Pareto-$\hat{k}$ as practical pre-asymptotic diagnostic of Monte Carlo estimates

Aki Vehtari

Aalto University

Finnish Center for Artificial Intelligence

Stan

ArviZ

with Andersen, Bürkner, Catalina, Dhaka, Gabry, Gelman, Huggins, Magnusson, Paananen, Piironen, Simpson, Welandawe & Yao

Relevance of this talk

- Practical diagnostic tool
  - Monte Carlo, MCMC, quasi MC, importance sampling, particle filtering
  - stochastic optimization, stochastic variational inference
  - estimating divergences
  - assessing distributional approximations

(Markov chain) Monte Carlo

$$\theta^{(s)} \sim p(\theta)$$

$$\mathrm{E}_p[h(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)})$$

- Consistent and unbiased (MCMC asymptotically)
- If variance is finite $\rightarrow$ central limit theorem (CLT)

$$\mathrm{Var}[\mathrm{E}(h(\theta))] \approx \mathrm{Var}[h(\theta)]/S$$

(Markov chain) Monte Carlo

$$\theta^{(s)} \sim p(\theta)$$

$$\mathrm{E}_p[h(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)})$$

- Consistent and unbiased (MCMC asymptotically)
- If variance is finite $\rightarrow$ central limit theorem (CLT)

$$\mathrm{Var}[\mathrm{E}(h(\theta))] \approx \mathrm{Var}[h(\theta)]/S$$

In case of MCMC effective sample size (ESS) takes into account the within and between chain dependencies (see, e.g. Vehtari et al., 2021)

(Self-normalized) Importance sampling

$$\theta^{(s)} \sim g(\theta)$$

$$\mathrm{E}_p[h(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}) w^{(s)}, \quad \text{where} \quad w^{(s)} = \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

- IS estimate is consistent and unbiased

(Self-normalized) Importance sampling

$$\theta^{(s)} \sim g(\theta)$$

$$\mathrm{E}_p[h(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}) w^{(s)}, \quad \text{where} \quad w^{(s)} = \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

- IS estimate is consistent and unbiased

Self-normalized

$$\mathrm{E}_p[h(\theta)] \approx \frac{\sum_{s=1}^{S} h(\theta^{(s)}) w^{(s)}}{\sum_{s=1}^{S} w^{(s)}}$$

- Self-normalized IS estimate is consistent with bias $O(1/S)$

(Self-normalized) Importance sampling

$$\theta^{(s)} \sim g(\theta)$$

$$\mathrm{E}_p[h(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S} h(\theta^{(s)}) w^{(s)}, \quad \text{where} \quad w^{(s)} = \frac{p(\theta^{(s)})}{g(\theta^{(s)})}$$

- IS estimate is consistent and unbiased

Self-normalized

$$\mathrm{E}_p[h(\theta)] \approx \frac{\sum_{s=1}^{S} h(\theta^{(s)}) w^{(s)}}{\sum_{s=1}^{S} w^{(s)}}$$

- Self-normalized IS estimate is consistent with bias $O(1/S)$
- If $h(\theta)w$ and w have finite variance $\rightarrow$ CLT
  - variance goes down as $1/S$
  - ESS takes into account the variability in the weights

Some uses of importance sampling

- Fast leave-one-out cross-validation
- Fast bootstrapping
- Fast prior and likelihood sensitivity analysis
- Particle filtering
- Improving distributional approximation (e.g VI)

Estimating divergences

- f-divergences can be presented as expectations of the density ratio $w(\theta)$

$$\mathcal{L}_f(p \parallel g) := \mathbb{E}_{\theta \sim g}[f(w(\theta))] \approx \frac{1}{S}\sum_{s=1}^{S} f(w^{(s)})$$

Estimating divergences

- f-divergences can be presented as expectations of the density ratio $w(\theta)$

$$\mathcal{L}_f(p \parallel g) := \mathbb{E}_{\theta \sim g}[f(w(\theta))] \approx \frac{1}{S}\sum_{s=1}^{S} f(w^{(s)})$$

Objective	$f(w)$
Exclusive KL	$\log(w)$
Inclusive KL	$w \log(w)$
χ^2	$(w^2 - w)/2$
α-divergence	$(w^\alpha - w)/(\alpha(\alpha - 1))$

Estimating divergences

- f-divergences can be presented as expectations of the density ratio $w(\theta)$

$$\mathcal{L}_f(p \parallel g) := \mathbb{E}_{\theta\sim g}[f(w(\theta))] \approx \frac{1}{S}\sum_{s=1}^{S} f(w^{(s)})$$

Objective	$f(w)$
Exclusive KL	$\log(w)$
Inclusive KL	$w\log(w)$
χ^2	$(w^2 - w)/2$
α-divergence	$(w^\alpha - w)/(\alpha(\alpha-1))$

- Basis of stochastic variational inference
  - $w(\theta)$ connects IS and SVI

Central limit theorem

- We would like to have finite variance and CLT
  - sometimes these can be guaranteed by construction, e.g., by choosing $g(\theta)$ so that $w(\theta)$ is bounded
  - generally not trivial

Central limit theorem

- We would like to have finite variance and CLT
  - sometimes these can be guaranteed by construction, e.g., by choosing $g(\theta)$ so that $w(\theta)$ is bounded
  - generally not trivial
- If variance is infinite, but mean is finite
  $\rightarrow$ *generalized CLT and asymptotic consistency*

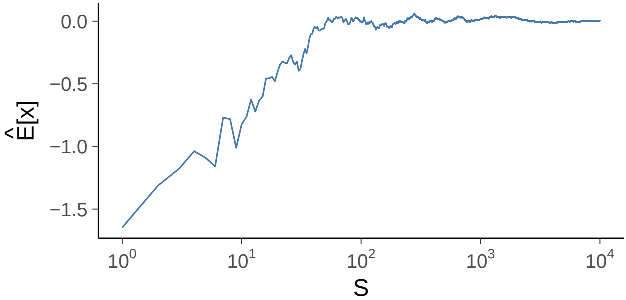
Central limit theorem

- We would like to have finite variance and CLT
  - sometimes these can be guaranteed by construction, e.g., by choosing $g(\theta)$ so that $w(\theta)$ is bounded
  - generally not trivial
- If variance is infinite, but mean is finite
  $\rightarrow$ *generalized CLT and asymptotic consistency*
- Pre-asymptotic and asymptotic behavior can be really different!

