State-Dependent Importance Sampling Schemes via Minimum Cross-Entropy

Thomas Taimre

ttaimre@maths.uq.edu.au

Department of Mathematics

The University of Queensland

Australia

Outline of Talk

- Introduction
- Minimum Cross-Entropy
- Examples & Numerics
- Discussion

Importance Sampling — Notation

- d-dimensional state space $\mathcal{X}$.
- Reference density f on $\mathcal{X}$.
- Performance function $H(\cdot;\gamma) : \mathcal{X} \to \mathbb{R}$.
- Interested in computing

$$\ell = \mathbb{E}_f \left[H(\mathbf{X};\gamma) \right] \, .$$

IS Notation Continued

- Denote our IS density as g.
- Quantity of interest can be expressed as

$$\ell = \mathbb{E}_g \left[H(\mathbf{X};\gamma) \frac{f(\mathbf{X})}{g(\mathbf{X})} \right] .$$

- We will estimate ℓ using the likelihood ratio estimator: Given $\mathbf{X}_1, \ldots \mathbf{X}_N \overset{\text{i.i.d.}}{\sim} g$

$$\widehat{\ell}_{\text{LR}} = \frac{1}{N} \sum_{k=1}^{N} H(\mathbf{X}_k;\gamma) \frac{f(\mathbf{X}_k)}{g(\mathbf{X}_k)} .$$

IS Continued

- Recall the *minimum variance IS density*:

$$g^*(\mathbf{x}) = \frac{|H(\mathbf{x};\gamma)|\ f(\mathbf{x})}{\mathbb{E}_f\left[|H(\mathbf{X};\gamma)|\right]}\ .$$

- In this talk, g^* will be the *target* IS density.
- Usually, g^* is unattainable directly.
- Can think of g as our best proxy for g^*.
- Often, g is restricted to some manageable parametric family (cf. Cross–Entropy method).

Minimum Cross-Entropy

Generic minimum cross-entropy (MCE) program:

$$\inf_g \mathbb{E}_g \left[\ln \left(\frac{g(\mathbf{X})}{f(\mathbf{X})} \right) \right]$$

subject to

$$\mathbb{E}_g \left[C_j(\mathbf{X}) \right] = c_j, \; j = 1, 2, \ldots, m\,,$$
$$\mathbb{E}_g \left[C_j(\mathbf{X}) \right] \geqslant c_j, \; j = m+1, m+2, \ldots, M\,,$$

and

$$\int g(\mathbf{x}) \mu(d\mathbf{x}) = 1\,.$$

MCE Solution

Solution given by

$$g(\mathbf{x}) = f(\mathbf{x})\mathrm{e}^{\lambda_0 + \sum_{i=1}^{M} \lambda_i C_i(\mathbf{x})}\,,$$

where the $\{\lambda_i\}$ solve the dual program

$$\sup_{\lambda_0,\lambda_1,\ldots,\lambda_M} \left[\lambda_0 + \sum_{i=1}^{M} \lambda_i c_i - \mathrm{e}^{\lambda_0}\mathbb{E}_f\left[\mathrm{e}^{\sum_{j=1}^{M} \lambda_j C_j(\mathbf{X})}\right]\right]$$

subject to the constraints $\lambda_j \geqslant 0$ for $j = m+1, \ldots, M$.

Sequential IS

- For certain models f, it is natural to consider $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots)$ as a sequence of states (eg. discrete-time Markov processes).
- In such cases, it is easy to think of g as a sequence of IS densities, each acting on the current state and possibly depending on the entire history.
- Via the chain rule, can write

$$g(\mathbf{x}) = g(\mathbf{x}_1)g(\mathbf{x}_2|\mathbf{x}_1)g(\mathbf{x}_3|\mathbf{x}_2, \mathbf{x}_1) \cdots g(\mathbf{x}_n|\mathbf{x}_{n-1}, \dots, \mathbf{x}_1) \,.$$

- Now, we obtain this sequence of conditional IS densities via MCE.

Sequential MCE

- The idea is to sample each state $\mathbf{X}_k$ sequentially; and:
- To *re-solve* the MCE program *conditional* on the entire sampling history, $\mathbf{x}_1, \ldots, \mathbf{x}_k$.
- This in turn updates g, given the current sample path.

