Simulation to solve Bayesian sequential clinical trial problem using parametric decision boundaries*

Erwann Rogard

2022-04-04

Abstract

We revisit a Bayesian sequential clinical trial problem[1] by proposing efficient simulation schemes for finding the optimal parameters of the continuation region (equivalently stopping rule). This is a re-edition of a note from 2006-10-03[1].

1 Introduction

In section 2, we introduce the problem formulation for a simple model and utility framework, starting from the intuitive (but practically not recommended) Bellman recursion, and translating it into a parametric stopping rule (SR) approach, which is the focus of our study.

Under SR we have to maximize an expected utility over decision boundaries that have a parametric representation. Without a closed form for the expected utility, it has to be approximated from simulation samples. Stochastic optimization methods are therefore considered.

In general, for this type of problem, there is no straightforward "interchange of differentiation and expectation" formula that would allow to incorporate derivative information into the algorithm directly. To address this issue, in Section 3, we present a finite difference approach to the problem, namely the Kiefer–Wolfowitz algorithm. The challenge, here, is variance reduction, as the use of finite difference induces a particular "importance region". In Section 4, we re–express the gradient as a multidimensional integral restricted to a hypercube region of the space. By treating the integrand times the indicator for the hypercube as an un–normalized density, the problem becomes amenable to importance sampling or MCMC.

In Section 5, we consider more realistic models, implying more complicated and unknown policy spaces.

*Revisions: `https://github.com/erwannr/statistics/commits/main/ratio_nc`

[1] `https://github.com/erwannr/statistics/tree/main/carlin_revis`

In Section 6, we address convergence issues from a theoretical standpoint, and bridge our work with established sequential trials' methods/ and or an application

2 Problem formulation

Consider a sequential clinical trial extending over a maximum of n periods to assess the effect θ of a treatment, T, in comparison to the placebo, P. Suppose that at period i, $0 \leq i \leq n$, we have accumulated data $\mathbf{y}_i = (y_0, \ldots, y_i)$, with y_0 as prior. In this situation, our action space is $\mathcal{A} = \{T, P, C\}$, where C means continue to $i+1$, which incurs a sampling cost c_{i+1}. Let $\mathbf{a}_i = \{a_j\}_{j=0}^{i}$ denote the sequence of actions up to period i. For that period, our utility (or negative loss) is

$$u(i, \theta, a) = \begin{cases} -c_{i,T}(c_P - \theta)^+ - c_j, & a = T \\ -c_{i,P}(\theta - c_T)^+ - c_j, & a = P \\ -\infty 1\{i = n\}, & a = C \end{cases} \quad (1)$$

where the $c_{i,a}$'s denote penalties, and $[c_P, c_T]$ an indifference zone. The cost of forming the prior, c_0, does not affect the subsequent decisions, so we assume $c_0 = 0$. If we know that $a_{i-1} = C$ and $a_i \neq C$, then the relevant action space at period i is $\mathcal{A}^{\mathrm{s}} = \{T, P\}$, and for $j > i$, $\mathcal{A}_j = \{\varnothing\}$ and $u(j, \theta, \varnothing) = 0$. The utility up to period i is $u(\theta, \mathbf{a}_i) = \sum_{j=0}^{i} \beta^j u(j, \theta, a_j)$ where $0 \leq \beta \leq 1$ is a discount factor. For now we assume $\beta = 1$.

Our problem is to characterize a policy, $a : \boldsymbol{\mathcal{Y}}_i \to \mathcal{A}$, which maximizes the expected utility. Let $u(\mathbf{y}_i, a) = \mathrm{E}[u(\theta, \mathbf{a}_i)|\mathbf{y}_i, a_{i-1} = C, a_i \neq C]$. Within this restricted action space, we have to solve $\max_{a \in \mathcal{A}^{\mathrm{s}}} u(\mathbf{y}_i, a)$, whose expected utility and solution we denote $u(\mathbf{y}_i)$ and $a^{\mathrm{s}}(\mathbf{y}_i)$, respectively.

We will first study the model

$$p(\theta, \mathbf{y}_n) = N(\theta|\mu, \sigma_0)\Pi_{i=1}^{n} N(y_i|\theta, \sigma_i) \quad (2)$$

where $\mu, \sigma_0, ..., \sigma_n$ are given parameters. Under this model, the policy for $\mathcal{A}^{\mathrm{s}}$ has a simple characterization:

$$a^{\mathrm{s}}(\mathbf{y}_i) = \mathrm{T} \Leftrightarrow \mu(\mathbf{y}_i) \geq \mu_i \quad (3)$$

where $\mu_i = \frac{c_{i,T}c_P - c_{i,P}c_T}{c_{i,T} + c_{i,P}}$ and $\mu(\mathbf{y}_i) = E[\theta|\mathbf{y}_i]$.

The sequential problem is characterized by the Bellman recursion (BR):

$$u_*(\mathbf{y}_n) = u(\mathbf{y}_n) \quad (4)$$

$$\upsilon(\mathbf{y}_i) = \mathrm{E}[u_*(\mathbf{y}_{i+1})|\mathbf{y}_i] \quad (5)$$

$$u_*(\mathbf{y}_i) = \max(u(\mathbf{y}_i), \upsilon(\mathbf{y}_i)) \quad (6)$$

for $i = i_{\max -1}, ..., 0$, where $\upsilon(\mathbf{y}_i)$ and $u_*(\mathbf{y}_{i+1})$ are the expected utility, from moving to $i+1$ (and pursuing an optimal strategy thereafter), and from making

the overall best decision. Clearly, $u_*(\mathbf{y}_i) = v(\mathbf{y}_i) \Rightarrow a(\mathbf{y}_i) = C$, otherwise, $a(\mathbf{y}_i) = a^{\mathrm{s}}(\mathbf{y}_i)$.

The above problem is solved in two steps:

Algorithm 1 (BR)**.**

1. Fix m, and starting from $\mathbf{y}_0$, apply the recursion $y_{i+1}^{j_1,\ldots,j_i,j_{i+1}} \overset{\text{iid}}{\sim} p(y_{i+1}|\mathbf{y}_i^{j_1,\ldots,j_i})$, $j_{i+1} = 1, ..., m$, $0 \leq i < n$, resulting in a tree structure rooted at $\mathbf{y}_0$ and m^n terminal nodes

2. At all the nodes at $i = n$ compute (4) and for every other node, in the order $i = n-1, ..., 0$ compute (5–6)with (5) approximated as follows:

$$\hat{v}(\mathbf{y}_i^{j_1,\ldots,j_i}) = \frac{1}{m}\sum_{j=1}^{m} \hat{u}_*(\mathbf{y}_{i+1}^{j_1,\ldots,j_i,j}) \tag{7}$$

The above method has the following characteristic:

1. The simulated tree implies a computational demand exponential in n

2. The relevant distribution, at each node, is the predictive density i.e. $p(y_{i+1}|\mathbf{y}_i)$.

3. A trivial optimization problem, (6), is solved at each node of the simulated tree

4. The method is independent of the problem's structure (model and loss)

5. The bias of (7) is positive for finite m, and zero in the limit as $m \to \infty$

In the iid case, $p(y_{i+1}|\mathbf{y}_i) = \mathrm{E}[p(y_{i+1}|\theta)|\mathbf{y}_i]$ which in general is not easy to sample.