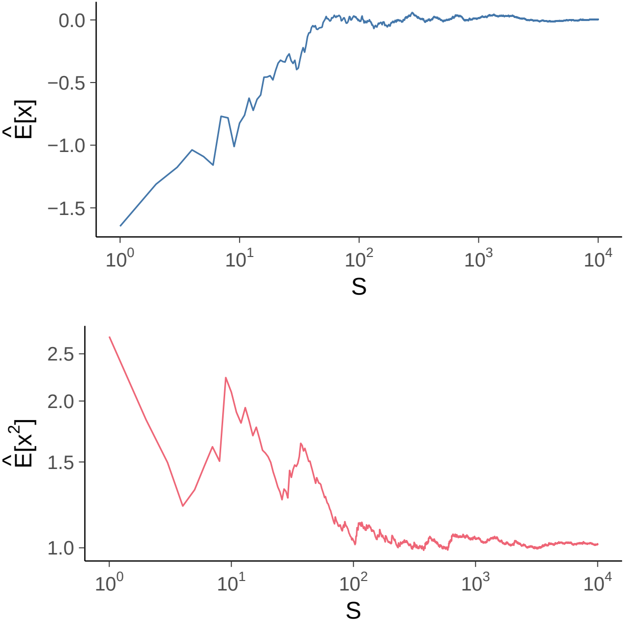
Simple example: $x \sim N, \quad t_4, \quad t_2, \quad t_1, \quad t_{1/2}$

- N has all moments finite
- t_ν has less than ν fractional moments

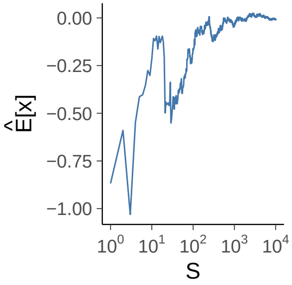
Simple example: $x \sim N$



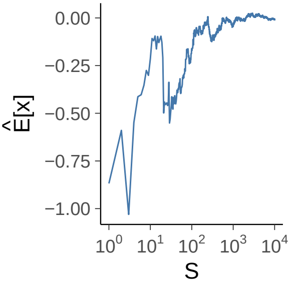
Simple example: $x \sim N$

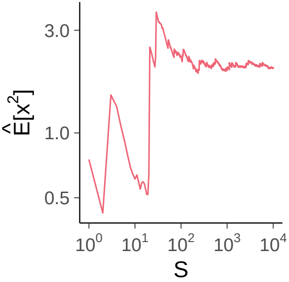


Simple example: $x \sim t_4, \quad t_2, \quad t_1$

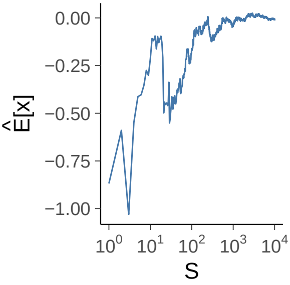


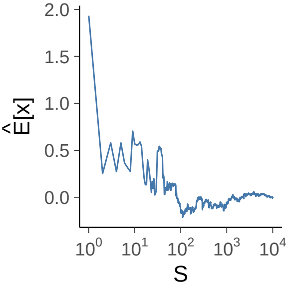
Simple example: $x \sim t_4, \quad t_2, \quad t_1$

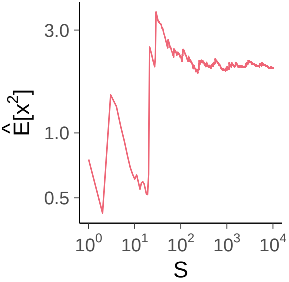




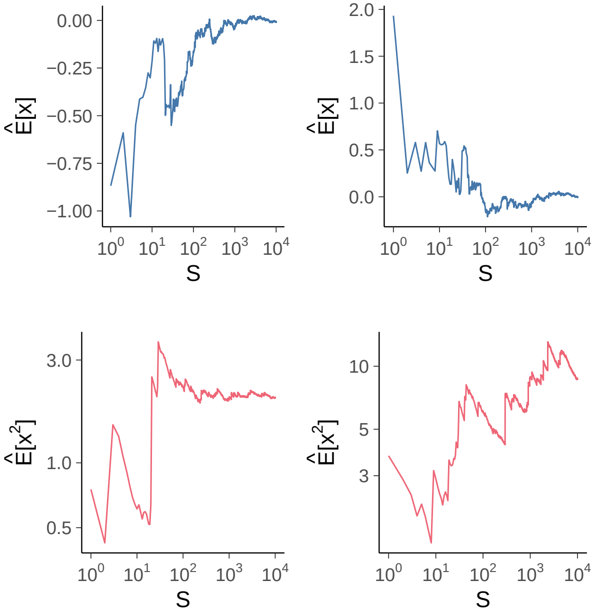
Simple example: $x \sim t_4, \quad t_2, \quad t_1$