Sequential MCE

- Suppose that we have sampled $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}$, so that the current state to be realised is $\mathbf{X}_k$.
- We solve the MCE program for $g(\mathbf{x}_k, \ldots, \mathbf{x}_n \,|\, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_1)$. Note that the constraints in the MCE program now incorporate $\mathbf{x}_{k-1}, \ldots, \mathbf{x}_1$.
- Via the chain rule,

$$g(\mathbf{x}_k, \ldots, \mathbf{x}_n \,|\, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_1) = g(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_1) \times g(\mathbf{x}_{k+1}, \ldots, \mathbf{x}_n \,|\, \mathbf{x}_k, \ldots, \mathbf{x}_1)\,.$$

- We sample from $g(\mathbf{x}_k \,|\, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_1)$, and then update the MCE program and repeat the process.

Example: I.I.D. Sums

- Let $\{X_k\}$, $k = 1, 2, \ldots$ be a collection of i.i.d. random variables with common pdf f.
- Define $S_n = \sum_{k=1}^{n} X_k$ for $n = 1, 2, \ldots,$ with $S_0 = 0$.
- Problem is to estimate tail probabilities of the form

$$\ell = \mathbb{P}_f(S_n > \alpha n)\,,$$

for *fixed* α and different n.

- In this case $H(\mathbf{X}; n) = I_{\{\sum_{k=1}^{n} X_k > \alpha n\}}$.
- Hence g^* is the density f conditional on $\{S_n > \alpha n\}$.

MCE for the Example

- We will impose a single *inequality* constraint in the MCE program, namely

$$\mathbb{E}_g\,[C(\mathbf{X})] \geqslant \alpha n\,,$$

where

$$C(\mathbf{X}) = \sum_{k=1}^{n} X_k\,.$$

- Hence, the MCE program finds g as close as possible to f in the Kullback-Leibler CE sense, while ensuring that $\mathbb{E}_g[S_n] \geqslant \alpha n$.

MCE Solution for the Example

- Corresponding dual program given by

$$\sup_{\lambda_0,\lambda_1} \left[\lambda_0 + \lambda_1(\alpha n - s_{k-1}) - \mathrm{e}^{\lambda_0}\mathbb{E}_f \left[\mathrm{e}^{\lambda_1(X_k+\cdots+X_n)}\right]\right]$$

subject to the constraint that $\lambda_1 \geqslant 0$.

- Solution to the MCE program given by

$$g(x_k,\ldots,x_n \,|\, x_{k-1},\ldots,x_1) = f(x_k,\ldots,x_n)\mathrm{e}^{\lambda_0+\lambda_1\sum_{j=k}^n x_j} \;.$$

- We will sample from the (ET) conditional

$$g(x_k \,|\, x_{k-1},\ldots,x_1) = f(x_k)\mathrm{e}^{\widetilde{\lambda}_0+\lambda_1 x_k} \;.$$

Example: Gaussian Case

- If the X_k are i.i.d. $\mathsf{N}(\mu, \sigma^2)$ distributed, the MGF of X_k is given by

$$\mathbb{E}_f\left[\mathrm{e}^{\lambda_1 X_k}\right] = \mathrm{e}^{\frac{1}{2}\lambda_1(\lambda_1\sigma^2 + 2\mu)}\,.$$

- Hence the appropriate dual is given by

$$\sup_{\lambda_0,\lambda_1}\left[\lambda_0 + \lambda_1(\alpha n - s_{k-1}) - \mathrm{e}^{\lambda_0}\left(\mathrm{e}^{\frac{1}{2}\lambda_1(\lambda_1\sigma^2+2\mu)}\right)^{n-k+1}\right],$$

subject to $\lambda_1 \geqslant 0$.

Gaussian Case Continued

- The solution yields that the conditional distribution corresponding to the next increment, X_k, is Gaussian with mean

$$\begin{cases} \frac{\alpha n - s_{k-1}}{n-k+1} & \frac{\alpha n - s_{k-1}}{n-k+1} \geqslant \mu \\ \mu & \text{otherwise} \end{cases}$$

and variance σ^2.

- Interpretation: change of measure places next increment's mean on line connecting current state to target level αn, $unless$ expected trajectory from the current point is already $\geqslant \alpha n$, in which case no change of measure is performed.