The first characteristic means that algorithm (1) is impractical beyond 2 or 3 periods. However, the memory requirement can be made linear in n by depth first processing:

2.1 independent paths alternative

To remedy the first limitation of algorithm 1 we would like replace the simulated tree in the first tree by independent paths and correct with importance sampling. Consider the formulation of (5) as

$$\mathrm{E}[u_*(\mathbf{y}_{i+1})|\mathbf{y}_i] = \mathrm{E}_*[u_*(\mathbf{y}_{i+1}^*)\tfrac{p(\mathbf{y}_{i+1}^*|\mathbf{y}_i^*=\mathbf{y}_i)}{p(\mathbf{y}_{i+1}^*)}] \tag{8}$$

$$= \mathrm{E}_*[u_*(\mathbf{y}_{i+1}^*)\tfrac{p(\mathbf{y}_{i+1}^*|\mathbf{y}_i^*=\mathbf{y}_i)}{\mathrm{E}_{**}[p(\mathbf{y}_{i+1}^*|\mathbf{y}_i^{**}=\mathbf{y}_i^*)]}] \tag{9}$$

where $\mathbf{y}_i^*$ indicates a variable independent of $\mathbf{y}_i$ but with the same distribution. In order for the above to make sense, we have to take $\mathbf{y}_i$ as meaning the sufficient statistic for θ of the data accumulated up to period i, and assume[2]

$$p(\mathbf{y}_{i+1}|\mathbf{y}_{i-1}, ..., \mathbf{y}_1) = p(\mathbf{y}_{i+1}|\mathbf{y}_{i-1}). \tag{10}$$

The new procedure is:

Algorithm 2 (BR–IS)**.**

1. Simulate m iid paths with transition $p(\mathbf{y}_{i+1}|\mathbf{y}_i), \quad i = 0, ..., n-1$
2. At all the nodes at $i = n$ compute (4) and for every other node, in the order $i = n-1, ..., 0$ compute (5–6) with the (5) approximated as follows:

$$\hat{\upsilon}(\mathbf{y}_i^j) = \frac{1}{m} \sum_{k=1}^{m} w(\mathbf{y}_{i+1}^k|\mathbf{y}_i^j) \hat{u}_*(\mathbf{y}_{i+1}^k) \tag{11}$$

and the importance weights computed as either of the two forms[3]:

$$w(\mathbf{y}_{i+1}^k|\mathbf{y}_i^j) = \frac{p(\mathbf{y}_{i+1}^k|\mathbf{y}_i^j)}{p(\mathbf{y}_{i+1}^k)} \tag{12}$$

$$w(\mathbf{y}_{i+1}^k|\mathbf{y}_i^j) = \frac{p(\mathbf{y}_{i+1}^k|\mathbf{y}_i^j)}{\frac{1}{m}\sum_{l=1}^{m} p(\mathbf{y}_{i+1}^k|\mathbf{y}_i^l)} \tag{13}$$

2.2 stopping rule formulation

We define

$$\tau_* \triangleq \min\{i : u(\mathbf{y}_i) > \upsilon(\mathbf{y}_i)\} \tag{14}$$

and $u(\theta, \mathbf{y}_i) \triangleq u(i, \theta, a(\mathbf{y}_i))$. In view of (4), (5) and (6), $\upsilon(\mathbf{y}_i) = \mathrm{E}[u(\theta, \mathbf{y}_{\tau_*})|\mathbf{y}_i]$, and in particular,

$$\upsilon \triangleq \upsilon(\mathbf{y}_0) = \mathrm{E}[u(\theta, \mathbf{y}_{\tau_*})] \tag{15}$$

where E[.] is understood as conditioning on $\mathbf{y}_0$ to alleviate subsequent notation. With the above stopping rule, a given path $\mathbf{y}$ is either stopped before i, at i or beyond i. This defines a partition $\boldsymbol{\mathcal{Y}}_i = \mathbf{C}_{i-1}^{\mathrm{c}} \cup \mathbf{S}_i \cup \mathbf{C}_i$. The relationships between these sets are $\mathbf{C}_i = \mathbf{C}_{i-1} \cup C_i$, for $i = 1, ..., n-1$, $\mathbf{C}_0 = C_0$ and $\mathbf{S}_i = \mathbf{C}_{i-1} \cap C_{i-1}^{\mathrm{c}}$ where

$$C_i = \{\mathbf{y}_i : \upsilon(\mathbf{y}_i) > u(\mathbf{y}_i)\}, \quad i = 0, ..., n-1 \tag{16}$$

$$C_n = \{\emptyset\} \tag{17}$$

[2] see Glasserman, 2000, Section 8.5.1 for all conditions

[3] see the discussion of Glasserman, 2000, Section 8.5.2 for a discussion of the superiority of the second form

We can estimate the stopping rule as follows. Suppose have generated $j = 1, ..., m$ paths and computed $(u(\mathbf{y}_i^j), \hat{v}(\mathbf{y}_i^j))$ by BR–IS, for each node (i, j). For any given $\mathbf{y}_i^*$, independent of the previous draws, we can estimate $v(\mathbf{y}_i^*)$ by plugging $\mathbf{y}_i^*$ in place of $\mathbf{y}_i^j$ into (11). By plugging the estimate, $\hat{v}(\mathbf{y}_i^*)$, into (14) results in an estimate $\hat{\tau}$. By plugging $\hat{\tau}$ in place of τ_* in (15) we obtain:

$$v(\hat{\tau}) \triangleq \mathrm{E}[u(\theta, \mathbf{y}_{\hat{\tau}})] \tag{18}$$

$$< v \tag{19}$$

As we recall the BR estimator (and therefore the BR–IS) of v a positive bias, whereas (18) has a negative bias. By combining the two estimators, we can therefore obtain an interval which contains the true value v for a given confidence level.

3 Parametric decision boundaries

The second way to estimate the stopping rule is to model it. Let $C = (C_1, ..., C_{n-1})$ and suppose we postulate $\mathbf{C} : \Gamma \to \mathcal{Y}^{n-1}$. As we saw, the continuation regions determine the stopping rule, which in turn determines a continuation value:

$$\tau(\gamma) = \min\{i : \mathbf{y}_i \notin C_i(\gamma)\} \tag{20}$$

$$v(\gamma) = \mathrm{E}[u(\theta, \mathbf{y}_{\tau(\gamma)})] \tag{21}$$

Our objective is now the maximization of (21). Clearly, we should expect

$$\sup_{\gamma \in \Gamma} v(\gamma) \leq v \tag{22}$$

Recalling the characteristics of BR at the end of the previous Section, let us now contrast them with those of SR:

1. Estimating the expectation in (21) requires parallel paths
2. The relevant distribution is $p(\theta, \mathbf{y}_\tau)$
3. We have to optimize over all entries of γ simultaneously.
4. The structure of the optimization problem depends on the distribution and utility under consideration

Assuming iid data $p(\theta, \mathbf{y}_\tau) = p(\theta)\Pi_{j=1}^{\tau} p(y_j)$, so it is essential that $p(\theta)$ be "proper" for sampling purposes. Practically, this often means restricting the search to conjugate priors, or otherwise incorporating some actual data into the prior (see Gelman et al., Section 4.3).