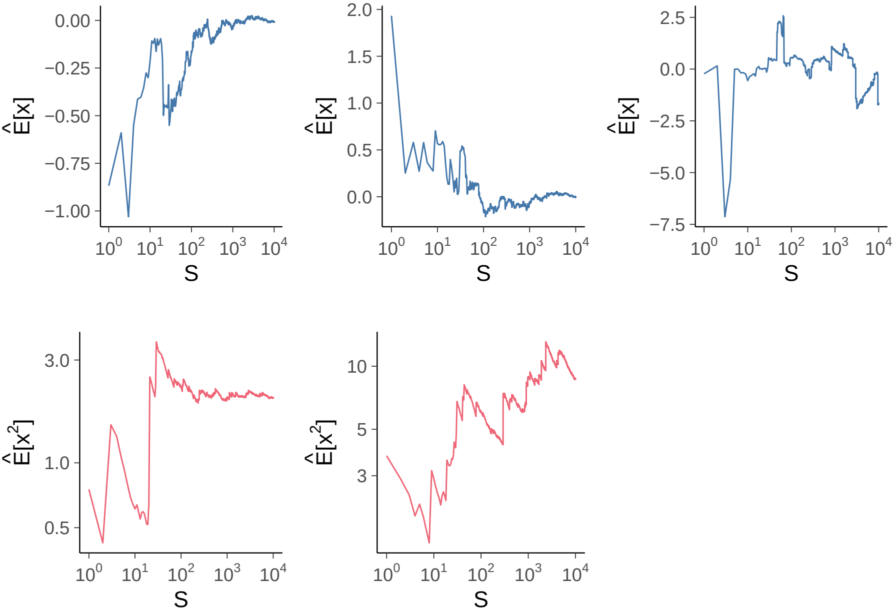




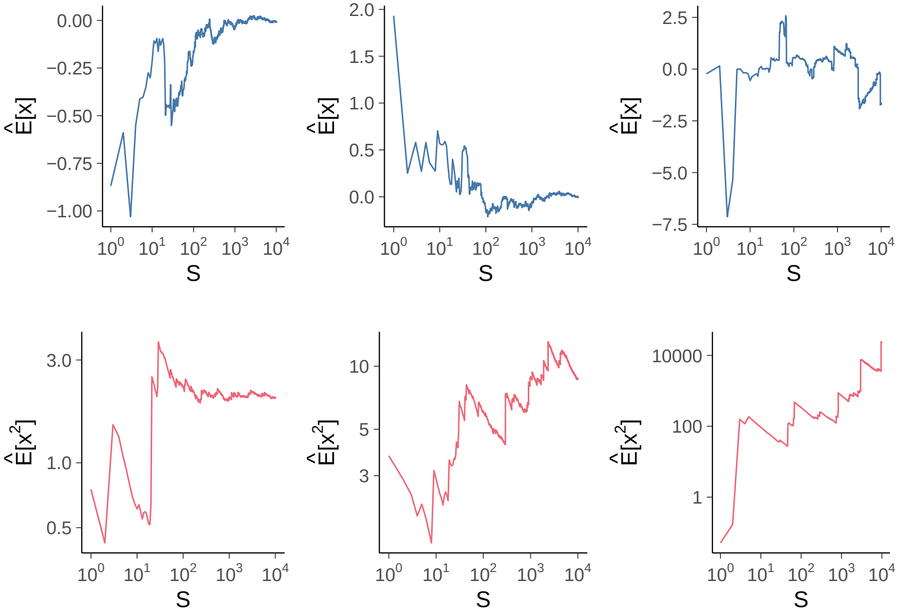
Simple example: $x \sim t_4, \quad t_2, \quad t_1$



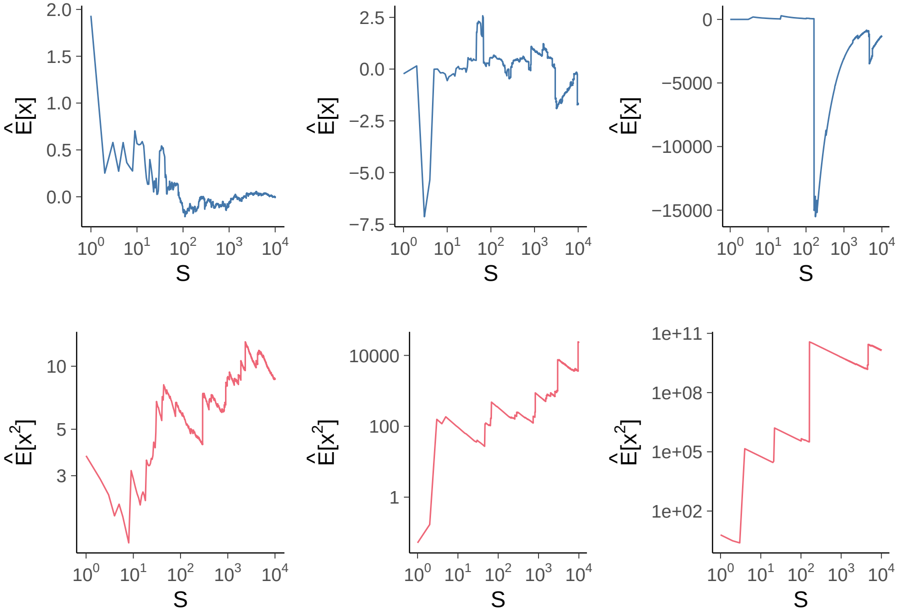
Simple example: $x \sim t_4, \quad t_2, \quad t_1$



