step. It has been shown that an alternative weighting, i.e., deterministic mixture (DM) weighting, achieves better results by considering *all* the proposals in the weighting of each sample (Elvira et al., 2019). The unnormalized DM weight of the sample $x_t^{(m)}$ reads

$$\widetilde{w}_t^{(m)} = \frac{\widetilde{\pi}(x_t^{(m)})}{\frac{1}{T}\sum_{\tau=1}^{T} q_\tau(x_t^{(m)})}. \quad (5)$$

The DM weighting is notably used by the adaptive multiple importance sampling (AMIS) algorithm proposed in Cornuet et al. (2012), where at each iteration, the proposal is adapted using all the past samples using DM weights. Cornuet et al. (2012) suggest to use the DM weights to update the proposal such that its moments match the (estimated) moments of π. Consistent estimators can be constructed by combining weighted samples from all iterations $t = 1, \dots, T$ (Marin et al., 2019).

2.2 Escort Distributions and α-Divergence Minimization

We introduce now existing results about the minimization of statistical divergences over Student-t distributions which we will use to develop our new method.

Definition 1 (Multivariate Student-t)**.** *The multivariate Student-t distribution on $\mathbb{R}^d$ with $\nu > 0$ degrees of freedom, location parameter $\mu \in \mathbb{R}^d$, and positive-definite scale matrix $\Sigma \in \mathcal{S}_{++}^d$ has a pdf with respect to the Lebesgue measure of the form*

$$q_{\mu,\Sigma,\nu}(x) \propto \left(1 + \frac{1}{\nu}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}} \quad (6)$$

and is normalized by $Z_{\nu,\Sigma} = \frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+d}{2})}\left(\nu^d \pi^d \det(\Sigma)\right)^{\frac{1}{2}}$.

Student-t distributions recover Cauchy distributions when $\nu = 1$ and Gaussian distributions in the limit $\nu \to +\infty$. They have finite first moment for $\nu > 1$ and finite second moment for $\nu > 2$. Next, we introduce the concepts of escort distribution and α-divergence, which will be used throughout Section 3.

Definition 2 (Escort version of a pdf)**.** *Consider $\alpha > 0$ and a pdf p. The* escort *version of p (Tsallis, 2009) with exponent α is the pdf $p^{(\alpha)}$ defined by*

$$p^{(\alpha)}(x) = \frac{p(x)^\alpha}{\int p(x)^\alpha dx}, \quad (7)$$

assuming that the normalizing constant is finite.

Definition 3 (α-divergence)**.** *The α-divergence is defined for $\alpha > 0$ and $\alpha \neq 1$ as*

$$\mathrm{D}_\alpha(\pi, q) = \frac{1}{\alpha(\alpha-1)}\left(\int \pi(x)^\alpha q(x)^{1-\alpha} dx - 1\right). \quad (8)$$

Its discrete counterpart $\mathrm{D}_\alpha^M(\cdot,\cdot)$ is defined similarly on the simplex of $\mathbb{R}^M$, denoted by Δ_M.

The α-divergence generalizes many well-known divergences such as $\mathrm{KL}(\pi, q)$ $(\alpha \to 1)$ and $\chi^2(\pi, q)(\alpha = 2)$. The KL divergence is such that $\theta \longmapsto \mathrm{KL}(\pi, q_\theta)$ is minimized under a moment-matching property when the pdf q_θ form an exponential family (Bishop, 2006, Equation (10.187)). This is the case of Gaussian distributions and hence, the optimal KL approximation of π is the Gaussian pdf with same first and second order moments as π. The above result has been generalized beyond this setting under a specific relationship between the parameter α of the divergence and the degree of freedom parameter ν of the Student-t family, as the next result shows.

Proposition 1. *(Guilmeau et al., 2023) Consider a target pdf π and the family of Student-t distributions*

with $\nu > 0$ degrees of freedom. If the escort pdf $\pi^{(\alpha)}$ with $\alpha = 1 + \frac{2}{\nu+d}$ exists and has finite first and second-order moments, then the parameters $(\mu_\nu^\star, \Sigma_\nu^\star)$ such that

$$\begin{cases} \mu_\nu^\star = \int x\, \pi^{(\alpha)}(x) dx \\ \Sigma_\nu^\star = \int x x^\top \pi^{(\alpha)}(x) dx - \mu_\nu^\star \mu_\nu^{\star\top} \end{cases} \tag{9}$$

minimize $(\mu, \Sigma) \longmapsto \mathrm{D}_\alpha(\pi, q_{\mu,\Sigma,\nu})$.

3 ADAPTIVE HEAVY-TAILED IMPORTANCE SAMPLING

We now present our proposed AIS framework, `AHTIS`, for handling target distributions with heavy tails and potentially undefined moments based on α-divergence minimization. Our framework is summarized in Algorithm 1, which we describe next. In Section 3.2, we show that the so-called α-ESS can be used to approximate the α-divergence. We exploit this insight to propose our tail parameter ν adaptation in Section 3.3.

3.1 Step-by-step Breakdown of `AHTIS` and Justification.

First, as input to Algorithm 1 we require initial location, scale, and tail parameters for the proposal, i.e., (μ_0, Σ_0, ν_0) respectively. The algorithm follows the following steps for $T > 0$ iterations. First, we generate samples from q_{μ_t,ν_t,Σ_t} (**step 2**). Then, tail adaptation (**step 3**) finds ν_{t+1} (and therefore α_{t+1}) with Bayesian optimization (BO), which we detail in Section 3.3. This step can be performed in an automatic way using existing implementations (see Appendix D for more details). The weighting (**step 4**) uses the DM approach described in Section 2 allowing the proposal to learn from *all* the generated samples. Note that the numerator involves the escort version of the target, $\pi^{(\alpha_{t+1})}$. Notably, this means that when the variance of the weight with respect to the true target π is infinite (as it would be the case for existing AIS algorithms, and is common), since $\alpha_t > 1$, the variance of Eq. (10) may still be finite. Finally, the escort moment-matching (**step 5**) minimizes $(\mu, \Sigma) \longmapsto \mathrm{D}_{\alpha_{t+1}}(\pi, q_{\mu,\Sigma,\nu_{t+1}})$ as explained in Section 2. `AHTIS` is motivated by the minimization of the α-divergence $\mathrm{D}_\alpha(\pi, q)$ between target and proposal, which is known to exhibit favourable properties for heavy-tailed distributions (Birrell et al., 2021), as well as for robust approximate inference with generalized VI on misspecified models in Bayesian statistics (Knoblauch et al., 2022; Boustati et al., 2020). More precisely, Algorithm 1 addresses the following joint optimization problem involving (μ, Σ, ν),

$$\mu^\star, \Sigma^\star, \nu^\star = \arg\min_{\mu,\Sigma,\nu} \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\Sigma,\nu}). \tag{13}$$

Algorithm 1 `AHTIS`

Require: $\nu_0 > 0$, $\mu_0 \in \mathbb{R}^d$, $\Sigma_0 \in \mathcal{S}_{++}^d$

1: **for** $t = 0, ..., T$ **do**

2: **Sampling:** $\{x_t^{(m)}\}_{m=1}^M \overset{\text{i.i.d.}}{\sim} q_{\mu_t,\Sigma_t,\nu_t}$.

3: **Tail adaptation with BO:**
- If $t = 0$, $\nu_1 = \nu_0$, else, set ν_{t+1} with Algorithm D.1 in Appendix D
- Set $\alpha_{t+1} = 1 + \frac{2}{\nu_{t+1}+d}$.

4: **Temporal DM weighting:** For $m = 1, \ldots, M$ and $\tau = 0, \ldots, t$, compute the unnormalized importance weights using the unnormalized escort target as

$$\widetilde{w}_\tau^{(m)} = \frac{\left(\widetilde{\pi}(x_\tau^{(m)})\right)^{\alpha_{t+1}}}{\frac{1}{t+1}\sum_{k=0}^t q_{\mu_k,\Sigma_k,\nu_k}(x_\tau^{(m)})} \tag{10}$$

and normalize to obtain $\overline{w}_\tau^{(m)} = \widetilde{w}_\tau^{(m)} / \sum_{\ell=1}^M \widetilde{w}_\tau^{(\ell)}$.

5: **Escort moment matching:** Set $(\mu_{t+1}, \Sigma_{t+1})$ with the updates

$$\mu_{t+1} = \sum_{\tau=0}^t \sum_{m=1}^M \overline{w}_\tau^{(m)} x_\tau^{(m)} \tag{11}$$

$$\Sigma_{t+1} = \sum_{\tau=0}^t \sum_{m=1}^M \overline{w}_\tau^{(m)} x_\tau^{(m)} x_\tau^{(m)\top} - \mu_{t+1}\mu_{t+1}^\top \tag{12}$$

6: **end for**

7: **Return:** $\{\overline{w}_t^{(m)}, x_t^{(m)}\}_{t=1,m=1}^{T,M}$, $\{\mu_t, \Sigma_t\}_{t=1}^T$

Recall from Section 2 that the value $\alpha(\nu)$ in (13) is such that $\alpha(\nu) = 1 + \frac{2}{\nu+d}$, where d is the dimension of x. Hence, we are not minimizing a fixed α-divergence, rather jointly adapting the α-divergence parameter and the approximating family's degree of freedom parameter ν. We now establish in Proposition 2 that when π is a Student-t distribution, the optimization problem in Eq. (13) is solved when the proposal recovers π, illustrating the rationale of our approach. However, we remark that our algorithm `AHTIS` is not restricted to Student-t targets.

Proposition 2 (Well-posedness of tail-adaptation)**.** *Suppose that the target π is a Student-t pdf with $\nu_\pi > 0$ degrees of freedom. Then, Problem (13) is solved by $(\mu^\star, \Sigma^\star, \nu^\star)$ such that $\nu^\star = \nu_\pi$ and $q_{\mu^\star,\Sigma^\star,\nu^\star} = \pi$.*

The proof is postponed to Appendix C. To obtain a practical algorithm to minimize the problem in Eq. (13), we propose to consider (μ, Σ) and ν separately, and equivalently reformulate Eq. (13) as

$$\nu^\star = \arg\min_\nu \min_{\mu,\Sigma} \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\Sigma,\nu}). \tag{14}$$

This is motivated by the fact that for a given $\nu > 0$, $\min_{\mu,\Sigma} \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\Sigma,\nu}) = \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu^\star_\nu,\Sigma^\star_\nu,\nu})$, with $(\mu^\star_\nu, \Sigma^\star_\nu)$ satisfying Eq. (9). The behaviour of $\nu \longmapsto \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu^\star_\nu,\Sigma^\star_\nu,\nu})$ is illustrated in Fig. 2.

