Improving the Convergence of Iterative Importance Sampling for Computing Upper and Lower Expectations

Thomas Fetz, Unit for Engineering Mathematics, University of Innsbruck, Austria

universität innsbruck

t.math

Problem statement

Given: A function $h : A \subseteq \mathbb{R}^d \to \mathbb{R}$ which is **expensive** to evaluate and a family $\{f_t\}_{t\in\mathcal{T}}$ of density functions.

Aim: Efficient computation of **lower expectation** $\theta_* = \min_{t\in\mathcal{T}} \theta(t)$, optimising w.r.t. $\theta(t) = \int_A h(x)\, f_t(x)\mathrm{d}x$.

Method: Using **Monte-Carlo simulation** and **importance sampling** to get an estimate $\hat{\theta}$ which is **cheap to evaluate** in the optimising algorithm. The estimate $\hat{\theta}_*$ of θ_* is improved applying fixed point iteration. Here, we are focussed on the **improvement of the convergence of the fixed point iteration reusing previous results of the iteration**.

Simple numerical example for visualising

- **Function** h: $h(x) = \mathbb{1}_D(x)$, $D = (-\infty,-2]\cup[2,\infty)$.
- **Family of density functions:** $f_t \sim N(\mu(t),\sigma^2(t))$, mean $\mu(t) = t$, variance $\sigma^2(t) = 4$, $t \in \mathcal{T} = [-7,7]$.
- **Exact result:** $\theta_* = 0.3173$ at $t_* = 0$. (plot of θ below)

Methods for estimating function θ

Three steps of Monte Carlo simulation

1. Set of **random numbers** $\Omega_t = \{U_1, U_2, \dots, U_n\}$, $U_i = (V_i, W_i)$, $V_i, W_i \sim \mathcal{U}([0,1])$ i.i.d.
2. **Sample points** $\sim N(\mu(t),\sigma^2(t))$ (Box-Muller)
   $x_t(U_i) = x_t(V_i,W_i) = \mu(t) + \sigma(t)\cdot\sqrt{-2\ln V_i}\cdot\cos(2\pi W_i)$.
3. **Monte Carlo simulation w.r.t.** Ω_t**:**
   $\hat{\theta}_{\Omega_t}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_D(x_t(V_i,W_i))$, estimate of $\theta(t)$.

Approach 1, expensive

- **Different** sets of random numbers (V_i,W_i) for each $t \in \mathcal{T}$.
- **Different** sets of n sample points $x_t(V_i,W_i)$ for each $t \in \mathcal{T}$.
- Resulting in a **noisy** estimate $\hat{\theta}_{\Omega(\cdot)}$ of θ using MC.
- **Expensive** because of the evaluation of h for different sets of sample points. **Bad** for optimising.

Approach 2, only a little bit better

- **One single** set of n random numbers $U_i = (V_i,W_i)$ for all parameter values $t \in \mathcal{T}$, $\Omega_t = \Omega$.
- **Different** sets of n sample points $x_t(V_i,W_i)$ for each $t \in \mathcal{T}$.
- Estimate $\hat{\theta}_\Omega$ of θ is now a **step function**.
- Again expensive and difficult to optimize.

Approach 3, cheap but inaccurate

- **One single** set of n random numbers $U_i = (V_i,W_i)$.
- **One single** set of n sample points $x_t(V_i,W_i)$ for all parameter value t. **Reweighting the sample** using
  $$\int_A \mathbb{1}_D(x)\, f_t(x)\mathrm{d}x = \int_A \mathbb{1}_D(x)\,\frac{f_t(x)}{f_s^R(x)}\, f_s^R(x)\mathrm{d}x.$$
- **Importance sampling density** f_s^R for f_s, $s \in \mathcal{T}$. (here in the example: $f_s^R := f_s$ and $s = 6$)
- **Weights** $w_{st}(x) = f_t(x)/f_s^R(x)$. (classical: $w_s(x) = f_s(x)/f_s^R(x)$).

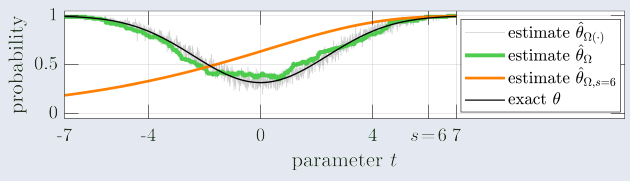
  $f_t \to f_t^R$ ↘ $f_s \to f_s^R$
- **MC:** $\hat{\theta}_{\Omega,s}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_D(x_s(V_i,W_i))\cdot w_{st}(x_s(V_i,W_i))$. (classical importance sampling if $s = t$)
- **Normalisation:** $\hat{\theta}_{\Omega,s}$ divided by $\sum_{i=0}^{n} w_{st}(x_s(V_i,W_i))$
- Estimate $\hat{\theta}_{\Omega,s}$ is **continuous**. Good for optimising!
- **Very cheap** to evaluate $\hat{\theta}_{\Omega,s^{(i)}}$.
- Minima depending on s: $\tau_{*\Omega}(s) = \arg\min_{t\in\mathcal{T}} \hat{\theta}_{\Omega,s}(t)$.