Simple example: $x \sim t_4, \quad t_2, \quad t_1$

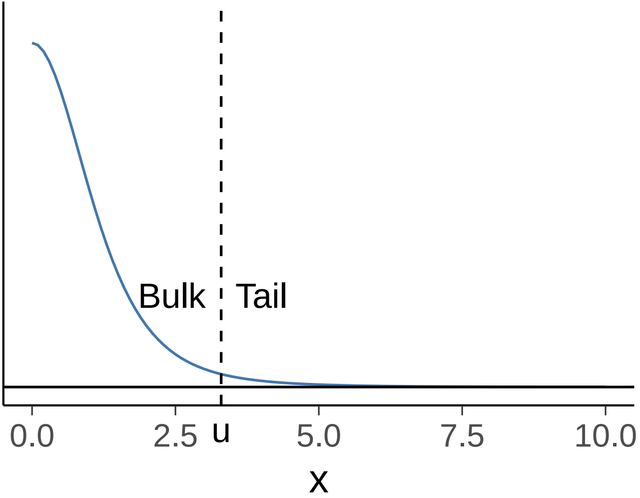


Simple example: $t_2, \quad t_1, \quad t_{1/2}$



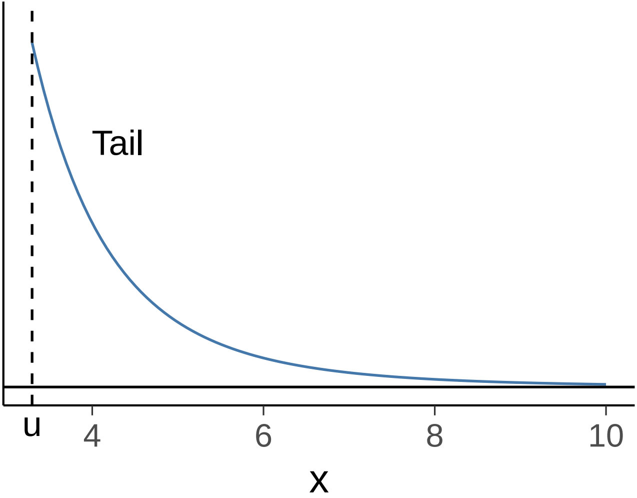
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)



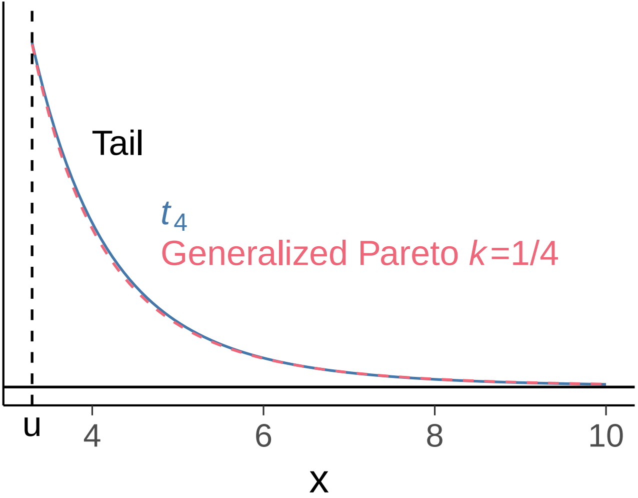
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)



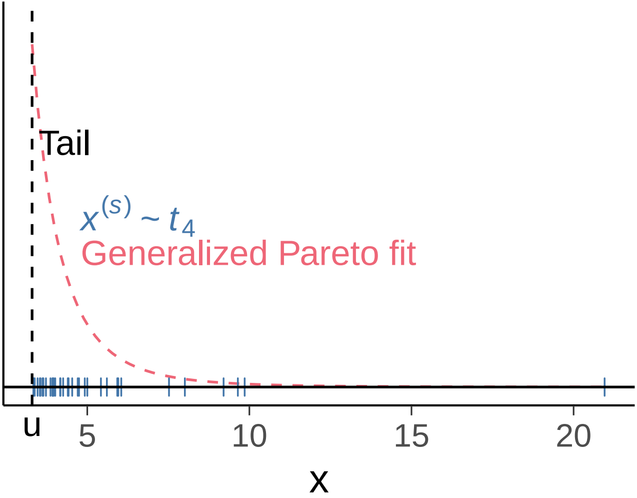
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)



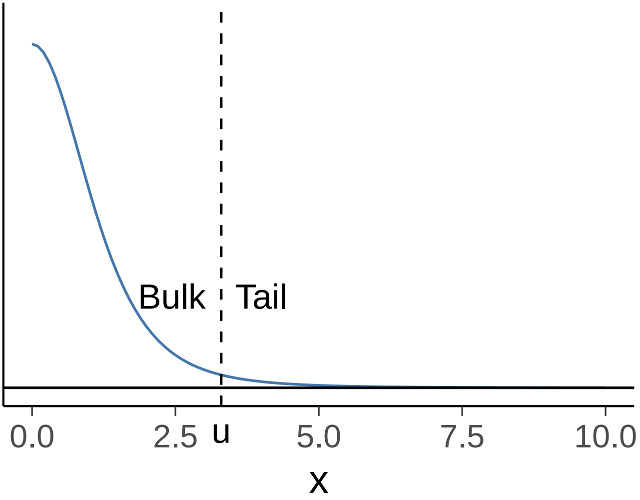
Pareto-$\hat{k}$ diagnostic

Pickands (1975): many distributions have tail ($x > u$) that is well approximated with Generalized Pareto distribution (GPD)