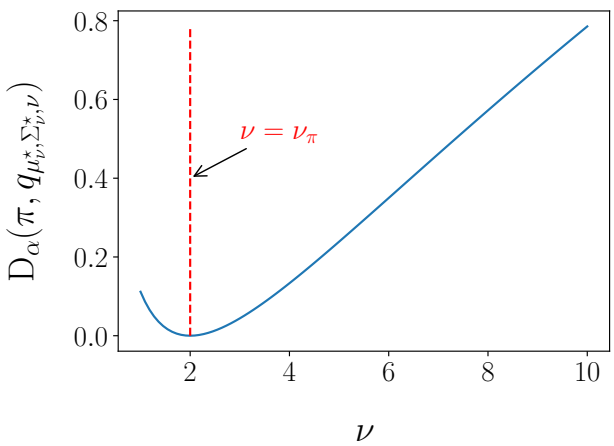


Figure 2: Optimal α-divergence value as a function of $\nu > 0$ from Proposition 1 when π is a Student-t distribution in dimension $d = 5$ and degree of freedom parameter $\nu_\pi = 2$ (vertical dotted red line).

Next, we propose an approach to solve this optimization problem within **step (3)** of Algorithm 1 without using any additional samples. This requires to evaluate the objective in Eq. (14), which we address now.

3.2 Connecting VI and IS with the α-ESS

A challenge is that, for realistic target distributions π, one cannot evaluate the cost function $\mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\nu,\Sigma})$ appearing in the minimization problem of Eq. (14). We now show that a SNIS approximation of $\mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\nu,\Sigma})$ is related in a precise way to an existing generalized ESS, the α-ESS, which belongs to the Huggins-Roy family of ESS metrics (Martino et al., 2017a; Huggins and Roy, 2019). This result connects further VI and IS and allows us to obtain a practical way to approximate $\mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\nu,\Sigma})$, that we will use to adapt the tail parameter ν in Section 3.3. The α-ESS is defined over the simplex Δ_M as:

$$\widehat{\mathrm{ESS}}_\alpha(\overline{w}) = \left(\sum_{m=1}^{M} \left(\overline{w}^{(m)}\right)^\alpha \right)^{\frac{1}{1-\alpha}}, \forall \overline{w} \in \Delta_M. \quad (15)$$

We now show our main result connecting $\widehat{\mathrm{ESS}}_\alpha$ and $\mathrm{D}_\alpha(\pi, q)$ for general target and proposal distributions.

Proposition 3 (Almost sure convergence)**.** *Consider a target π and a proposal q with normalized importance weights $\{\overline{w}^{(m)}\}_{m=1}^M$ associated with i.i.d. samples from q. Then, the discrete α-divergence between the weights $\{\overline{w}^{(m)}\}_{m=1}^M$ and the uniform weights $\{1/M\}_{m=1}^M$ is related to $\widehat{\mathrm{ESS}}_\alpha$ as follows:*

$$\mathrm{D}_\alpha^M(\{\overline{w}^{(m)}\}_{m=1}^M, \{1/M\}_{m=1}^M) = \frac{M^{\alpha-1}}{\alpha(\alpha-1)} \left(\widehat{\mathrm{ESS}}_\alpha(\{\overline{w}^{(m)}\}_{m=1}^M)^{1-\alpha} - M^{1-\alpha} \right). \quad (16)$$

Moreover, D_α^M converges to $\mathrm{D}_\alpha(\pi, q)$, i.e.,

$$\mathrm{D}_\alpha^M(\{\overline{w}^{(m)}\}_{m=1}^M, \{1/M\}_{m=1}^M) \xrightarrow[M\to+\infty]{a.s.} \mathrm{D}_\alpha(\pi, q) \quad (17)$$

in an almost sure sense when $\pi(x) > 0 \Rightarrow q(x) > 0$.

The proof is provided in Appendix C. The quantity $\widehat{\mathrm{ESS}}_\alpha$ can be cheaply computed. Further, since our derivation shows that $\widehat{\mathrm{ESS}}_\alpha$ is specifically a SNIS estimator, we obtain a central limit theorem (CLT) by extending standard SNIS results (Chopin and Papaspiliopoulos, 2020), which allows to quantify uncertainty using asymptotic confidence intervals.

Proposition 4 (CLT)**.** *If $\pi(x) > 0 \Rightarrow q(x) > 0$ and $\mathbb{V}_q\left[\left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^\alpha\right] < +\infty$, the estimator D_α^M of the α-divergence is $\sqrt{M}$-asymptotically normal, i.e.,*

$$\sqrt{M}\left(\mathrm{D}_\alpha^M(\{\overline{w}^{(m)}\}_{m=1}^M, \{1/M\}_{m=1}^M) - \mathrm{D}_\alpha(\pi, q)\right) \xrightarrow[N\to+\infty]{d} \mathcal{N}\left(0, \sigma^2\right), \quad (18)$$

with variance

$$\sigma^2 = \left(\frac{\int \widetilde{\pi}(x)^{2\alpha} q(x)^{1-2\alpha} dx}{\left(\alpha(\alpha-1)\int \widetilde{\pi}(x)^\alpha q(x)^{1-\alpha} dx\right)^2} - 1 \right). \quad (19)$$

See Appendix C for a proof. Next, we detail **step 3** of Algorithm 1, which relies on $\widehat{\mathrm{ESS}}_\alpha$.

3.3 Tail Adaptation with Bayesian Optimization

We now describe how to adapt without generating additional samples the parameter ν within the optimization problem in Eq. (13) (the procedure is further detailed in Appendix D).

The outer problem on ν consists in minimizing the function $\nu \longmapsto \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu^\star_\nu,\Sigma^\star_\nu,\nu})$, with $(\mu^\star_\nu, \Sigma^\star_\nu)$ satisfying Eq. (9). Although one-dimensional, this problem is difficult as it involves intractable integrals and inner optimization. We propose a Bayesian optimization (BO) approach (Garnett, 2023). BO algorithms do not require the computations of derivatives and can cope with noisy estimations of the objective function. Further, they only require a small number of these noisy evaluations, which fits well within our context, since in AMIS (Cornuet et al., 2012), the value of T does not need to be large (see Section 5 for details).

To solve (14) with BO, the main challenge is to approximate at every iteration $t = 1, \ldots, T$ the quantity $\mathrm{D}_{\alpha(\nu_t)}(\pi, q_{\mu^\star_\nu, \Sigma^\star_\nu, \nu_t})$. To do so, we first remark that

$$(\mu_t, \Sigma_t) \approx \arg\min_{\mu,\Sigma} \mathrm{D}_{\alpha_t}(\pi, q_{\mu,\Sigma,\nu_t}), \tag{20}$$

in the sense that (μ_t, Σ_t) are constructed following (11)-(12) which are estimators of the optimality conditions (9). Then, the quantity $\mathrm{D}_{\alpha_t}(\pi, q_{\mu^\star_{\nu_t}, \Sigma^\star_{\nu_t}, \nu_t})$ is approximated by computing the α_t-ESS with target π and proposal q_{μ_t,Σ_t,ν_t}, following our Proposition 3.

BO algorithms construct a probabilistic model of the function $\nu \longmapsto \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu^\star_\nu, \Sigma^\star_\nu, \nu})$ in the form of a Gaussian process (GP). At every iteration, the GP is updated with the data $\{\nu_\tau, \alpha_\tau\text{-ESS}\}_{\tau=1}^t$, where the values α_τ-ESS are seen as noisy observations of the α_τ-divergence. Then, an acquisition function, which governs the trade-off between exploration and exploitation, is maximized, yielding the next value ν_{t+1}. We use Upper Confidence Bound (UCB) as the acquisition function, which offers theoretical guarantees on cumulative regret by balancing exploration and exploitation with a logarithmic regret bound (Garnett, 2023, Chapter 10). As kernel for the GP, we use a standard radial-basis function (RBF) kernel with default parameters. For more details on the BO procedure, see Appendix D.

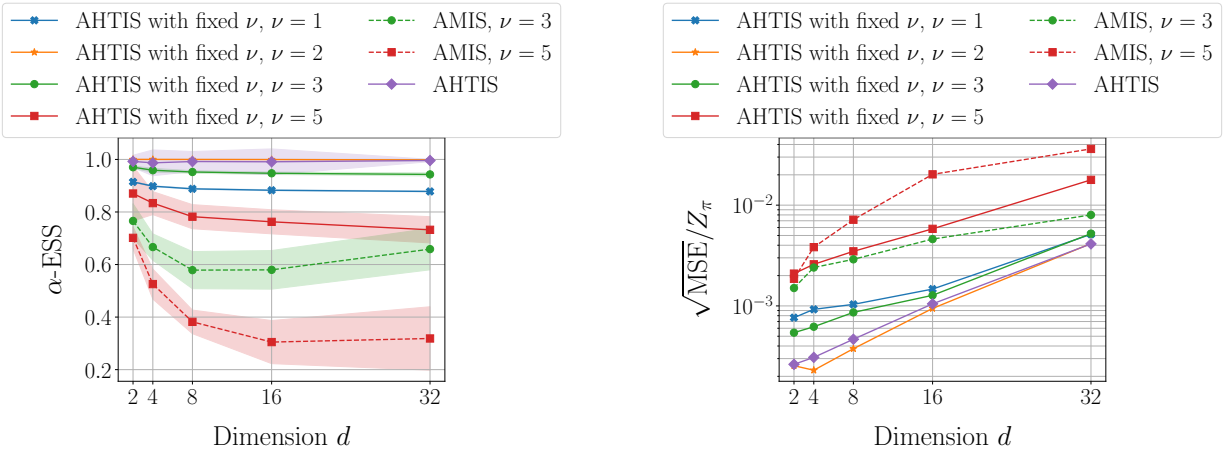
Computational complexity of `AHTIS`. The complexity of Algorithm 1 can be analysed by the one of AMIS, and the added complexity given by **step (3)**, the tail adaptation based on BO. Like AMIS, the runtime of `AHTIS` (in terms of number of proposal evaluations) is $\Theta(MT^2)$ due to the use of deterministic mixture weighting. The values of M and T are fixed and their influence on the final error is not well-understood in the AIS literature. We find consistent results with the original AMIS paper (Cornuet et al., 2012) where T does not need to be very large (between 20 and 30 in both our examples and theirs) while M is sufficiently large. This implies that the BO procedure (see Appendix D) is usually not expensive in practice, in our experiments, even if cubic in T in theory since the GP is fitted on T examples. Recall that the dimension of our BO problem is fixed to 1, since ν is a scalar. To summarize, the complexity of BO is driven by **(i)** sequentially updating the GP and **(ii)** maximizing the acquisition function. Many works and active research in the BO literature aim to reduce these costs, see e.g. (Garnett, 2023, Chapters 9.1, 9.2). In our case, UCB is one of the cheapest acquisitions to maximize (Wilson et al., 2018). Finally, previous work has also managed to reduce AMIS complexity to $\Theta(MTK)$ (for some constant $K < T$) with good tradeoffs in estimator variance (El-Laham et al., 2019) whose techniques also straightforwardly apply to `AHTIS`.