Approach 4, fixed point iteration

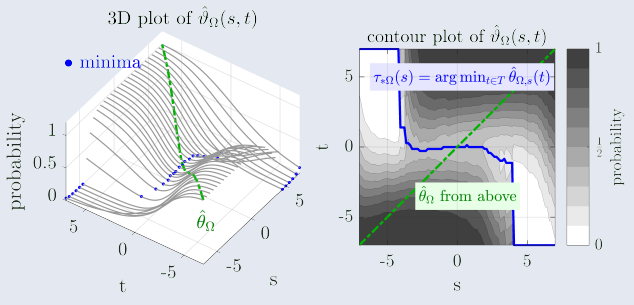
$$s^{(k+1)} = \tau_{*\Omega}(s^{(k)}) = \arg\min_{t\in\mathcal{T}} \hat{\theta}_{\Omega,s^{(k)}}(t), \quad k = 1,2,\dots$$

Visualising estimates

- Exact θ and estimates $\hat{\theta}_{\Omega(\cdot)}$, $\hat{\theta}_\Omega$, $\hat{\theta}_{\Omega,s=6}$ for $n = 100$:



- $\hat{\vartheta}_\Omega(s,t) = \hat{\theta}_{\Omega,s}(t)$, $\hat{\theta}_\Omega$ and minima $\tau_{*\Omega}$ for $n = 1000$:



Combining previous results of iteration

Observations

- Estimates $\hat{\theta}_{\Omega,s^{(i)}}(t)$ are **bad for** t **far from** $s^{(i)}$
  → wrong minimum at $\tau_{*\Omega}(s^{(i)})$ far from $s^{(i)}$
  → leading away from fixed point → circling.
- Exact function ϑ is **constant** in s-direction.

Idea

- **Weighted combination** of previous results $\hat{\theta}_{\Omega,s^{(i)}}$.
- **High** weights for t **close to** $s^{(i)}$ (good estimates).
- **Low** weights for t **far from** $s^{(i)}$ (bad estimates).

New iteration scheme

$$s^{(k+1)} = \tau_{*\Omega}^{(k)}(s^{(k)},\dots,s^{(1)}) = \arg\min_{t\in\mathcal{T}} \sum_{i=1}^{k} \varphi_{s^{(i)}}^{(k)}(t)\cdot\hat{\theta}_{\Omega,s^{(i)}}(t).$$

Weights $\varphi_{s^{(i)}}^{(k)}(t)$ with $\varphi_{s^{(i)}}^{(k)}(t) \geq 0$ & $\sum_{i=1}^{k}\varphi_{s^{(i)}}^{(k)}(t) = 1$.

Updating $\hat{\vartheta}_\Omega$ for visualising

$$\hat{\vartheta}_\Omega^{(k)}(s,t) = \varphi_s^{(k)}(t)\cdot\hat{\vartheta}_\Omega(s,t) + \sum_{i=1}^{k-1}\varphi_{s^{(i)}}^{(k)}(t)\cdot\hat{\theta}_{\Omega,s^{(i)}}(t).$$

Three approaches for weighting functions

Exponential functions

Norm $\|t - s^{(i)}\|_D^2 = (t-s^{(i)})^\mathsf{T} D^2 (t-s^{(i)})$ measures the distance between t and $s^{(i)}$.

The diagonal matrix $D = \mathrm{diag}(d_1,\dots,d_m)$ is used for scaling and for defining what is "near" to $s^{(i)}$.

Weighting functions $\tilde{\varphi}_{s^{(i)}}^{(k)}(t) = \mathrm{e}^{-\|t-s^{(i)}\|_D^2}$ becomes smaller for increasing distance between t and $s^{(i)}$.

Normalisation: $\varphi_{s^{(i)}}^{(k)}(t) = \tilde{\varphi}_{s^{(i)}}^{(k)}(t)/\sum_{j=1}^{k}\tilde{\varphi}_{s^{(j)}}^{(k)}(t)$.

In case where $\mathcal{T} = [\underline{t}_1,\bar{t}_1]\times\cdots\times[\underline{t}_m,\bar{t}_m]$ one could define $d_j = c/(\bar{t}_j - \underline{t}_j)$ with a constant $c > 0$.

Disadvantage: We have to find an appropriate D.