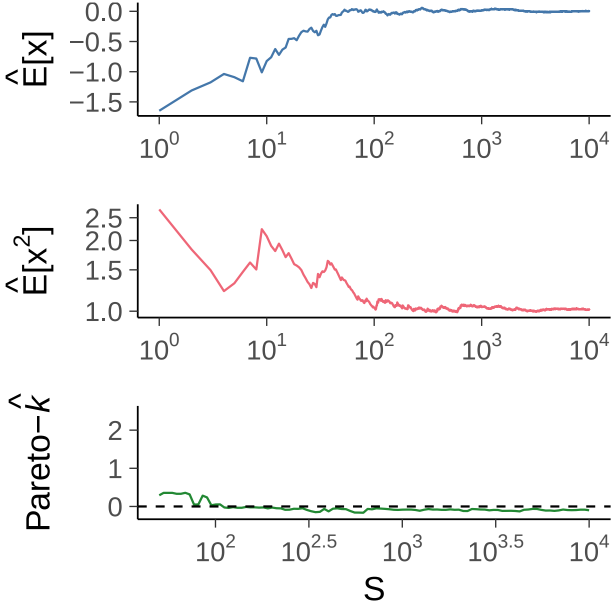


Pareto-$\hat{k}$ diagnostic

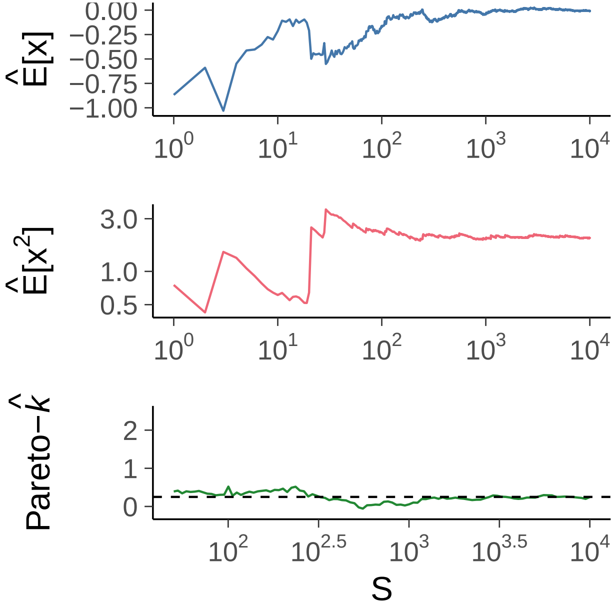
GPD has a shape parameter k,
and $1/k$ finite fractional moments



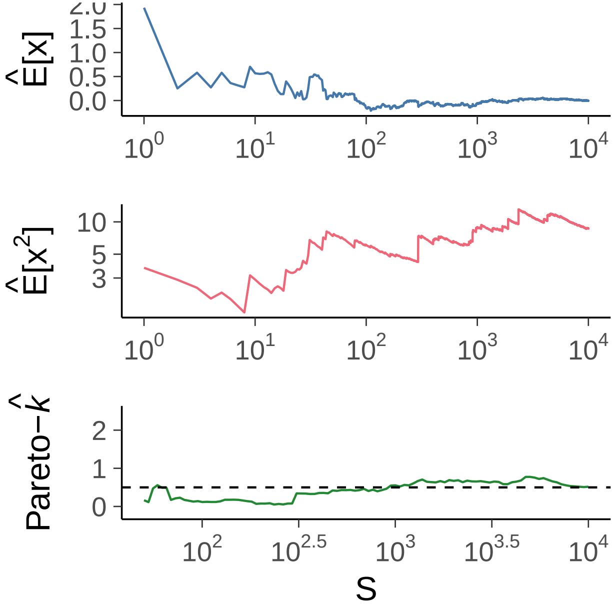
Pareto-$\hat{k}$ diagnostic: $x \sim N$



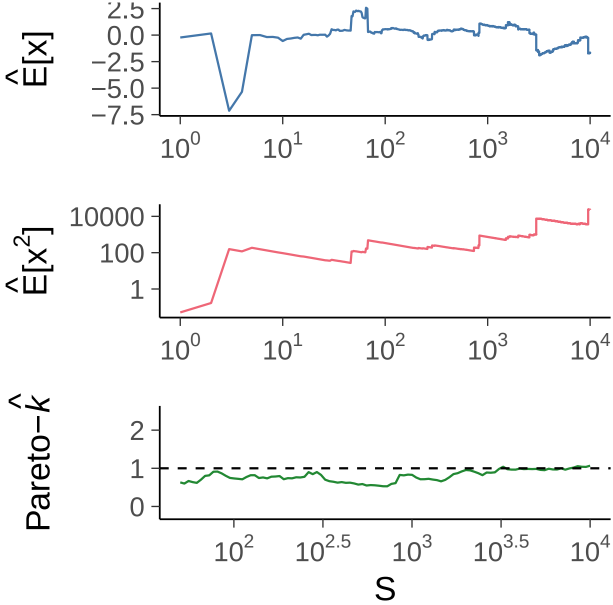
Pareto-$\hat{k}$ diagnostic: $x \sim t_4$



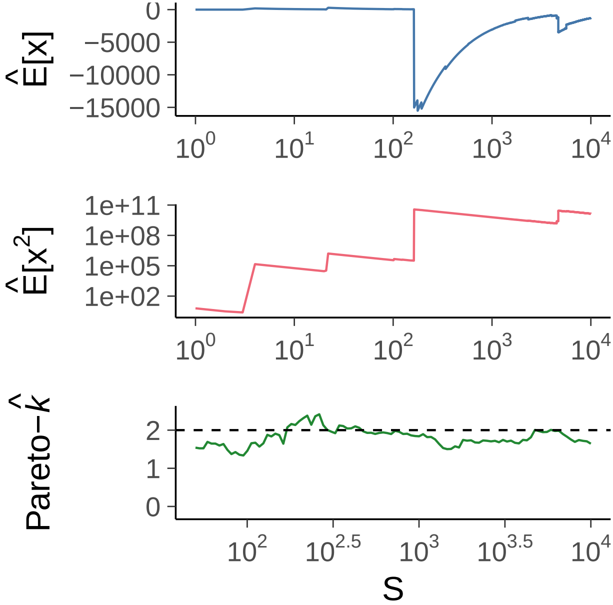
Pareto-$\hat{k}$ diagnostic: $x \sim t_2$