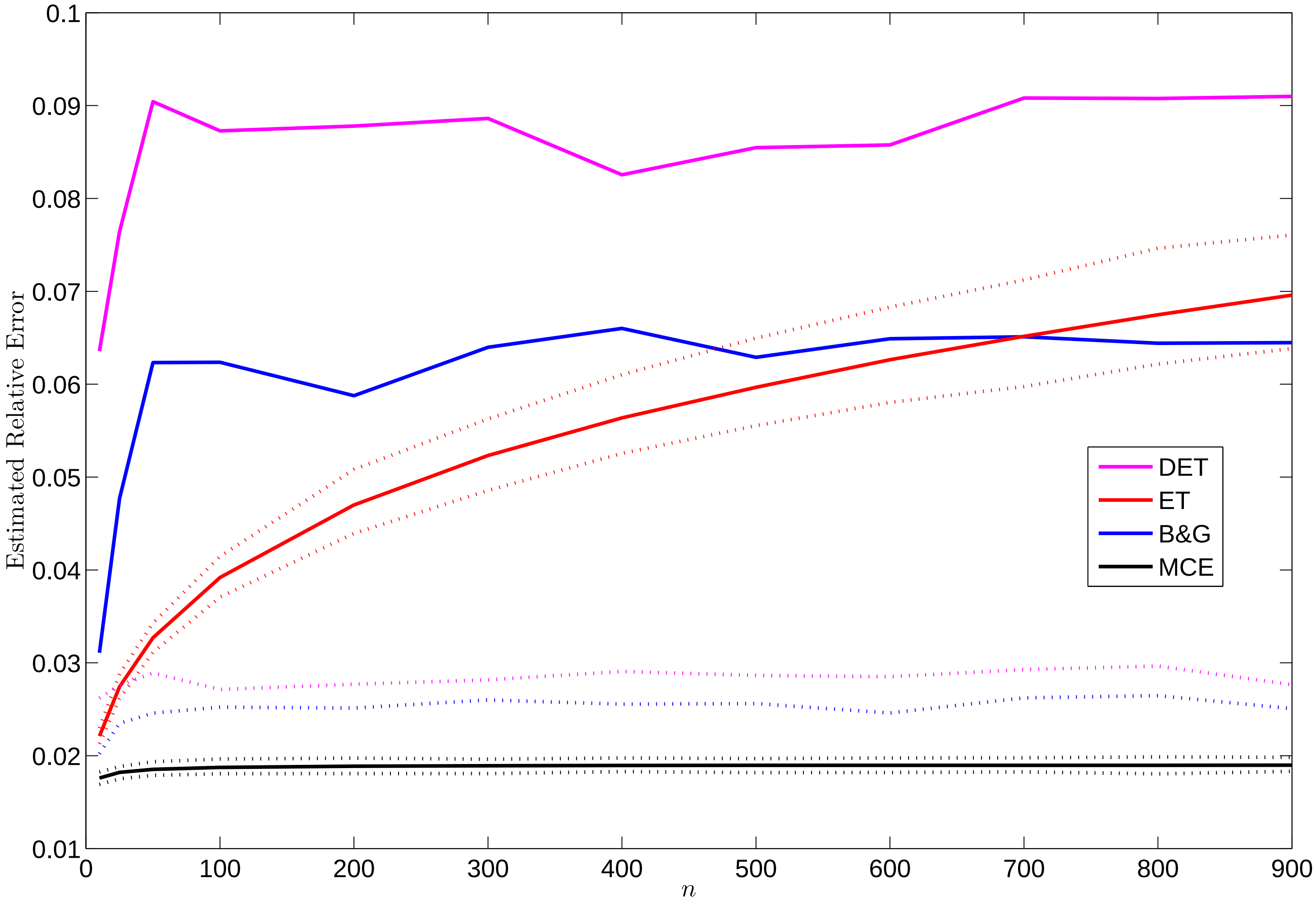
Gaussian Case: Numerics

- Suppose X_k under f are standard Normal increments ($\mu = 0$, $\sigma = 1$).
- Level to be reached: $\alpha = \frac{2}{3}$; so $\ell = \mathbb{P}_f(S_n > \frac{2}{3}n)$.
- Compare sequential MCE with *inequality* constraint to:
  - MCE with *equality* (i.i.d. ET). (Sets $\mathbb{E}_g[X_k] = \alpha$.)
  - sequential MCE with *equality* constraint (dynamic ET).
  - Algorithm of Blanchet & Glynn (2006) (on next slide).
- Use $N = 5 \cdot 10^3$ samples per LR estimate, $\widehat{\ell}_{\text{LR}}$.
- Obtain $1{,}000$ independent estimates. Give min, mean, and max statistics for RE and logarithmic efficiency.