Under (1) and (2), Carlin et al. (1998) have derived the true model, i.e. that which achieves equality for (22). It is characterized as follows:

$$\gamma = (\gamma_1, ..., \gamma_{2(n-1)}) \tag{23}$$

$$C_i(\gamma) = \{\mathbf{y}_i : \gamma_i^- \leq \mu(\mathbf{y}_i) \leq \gamma_i^+\}, \quad 1 \leq i \leq n-1 \tag{24}$$

$$\Gamma = \{\gamma : \gamma_k \in \mathbb{R}, \gamma_i^- \leq \mu_i \leq \gamma_i^+\} \tag{25}$$

where $\gamma_i^- \triangleq \gamma_{2(i-1)+1}$, $\gamma_i^+ \triangleq \gamma_{2i}$. We define $i(k) = \lfloor k+1 \rfloor/2$.

From (23–25), and recalling (3), we see that $\mu(\mathbf{y}_i)$ summarizes the observable state variables. Moreover, $\mu(\mathbf{y}_i) = \mathbf{s}_i\mu(\mathbf{y}_{i-1}) + (1-\mathbf{s}_i)y_{i+1}$ with $\mu(\mathbf{y}_0) = \mu$ and $y_{i+1} = \theta + \epsilon_{i+1}$ so that conditional on θ, $\{\mu(\mathbf{y}_i)\}_{i=0}^n$ evolves as a Markov chain.

3.1 The optimization problem

The algorithms that we will consider to solve (22) start from an initial guess $\gamma^{(0)}$, and update it in an iterative fashion so as maximize a local approximation to (21):

$$\gamma^{(b+1)} \quad = \quad \mathrm{Proj}_\Gamma(\gamma^{(b)} - \alpha^{(b)} {H^{(b)}}^{-1} \widehat{\nabla}^{(b)} \upsilon) \tag{26}$$

where $\alpha^{(b)}$, $H^{(b)}$, are the step size, and scaling matrix (ideally the the hessian). We will consider different ways to compute $\widehat{\nabla}\upsilon^{(b)} \approx \nabla\upsilon(\gamma^{(b)})$ using averages over simulation samples. The particular distribution that we choose for $\widehat{\nabla}\upsilon$ at iteration (b), say d, induces a distribution, say d for $\gamma^{(b+1)}$. Our objective, therefore, can be expressed as

$$\min_d \mathrm{E}_d[||\gamma_*^{(b+1)} - \gamma^{(b+1)}||] \tag{27}$$

where $\gamma_*^{(b+1)}$ is optimal in some sense, for example the solution of (26) assuming perfect knowledge of $\nabla_2\upsilon(.)$ and $\nabla\upsilon(.)$ [4]. This expression is intricate. A reasonable intermediary objective, however, is the minimization of

$$\Delta_d^{(b)} = \mathrm{E}_d[||\nabla\upsilon(\gamma^{(b)}) - \widehat{\nabla}^{(b)}\upsilon||] \tag{28}$$

(TODO: what norm is appropriate? $||.||_{H^{-1}}$?)

Consider the following decomposition:

$$\upsilon(\boldsymbol{\gamma}_i) = \mathrm{E}[u(\theta, \mathbf{y}_i) 1\{\tau(\gamma) = i\}] \tag{29}$$

$$\upsilon(\gamma) = \upsilon(\boldsymbol{\gamma}_1) + ... + \upsilon(\boldsymbol{\gamma}_{n-1}) \tag{30}$$

where $\boldsymbol{\gamma}_i = \{\gamma_k\}_{k=1}^{2i}$. Under (21), the quantity (29) has the representation

$$\upsilon(\boldsymbol{\gamma}_i) = -\mathbf{c}_{i-1} + \int_{\mathbb{R}} \int_{\mathbf{R}_{i-1}} \cdots \int \Big[\int_{-\infty}^{\gamma_i^-} u(i,\theta,T) p(\theta, \mathbf{y}_i) dy_i + \int_{\gamma_i^+}^{\infty} u(i,\theta,P) p(\theta, \mathbf{y}_i) dy_i \Big] d\mathbf{y}_{i-1} d\theta \tag{31}$$

with $\gamma_n^{\pm} = \mu_n$ and $\mathbf{c}_i = \sum_{j=1}^i c_j$. Let

$$\Upsilon(\boldsymbol{\gamma}_i) = \mathbf{1}\{\mathbf{y}_{i-1} \in \mathbf{R}_{i-1}, \}(\mathbf{1}\{\mu(\mathbf{y}_i) \leq \gamma_i^-\} u(i,\theta,T) + \mathbf{1}\{\mu(\mathbf{y}_i) \geq \gamma_i^+\} u(i,\theta,P)) p(\theta, \mathbf{y}_i) \tag{32}$$

we have $\upsilon(\boldsymbol{\gamma}_i) = \mathrm{E}[\Upsilon(\boldsymbol{\gamma}_i)]$. However, the presence of indicator functions implies $\nabla\upsilon(\boldsymbol{\gamma}_i) \neq \mathrm{E}[\nabla\Upsilon(\boldsymbol{\gamma}_i)]$, which precludes a pathwise simulation approximation to the derivative. To resolve this difficulty, we will consider two methods.

[4] In this case, $\Delta_q = \mathrm{E}_q[||\nabla_2^{-1}\upsilon(\gamma^{(b)})\nabla\upsilon(\gamma^{(b)}) - {H^{(b)}}^{-1}\widehat{\nabla}^{(b)}\upsilon_{\mathrm{sim}}||]$

The first is the finite difference method. This involves only evaluating (30) at $\boldsymbol{\gamma}_i$ and $\boldsymbol{\gamma}_i + \Delta$ for some small Δ. As already noted, simulating from $p(\theta, \mathbf{y}_i)$ should not present major difficulties in the iid case. However, we are not interested in simulating $p(.)$, rather a truncated version of it. Importance sampling solutions will be proposed to improve efficiency.

The second method is to take partial derivatives of (31) with respect to γ_j. The resulting expression is no longer an expectation with respect to a well defined distribution. However, we can treat this expression as a un un–normalized density, and since it can be evaluated, the problem amenable to importance sampling or Markov chain simulation.