4 RELATED WORKS

In general, AIS methods do not specifically handle heavy-tailed targets with undefined moments. Although some works use heavy-tailed proposals, to the best of our knowledge, no existing AIS work adapts the tail parameter of a heavy-tailed proposal as in Algorithm 1, while some works in VI do so.

AIS. Wang and Swartz (2022) in the context of AIS match the first three moments of skew-Student proposals with the target's moments for adaptation, without adapting ν, requiring $\nu > 3$, and with no connection with α-divergences. Korba and Portier (2022) introduce an AIS scheme using a mixture of an iteratively adapted kernel density estimator and a safe heavy-tailed distribution, however without detailing the latter's construction. Other AIS works using moment matching mention the use of Student-t distributions, but do not adapt the tail parameter ν (Cornuet et al., 2012; Portier and Delyon, 2018). El-Laham et al. (2020) minimize Renyi divergence, which is related to the α divergence, but incorporate the use of Markov chain MC steps.

VI. Daudel et al. (2023) propose a general VI framework that allows in particular to minimize a fixed α-divergence over a mixture of Student-t distributions. The location, scale, and tail parameters of the Student-t distributions are adapted. While we adapt ν using a BO algorithm, they do so by solving a non-linear equation. However, their procedure may not be able to reach low value of ν, contrary to ours (see Appendix D.2 for a justification), and they did not implement a practical scheme showing experimental results . The work of Wang et al. (2018) proposes to minimize an f-divergence that is implicitly defined at each iteration by the importance weights of the samples. This is connected with the dependence of the α-divergence we minimize on the degree of freedom parameter. However, their goal diverges from ours by focusing on obtaining mass-covering proposals. The minimization of an α-divergence (or a Rényi divergence) is also considered in (Hernandez-Lobato et al., 2016; Li and Turner, 2016). In these works, the resulting optimization problem is solved by stochastic gradient descent on a general proposal family, while here we exploit the Student-t assumption to obtain direct optimality conditions. Further, note that VI methods **(i)** do not use recycling of past samples, **(ii)** usually yield only a lower bound of Z_π. This is in contrast with the AIS literature, where samples recycling strategies such as DM weighting have been used (Marin et al., 2019), allowing to construct $\mathcal{O}(1/M)$-consistent estimates of Z_π.



(a) α-ESS (mean $\pm$ one standard deviation, higher is better) for various dimensions d. `AHTIS` outperforms AMIS for any ν, sometimes by an order of magnitude, and the ν-adaptive version converges to the true value $\nu_\pi = 2$.

(b) Relative square root MSE (lower is better) for various dimensions d. Note that $Z_\pi = Z_{\nu_\pi,\Sigma_\pi}$ is the true normalizing constant, which is available. `AHTIS` outperforms AMIS for any ν and the ν-adaptive version converges to $\nu_\pi = 2$.

Figure 3: Results for Section 5.1. All algorithms are run for $T = 20$ iterations, with $M = 10^4$ samples per iteration and results are averaged over 100 replications. A dashed line identifies AMIS, while solid line is `AHTIS`, and same marker/color indicates same ν. Recall that AMIS updates are not defined for $\nu \in \{1, 2\}$ as the proposal moments are undefined.

5 EXPERIMENTS

We demonstrate the benefits of `AHTIS` first on a controlled scenario with synthetic heavy-tailed targets (Student-t distributions of varying dimensions), second on a posterior distribution arising from a Bayesian robust regression problem on clinical trial data. Code for reproducibility and additional results are publicly available at https://github.com/nicola144/ais-heavy-tails.

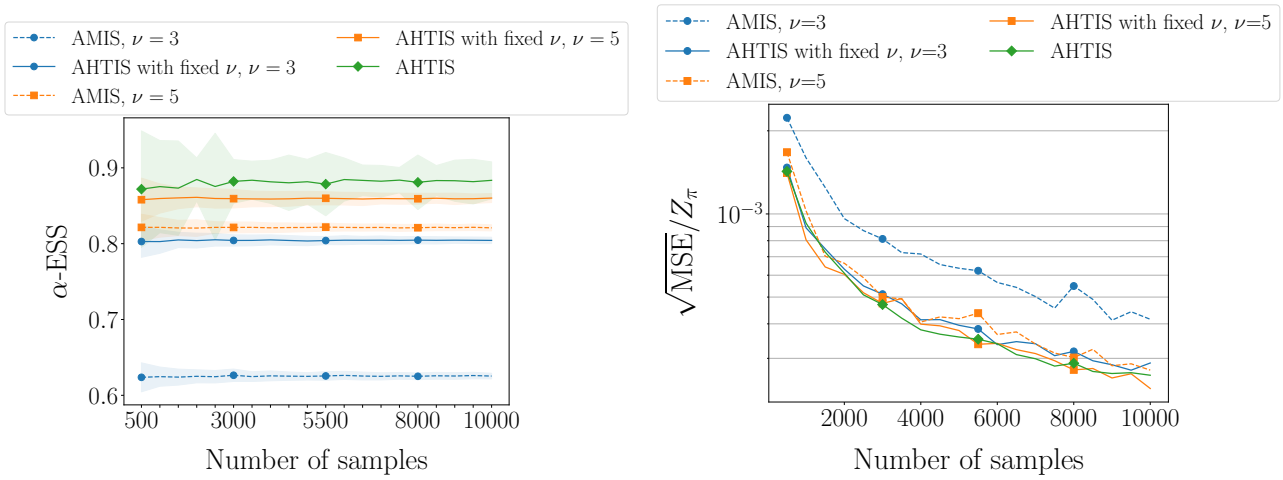
We evaluate the algorithms using the α-ESS metric, shown in Section 3.2 to be a theoretically sounded approximation of D_α, and the MSE on the estimation of the normalizing constant Z_π, a key distinguishing feature of (A)IS algorithms (Llorente et al., 2023).

5.1 Controlled Scenario with Varying Dimension Student-t Targets

We start with the problem of approximating integrals involving a heavy-tailed Student-t target π with $\nu_\pi \in \{2, 5\}$. Note that the second-order moments of π are not defined when $\nu_\pi = 2$. The sought target has a location parameter sampled in $\mathrm{Uniform}[-1, 1]^d$. Moreover, its scale matrix Σ_π is built so as to reach a condition number $\kappa = 5$, following (Moré and Toraldo, 1989, Sec. 5). We consider dimensions $d \in \{2, 4, 8, 16, 32\}$.

We run `AHTIS` and AMIS algorithms for $T = 20$ iterations, with $M = 10^4$ samples per iteration, following the guidelines from (Cornuet et al., 2012). In the spirit of an ablation study, we analyze the benefits of the tail adaptation in `AHTIS`. That is, we also run `AHTIS` without **step (3)** of Algorithm 1, ν being fixed and possibly different from ν_π. All algorithms are initialized with μ_0 sampled in $\mathrm{Uniform}[-5, 5]^d$ and $\Sigma_0 = 10 I_d$. For `AHTIS` with **step (3)**, the value $\nu_0 = 1$ is used. Else, the degrees of freedom $\nu \in \{1, 2, 3, 5\}$ are considered for the algorithms without tail adaptation. Note that in the case $\nu \leq 2$ the updates of AMIS are not defined.

Results. The results in terms of the considered metrics are shown in Fig. 3a-3b. The best performance in both metrics are reached by the ν-adaptive `AHTIS` and by `AHTIS` with $\nu = \nu_\pi$. This shows that the ν-adaptive `AHTIS` is able to capture the tail behaviour of the target and confirms the result of Proposition 2. When ν is fixed, `AHTIS` outperforms AMIS in both metrics when $\nu > 2$, and allows in addition to use heavy-tailed proposals with $\nu \leq 2$. Such proposals yield better performance on this heavy-tailed target. We show additional results in Appendix E, including the case $\nu_\pi = 5$ revealing similar behaviours, and $\nu_\pi = 50$ where the target has lighter tails than the considered proposals. We also provide an analysis of the adaptation of ν of `AHTIS` on these different targets, as well as bootstrap confidence intervals of the estimators of Z_π returned by the algorithms.



(a) α-ESS (mean $\pm$ one standard deviation, higher is better). `AHTIS` outperforms AMIS with fixed ν. The ν-adaptive `AHTIS` yields better mean, but exhibits a larger variance.

(b) In terms of relative square root MSE (lower is better), `AHTIS` outperforms AMIS for any ν and in this case `AHTIS` reaches the best performance.

Figure 4: Results for the creatinine dataset, Section 5.2. All algorithms are run for $T = 25$ iterations, with varying number of samples and results are averaged over 250 replications. A dashed line identifies AMIS, while solid line is `AHTIS`, and same marker/color indicates same ν.

5.2 Application to Bayesian Student-t regression on real data

We apply our methodology using, as the target π, the posterior resulting from a robust regression model on the creatinine dataset (Liu and Rubin, 1995).[1] This dataset has been used to benchmark state-of-the-art VI and MCMC algorithms (Xu et al., 2023). It contains the results of a clinical trial on $N = 34$ male patients. Such a small number of datapoints makes the inference task challenging since the target is likely outside the Bernstein von-Mises regime, i.e., not close to Gaussian. The regression model assumed in (Xu et al., 2023) to tackle this dataset is a Bayesian Student-t regression for scalar observations $\{y_n\}_{n=1}^N$ representing endogenous creatinine clearance (CR); the covariates $X_n \in \mathbb{R}^3$ represent body weight in kg, serum creatinine concentration, and age in years. The goal is to predict CR of the patients. Therefore, the model (which includes an intercept) is given by

$$y_n \mid X_n, \beta \overset{\text{i.i.d.}}{\sim} \mathcal{T}([X_n, 1]^\top \beta, I_4, 5), \tag{21}$$

where $\beta \in \mathbb{R}^4$ follows the prior $p_0 = \mathcal{T}(0, I_4, 1)$, and $\mathcal{T}(\mu, \Sigma, \nu)$ is the Student-t distribution with location μ, scale Σ, and ν degrees of freedom. The posterior pdf π, with likelihood p and prior pdf p_0, is such that

$$\pi(\beta|\{X_n, y_n\}_{n=1}^N) \propto \left(\prod_{n=1}^N p(y_n|X_n, \beta)\right) p_0(\beta). \tag{22}$$

[1] publicly available at https://github.com/faosorios/heavy/blob/master/data/creatinine.rda

The normalizing constant of π, Z_π, is of practical importance as it can be used for model selection and is known as model evidence (Mackay, 1992).