Pareto-$\hat{k}$ diagnostic: $x \sim t_1$

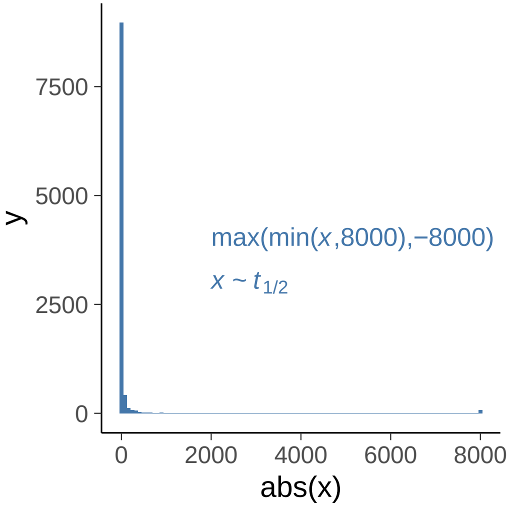


Pareto-$\hat{k}$ diagnostic: $x \sim t_{1/2}$

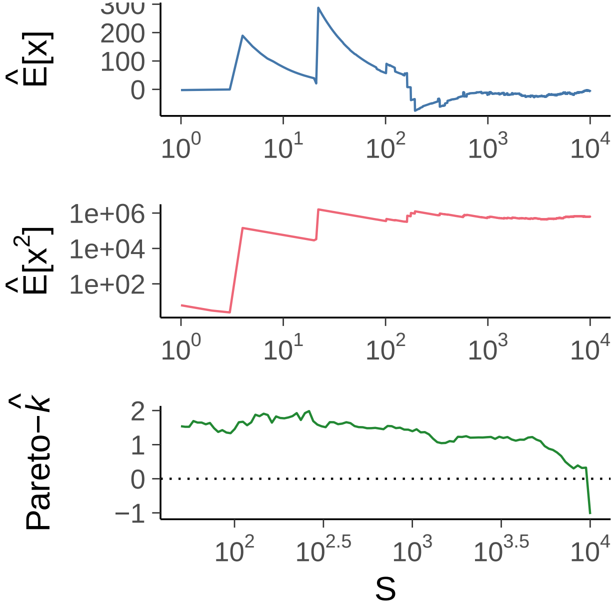


Pareto-$\hat{k}$ diagnostic is pre-asymptotic diagnostic

We can make estimates only based on what we have observed



Pareto-$\hat{k}$ diagnostic: thick-tailed bounded distribution



Thick-tailed bounded distributions in practice

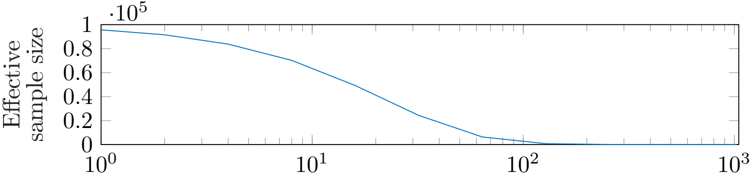
- Thick-tailed distributions are common in importance sampling and divergence estimation
  - if $g(\theta)$ has thinner tails than $p(\theta)$
    $\rightarrow w(\theta)$ is likely to have thick tails
  - if $g(\theta)$ has thicker tails than $p(\theta)$
    $\rightarrow w(\theta)$ is bounded, but that bound can be far

High-dimensional spaces are scary

$p(\theta) = N$, $g(\theta) = t_7$ which has thicker tails than normal, and thus ratios $w(\theta)$ are bounded. $S = 10^5$. D varies. Estimating the normalization.

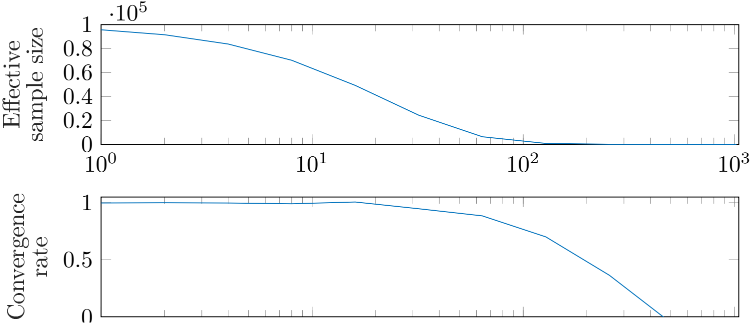
High-dimensional spaces are scary

$p(\theta) = N$, $g(\theta) = t_7$ which has thicker tails than normal, and thus ratios $w(\theta)$ are bounded. $S = 10^5$. D varies. Estimating the normalization.



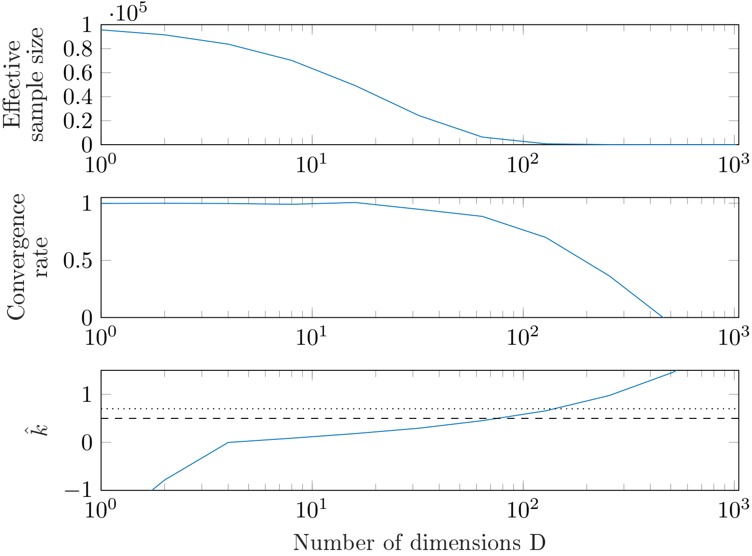
High-dimensional spaces are scary

$p(\theta) = N$, $g(\theta) = t_7$ which has thicker tails than normal, and thus ratios $w(\theta)$ are bounded. $S = 10^5$. D varies. Estimating the normalization.