4 Method 1

4.1 Procedure

We now take $\widehat{\nabla}\upsilon$ to be a simulation approximation to finite difference gradient, $\widehat{\nabla}_{\text{FD}}\upsilon$. In this case, the recursion (26) is due to Kiefer–Wolfowitz. Let us start from the classic finite difference expression for $\nabla_{\text{FD}}\upsilon$. Fix (b), so that $\tau = \tau(\gamma^{(b)})$, and define $\tau_k = \tau(\gamma + e_k h_k)$. Throughout, we will assume a sequential importance form for the finite difference gradient:

$$\nabla_{\text{FD}}\upsilon = \mathrm{E}_q[(w \times \nabla_{\text{FD}} u)(\theta, \mathbf{y}_\tau)] \tag{33}$$

The k^{th} element of the integrand is

$$(w \times \nabla_{\text{FD}} u(\theta, \mathbf{y}_\tau))_k = \frac{(w \times u)(\theta, \mathbf{y}_{\tau_k}) - (w \times u)(\theta, \mathbf{y}_\tau)}{h_k} \tag{34}$$

where, $q(.)$ is a distribution of our choice within the class $q \gg p$, $w(\theta, \mathbf{y}_\tau) = p(\theta, \mathbf{y}_\tau)/q(\theta, \mathbf{y}_\tau)$, and $\times$ is the product operator i.e. $(f \times g)(.) = f(.)g(.)$. The resulting expression for (33) will equal $\nabla\upsilon$ plus a residual term that converges to zero as $||h|| \to 0$ where $h = (h_1, \ldots, h_{2(n-1)})$.

The procedure to estimate (33) is to sample n iid paths from $q(\theta, \mathbf{y}_{\max(\tau, \tau_1, \ldots, \tau_{2(n-1)})})$ and evaluate the sample average $\widehat{\nabla}_{\text{FD}}\upsilon = \widehat{\mathrm{E}}_q[(w \times \nabla_{\text{FD}} u)(\theta, \mathbf{y}_\tau)]$. Each term of difference in (34) is evaluated with the *same* path, which reduces variance, provided there is positive correlation between the two terms. For each iteration (b), a new set of n paths is simulated, which is indicated in (26) by the superscript(b) of $\widehat{\nabla}^{(b)}_{\text{FD}}\upsilon$. This is part of the conditions for convergence and stems from the fact that averaging of the $\gamma^{(b)}$'s is already present, as in implicit in (26). If it is felt that the behavior of the algorithm is erratic, a larger value for n or a smaller value for $\alpha^{(0)}$ is recommended.

We can weaken the above formulation to $w(\theta, \mathbf{y}_\tau) \propto p(\theta, \mathbf{y}_\tau)/q(\theta, \mathbf{y}_\tau)$, with an arbitrary normalizing constant, which imposes the requirement that the sample w's be normalized to one. This may be a matter of design (to make the sample average equivariant, see Hestenberg, 1995) or necessity, in case either of

$q(.)$ or $p(.)$'s normalizing constant is not known. Unbiasedness of $\widehat{\nabla}_{\text{FD}}\upsilon$, relative to $\nabla_{\text{FD}}\upsilon$,is only preserved in the limit as $n \to \infty$. If we explicitly need to distinguish this method from the standard one, we will write $\tilde{w}$ and $\widetilde{\mathrm{E}}_q[.]$.

In view of (28), $\Delta_d \triangleq \Delta_{q,h}$ defined as

$$\Delta_{q,h} = \mathrm{E}_q[||\nabla\upsilon - \widehat{\nabla}^h_{\text{FD}}\upsilon||] \tag{35}$$

$$= \mathrm{E}_q[||\nabla\upsilon - \nabla^h_{\text{FD}}\upsilon|| + ||\widehat{\nabla}^h_{\text{FD}}\upsilon - \nabla^h_{\text{FD}}\upsilon||] \tag{36}$$

where the finite difference increment h is now shown explicitly in superscript. One problem with minimizing this expression is the interaction between $q(.)$ and h. A first step towards simplification is to fix h, thereby fixing $bias(h) = ||\nabla^h_{\text{FD}}\upsilon - \nabla\upsilon||$, and focusing on $q(.)$ to reduce some variance measure (TODO: define it). This approach has practical interest: often we have a prior belief (say from initial trials) about $d_k^{(0)} = ||\gamma_k^{(0)} - \gamma_k^*||$, and it seems reasonable to set $h \leq \epsilon d_k^{(0)}$, say $h_k^{\max}$. In our own experimentation, for $q(.) = p(.)$, the value of h that minimizes $||\nabla_{\text{FD}}\upsilon - \nabla\upsilon||$ can be greater than $d_k^{(0)}$ than itself, if the curvature of $\upsilon(\gamma^{(0)})$ is small, which underlines the need to fix an upper bound. Our procedure for initializing the algorithm, therefore,

1. decide on a appropriate value for $h_{\max}$
2. solve $\min_q \Delta_{q,h_{\max}}$, yielding q_*
3. fine tune the finite difference increment, $\min_{h \leq h^{\max}} \Delta_{q_*,h}$.

It may be possible to perform $\min_q \Delta_{q,h^{(b)}}$ throughout the course of the algorithm, where $\{h^{(b)}\}$ is an appropriately diminishing sequence but in this case a proper convergence analysis is required.

The discussion that follows is also valid for variants of the finite difference method, with a few adjustments. In particular, if $dim(\Gamma)$ is large, it may be worthwhile perturbing all elements of γ as opposed to each separately. Let $\tau_\epsilon = \tau(\gamma + h\epsilon)$,and

$$(\nabla_{\text{FD}} u(\theta, \mathbf{y}_\tau))_k = \frac{u(\theta, \mathbf{y}_{\tau'}) - u(\theta, \mathbf{y}_\tau)}{h\epsilon_k} \tag{37}$$

where $\epsilon = \{\epsilon_k\}_{k=1}^{2(n-1)}$ is a random perturbation vector. This finite difference alternative is termed "simultaneous perturbation" (SP). Compared with the previous formulation, only two evaluations of $u(.)$ are needed, a gain of to $2(n-1)-1$ evaluations. Under reasonable conditions, this gain more than offsets the loss in efficiency. Among these conditions, $E[|1/\epsilon_k|] \leq \infty$, which is verified for independent ϵ_k's distributed as ± 1 Bernoulli's. The search for an optimal h can be derived from that of the standard finite difference method (see Section tuning the gains parameters)

4.2 Importance region

The quantity (34) has value 0 outside of the set

$$\mathbf{A}_k = \{\mathbf{y}_{i(k)} : \tau \neq \tau_k\} = \mathbf{R}_{i(k)-1} \cap A_k \tag{38}$$

$$A_k = \{\mathbf{y}_{i(k)} : \gamma_k \leq \mu(\mathbf{y}_{i(k)}) \leq \gamma_k + e_k h_k\} \tag{39}$$

where $i(k) = \lfloor k+1 \rfloor/2$. From a computation standpoint, if we simulate along paths in a stepwise fashion, we may record 0 for (34) for each k as soon as we observe $\tau < i(k)$.

From an efficiency perspective, the optimal sampling density is $q_*(.) \propto p(.) \times |(\nabla_{\text{FD}} u)_k(.)|$. One way to get closer to this goal is to take $q(\theta, \mathbf{y}_{i(k)}) \approx p(\theta, \mathbf{y}_{i(k)}|\mathbf{A}_k)$, so that $w(\theta, \mathbf{y}_{i(k)}) \approx p(\mathbf{A}_k)$. This objective is all the more critical that $\lim_{b\to\infty} A_k^{(b)} = \{\mathbf{y}_{i(k)} : \mu(\mathbf{y}_{i(k)}) = \gamma_k\}$ implying $\lim_{b\to\infty} p(\mathbf{A}_k^{(b)}) = 0$ i.e. $\mathbf{A}_k$ is asymptotically a rare event. Note also, that $\lim_{k\to\infty} p(\mathbf{A}_k^{(b)}) = 0$ (curse of dimensionality TODO: verify).