We use `AHTIS` and AMIS to approximate Z_π. We use $T = 25$ iterations and varying number of samples. In order to obtain a better adaptation of the degree of freedom parameter ν by `AHTIS` in this case, we optimize the Gaussian process hyperparameters, with regularized maximum likelihood (full details in Appendix D). Algorithms are initialized with μ_0 sampled in Uniform$[-5, 5]^d$, $\Sigma_0 = \sigma^2 \cdot I_d$, where $\sigma^2 = 4$ (here, $d = 4$). `AHTIS` with adaptation of ν is initialized with $\nu_0 = 1$ while the algorithms with fixed ν use $\nu \in \{3, 5\}$. A large enough value of σ^2 spreads the mass more and is generally safer, although too large a value can create excessive discrepancy w.r.t. the posterior.

As before, we evaluate the α-ESS and the MSE on the estimation of Z_π. Since we do not have access to the true value of Z_π, we estimate the ground truth using AMIS with 10^5 samples for $T = 25$ iterations and initialised with the Laplace approximation of π (MacKay, 1992). Namely, we run AMIS with degree of freedom $\nu = 5$ and initial values $\mu_0 = \arg\max_\beta \pi(\beta|\{X_n, y_n\}_{n=1}^N)$ and $\Sigma_0 = -\left[\frac{\partial^2}{\partial \beta^2} \log \pi(\beta = \mu_0, \{X_n, y_n\}_{n=1}^N)\right]^{-1}$.

Results. In Fig. 4a-4b, we display the α-ESS and the square root relative MSE as functions of the number of samples M. In this experiment, there is no obvious true value for ν, due to the intractable π. The ν-adaptive `AHTIS` shows the best mean α-ESS values,

albeit with a larger variance. We expect this to be the case, since AHTIS has to learn ν adaptively through the minimization of a complicated objective function which does not fit in the hypotheses of the convergence results in (Garnett, 2023, Chapter 10) nor in the simpler setting of Student-t π that we described in Figure 2 and Proposition 2. Our adaptation procedure for ν is thus subject to non-convexities, such as local minima or plateaus, that may complicate reaching the global minimizer. In terms of MSE, the best performance is reached by the ν-adaptive AHTIS, and second best by AHTIS with $\nu = 5$ (which motivated using this ν for the ground truth). Note that when ν is fixed, AHTIS reaches better performance in both metrics than AMIS. We report results with more values of ν in Appendix E, with qualitatively similar findings.

6 CONCLUSIONS

We have proposed AHTIS, an AIS framework specifically suited for heavy-tailed target distributions π, being the first to do so explicitly in the AIS literature. AHTIS allows for the adaptation of location, scale, and tail parameter of a Student-t proposal, hereby differing from most previous AIS works. We also explicitly minimize an α-divergence between the target and the proposal, in the spirit of VI methods. We showed that the α-divergence can be approximated by a quantity involving the α-ESS, connecting further AIS and VI algorithms and allowing us to design our tail adaptation method.

Our framework is compatible with the use of mixture proposals when the target is suspected to be multimodal, and an extension towards this direction is interesting future work. Further, the computational efficiency of the tail adaptation procedure, when a good ν is not known in advance, could benefit from existing works in the BO literature.

Acknowledgements

T. Guilmeau and E. Chouzenoux acknowledge support from the European Research Council under Starting Grant MAJORIS ERC-2019-STG-850925. The work of V. E. is supported by ARL/ARO under grant W911NF-22-1-0235.

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

Agarwal, M., Vats, D., and Elvira, V. (2022). A principled stopping rule for importance sampling. *Electronic Journal of Statistics*, 16(2):5570–5590.

Akyildiz, Ö. D. and Míguez, J. (2021). Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31:1–17.

Amrouche, M., Carfantan, H., and Idier, J. (2022). Efficient sampling of Bernoulli-Gaussian-mixtures for sparse signal restoration. *IEEE Transactions on Signal Processing*, 70:5578–5591.

Birrell, J., Dupuis, P., Katsoulakis, M. A., Rey-Bellet, L., and Wang, J. (2021). Variational representations and neural network estimation of rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Botev, Z. I., L'Ecuyer, P., and Tuffin, B. (2013). Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing*, 23:271–285.

Boustati, A., Akyildiz, O. D., Damoulas, T., and Johansen, A. (2020). Generalised bayesian filtering via sequential monte carlo. *Advances in neural information processing systems*, 33:418–429.

Branchini, N. and Elvira, V. (2024). An adaptive mixture view of particle filters. *Foundations of Data Science*.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Miguez, J., and Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79.

Chib, S., Nardari, F., and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer.

Cornuet, J. M., Marin, J. M., Mira, A., and Robert, C. P. (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812.

Daudel, K., Douc, R., and Roueff, F. (2023). Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76.

Dhaka, A. K., Catalina, A., Welandawe, M., Andersen, M. R., Huggins, J., and Vehtari, A. (2021). Challenges and opportunities in high dimensional variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7787–7798.

Domke, J. and Sheldon, D. R. (2018). Importance weighting and variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4470–4479.

Doucet, A., Moulines, E., and Thin, A. (2023). Differentiable samplers for deep latent variable models. *Philosophical Transactions of the Royal Society A*, 381(2247):20220147.

El-Laham, Y., Djurić, P. M., and Bugallo, M. F. (2020). Enhanced mixture population Monte Carlo via stochastic optimization and Markov chain Monte Carlo sampling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5475–5479.

El-Laham, Y., Martino, L., Elvira, V., and Bugallo, M. F. (2019). Efficient adaptive multiple importance sampling. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Elvira, V. and Chouzenoux, E. (2022). Optimized population Monte Carlo. *IEEE Transactions on Signal Processing*, 70:2489–2501.

Elvira, V., Chouzenoux, E., Akyildiz, O. D., and Martino, L. (2023). Gradient-based adaptive importance samplers. *Journal of the Franklin Institute*, 360:9490–9514.

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. F. (2019). Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155.

Elvira, V., Martino, L., Luengo, D., and Corander, J. (2015). A gradient adaptive population importance sampler. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4075–4079.

Elvira, V., Martino, L., and Robert, C. P. (2022). Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550.

Fernández, C. and Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Finke, A. and Thiery, A. H. (2019). On importance-weighted autoencoders. https://arxiv.org/abs/1509.00519.

Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459.

Guilmeau, T., Chouzenoux, E., and Elvira, V. (2023). On variational inference and maximum likelihood estimation with the λ-exponential family. https://arxiv.org/abs/2310.05781.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. (2016). Black-box alpha divergence minimization. In *International Conference on Machine Learning (ICML)*, pages 1511–1520.

Huggins, J. H. and Roy, D. M. (2019). Sequential Monte Carlo as approximate sampling: bounds, adaptive resampling via ∞-ESS, and an application to particle Gibbs. *Bernoulli*, 25(1):584–622.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37:183–233.

Knoblauch, J., Jewson, J., and Damoulas, T. (2022). An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(1):5789–5897.

Kong, A. (1992). A note on importance sampling using standardized weights. University of Chicago, Dept. of Statistics, Tech. Rep.

Korba, A. and Portier, F. (2022). Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 11503–11527.

Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392.

Kviman, O., Melin, H., Koptagel, H., Elvira, V., and Lagergren, J. (2022). Multiple importance sampling ELBO and deep ensembles of variational approximations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 10687–10702.

L'Ecuyer, P. (2004). Quasi-Monte Carlo methods in finance. In *Winter Simulation Conference*, pages 1645–1655.

Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in neural information processing systems (NeurIPS)*.

Liu, C. and Rubin, D. B. (1995). ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39.

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2023). Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *SIAM Review*, 65(1):3–58.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Mackay, D. J. C. (1992). *Bayesian methods for adaptive models*. California Institute of Technology.

Marin, J.-M., Pudlo, P., and Sedki, M. (2019). Consistency of adaptive importance sampling and recycling schemes. *Bernoulli*, 25(3):1977 – 1998.

Marnissi, Y., Zheng, Y., Chouzenoux, E., and Pesquet, J.-C. (2017). A variational Bayesian approach for image restoration. Application to image deblurring with Poisson-Gaussian noise. *IEEE Transactions on Computational Imaging*, 3(4):722–737.

Martino, L., Elvira, V., and Louzada, F. (2017a). Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401.

Martino, L., Elvira, V., Luengo, D., and Corander, J. (2017b). Layered adaptive importance sampling. *Statistics and Computing*, 27:599–623.

Mattei, P.-A. and Frellsen, J. (2022). Uphill roads to variational tightness: Monotonicity and Monte Carlo objectives. https://arxiv.org/abs/2201.10989.

Mnih, A. and Rezende, D. (2016). Variational inference for monte carlo objectives. In *International Conference on Machine Learning*, pages 2188–2196. PMLR.

Moré, J. J. and Toraldo, G. (1989). Algorithms for bound constrained quadratic programming problems. *Numerische Mathematik*, 55(4):377–400.

Orsak, G. C. and Aazhang, B. (1991). Constrained solutions in importance sampling via robust statistics. *IEEE Transactions on Information Theory*, 37(2):307–316.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/.

Paananen, T., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, 31(2):16.

Paleyes, A., Mahsereci, M., and Lawrence, N. D. (2023). Emukit: A Python toolkit for decision making under uncertainty. In *Python in Science Conference*, pages 68–75.

Paleyes, A., Pullin, M., Mahsereci, M., McCollum, C., Lawrence, N., and González, J. (2019). Emulation of physical processes with Emukit. In *Second Workshop on Machine Learning and the Physical Sciences, NeurIPS*.

Portier, F. and Delyon, B. (2018). Asymptotic optimality of adaptive importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3134–3144.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer.

Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

Roy, V. and Hobert, J. P. (2010). On Monte Carlo methods for Bayesian multivariate regression models with heavy-tailed errors. *Journal of Multivariate Analysis*, 101(5):1190–1202.

Sakaya, J. and Klami, A. (2017). Importance sampled stochastic optimization for variational inference. In *33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Sanz-Alonso, D. (2018). Importance sampling and necessary sample size: an information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):867–879.

Sanz-Alonso, D. and Wang, Z. (2020). Bayesian update with importance sampling: Required sample size. *Entropy*, 23(1):1–22.