Effective sample size

$\tilde{\varphi}_{s^{(i)}}^{(k)}(t) = n_{\mathrm{eff},s^{(i)}}(t)$ which is the effective sample size

$$n_{\mathrm{eff},s}(t) = \frac{(\sum_{k=1}^{n} w_{st}(x_s(V_k)))^2}{\sum_{k=1}^{n} w_{st}(x_s(V_k))^2}.$$

$n_{\mathrm{eff},s^{(i)}}$ is equal to the sample size for $t = s^{(i)}$ (standard sampling, $\hat{\theta}_\Omega$) and becomes smaller for t far from $s^{(i)}$.

Normalisation: $\varphi_{s^{(i)}}^{(k)}(t) = \tilde{\varphi}_{s^{(i)}}^{(k)}(t)/\sum_{j=1}^{k}\tilde{\varphi}_{s^{(j)}}^{(k)}(t)$.

Advantage: Cheap and parameter free!

Piecewise multi-linear interpolation

We use basis functions $\varphi_{s^{(i)}}^{(k)}$ as in the finite element method which have the following properties:

$\varphi_{s^{(i)}}^{(k)}(s^{(j)}) = \delta_{ij}$, no disturbance by other $\varphi_{s^{(j)}}^{(k)}$ at $s^{(i)}$.

$\varphi_{s^{(i)}}^{(k)}(t) = 0$ outside the elements surrounding $s^{(i)}$.

Weighting functions (1D version):

For $\mathcal{T} = [\underline{t},\bar{t}]$: $(s^{(i_1)},\dots,s^{(i_k)}) = \mathrm{sort}(s^{(1)},\dots,s^{(k)})$.

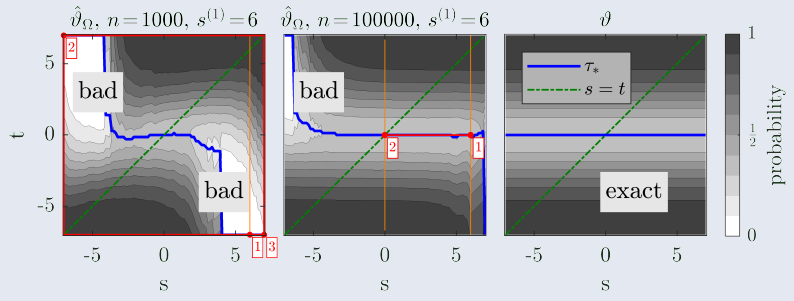
The intervals $[\underline{t},s^{(i_1)}],[s^{(i_1)},s^{(i_2)}],\dots,[s^{(i_k)},\bar{t}]$ are the elements and the $\varphi_{s^{(i_j)}}^{(k)}$ corresponds to nodes $s^{(i_j)}$:

$$\varphi_{s^{(i_j)}}^{(k)}(t) = \begin{cases} \dfrac{t - s^{(i_{j-1})}}{s^{(i_j)} - s^{(i_{j-1})}} & t \in [s^{(i_{j-1})}, s^{(i_j)}],\ j = 2,\dots,k, \\ \dfrac{s^{(i_{j+1})} - t}{s^{(i_{j+1})} - s^{(i_j)}} & t \in [s^{(i_j)}, s^{(i_{j+1})}],\ j = 1,\dots,k-1, \\ 1 & t \in [\underline{t}, s^{(i_1)}],\ j = 1, \\ 1 & t \in [s^{(i_k)}, \bar{t}],\ j = k, \\ 0 & \text{otherwise.} \end{cases}$$

Disadvantage: Difficult in higher dimensions.

Examples: standard fixed point iterations & exact ϑ

- Contour plots of $\hat{\vartheta}_\Omega$, iteration paths of the fixed point iterations starting at $s^{(1)} = 6$ for $n = 1000, 100000$, $s = t$ ($\hat{\theta}_\Omega$ from above) and $\tau_{*\Omega}$.
- Bad estimates $\hat{\theta}_{\Omega,s^{(i)}}$ for t far from $s^{(i)}$ → may cause **divergence**.