If the weights are normalized, $\tilde{w}$, which is common in Bayesian analysis, the optimal density is $q_*(.) \propto p(.)|(\nabla_{\text{FD}} u)_k(.) - (\nabla_{\text{FD}} v)_k|$. In this case, we need to sample from *both* $\mathbf{A}_k$ and its complement. Sampling within $\mathbf{A}_k$, however, remains just as critical and difficult.

Based on the preceding remarks, one direction is stratification, i.e. controlling the number of paths that fall in $\mathbf{A}_k$ and those falling outside (if any are needed)so that across strata variance –which in this case is large– is neutralized. Specifically, under (2), conditional on θ, $\mathbf{x}_i \triangleq \mu(\mathbf{y}_i) = \mathbf{s}_i \mu(\mathbf{y}_{i-1}) + (1-\mathbf{s}_i)\theta + (1-\mathbf{s}_i)\epsilon_i$ so there exists a vector μ_i such that $\mu(\mathbf{y}_i) = \mu_i^{\mathrm{T}} \epsilon_i$. Fix k and let $\tilde{\gamma}_j^{\pm} = \gamma_j^{\pm}$ for $j < i(k)$, $\tilde{\gamma}_{i(k)}^{-} = \gamma_k$ and $\tilde{\gamma}_{i(k)}^{+} = \gamma_k + e_k h_k$. Then $\mathbf{A}_k = \{\mathbf{y}_{i(k)} : \tilde{\gamma}_j^{-} \leq \mu(\mathbf{y}_j) \leq \tilde{\gamma}_j^{+}, j = 1, ..., i(k)\}$ which translates into a hypercube for $\mathbf{x}_i \sim N(\theta \mathbf{1}_i, \Sigma)$. Please read the relevant section of the appendix. We should expect high correlation between the x_i's and therefore sampling from truncated univariate normals will be inefficient. The alternative method is therefore highly relevant here.

So far we have discussed a single k in isolation, whereas the search for $q(.)$ should take their joint contribution into account (Veach and Guibas, 1995). In particular, the $\mathbf{A}_k$'s overlap in a non trivial way[5], leading to a complicated resource allocation problem. At the very least, however, we should expect stratification to work well when γ is altered only in one direction at each step (b).

A second approach is to take $q(.)$ and $p(.)$ identical, except, for the θ portion of $(\theta, \mathbf{y}_\tau)$, leading to a simple formulation for the importance ratio in the iid case: $w(\theta, \mathbf{y}_\tau) \propto p(\theta)/q(\theta)$. Moreover, there are no restrictions of the kind mentioned for stratification. A convenient approach, here, is to define a family of sampling distributions indexed by λ: $\{q_\lambda(.), \lambda \in \Lambda\}$ and select λ to minimize some variance criterion. In general $p(.)$ should be incorporated into $q(.)$ as a defensive sampling measure (Owen and Zhou, 1998) and also to

[5] $\mathbf{A}_{2(i-1)+1}$ and $\mathbf{A}_{2(i-1)+2}$ have $\mathbf{R}_{i-1}$ in common and A_k and A_{k+2} have $\mathrm{R}_{i(k)-1}$ in common.

"learn" the optimal λ. Please read the relevant section in the appendix with $q_*(.) \propto p(.)|(\nabla_{\text{FD}} u)_k(.) - c)_k|$ where c is 0 and $\nabla_{\text{FD}} \upsilon$, if the weights w and $\tilde{w}$ are used, respectively. In view of our previous discussion, we would like the minimization step to incorporate a constraint of the form $q_\lambda(\mathbf{A}_k) > \epsilon$.

We should note that $\{\nabla \upsilon\}_{\text{odd}} = \{(\nabla \upsilon)_{2(i-1)+1}\}_{i=1}^{n-1}$ and $\{\nabla \upsilon\}_{\text{even}} = \{(\nabla \upsilon)_{2(i-1)+2}\}_{i=1}^{n-1}$ form two homogeneous groups in the following way: the expressions for each involve $a = T$ and $a = P$, respectively. Provided the cost structure, for either of $a \in \{T, P\}$, changes smoothly enough from i to $i+1$, we may expect some positive correlation between the estimates of members of each group. However, a key additional assumption is required: that the members of γ_{odd} and γ_{even} are also smooth along $i = 1, ..., n$. We should therefore impose this requirement to set the starting value $\gamma^{(0)}$. Since we can expect that this condition will be verified for $\gamma^\infty = \gamma_*$, provided the algorithm converges, some level of smoothness should be enforced for each (b). Moreover, neighboring members of each group should be close in expectation. Several implications follow:

1. This suggests defining $q_\lambda(.)$ as mixture of distributions, one "taking care" of each group.

2. For a given computational budget, particularly if $\beta < 1$, it may be better to favor paths that stop early, as the later periods contribute a diminishing portion to υ. Accordingly, we would set $q_\lambda(.)$ to meet that objective. TODO: define the resource allocation problem.

3. Smoothing might be appropriate across neighboring k's, within each of $\{\nabla \upsilon\}_{\text{odd}}$ and $\{\nabla \upsilon\}_{\text{even}}$ or alternatively $\{\gamma\}_{\text{odd}}$ and $\{\gamma\}_{\text{even}}$. This form of smoothing should ne confused with smoothing across (b)'s, which –as already mentioned– is not valid for estimating $\nabla \upsilon$ as part of (26).

4.3 Tuning the gains sequences

Fix q and k. The objective, is to balance the bias and variance of $(\widehat{\nabla}_{\text{FD}} \upsilon)_k$, as part of the system (26). It is a classical result that the system

$$\alpha^{(b)} = \alpha_0/(b+1+B)^{\alpha_{\text{rate}}} \tag{40}$$

$$h_k^{(b)} = h_k^0/(b+1)^{h_{\text{rate}}} \tag{41}$$

with the choice of parameters $\alpha_{\text{rate}} = 1$ and $h_{\text{rate}} = 1/6$, is asymptotically (in b) optimal, and yield a convergence rate for (26) of $(\alpha_{\text{rate}} - 2h_{\text{rate}})/2 = 1/3$. As the iteration proceeds, taking smaller steps h_k increase the variance and decreases the bias of $(\nabla_{\text{FD}} \upsilon_{\text{sim}})_k$ which is offset by the implicit averaging in (26) and the increasing damping in (40).

The choice of $\gamma^{(0)}$, B, α_0 and h_k^0 rests upon us. Suppose the first three are fixed; for h_k^0, given $\gamma^{(0)} = \gamma$, we suggest a value proportional to that which minimizes the asymptotic (in n) mean square error of (33). This is

$$h_k^* = \left(\frac{2\sigma^2(\gamma)}{n \upsilon_k''(\gamma)^2} \right)^{1/3} \tag{42}$$

where n is the simulation size, and $\sigma^2(\gamma) = \mathrm{Var}_q[(w \times u)(\theta, \mathbf{y}_\tau)]$. This takes into account the use of common random numbers for evaluating $(w \times u)(\theta, \mathbf{y}_{\tau_k})$ and $(w \times u)(\theta, \mathbf{y}_\tau)$ in (33). TODO: think about the lack of smoothness of $\nabla_{\mathrm{FD}} u_k(\theta, \mathbf{y}_\tau)$ due to the fact that it is zero outside $\mathbf{A}_k$.