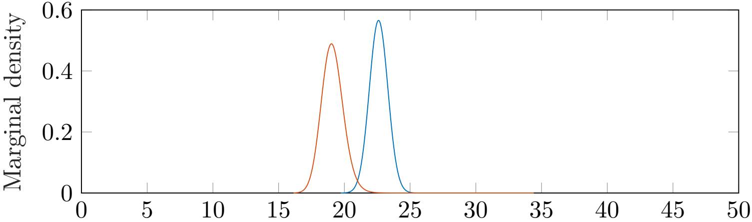
High-dimensional spaces are scary

$p(\theta) = N$, $g(\theta) = t_7$ which has thicker tails than normal, and thus ratios $w(\theta)$ are bounded. $S = 10^5$. D varies. Estimating the normalization.



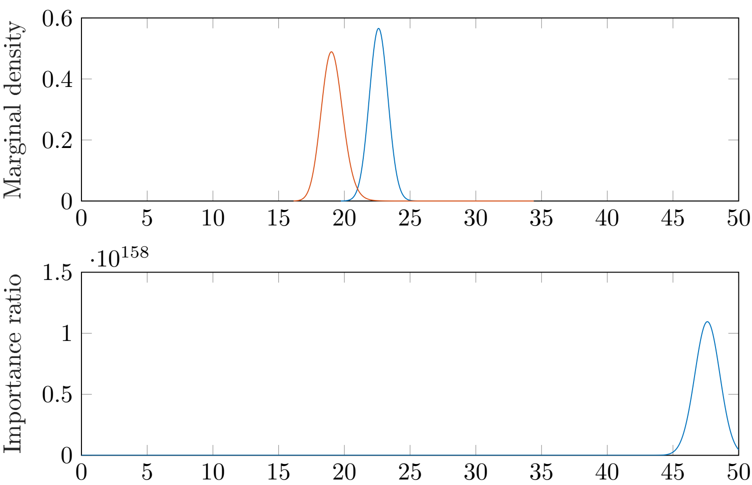
Concentration of measure and typical sets

Example continued: $p(\theta)$ = N (blue), $g(\theta) = t_7$ (red) with equal variance and thicker tails, and thus importance ratios are bounded. $S = 10^5$, $D = 512$.



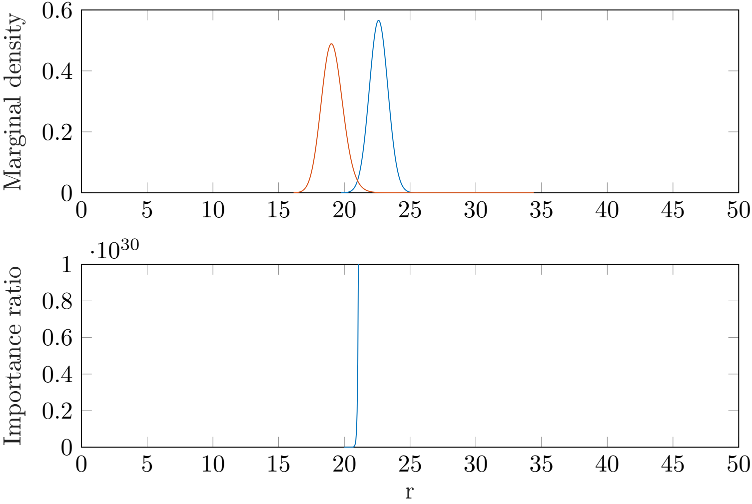
Concentration of measure and typical sets

Example continued: $p(\theta)$ = N (blue), $g(\theta) = t_7$ (red) with equal variance and thicker tails, and thus importance ratios are bounded. $S = 10^5$, $D = 512$.



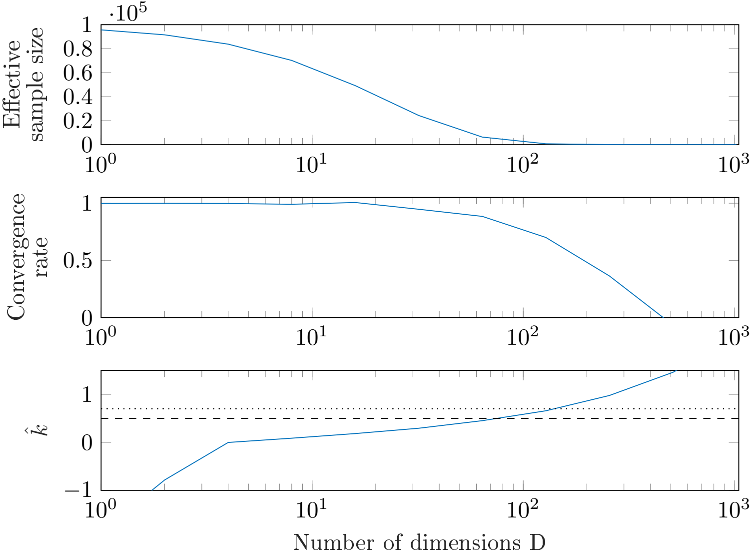
Concentration of measure and typical sets

Example continued: $p(\theta)$ = N (blue), $g(\theta) = t_7$ (red) with equal variance and thicker tails, and thus importance ratios are bounded.
$S = 10^5$, $D = 512$.



High-dimensional spaces are scary

$p(\theta) = N$, $g(\theta) = t_7$ which has thicker tails than normal, and thus ratios $w(\theta)$ are bounded. $S = 10^5$. D varies.



Pareto-$\hat{k}$ and convergence rate

- CLT says that to half the MCSE, need 4 times bigger S

Pareto-$\hat{k}$ and convergence rate

- CLT says that to half the MCSE, need 4 times bigger S
- If Pareto-$\hat{k} \approx 0.7$, to half the MCSE, need 10 times bigger S

Pareto-$\hat{k}$ and convergence rate

- CLT says that to half the MCSE, need 4 times bigger S
- If Pareto-$\hat{k} \approx 0.7$, to half the MCSE, need 10 times bigger S
- If Pareto-$\hat{k} > 1$, to half the MCSE, nothing helps

Pareto smoothed importance sampling (PSIS)

- Replace the largest observed ratios with expected ordered statistics of the fitted Pareto distribution
  - corresponds to modeling of the tail, and as usual, modeling reduces the noise

How many fractional moments are needed?