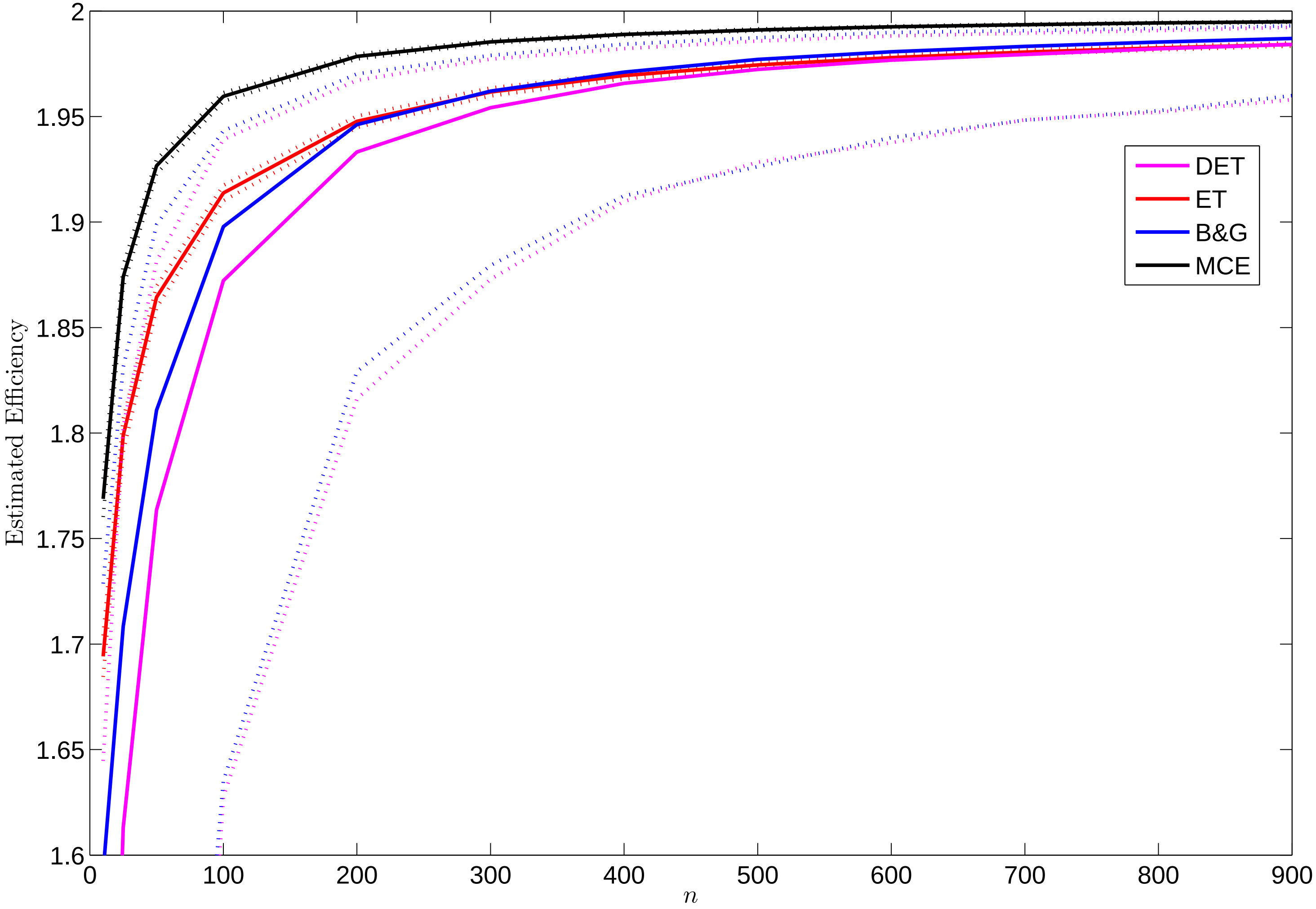
Gaussian Case: Algorithm of B&G

- Blanchet & Glynn (2006) algorithm (for $X_k \sim \mathsf{N}(0,1)$).
  - Set $k = 1$ and $s_{k-1} = 0$.
  - If $k < n$, sample X_k from $\mathsf{N}\left(\frac{\alpha n - s_{k-1}}{n-k}, 1 + \frac{1}{n-k}\right)$.
    Set $s_k = s_{k-1} + x_k$, $k = k + 1$, and repeat.
  - Otherwise if $k = n$, sample directly from the distribution of X_n given $\{X_n + s_{n-1} > \alpha n\}$.
- This was shown to give bounded relative error as $n \to \infty$.
- In contrast, we have not yet shown optimality, despite the following suggestive numerics.