We now explain how to estimate (42) from simulation data with a bootstrap methodology. Start with an initial guess for h_k^*, h_k, and update it as follows:

$m \leftarrow 1$
$\bar{\sigma}^2 \leftarrow 0$
$\bar{v}_k'' \leftarrow 0$
repeat
  Generate $s = \{(\theta, \mathbf{y}_n)^{(j)}\}_{j=1}^n$ from q
  $\hat{\sigma}^2 \leftarrow \widehat{\mathrm{Var}}_q[(w \times u)(\theta, \mathbf{y}_\tau)]$
  $\bar{\sigma}^2 \leftarrow ((m-1)\bar{\sigma}^2 + \hat{\sigma}^2)/m$
  Compute $\hat{v}_k''$ by finite difference approximation with increment h_k, based on s
  $\bar{v}_k'' \leftarrow ((m-1)\bar{v}_k'' + \hat{v}_k'')/m$
  $h_k \leftarrow h_k(\bar{\sigma}^2, \bar{v}_k'')$ by (42)
  $\hat{\sigma}_{h_k} \leftarrow \widehat{\mathrm{Var}}[h_k]$
  $m \leftarrow m + 1$
until $m > (\Phi^{-1}(1-\delta/2)\hat{\sigma}_{h_k}/(|h_k|\epsilon))^2$

The if statement ensures that the error for each k is less than ϵ, with $1-\delta$ probability based on a normal approximation, resulting in a formula, whose parameters h_k^* and σ_{h_k} are replaced by their estimate h_k and $\hat{\sigma}_{h_k}$.

A byproduct of the above procedure is that we can reuse the estimate of $\{v''(\gamma)\}_{k=1}^{2(n-1)}$ as the diagonal of H, provided it is positive definite.

5 Method 2

Recall (29). Without loss of generality we will now ignore $\mathbf{c}_j$ and the upper bounds and focus on a given i:

$$v(\boldsymbol{\gamma}_i) = \int_{\mathbb{R}} \int_{\mathbf{R}_{i-1}} \cdots \int \Big[\int_{-\infty}^{\gamma_i} u(i, \theta, T) p(\theta, \mathbf{y}_i) dy_i d\mathbf{y}_{i-1} d\theta \tag{43}$$

Let $v_j(\boldsymbol{\gamma}_i) = \partial v(\boldsymbol{\gamma}_i)/\partial \gamma_j$:

$$v_j(\boldsymbol{\gamma}_i) = \begin{cases} 0, & j > i \\ \int_{\mathbb{R}} \int_{\mathbf{R}_{i-1}} \cdots \int \Big[u(i, \theta, T) p(\theta, \mathbf{y}_{i-1}, \gamma_i) \Big] d\mathbf{y}_{i-1} d\theta, & j = i \\ \int_{\mathbb{R}} \int_{\mathbf{R}_{i-1}} \cdots \int \Big[\int_{-\infty}^{\gamma_i} u(i, \theta, T) p(\theta, \mathbf{y}_{j-1}, \gamma_j, \mathbf{y}_{j<l\le i}) \Big] d\mathbf{y}_{j-1} d\mathbf{y}_{j<l\le i} d\theta, & j < i \end{cases} \tag{44}$$

The third equality is similar to the second, so our focus will be on the later:

$$\upsilon_i(\boldsymbol{\gamma}_i) = \int_{\mathbb{R}} \int_{\mathbf{R}_{i-1}} \cdots \int \Big[u(i,\theta,T) p(\theta, \mathbf{y}_{i-1}, \gamma_i) \Big] d\mathbf{y}_{i-1} d\theta \tag{45}$$

Since the above quantity is bounded and $u(i,\theta,T) \geq 0$ we may define the density

$$L(\theta, \mathbf{y}_{i-1}) = \mathbf{1}\{\mathbf{y}_{i-1} \in \mathbf{R}_i\} u(i,\theta,T) p(\theta, \mathbf{y}_{i-1}, \gamma_i) / \upsilon_i(\boldsymbol{\gamma}_i) \tag{46}$$

$$= \mathbf{1}\{\mathbf{y}_{i-1} \in \mathbf{R}_i, \theta < \gamma_i\} \theta p(\theta, \mathbf{y}_{i-1}, \gamma_i) / \upsilon_i(\boldsymbol{\gamma}_i) \tag{47}$$

and call $L_{\text{un}}(.)$ its un–normalized version. Given $q \gg p$, the importance sampling representation for this problem is

$$\mathrm{E}_q[L_{\text{un}}(\theta, \mathbf{y}_{i-1}) / q(\theta, \mathbf{y}_{i-1})] = \mathrm{E}_q[L(\theta, \mathbf{y}_{i-1}) / q(\theta, \mathbf{y}_{i-1})] \upsilon_j(\boldsymbol{\gamma}_i) \tag{48}$$

$$= \upsilon_j(\boldsymbol{\gamma}_i) \tag{49}$$

or,

$$\mathrm{E}_q[L_{\text{un}}(\theta, \mathbf{y}_{i-1}) / q_{\text{un}}(\theta, \mathbf{y}_{i-1})] = \upsilon_j(\boldsymbol{\gamma}_i) / C_q \tag{50}$$

$$E_q[q_{\text{un}}(\theta, \mathrm{y}_{i-1})] = C_q \tag{51}$$

Alternatively, we may determine a Markov chain $\{z_n, n \in \mathbb{N}\}$ with initial distribution K_0 and transition probabilities $\{K_n(.,.)\}_{n \in \mathbb{N}}$ which converges to $L(.)$ in the limit as $n \to \infty$. Specifically, $\eta_n(A) = \Pr(z_n \in A)$ is defined recursively as follows:

$$\eta_0(A) \triangleq K_0(A) \tag{52}$$

$$\eta_n(A) \triangleq \mathrm{E}_{\eta_{n-1}}[K_n(z, A)] \tag{53}$$

We want to find $K_0(.)$ and $\{K_n(.,.)\}_{n \in \mathbb{N}}$ such that $\eta_\infty(A) = L(A)$. Given a choice of proposal initial and transition distributions, $Q_0(.)$ and $Q_n(.)$, the Metropolis–Hastings algorithm is appropriate, and is defined by

$$\alpha(z_{n-1}, z) = \min\{\frac{Q_n(z_{n-1}, z) L(z)}{Q_n(z, z_{n-1}) L(z_{n-1})}, 1\} \tag{54}$$

$$K_n(z_{n-1}, dz_n | z) = \delta_{\{z\}}(z_n) \alpha(z_{n-1}, z) + (1 - \alpha(z_{n-1}, z)) \delta_{\{z_{n-1}\}}(z_n) \tag{55}$$

$$K_n(z_{n-1}, dz_n) = E_{Q_n}[K_n(z_{n-1}, dz_n | z)] \tag{56}$$

The ratio formulation of $\alpha(.,.)$ ensures that it is insensitive to replacing $L(.)$ $L_{\text{un}}(.)$ or $Q_n(.)$ by $Q_{n,\text{un}}(.)$. Loosely speaking, we must ensure that $(Q_0, \{Q_n\}_{n \in \mathbb{N}})$ will reach all regions A such that $\Pr(z \in A) > 0$ from an arbitrary starting point, within finite time. These conditions are usually easy to meet. Our estimation procedure will be based on

$$E_{\eta_\infty}[1 / L_{\text{un}}(z)] = \upsilon_j(\boldsymbol{\gamma}_i) \tag{57}$$