Särkkä, S. and Svensson, L. (2023). *Bayesian filtering and smoothing*. Cambridge university press.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. https://arxiv.org/abs/0912.3995.

Thin, A., Janati El Idrissi, Y., Le Corff, S., Ollion, C., Moulines, E., Doucet, A., Durmus, A., and Robert, C. P. (2021). Neo: non equilibrium sampling on the orbits of a deterministic transform. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 17060–17071.

Tipping, M. E. and Lawrence, N. D. (2005). Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69:123–141.

Tsallis, C. (2009). *Introduction to nonextensive statistical mechanics: approaching a complex world*. Springer.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Wang, D., Liu, H., and Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5737–5747.

Wang, S. and Swartz, T. (2022). Moment matching adaptive importance sampling with skew-Student proposals. *Monte Carlo Methods and Applications*, 28(2):149–162.

Wilson, J., Hutter, F., and Deisenroth, M. (2018). Maximizing acquisition functions for bayesian optimization. In *Advances in neural information processing systems (NeurIPS)*.

Wong, T.-K. L. and Zhang, J. (2022). Tsallis and Rényi deformations linked via a new λ-duality. *IEEE Transactions on Information Theory*, 68(8):5353–5373.

Xu, Z., Chen, N., and Campbell, T. (2023). Mixflows: principled variational inference via mixed flows. In *International Conference on Machine Learning (ICML)*, pages 38342–38376.

Zhang, L., Carpenter, B., Gelman, A., and Vehtari, A. (2022). Pathfinder: Parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, 23(1):13802–13850.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. https://github.com/nicola144/ais-heavy-tails

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes**, see our Supplementary
   (b) Complete proofs of all theoretical results. **Yes**, see our Supplementary
   (c) Clear explanations of any assumptions. **Yes**, see our Supplementary

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Not Applicable**
   (b) The license information of the assets, if applicable. **Not Applicable**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable**
   (d) Information about consent from data providers/curators. **Not Applicable**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable**

Thomas Guilmeau[⋆,◇], **Nicola Branchini**[†,◇], **Emilie Chouzenoux**[⋆], **Víctor Elvira**[†]

Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization: Supplementary Materials

In Appendix A, we give an example of the construction of an escort probability density that has lighter tails than the original. In Appendix B, we study the well-posedness of our variational formulation of the adaptation of the location, scale, and tail parameters of the proposal. In Section Appendix C, we give the proofs of our results about the sampling estimation of α-divergences. We detail our tail adaptation procedure in Appendix D, as well as another tail adaptation procedure proposed recently in the VI literature. Finally, we provide additional numerical experiments in Appendix E.

We run the synthetic experiments on a personal laptop with $7,6$ GB RAM and with 8 Intel Core i5 $-$ 8265U cores. We run the real data experiments on a a personal laptop (MacBook Pro) with 8 cores, M1 Apple Pro chip and 16 GB RAM. Our code is available at `https://github.com/nicola144/ais-heavy-tails`.

A Illustrative example of escort distribution

To illustrate how the escort version of a pdf makes the tails lighter with a concrete example, we show how the parameters of a Student-t distribution change when considering their escort version. In particular, the following proposition shows that it is possible to construct the escort pdf of a Student-t pdf such that the escort has a higher degree of freedom parameter than the original, and hence a lighter tail.

Proposition A.1. *(Guilmeau et al., 2023) Consider two Student-t families in dimension d with ν_q and ν degrees of freedom, respectively. Then the escort $q^{(\alpha)}_{\mu_q,\Sigma_q,\nu_q}$ of q_{μ_q,Σ_q,ν_q} with $\alpha = 1 + \frac{2}{\nu+d}$, is a Student-t distribution with $\nu^{(\alpha)}$ degrees of freedom, location $\mu^{(\alpha)}$, and shape $\Sigma^{(\alpha)}$ such that*

$$\begin{cases} \nu^{(\alpha)} = \nu_q + 2\frac{\nu_q+d}{\nu+d}, \\ \mu^{(\alpha)} = \mu_q, \\ \Sigma^{(\alpha)} = \frac{\nu_q}{\nu^{(\alpha)}}\Sigma_q. \end{cases} \tag{A.1}$$

B Divergence at the optimum for Student-t targets

We now study the properties of the optimization problem (13) when the target is a Student-t distribution. In particular, we give the proof of Proposition 2. We also describe in this case the inner problem in (14) and give an explicit expression of its optimum value, leading to the plot in Figure 2.

Proof of Proposition 2. The α-divergence is such that $D_\alpha(p,q) \geq 0$ with equality if and only if $p = q$ almost everywhere (for $\alpha > 0$ and $\alpha \neq 1$). Moreover, for any $\nu > 0$, $\alpha(\nu) = 1 + \frac{2}{\nu+d} > 1$. This implies that if (μ, Σ, ν) is such that

$$D_{\alpha(\nu)}(\pi, q_{\mu,\Sigma,\nu}) = 0, \tag{B.1}$$

then (μ, Σ, ν) is a solution of Problem (13).

Since π is a Student-t distribution, there exists $(\mu^\star, \Sigma^\star, \nu^\star)$ such that $q_{\mu^\star,\Sigma^\star,\nu^\star} = \pi$. In particular, $\nu^\star = \nu_\pi$. This implies that

$$D_{\alpha(\nu^\star)}(\pi, q_{\mu^\star,\Sigma^\star,\nu^\star}) = 0, \tag{B.2}$$

and hence the result. □

We now detail how to compute the function $\nu \longmapsto \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu^\star_\nu,\Sigma^\star_\nu,\nu})$ when π is a Student-t distribution, as it is plotted in Figure 2. To this end, we need to introduce the Rényi entropy of a pdf p that is defined by

$$H_\alpha(p) := \frac{1}{1-\alpha}\log\left(\int p(x)^\alpha dx\right). \tag{B.3}$$

We then use this notion to give an explicit expression of our quantity of interest.

Proposition B.1. *Consider a target distribution π and the family of Student-t distribution with ν degrees of freedom with $\alpha = 1 + \frac{2}{\nu+d}$. Consider $q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}$ such that Equation (9) is satisfied. Then we have*

$$\alpha(\alpha-1)\mathrm{D}_\alpha(\pi, q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}) = \exp\left((\alpha-1)\left(H_\alpha(q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}) - H_\alpha(\pi)\right)\right) - 1. \tag{B.4}$$

Proof. We can see from Proposition A.1 that Equation (9) implies that $\int x\,\pi^{(\alpha)}(x)dx = \int x\,q^{(\alpha)}_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}(x)dx$ and $\int xx^\top\,\pi^{(\alpha)}(x)dx = \int xx^\top\,q^{(\alpha)}_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}(x)dx$. We can deduce from that, and using (Wong and Zhang, 2022, Equation (3.17)), that the α-divergence RD_α between π and $q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}$ is such that

$$\mathrm{RD}_\alpha(\pi, q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}) = H_\alpha(q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}) - H_\alpha(\pi), \tag{B.5}$$

where RD_α is the Rényi divergence with parameter α. The result follows from the link between the α-divergence and the Rényi divergence (see (van Erven and Harremoës, 2014) for the definition and properties of the Rényi divergence). □

Proposition B.1 shows that, in order to compute the quantity plotted in Figure 2, we need to compute explicitly the Rényi entropy of a Student-t distribution and compute explicitly the parameters of $q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}$. We do so in the following two propositions.

Proposition B.2. *Consider two degree of freedom parameters $\nu, \nu_q > 0$, a dimension d, and set $\alpha = 1 + \frac{2}{\nu+d}$. Then, for any q_{μ_q,Σ_q,ν_q}, we have that*

$$H_\alpha(q_{\mu_q,\Sigma_q,\nu_q}) = -\frac{\nu+d}{2}\left(\log Z_{\nu^{(\alpha)},\Sigma^{(\alpha)}} - \alpha \log Z_{\nu_q,\Sigma_q}\right) \tag{B.6}$$

with $\nu^{(\alpha)} = \nu_q + 2\frac{\nu_q+d}{\nu+d}$ and $\Sigma^{(\alpha)} = \frac{\nu_q}{\nu^{(\alpha)}}\Sigma_q$.

Proof. Using the result of Proposition A.1, we first compute that for any $x \in \mathbb{R}^d$,

$$q_{\mu_q,\Sigma_q,\nu_q}(x)^\alpha = \frac{Z_{\nu^{(\alpha)},\Sigma^{(\alpha)}}}{Z^\alpha_{\nu_q,\Sigma_q}} q_{\mu^{(\alpha)},\Sigma^{(\alpha)},\nu^{(\alpha)}}(x) \tag{B.7}$$

From there, we deduce that

$$\begin{aligned} H_\alpha(\mu_q,\Sigma_q,\nu_q) &= \frac{1}{1-\alpha}\log\left(\int q_{\mu_q,\Sigma_q,\nu_q}(x)^\alpha dx\right) \\ &= \frac{1}{1-\alpha}\left(\log Z_{\nu^{(\alpha)},\Sigma^{(\alpha)}} - \alpha\log Z_{\nu_q,\Sigma_q}\right) \end{aligned}$$

which gives the result. □

Proposition B.3. *Consider two degree of freedom parameters $\nu, \nu_\pi > 0$, a dimension d, and set $\alpha = 1 + \frac{2}{\nu+d}$. Then, for any $\pi = q_{\mu_\pi,\Sigma_\pi,\nu_\pi}$, the Student-t distribution $q_{\mu_\nu^\star,\Sigma_\nu^\star,\nu}$ minimizing $(\mu,\Sigma) \longmapsto \mathrm{D}_\alpha(\pi, q_{\mu,\Sigma,\nu})$ is such that*

$$\begin{cases} \mu_\nu^\star = \mu_\pi, \\ \Sigma_\nu^\star = \frac{\nu_\pi}{\nu^{(\alpha)}-2}\Sigma_\pi, \end{cases} \tag{B.8}$$

provided that $\nu^{(\alpha)} > 2$.

Proof. This comes from the optimality result of Proposition 1, the characterization of $\pi^{(\alpha)}$ as a Student-t distribution with parameters given in Proposition A.1, and the fact that for any $q_{\mu,\Sigma,\nu}$ with $\nu > 2$, $q_{\mu,\Sigma,\nu}(x) = \mu$ and $q_{\mu,\Sigma,\nu}(xx^\top) = \frac{\nu}{\nu-2}\Sigma$. □

Gathering these three results, we can then get a closed-form expression for the function

$$\nu \longmapsto \min_{\mu,\Sigma} \mathrm{D}_{\alpha(\nu)}(\pi, q_{\mu,\Sigma,\nu}) \tag{B.9}$$

when π is Student-t distribution for some $\nu_\pi > 0$. Then, one can use it to draw Figure 2.