For finite variance

Objective	$f(w)$	Moments of w needed
IS normalization	w	2
Exclusive KL	$\log(w)$	δ
Inclusive KL	$w \log(w)$	$2 + \delta$
χ^2	$(w^2 - w)/2$	4
α-divergence	$(w^\alpha - w)/(\alpha(\alpha - 1))$	2α

How many fractional moments are needed?

For finite variance

Objective	$f(w)$	Moments of w needed
IS normalization	w	2
Exclusive KL	$\log(w)$	δ
Inclusive KL	$w\log(w)$	$2+\delta$
χ^2	$(w^2-w)/2$	4
α-divergence	$(w^\alpha-w)/(\alpha(\alpha-1))$	2α

For small error with practical sample sizes and Pareto smoothing

Objective	$f(w)$	Moments of w needed
IS normalization	w	1.4
Exclusive KL	$\log(w)$	δ
Inclusive KL	$w\log(w)$	$1.4+\delta$
χ^2	$(w^2-w)/2$	2.8
α-divergence	$(w^\alpha-w)/(\alpha(\alpha-1))$	1.4α

Estimating Pareto-$\hat{k}$

- Fast empirical profile Bayes quadrature estimate by Zhang and Stephens (2009)
  - excellent accuracy compared to exact Bayesian inference
  - see more in Vehtari, Simpson, Gelman, Yao & Gabry (2019)

Pareto-$\hat{k}$ diagnostic use cases

- Importance sampling
  - leave-one-out cross-validation (Vehtari et al., 2016, 2017; Bürkner at al, 2020)
  - Bayesian stacking (Yao et al., 2018, 2021, 2022)
  - leave-future-out cross-validation (Bürkner et al., 2020)
  - Bayesian bootstrap (Paananen et al, 2021, online appendix)
  - prior and likelihood sensitivity analysis (Kallioinen et al., 2021)
  - improving distributional approximations (Yao et al., 2018; Zhang et al., 2021; Dhaka et al., 2021)
  - implicitly adaptive importance sampling (Paananen et al., 2021)
- Stochastic optimization (Dhaka et al., 2020)
- Divergences and gradients in VI (Dhaka et al., 2021)
- MCMC (Paananen et al., 2021)

Co-authors and references

The main reference

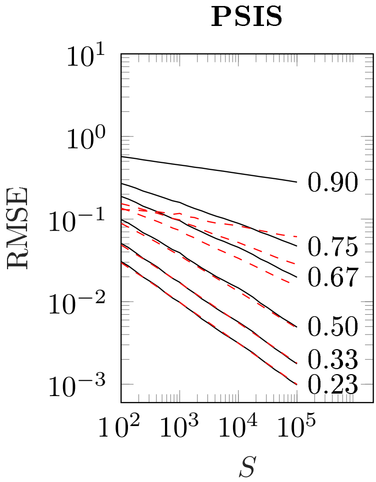
- Vehtari, Simpson, Gelman, Yao, and Gabry (2019). Pareto smoothed importance sampling. *arXiv:1507.02646v6*.

Use cases

- Bürkner, Gabry & Vehtari (2020). Approximate leave-future-out cross-validation for time series models. *J Stat Comp and Simul*, 90(14):2499–2523.
- Dhaka, Catalina, Andersen, Magnusson, Huggins & Vehtari (2020). Robust, accurate stochastic optimization for variational inference. *NeurIPS 2020*, 33:10961–10973.
- Dhaka, Catalina, Welandawe, Andersen, Huggins & Vehtari (2021). Challenges and opportunities in high-dimensional variational inference. *NeurIPS 2021*, to appear.
- Kallionen, Paananen, Bürkner & Vehtari (2021). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *arXiv preprint arXiv:2107.14054*
- Paananen, Piironen, Bürkner, and Vehtari (2021). Implicitly adaptive importance sampling. *Statistics and Computing*, 31, 16.
- Vehtari, Gelman, and Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.
- Yao, Vehtari, Simpson, and Gelman (2018). Yes, but Did It Work?: Evaluating Variational Inference. *35th ICML, PMLR*, 80:5577–5586.
- Yao, Vehtari, Simpson & Gelman (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917-1003,
- Yao, Pirš, Vehtari & Gelman (2021). Bayesian hierarchical stacking: Some models are (somewhere) useful. Bayesian Analysis, doi:10.1214/21-BA1287.
- Yao, Vehtari & Gelman (2022). Stacking for non-mixing Bayesian computations: The curse and blessing of multimodal posteriors. *JMLR*, accepted for publication.

Pareto smoothed importance sampling (PSIS)

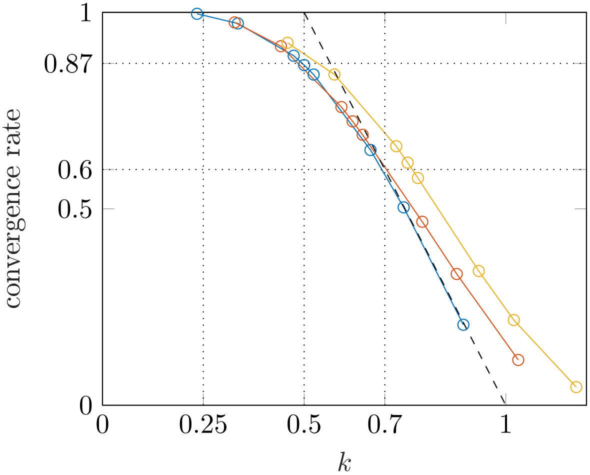
Empirical comparison to the theory



black line = RMSE, red dashed line = MCSE estimate

Pareto-$\hat{k}$ and convergence rate

Variance of the estimate goes down as $S^{-\alpha}$, where α is convergence rate



blue = 0th moment, red = 1st moment, yellow = 2nd moment