The difficulty in either of importance sampling or MCMC is to obtain reasonable convergence rates. The challenge, here is 1) to find a distribution that

approximates $\theta p(\theta, \mathbf{y}_{i-1}, \gamma) \propto \theta p(\theta, \mathbf{y}_{i-1}) p(\gamma_1|\gamma)$ and 2) to sample from the latter on the hypercube $\mathbf{1}\{\mathbf{y}_{i-1} \in \mathbf{R}_i, \theta < \gamma_i\}$. In the case of an independence chain for MCMC or importance sampling, please read the relevant section in the appendix. In the MCMC case for random walk chain, an adaptive procedure to maximize the jumping distance (weighted by the probability of acceptance $\alpha(.,.)$) is presented in Pasarica and Gelman, 2005).

So far, we have focused on $\upsilon_j(\boldsymbol{\gamma}_i)$ for a particular (i, j). We have to think about an efficient way to reuse the sample across (i, j)'s.

The above discussion should apply to the extended model of the next section (2), after condition on σ.

6 Model extensions

We begin by extending (2), which relies on known variances to the most immediate generalization:

$$\theta|\sigma^2 \sim N(y_0, \sigma^2/\kappa_0) \tag{58}$$

$$\sigma^2 \sim \text{Inv} - \chi(\nu_0, \sigma_0^2) \tag{59}$$

The parameter κ_0 can be interpreted as the number of prior measurements i.e. $\kappa_0 + \tau$ is the total number of measurements under a particular stopping rule.

Let us review standard properties:

$$p(\theta|\mathbf{y}_i) \sim t_{\nu_i}(\theta|\overline{\mathbf{y}}_i, \sigma^2(\mathbf{y}_i)/\kappa_i) \tag{60}$$

where $\overline{\mathbf{y}}_i = (\kappa_0 y_0 + \sum_{j=1}^{i} y_i)/(\kappa_0 + i)$, $\kappa_i = \kappa_0 + i$, $\nu_i = \nu_0 + i$ and $\nu_i \sigma^2(\mathbf{y}_i) = \nu_0\sigma_0^2 + (i-1)s(\mathbf{y}_{0<j\leq i})^2 + \frac{\kappa_0(\kappa_0+i)}{i}(\overline{\mathbf{y}}_i - y_0)^2$ and $s^2(.)$ is the the sample variance.

7 Appendix

7.1 Variance of importance sampling

Here, we consider transformation of stopped paths such as $(\theta, \mathbf{y}_\tau) \rightarrow (w \times u)(\theta, \mathbf{y}_\tau)$. Let $x_i \triangleq \mu(\mathbf{y}_i)\}_{i=0}^n$. We recall from Section 2 that conditional on θ, $\{x_i\}_{i=1}^n$ evolves as a Markov chain. Furthermore, remember that $u(\theta, \mathbf{y}_n) = u(\theta, \mathbf{y}_\tau)$ because $\tau > i \Rightarrow u(i, \theta, \mathbf{y}_i) = 0$. Therefore, assuming Assume $q(\theta, \mathbf{x}_i) = q(\theta)q(x_1|x_0)...q(x_i|x_{i-1})$,

$$\upsilon = \mathrm{E}_q[u(\theta, \mathbf{x}_n)w(\theta, \mathbf{x}_n)] \tag{61}$$

$$= \mathrm{E}_q[E_q[u(\theta, \mathbf{x}_\tau)w(\theta, \mathbf{x}_\tau)\mathrm{E}_q[w(\theta, \mathbf{x}_n)/w(\theta, \mathbf{x}_\tau)|\theta, \mathbf{x}_\tau]|\theta]] \tag{62}$$

$$= \mathrm{E}_q[u(\theta, \mathbf{x}_\tau)w(\theta, \mathbf{x}_\tau)] \tag{63}$$

Is this still valid with $\tilde{w}$, the normalized weight?

The advantage of the first equality is that we can work with $z = (\theta, \mathbf{y}_n)$ which is a vector variable of fixed size. The formula for this variance using normalized weights is:

$$\mathrm{Var}_q[\tilde{v}] = \mathrm{Var}_q[\sum_{j=1}^{n}(\tilde{w} \times u)(z_j)] \tag{64}$$

$$\to \frac{1}{n}\mathrm{E}_p[\tilde{w}(z)(u(z) - v)^2] \tag{65}$$

7.2 Bias variance decomposition of the finite difference method

To begin, let's assume that $q(.) = p(.)$, $n = 2$ and consider the univariate case by fixing $k = 1$. Fix $h > 0$, so that $\tau(\gamma + h) \leq \tau(\gamma)$ and let $u(\gamma) \triangleq u(\theta, \mathbf{y}_{\tau(\gamma)})$, so that $v(\gamma) = E[u(\gamma)]$:

$$\nabla^h_{\mathrm{FD}} u(\gamma) = \frac{u(\gamma + h) - u(\gamma)}{h} \tag{66}$$

$$= \begin{cases} 0 & , \tau(\gamma) = \tau(\gamma + h) \\ u(1, \theta, P) - u(2, \theta, P) & , \tau(\gamma) > \tau(\gamma + h) \quad \text{and} \quad a(\mathbf{y}_\tau) = P \\ u(2, \theta, P) - u(2, \theta, T) & , \tau(\gamma) > \tau(\gamma + h) \quad \text{and} \quad a(\mathbf{y}_\tau) = T \end{cases}. \tag{67}$$

$$\mathit{bias}(h) \triangleq \mathrm{E}[\nabla^h_{\mathrm{FD}} u(\gamma) - \nabla v(\gamma)] \tag{68}$$

$$= \nabla^h_{\mathrm{FD}} v(\gamma) - \nabla v(\gamma) \tag{69}$$

Provided $v(\gamma + h) = v(\gamma) + \nabla v(\gamma)h + \frac{1}{2}\nabla_2 v(\gamma)h^2 + o(h^2)$,

$$\mathit{bias}(h) = \frac{1}{2}\nabla_2 v(\gamma)h + o(h) \tag{70}$$

$$= O(|h|) \tag{71}$$

We need to show that

$$\mathit{var}(h) \triangleq \mathrm{Var}[\nabla^h_{\mathrm{FD}} u(\gamma)] \tag{72}$$

$$= O(1/|h|) \tag{73}$$

This is not guaranteed, given the discrete nature of (67) (TODO).