Gaussian Increments RE



Gaussian Increments Efficiency



Two Sided Example

- Let $\{X_k\}$, $k = 1, 2, \ldots$ be a collection of i.i.d. random variables with common pdf f.
- Define $S_n = \sum_{k=1}^{n} X_k$ for $n = 1, 2, \ldots$, with $S_0 = 0$.
- Problem is to estimate two-sided probabilities of the form

$$\ell = \mathbb{P}_f(\{S_n \geqslant \alpha n\} \cup \{S_n \leqslant -(1+\varepsilon)\alpha n\}),$$

for *fixed* (α, ε), and varying n.

MCE for the Example

- Augment the problem with independent $Y \sim \mathsf{Ber}(p)$ (under f).
- Again, we will impose a single inequality constraint in the MCE program:

$$\mathbb{E}_g\left[C(\mathbf{X})\right] \geqslant 0\,,$$

where

$$C(\mathbf{X}) = Y\left(S_n - \alpha n\right) - (1 - Y)\left(S_n + (1+\varepsilon)\alpha n\right)\;.$$

- As before, conditionals $g(x_k \,|\, x_{k-1}, \ldots, x_1, y)$ are ET.
- However, here twisting is toward the level determined by outcome of Y.

Example: Gaussian Case

- If $p = 1/2$, $X_k \sim \mathsf{N}(0,1)$, then under g, $Y \sim \mathsf{Ber}(\widetilde{p})$, where

$$\widetilde{p} = (1 + \mathrm{e}^{\varepsilon z^*})^{-1}$$

and z^* solves

$$(z + (1+\varepsilon)\alpha^2 n)\mathrm{e}^{\varepsilon z} + (z + \alpha^2 n) = 0\,.$$

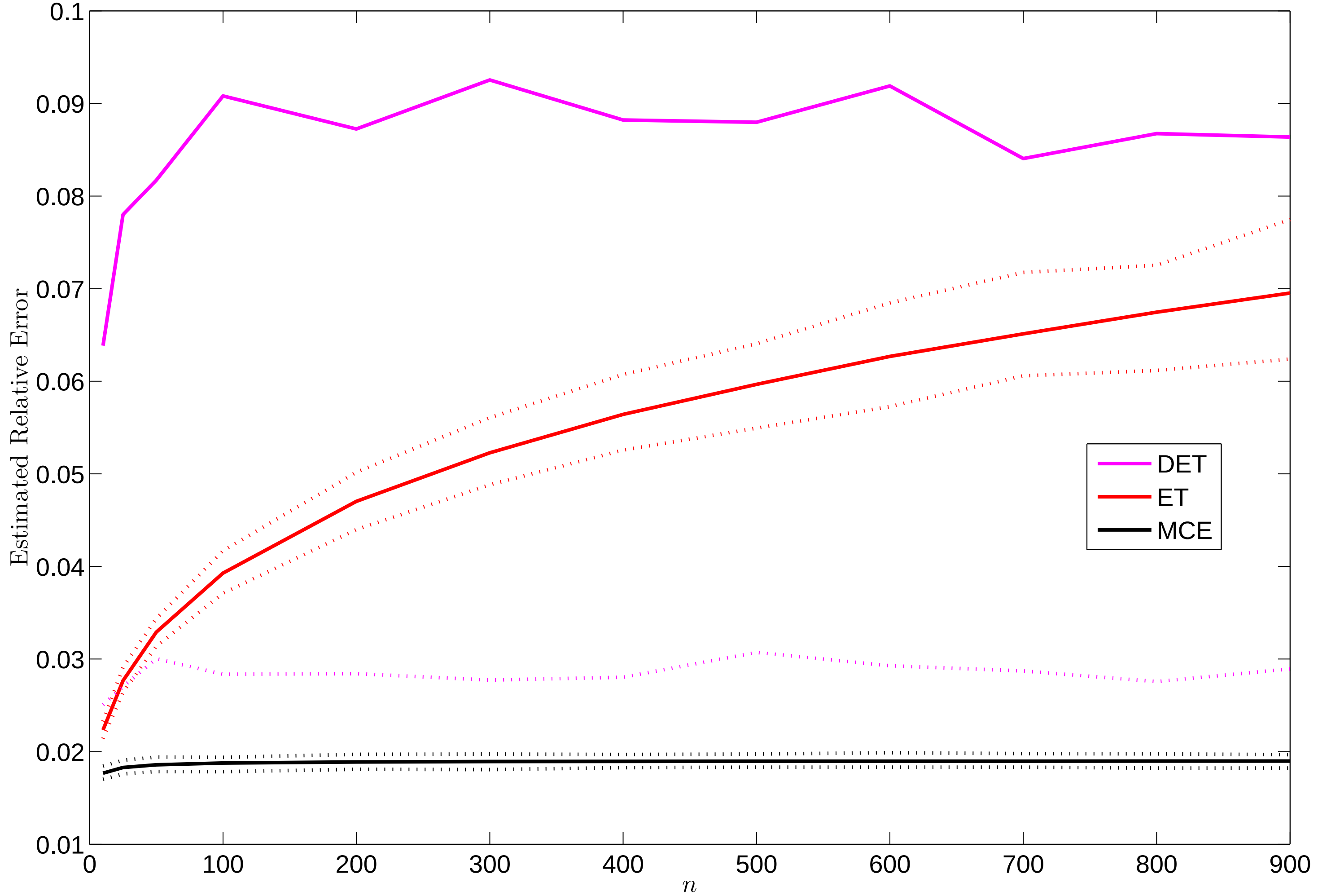
- The solution subsequently has: $X_k \sim \mathsf{N}(\widetilde{\mu}_k, \sigma^2)$, with