C Proofs of Section 3.2

We give below the proofs of Propositions 3 and 4, that describe the approximation of α-divergences by a self-normalized importance sampling estimator. This estimator is linked with the α-ESS and the (discrete) α-divergence between the normalized importance weights and the corresponding uniform weights.

Proof of Proposition 3. We derive the following self-normalized IS (SNIS) approximation of the α-divergence, making an ESS-like quantity appear:

$$\begin{aligned}
\mathrm{D}_\alpha(\pi, q) =& \frac{1}{\alpha(\alpha-1)} \left(\frac{\int \left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^\alpha q(x)dx}{Z_\pi^\alpha} - 1 \right) && \text{(C.1)}\\
\approx& \frac{1}{\alpha(\alpha-1)} \left(\frac{\frac{1}{M}\sum_{m=1}^M \left(\frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})}\right)^\alpha}{\left(\frac{1}{M}\sum_{m=1}^M \frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})}\right)^\alpha} - 1 \right) && \text{(C.2)}\\
=& \frac{1}{\alpha(\alpha-1)} \left(M^{\alpha-1} \sum_{m=1}^M \overline{w}_m^\alpha - 1 \right) && \text{(C.3)}\\
=& \frac{M^{\alpha-1}}{\alpha(\alpha-1)} \left(\sum_{m=1}^M \overline{w}_m^\alpha - M^{1-\alpha} \right) && \text{(C.4)}\\
=& \frac{M^{\alpha-1}}{\alpha(\alpha-1)} \left((ESS_\alpha)^{1-\alpha} - M^{1-\alpha} \right). && \text{(C.5)}
\end{aligned}$$

Notice the SNIS approximation from Eq. (C.1) to Eq. (C.2), i.e., the same set of samples is used to approximate a ratio of two integrals. Moreover, we can recognize from Eq. (C.3) that

$$\mathrm{D}_\alpha(\pi, q) \approx \mathrm{D}_\alpha^M(\{\overline{w}^{(m)}\}_{m=1}^M, \{1/M\}_{m=1}^M), \tag{C.6}$$

with the continuous α-divergence on the left and the discrete α-divergence on the right. We also have the almost sure convergence $\mathrm{D}_\alpha^M(\{\overline{w}^{(m)}\}_{m=1}^M, \{1/M\}_{m=1}^M) \xrightarrow[M\to+\infty]{a.s.} \mathrm{D}_\alpha(\pi, q)$ from standard SNIS results (see for instance (Owen, 2013, Theorem 9.2)). □

Proof of Proposition 4. We first compute the gap

$$D_\alpha^M(\{\overline{w}^{(m)}\}_{m=1}^M, \{1/M\}_{m=1}^M) - D_\alpha(\pi, q) = \frac{1}{\alpha(\alpha-1)} \left(\frac{\frac{1}{M}\sum_{m=1}^M \left(\frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})}\right)^\alpha}{\left(\frac{1}{M}\sum_{m=1}^M \frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})}\right)^\alpha} - \frac{\int \left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^\alpha q(x)dx}{\left(\int \widetilde{\pi}(x)dx\right)^\alpha} \right). \tag{C.7}$$

We now deal with the denominator. Due to our hypothesis $\pi(x) > 0 \Rightarrow q(x) > 0$, we have that $\frac{1}{M}\sum_{m=1}^M \frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})} \xrightarrow[M\to+\infty]{a.s.} Z_\pi$, from which we deduce the following almost sure convergence:

$$\left(\frac{1}{M} \sum_{m=1}^M \frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})} \right)^\alpha \xrightarrow[M\to+\infty]{a.s.} \left(\int \widetilde{\pi}(x)dx \right)^\alpha. \tag{C.8}$$

We now turn to the numerator. The quantity $\frac{1}{M}\sum_{m=1}^M \left(\frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})}\right)^\alpha$ is an unbiased Monte Carlo estimator of $\int \left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^\alpha q(x)dx$ with the variance of each term of the sum being equal to

$$\mathbb{V}_q\left[\left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^\alpha\right] = \int \widetilde{\pi}(x)^{2\alpha} q(x)^{1-2\alpha} dx - \left(\int \left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^\alpha q(x)dx \right)^2. \tag{C.9}$$

We have by the central limit theorem for Monte Carlo estimators that

$$\sqrt{M}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\widetilde{\pi}(x^{(m)})}{q(x^{(m)})}\right)^{\alpha}-\int\left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^{\alpha}q(x)dx\right)\xrightarrow[M\to+\infty]{d}\mathcal{N}\left(0,\mathbb{V}_q\left[\left(\frac{\widetilde{\pi}(x)}{q(x)}\right)^{\alpha}\right]\right). \tag{C.10}$$

Then, using Eq. (C.8)-(C.10) and Slutsky's theorem, we obtain that

$$\begin{aligned}&\sqrt{M}\left(D_{\alpha}^{M}(\{\overline{w}^{(m)}\}_{m=1}^{M},\{1/M\}_{m=1}^{M})-D_{\alpha}(\pi,q)\right)\\ &\xrightarrow[N\to+\infty]{d}\mathcal{N}\left(0,\left(\frac{\int\widetilde{\pi}(x)^{2\alpha}q(x)^{1-2\alpha}dx}{\left(\alpha(\alpha-1)\int\widetilde{\pi}(x)^{\alpha}q(x)^{1-\alpha}dx\right)^{2}}-1\right)\right),\end{aligned} \tag{C.11}$$

which yields the result. □

D Tail adaptation

D.1 Our tail adaptation procedure with Bayesian Optimization

We present below in full detail our proposed tail adaptation procedure.

Algorithm D.1 Tail adaptation with BO

Require:
- Current tail parameter and α-ESS, i.e., $\nu_t, \widehat{\mathrm{ESS}}_{\alpha_t}$
- Previous tail parameters and α-ESS values $\{\nu_\tau, \widehat{\mathrm{ESS}}_{\alpha_\tau}\}_{\tau=1}^{t-1}$
- Choice of parameterized kernel function $k(\nu,\nu';\theta)$. This defines the GP prior, i.e., the GP before observing any data.
- Choice of parameterized acquisition function $\mathrm{acq}(\nu;\psi;\mathcal{GP})$
- **(Optional):** Choice of prior distribution for θ, $p(\theta)$
- **(Optional):** Choice of prior distribution for the observation noise σ^2, $p(\sigma^2)$

1: Use transformed values of $\widehat{\mathrm{ESS}}_{\alpha_\tau}$ with the following monotonic transformation (thus preserving optima)

$$y_\tau=\log\left(1-\left(\frac{1}{M}\widehat{\mathrm{ESS}}_{\alpha_\tau}\right)\right),\quad \tau=1,\dots,t \tag{D.1}$$

2: Update Gaussian process $\mathcal{GP}_t$ posterior at current iteration t with new datapoint $\{\nu_t, y_t\}$. This involves calculating a new mean and covariance, see (Garnett, 2023, Chapter 2).

3: Obtain new tail parameter ν_{t+1} by maximizing the acquisition function,

$$\nu_{t+1}\leftarrow\arg\max_{\nu}\mathrm{acq}(\nu;\psi;\mathcal{GP}_t) \tag{D.2}$$

In Emukit, a popular BO package (Paleyes et al., 2019, 2023), several solvers are available depending on acq; we used a gradient-based solver.

4: **(Optional) Gaussian process hyperparameter optimization:** We model $\{y_\tau\}_{\tau=1}^{t}$ as *noisy* observations of the true (transformed) α-ESS from the $\mathcal{GP}$ with additive Gaussian noise with variance σ^2. Letting the observations (scalars) be $y=[y_1,\dots,y_t]$ and K_t be the $t\times t$ matrix with entries $k(\nu_\tau,\nu_{\tau'};\theta)$ for $(\tau,\tau')\in\{1,\dots,t\}\times\{1,\dots,t\}$, optimize θ and noise σ^2 to maximize the log-likelihood of the observed data, as

$$\theta,\sigma^2\leftarrow\arg\max_{\theta,\sigma^2}-\frac{1}{2}\log\left|\det\left[2\pi\left(K_t+\sigma^2I\right)\right]\right|-\frac{1}{2}y^{\top}\left(K_t+\sigma^2I\right)^{-1}y+\log p(\theta)+\log p(\sigma^2) \tag{D.3}$$

The optimization problem in Eq. (D.3) is addressed by Emukit (Paleyes et al., 2019, 2023) with gradient-based methods.

5: **Return:** ν_{t+1}

We describe below all implementation details regarding Algorithm D.1.

Kernel function. We used the perhaps most common kernel function in BO, i.e., the radial basis function (RBF) kernel, also known as exponentiated quadratic (EQ) or squared exponential (SE) (Garnett, 2023). For a one dimensional input as ν, the SE kernel has two scalar parameters, lengthscale l and function variance σ_f^2, i.e., $\theta = \{l, \sigma_f^2\}$, and its expression is given by

$$k_{\mathrm{SE}}(\nu, \nu'; \theta) = \sigma_f^2 \cdot \exp\left(-\frac{1}{2}\frac{(\nu - \nu')^2}{l^2}\right). \tag{D.4}$$

The lengthscale l indicates the typical distance between turning points in the function, while the intuition for σ_f is that by seeing a long enough horizontal stretch of the function, $\approx 2/3$ of the points would lie between $\pm\sigma_f$ of the GP mean.

Acquisition function. We experimented using the Gaussian process upper confidence bound (GP-UCB) (Srinivas et al., 2009), a parameterized (by ψ) acquisition function (perhaps the most well-known), acq, with only one scalar tuning parameter $\psi = \{\beta\}, \beta > 0$ given by

$$\nu_{t+1}^{\text{UCB-best}} = \underset{\nu \in [1, \nu_{\max}]}{\arg\max}\, \mu_{\mathcal{GP}_t}(\nu) + \beta^{1/2} v^2_{\mathcal{GP}_t}(\nu), \tag{D.5}$$

where, defining $k_t(\nu)$ as the vector-valued function $\nu \to [k(\nu, \nu_1), \dots, k(\nu, \nu_t)]$ (and omitting kernel parameters θ for brevity),

$$\mu_{\mathcal{GP}_t}(\nu) = k_t(\nu)^\top (K_t + \sigma^2 I)^{-1} y_t, \tag{D.6}$$

$$v^2_{\mathcal{GP}_t}(\nu) = k(\nu, \nu) - k_t(\nu)^\top (K_t + \sigma^2 I)^{-1} k_t(\nu). \tag{D.7}$$

The β parameter controls the typical exploration and exploitation tradeoff needed. To set β, we followed the theoretical guarantees described by (Garnett, 2023, Chapter 10, page 229); letting the search space for ν be $\mathcal{V} = [1, \nu_{\max}]$ and t for the BO iteration number (corresponding to t in our `AHTIS` algorithm), we selected

$$\beta_t^\star = \sqrt{2 \log\left(\frac{(t^2+1)\,|\mathcal{V}|}{\sqrt{2\pi}}\right)} \tag{D.8}$$

for the synthetic experiments. For the real data experiments, we used $\beta_t = 1.5 \cdot \beta_t^\star$ for higher exploration due to a much noisier and more challenging objective function. As search space for Eq. (D.5), we used $\nu_{\max} = 10$.