We may rewrite the order of the bias and variance, for an iid sample of size n as

$$\mathit{bias}(h) = bh + o(|h|) \tag{74}$$

$$\mathit{var}(h) = \frac{1}{nh}\sigma^2 + o(1/|h|) \tag{75}$$

with $b = \nabla_2 v(\gamma)$. Suppose $h = h_* n^{-\eta}$ and plug that expression into the above system. Then

$$mse(h,n) = bias^2(n,h) + var^2(n,h) \tag{76}$$
$$= b^2(h_* n)^{-2\eta} + \sigma^2 h_*^{\eta} n^{\eta-1} \tag{77}$$

This is minimized for $\eta = 1/3$. By plugging that value in the above, the new expression is minimized for

$$h_* = \left(\frac{2\sigma^2}{(\upsilon'')^2} \right)^{1/3} \tag{78}$$

This results derive from Glynn, 1989.

7.3 Stratification

Consider $\mathbf{x}^{\mathrm{T}} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\mathbf{x}$ has dimension n. Consider the linear restrictions:

$$\mathbf{R} = \{\mathbf{x} : \gamma_i^- \leq x_i \leq \gamma_i^+, \quad i \in\} \tag{79}$$

where the $\gamma_i^- = \infty$ and/or $\gamma_i^{\pm} = \pm\infty$ are allowed for some i, but for at least one i, $|\gamma_i^{\pm}| < \infty$. Our problem is to sample from $p_*(\mathbf{x}) \propto p(\mathbf{x})\mathbf{1}_{[\mathbf{R}]}(\mathbf{x})$.

Let's consider the unrestricted case first. A standard property of the normal distribution is:

$$x_i | \mathbf{x}_{-i} \sim N(\mu_{1|-1} | \Sigma_{1|-1}) \tag{80}$$

where

$$\mu_{1|-1} = \mu_i + \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}(\mathbf{x}_{-i} - \boldsymbol{\mu}_{-i}) \tag{81}$$
$$\Sigma_{1|-1} = \Sigma_{ii} - \Sigma_{i,-i}\Sigma_{-i,-i}^{-1}\Sigma_{-i,i} \tag{82}$$

By sampling $x_i \sim$, and iterating over all i's, we obtain a sample from the desired multivariate distribution. This is a particular application the Gibbs algorithm. It's efficiency is inversely related to n and the degree of positive correlation between the x_i's. However, it is of interest to us as it is relatively straightforward to sample from a truncated univariate normal, by the method of inverse cdf:

$\gamma_i^{\pm} \leftarrow (\gamma_i^{\pm} - \mu_{i|-i})/\Sigma_{i|-i}$
$u \sim \mathrm{Unif}[\Phi(\gamma_i^-), \Phi(\gamma_i^+)]$
$x_i^* \leftarrow \Sigma_{i|-i}\Phi^{-1}(u) + \mu_{i|-i}$

For the reasons previously mentioned, we need to consider a more efficient algorithm. Philippe and Robert, 2003 propose a perfect sampling scheme. Troughton and Godsill, propose a "Gaussian windowing" approach with intuitive appeal and simplicity. Consider as candidate density, $q(.)$

$$N(\mathbf{x}|\boldsymbol{\mu}_*, \Sigma_*) \propto N(\mathbf{x}|\boldsymbol{\mu}, \Sigma) N(\mathbf{x}|\bar{\boldsymbol{\gamma}}, \times\Lambda_\kappa) \tag{83}$$

where

$$\bar{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}^- + \boldsymbol{\gamma}^-)/2 \tag{84}$$

$$\Delta\boldsymbol{\gamma} = \boldsymbol{\gamma}^+ - \boldsymbol{\gamma}^- \tag{85}$$

$$\Lambda_\kappa = \kappa \times diag(\Delta\boldsymbol{\gamma}) \tag{86}$$

so that

$$\Sigma_* = (\Sigma^{-1} + \Lambda_\kappa^{-1})^{-1} \tag{87}$$

$$\mu_* = \Sigma_*(\Sigma^{-1}\boldsymbol{\mu} + \Lambda_\kappa^{-1}\bar{\boldsymbol{\gamma}}) \tag{88}$$

We are no longer restricted to sample from each conditional univariate distribution. Furthermore, the mean is a weighted average of the true mean and the center of the bounds. Decreasing κ increases the probability that the draws will fall in the hypercube; conditional on that event, however, $w(.) = q(.)/p(.)$ will be more variable. In this context, the adaptive importance sampling method described below seems highly relevant, likewise for MCMC.

7.4 Adaptive importance sampling

We recall that μ is the prior mean in $p(.)$. Here, we define a family of sampling distributions indexed by λ: $\{q_\lambda(.), \lambda \in \Lambda\}$. In the case of model (2), this will be, for example, $q_\lambda(.) = p(.|\mu = \lambda)$.

For a vector $z \sim q_*(z)$, and a quantity of interest $g = \mathrm{E}[z]$, the method traditionally proceeds as follows (see Oh and Berger, 1992). First we find Λ such that $\mathrm{E}_\lambda[\Lambda(z)] = \lambda$. Starting from and initial guess $\lambda = \lambda_0$, we update it as follows

$j \leftarrow 0$; $n_0 \leftarrow 0$
for $j = 1$ to J **do**
  Generate an iid sample $z_1, ..., z_{n_j}$ from $q_{\lambda_{j-1}}$
  $\lambda_j \leftarrow \hat{\mathrm{E}}_{\lambda_{j-1}}[w_{j-1}(z)\Lambda(z)]$
  $\hat{g} \leftarrow (N_{j-1}\hat{g} + n_j\hat{\mathrm{E}}_{\lambda_1}[w_1(z)z])/(N_{j-1} + n_j)$
  $N_{j-1} \leftarrow N_{j-1} + n_j$
end for

We would expect $\lambda \xrightarrow{p} \lambda_* = \mathrm{E}_*[\Lambda(z)]$ and therefore this method will be most effective if q_{λ_*} is a good approximation to $q_*(.)$. The dimensionality of Λ should be inversely related to expected variability of the data.

A possible variant is to update λ the same way we update g, that is by smoothing it across iterations. Moreover the coefficient of smoothing can be set according to the estimated variability in each sample. Finally, we can replace the moment matching step by a sample variance minimizing step, that is $\lambda_l = \min_\lambda \widehat{\mathrm{Var}}_{\lambda_{l-1}}[w_{l-1}(z)z]$ (see Pasarica and Gelman, 2005 for a successful implementation of this idea in a Markov chain context).

7.5 misc

Given $x \sim N(\mu, \sigma^2)$,

$$\begin{aligned} \mathrm{E}[x1\{a \leq x \leq b\}] = &- \sigma(\phi((b-\mu)/\sigma) \\ &- \phi((a-\mu)/\sigma) \\ &+ \mu(\Phi((b-\mu)/\sigma) \\ &- \Phi((a-\mu)\sigma)) \end{aligned} \tag{89}$$

Bibliography

[1] Bradley P. Carlin, Joseph B. Kadane, and Alan E. Gelfand. "Approaches for Optimal Sequential Decision Analysis in Clinical Trials". In: *Biometrics* 54 (1998), pp. 964–975.