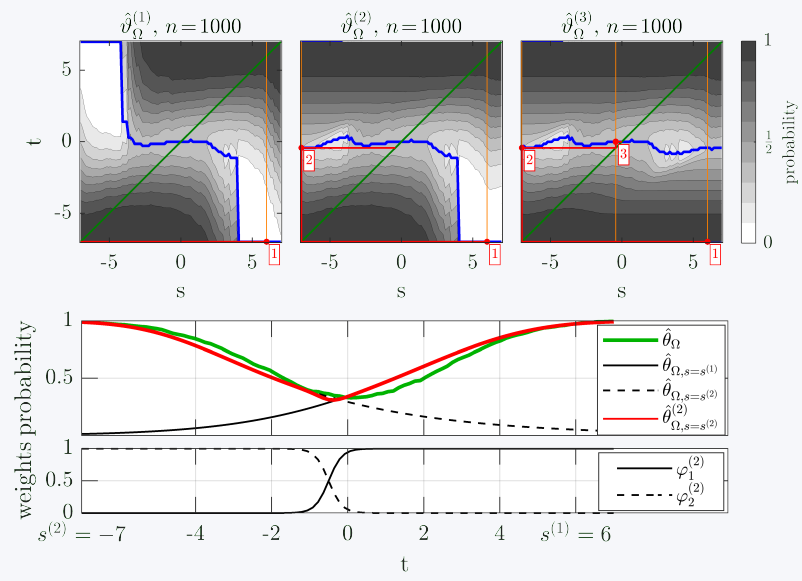


Examples: three combination methods

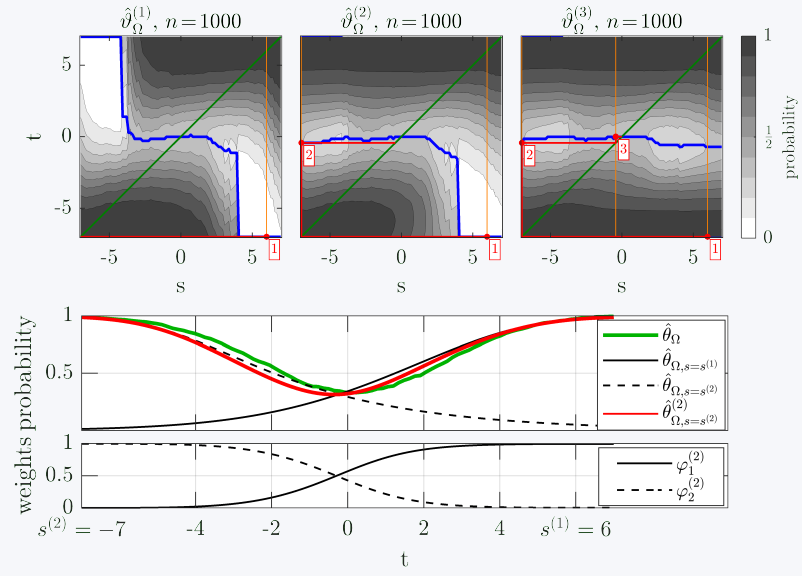
- Updated contour plots $\hat{\vartheta}_\Omega^{(k)}$, $k = 1,2,3$, iteration paths, $s = t$ and $\tau_{*\Omega}$.
- Single estimates $\hat{\theta}_{\Omega,s^{(1)}}$, $\hat{\theta}_{\Omega,s^{(2)}}$, combined estimate $\hat{\theta}_{\Omega,s^{(2)}}^{(2)}$ and $\hat{\theta}_\Omega$.

Combination method using exponential functions

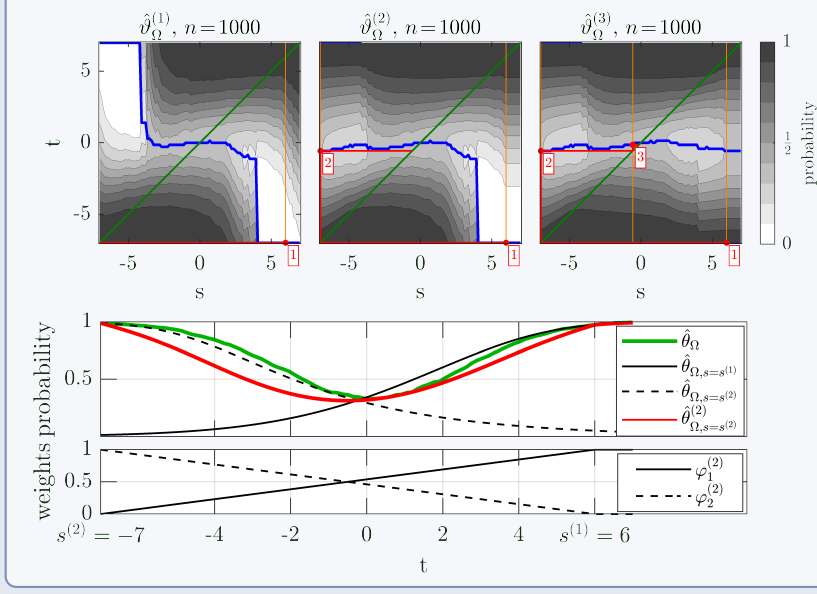
- For $c = 30$ we get scaling factor $d = c/(\bar{t} - \underline{t}) = 30/14 = 2.14$.



Combination method using effective sample size



Piecewise linear interpolation



Conclusion

- All three approaches **lead to convergence** after few iterations.
- The method with effective sample size is **cheap** and **easy to apply**.
- **No new evaluations** of expensive function h are needed.
- The methods are complementary to the increase of sample coverage.