Hyperparameter priors. As described in the main paper, for the real data experiments we optimized the GP hyperparameters at each iteration (**step (4)** of Algorithm D.1) using a prior for both $\theta = l, \sigma_f^2$ and σ. For all these parameters, we used an inverse Gamma prior,

$$p(\sigma^2 | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} e^{-\frac{\beta}{\sigma^2}}, \tag{D.9}$$

where $\Gamma(\cdot)$ is the gamma function, (omitting equivalent equations for l and σ_f^2) with α and β selected such that $\mathbb{E}[\sigma^2] = 3, \mathbb{V}[\sigma^2] = 2$; $\mathbb{E}[\sigma_f^2] = 5, \mathbb{V}[\sigma_f^2] = 2$; $\mathbb{E}[l] = 5, \mathbb{V}[l] = 2$.

Finally, to implement all of the above steps we used the Python library Emukit (Paleyes et al., 2019, 2023).

D.2 Another tail adaptation method

Daudel et al. (2023) propose a VI method for minimizing a fixed α-divergence (differently than us, as the α changes at each iteration) over a mixture of Student-t distributions in (Daudel et al., 2023, Example 5). For each component of the mixture, the location, scale, and tail parameters are all adapted. We now show that their tail-adaptation procedure is not able to produce degree of freedom parameters that are less than a constant $\nu_{\min} \in (2.5, 2.6)$.

In order to observe that, we consider the update (Daudel et al., 2023, Equation (70)). For simplicity, we consider the case where the mixture is reduced to one component, but our analysis still applies in this more general setting. In the simplified setting we consider, we have at iteration t that the next degree of freedom parameter ν_{t+1} satisfies

$$\kappa\left(\frac{\nu_{t+1}}{2}\right) = \int (z - \ln(z)) p_{\mu_t, \Sigma_t, \nu_t}(y, z, dy, dz), \tag{D.10}$$

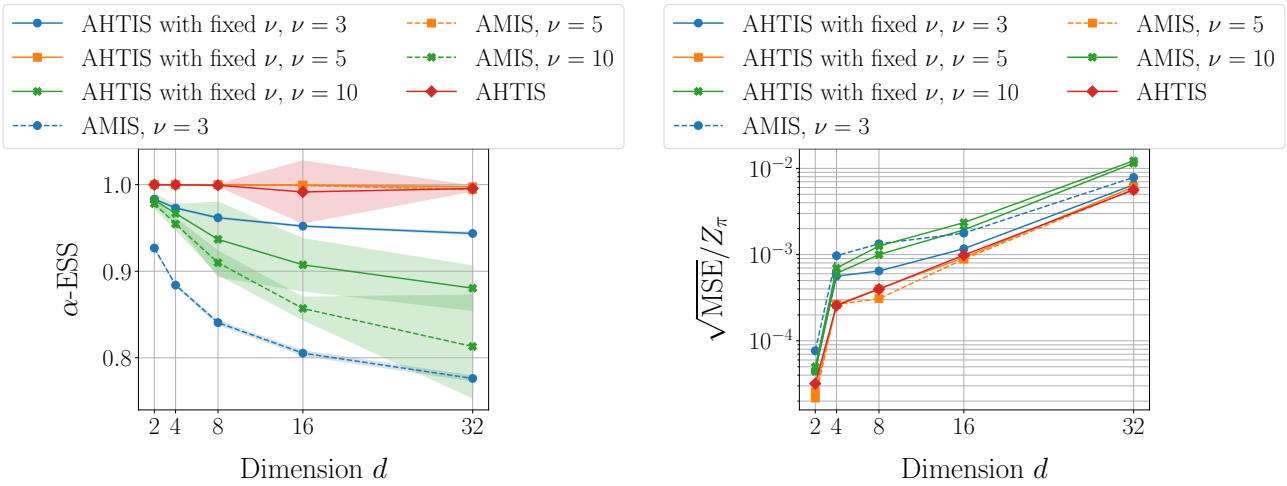
with $\kappa(x) = \ln(x) + \frac{\Gamma'(x)}{\Gamma(x)}$, $\alpha \in [0, 1)$, and a positive measure $p_{\mu_t, \Sigma_t, \nu_t}$ over $\mathbb{R} \times \mathbb{R}^d$. For any $z > 0$, we have $z - \ln(z) \geq 1$. We can thus check that the right-hand side of Eq. (D.10) is positive. The function κ is increasing and bijective from $(0, +\infty)$ to $\mathbb{R}$ from (Daudel et al., 2023, Lemma 13).

Now let us demonstrate that there exists a scalar $\nu_{\min} > 0$ such that $\nu_{t+1} > \nu_{\min}$ and give some bounds on $\nu_{\min}$. We define $\nu_{\min}$ such that $\kappa\left(\frac{\nu_{\min}}{2}\right) = 0$. The function κ is increasing and bijective from $(0, +\infty)$ to $\mathbb{R}$ from (Daudel et al., 2023, Lemma 13). We can check that that $\kappa(1.25) < 0$ and that $\kappa(1.3) > 0$. This means that the scalar $\nu_{\min}$ exists and satisfies $\nu_{\min} \in (2.5, 2.6)$. This shows that there are values of ν that cannot be attained by the algorithm of Daudel et al. (2023). Although this lower bound is reasonable, it may not yield optimal performance on heavy-tailed targets such as the one considered in Section 5.1.

E Further Numerical Experiments

E.1 Controlled Scenario with Varying Dimension Student-t Targets

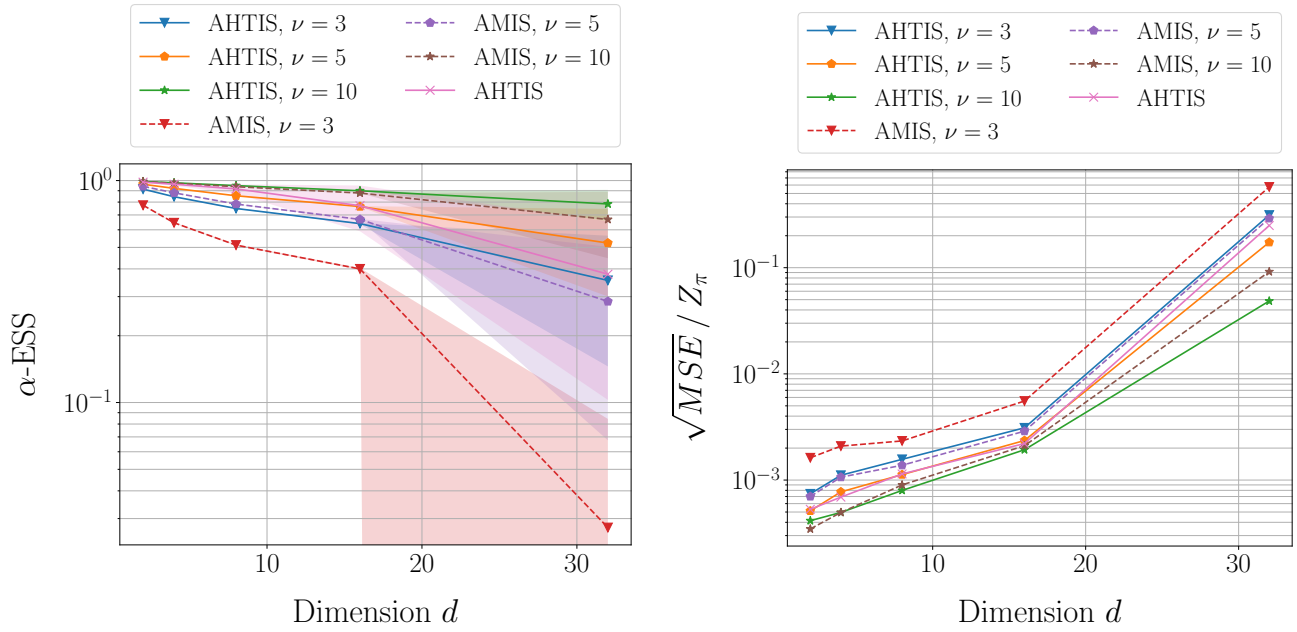
We give here supplementary numerical experiments in the case of a Student-t target distribution in varying dimension, that is described in Section 5.1. In addition to the results already presented in Section 5.1, we show in Fig. E.1 the α-ESS and square-root relative MSE on the normalizing constant of the target when the target has degree of freedom $\nu_\pi \in \{5, 50\}$. We also compute the bootstrap 90% confidence intervals to better understand the distributions of the estimates yielded by the algorithms. They are shown in Fig. E.3. Finally, we describe the final degree of freedom parameters reached by `AHTIS` with adaptation of ν when the target has degree of freedom parameter $\nu_\pi \in \{2, 5, 50\}$ in Table E.1.



(a) α-ESS (mean $\pm$ one standard deviation, higher is better) for various dimensions d. `AHTIS` outperforms AMIS for any ν, sometimes by an order of magnitude, and the ν-adaptive version converges to the true value $\nu_\pi = 5$.

(b) Relative square root MSE (lower is better) for various dimensions d. Note that $Z_\pi = Z_{\nu_\pi, \Sigma_\pi}$ is the true normalizing constant, which is available. `AHTIS` outperforms AMIS for any ν and the ν-adaptive version converges to $\nu_\pi = 5$.