$$\widetilde{\mu}_k = \begin{cases} \frac{\alpha n - s_{k-1}}{n-k+1} & y = 1,\ \frac{\alpha n - s_{k-1}}{(n-k+1)} \geqslant \mu \\ -\frac{(1+\varepsilon)\alpha n + s_{k-1}}{n-k+1} & y = 0,\ -\frac{(1+\varepsilon)\alpha n + s_{k-1}}{n-k+1} \leqslant \mu \\ \mu & \text{otherwise}\,. \end{cases}$$

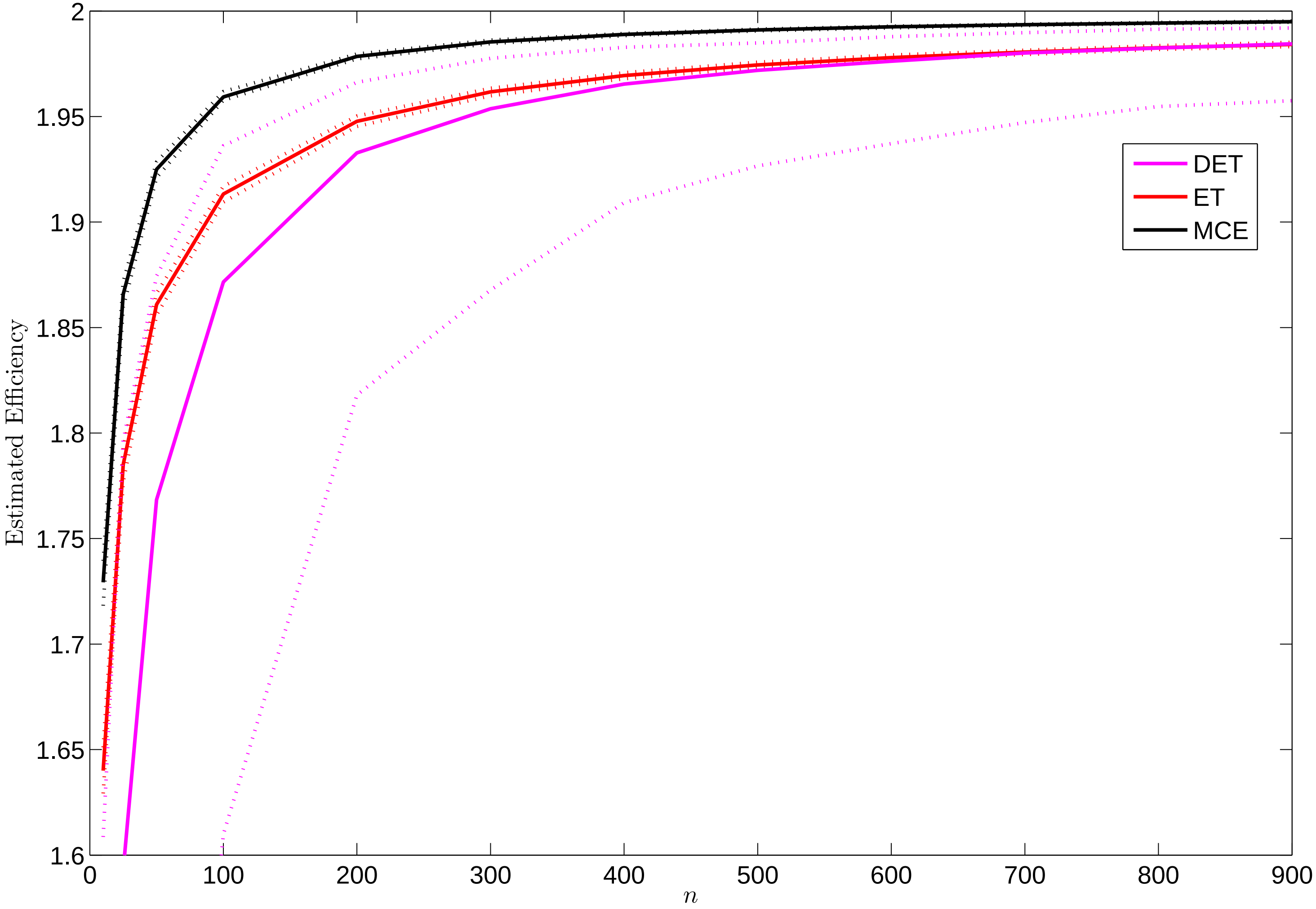
Gaussian Numerics II

- Again, take X_k as standard Normal ($\mu = 0$, $\sigma = 1$).
- Levels: $\alpha = \frac{2}{3}$, and $\varepsilon = 0.05$.
- Compare sequential MCE with *inequality* constraint to:
  - MCE with *equality* (mixture of i.i.d. ET).
  - sequential MCE with *equality* constraint (mixture of dynamic ET).
- Use $N = 5 \cdot 10^3$ samples per LR estimate, $\widehat{\ell}_{\mathrm{LR}}$.
- Obtain $1{,}000$ independent estimates. Give min, mean, and max statistics for RE and logarithmic efficiency.

Gaussian Increments RE II



Gaussian Increments Efficiency II



Discussion

- This MCE scheme only applies in cases where

$$\mathbb{E}_f \left[\mathrm{e}^{\lambda_k C_k(\mathbf{X})} \right]$$

  is defined for all constraints C_k for some corresponding λ_k.

- In particular, with $C(\mathbf{X}) = \sum_{k=1}^{n} X_k$ as in the examples, the program is only applicable when f is light-tailed, since the above involves the MGFs of the increments under f.
  - To overcome this, one could modify the constraints (eg. hazard rate twisting); or
  - Change the divergence measure from KL to some other.

Discussion Continued

- Solving the sequence of MCE programs gives a structured way to obtain state- and time-dependent IS schemes.
- Further, the use of inequality constraints ensures that each constraint is only imposed when necessary.

Acknowledgements

- Ad Ridder, Zdravko Botev, Dirk Kroese.
- ARC Centre of Excellence for Mathematics and Statistics of Complex Systems, and the Commonwealth Government of Australia, for funding.

References

- Blanchet, J. and Glynn, P. (2006) Strongly Efficient Estimators for Light-tailed Sums *Proc. Valuetools06*.
- de Boer, P. T. (2000) *Analysis and Efficient Simulation of Queueing Models of Telecommunication Systems*. PhD Thesis, Universiteit Twente, October 2000.
- Dupuis, P. and Wang, H. (2005) Dynamic Importance Sampling for Uniformly Recurrent Markov Chains. *Ann. Appl. Probab.* 15, 1–38.
- Rubinstein, R. Y. (2005) A Stochastic Minimum Cross-Entropy Method for Combinatorial Optimization and Rare-event Estimation. *Meth. Comp. Appl. Prob.* 7, 5–50.