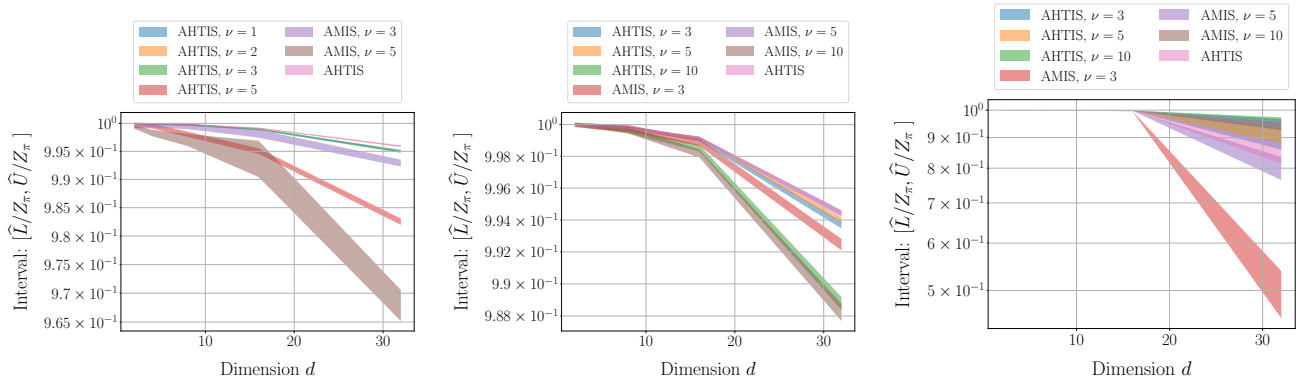
Figure E.1: Results for a synthetic Student-t target with $\nu_\pi = 5$. All algorithms are run for $T = 20$ iterations, with $M = 10^4$ samples per iteration and results are averaged over 100 replications. A dashed line identifies AMIS, while solid line is `AHTIS`, and same marker/color indicates same ν.



(a) α-ESS (mean $\pm$ one standard deviation, higher is better) for various dimensions d. `AHTIS` outperforms AMIS for small ν, but does not always yield the highest possible value of ν, for which the best results are obtained with the escort version of AMIS.

(b) Relative square root MSE (lower is better) for various dimensions d. Note that $Z_\pi = Z_{\nu_\pi,\Sigma_\pi}$ is the true normalizing constant, which is available. `AHTIS` outperforms AMIS for small ν, but does not always yield the highest possible value of ν, for which the best results are obtained with the escort version of AMIS.

Figure E.2: Results for a synthetic Student-t target with $\nu_\pi = 50$. All algorithms are run for $T = 20$ iterations, with $M = 10^4$ samples per iteration and results are averaged over 100 replications. A dashed line identifies AMIS, while solid line is `AHTIS`, and same marker/color indicates same ν.



(a) Normalized bootstrap 90% confidence intervals for a Student-t target with $\nu_\pi = 2$.

(b) Normalized bootstrap 90% confidence intervals for a Student-t target with $\nu_\pi = 5$.

(c) Normalized bootstrap 90% confidence intervals for a Student-t target with $\nu_\pi = 50$.

Figure E.3: Normalized bootstrap 90% confidence intervals for the final estimators of Z_π for a synthetic Student-t target π with $\nu_\pi \in \{2, 5, 50\}$ in dimension $d \in \{2, 4, 8, 16, 32\}$. The algorithms are run for $T = 20$ iterations with $M = 10^4$ samples per iterations, yielding an estimator of Z_π. This experiment has been replicated 100 times, and we have used these estimators to compute the bias-corrected bootstrap confidence intervals $[\widehat{L}, \widehat{U}]$ for each algorithm. The lower and upper bounds of the confidence have been normalized by dividing them by Z_π. For each algorithm, the shaded area represents the resulting normalized confidence interval. Good normalized bounds are thus closed to the value 1 (since we are trying to estimate Z_π and we are normalizing by this same value).

	$\nu_\pi = 2$	$\nu_\pi = 5$	$\nu_\pi = 50$
$d = 2$	2.05 ± 0.569	4.87 ± 0.142	8.38 ± 1.66
$d = 4$	2.16 ± 1.18	4.96 ± 0.168	8.17 ± 1.54
$d = 8$	1.98 ± 0.574	4.93 ± 0.293	8.13 ± 1.98
$d = 16$	2.03 ± 0.562	5.03 ± 0.978	6.71 ± 3.10
$d = 32$	2.04 ± 0.175	5.03 ± 0.504	5.62 ± 2.48

Table E.1: Final degree of freedom ν_T yielded by Algorithm 1 for a target with degree of freedom parameter ν_π in dimension d. The results are of the form mean $\pm$ one standard deviation, for $T = 20$ over 100 runs.

Results. Table E.1 reveals that the ν-adaptive `AHTIS` is able to correctly capture the tail behaviour of the target with good precision when $\nu_\pi \in \{2, 5\}$. In the case $\nu_\pi = 50$ where the tails of the target are lighter, we see that the adapted value of ν yielded by `AHTIS` is lower than ν_π. This is because ν is restricted to an interval $[1, \nu_{\max}]$ (see Eq. (D.5)). We have set $\nu_{\max} = 10$ in our experiments, thus excluding $\nu_\pi = 50$. This is not a problem in IS, as we aim to create proposals that, if mismatched, have heavier tails than the target (Owen, 2013, Chapter 9). Fig. E.1 shows a situation where AMIS and `AHTIS` with $\nu = 5$, and the ν-adaptive `AHTIS` are able to reach similar performance in terms of α-ESS and MSE. The fact that AMIS is now able to reach performance similar to `AHTIS` (in contrast with the results of Fig. 3) is because $\nu_\pi = 5$, meaning that π has well-defined first and second order moments and that AMIS with $\nu = \nu_\pi$ can be used. Note however that in the case of a mismatch $\nu \neq \nu_\pi$, AMIS is inferior to `AHTIS`. Fig. E.2a shows a situation the target π is significantly lighter-tailed than the considered proposals. Indeed, $\nu_\pi = 50$ while proposals have degree of freedom parameter $\nu \in \{3, 5, 10\}$ and `AHTIS` is constrained to $\nu \in [1, \nu_{\max}]$ (see Eq. (D.5)) with $\nu_{\max} = 10$. In this situation, the methods with fixed ν perform better, with the best performance obtained with the escort AMIS with $\nu = 10$. `AHTIS` does not always identify the highest value of ν as the best, but still outperforms the other methods when the proposals have $\nu \neq 10$. Fig. E.3 show that the estimators of Z_π tend to underestimate the true value of Z_π in high dimensions, since the bootstrap 95% intervals often gets below Z_π. This phenomenon is more pronounced, especially for the AMIS algorithm, when there is a mismatch between the degree of freedom parameter of the target and the one of the proposal, showcasing the effectiveness of our `AHTIS` methodology in this respect.

E.2 Application to Bayesian Student-t regression on real data

We include figures with additional results for $\nu = 4$ for all algorithms (excluded from the main paper for better readability of the main plots).

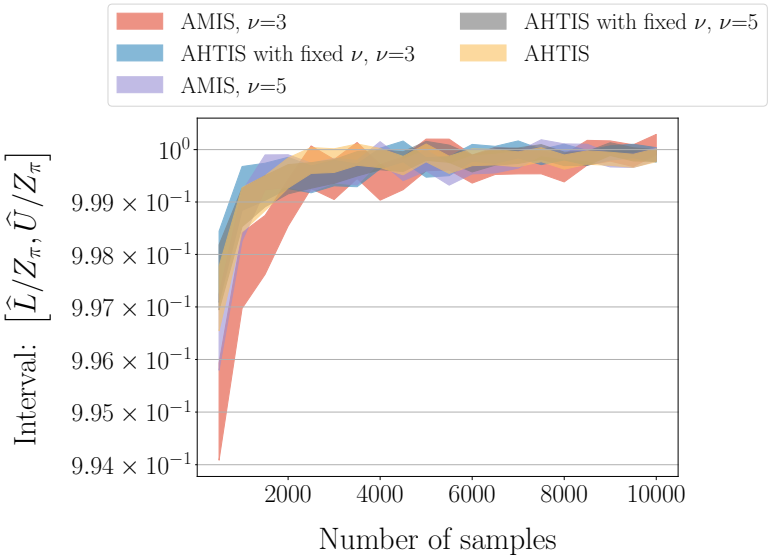
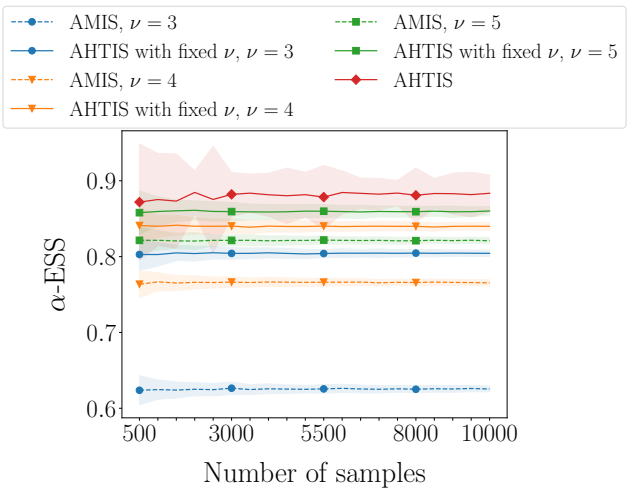
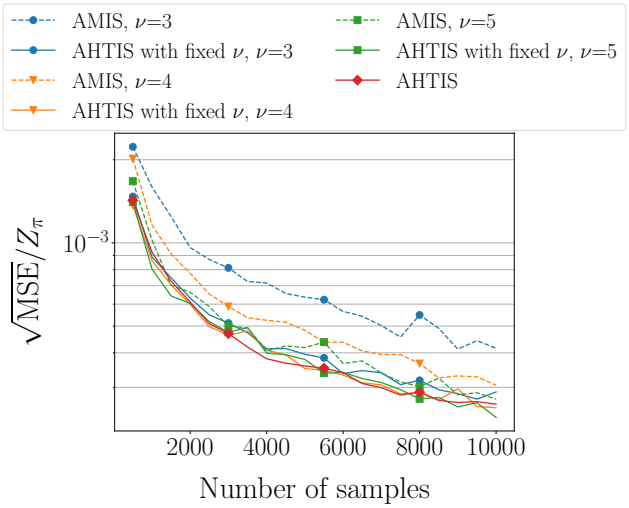


Figure E.5: Normalized bootstrap 90% confidence intervals for the final estimators of Z_π for the posterior of the creatinine dataset. Note that here since the dimension is fixed, we have number of samples on the x-axis.



(a) α-ESS (mean $\pm$ one standard deviation, higher is better) for the creatinine dataset experiments.



(b) Relative square root MSE (lower is better) for the creatinine dataset experiments.

Figure E.4: Results here are as in Figs. 4a and 4b (main paper), but with added $\nu = 4$. Recall that all algorithms are run for $T = 25$ iteration and results are averaged over 250 replications. A dashed line identifies AMIS, while solid line is `AHTIS`, and same marker/color indicates same ν.