Importance methods

Lecturer: Riccardo Corradin

Welcome back importance sampling

- **Importance sampling** is one of the most commonly used Monte Carlo methods, which often can simplify complex problem by resorting to instrumental distributions.

- We assume the usual setting, where we want to produce a sample from the posterior distribution of $\boldsymbol{\theta} \mid \boldsymbol{X}$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is the parameter of interest, $\boldsymbol{X} \in \mathbb{X}^n$ denotes the observed data, with $\mathbb{X} \in \mathbb{R}^d$.

- As usual $\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto \mathrm{L}(\boldsymbol{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$ the density function of the posterior distribution.

- Recall that, in a Monte Carlo setting, we want to estimate

$$\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left[g(\boldsymbol{\theta})\right] = \int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{\theta},$$

  by resorting to a sampling procedure

$$\frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}^{(r)}), \qquad \boldsymbol{\theta}^{(r)} \overset{iid}{\sim} \pi(\boldsymbol{\theta} \mid \boldsymbol{X}).$$

- Importance sampling, among other motivations, is a suitable sampling strategy when we can **easily evaluate the density function** $\boldsymbol{\theta} \mid \boldsymbol{X}$, but it is **difficult to sample** from such a distribution.

Rejection sampler, an old friend

- We recall one of the basic Monte Carlo sampler, the rejection sampler.
- Suppose we want to sample from $\pi(\boldsymbol{\theta} \mid X)$ using an auxiliary distribution $h(\boldsymbol{\theta})$. With rejection sampling we can produce a sample as follows.

At the rth sampling step

i) We generate $\boldsymbol{\theta}^{(r)} \overset{iid}{\sim} h(\boldsymbol{\theta})$ (independent of the previous state).

ii) Accept $\boldsymbol{\theta}^{(r)}$ with probability

$$\frac{\pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})}{Kh(\boldsymbol{\theta}^{(r)})}, \qquad \text{with} \qquad K \geq \max_{\boldsymbol{\theta}} \frac{\pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})}{h(\boldsymbol{\theta}^{(r)})},$$

otherwise go back to i).

- Here we still have to evaluate the likelihood function, we can extend this algorithm accommodating for synthetic data.
- The support of $h(\cdot)$ should cover the support of $\pi(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X})$.

Normalised importance sampling

- Importance sampling get advantage of knowing another density function, say $q(\boldsymbol{\theta})$, for which $\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) \ll q(\boldsymbol{\theta})$.

- In practice, we use the distribution $q(\boldsymbol{\theta})$ to sample, but we also exploit its relation with the targed distribution.

- The crucial part associated with importance sampling (IS) is summarized in the following identity

$$\int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{\theta} = \int_{\Theta} g(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{X})}{q(\boldsymbol{\theta})}q(\boldsymbol{\theta})\mathrm{d}\theta,$$

  where $w(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}|\boldsymbol{X})}{q(\boldsymbol{\theta})}$ is called the importance weight function.

- The previous identity holds if $q(\boldsymbol{\theta}) > 0$ whenever $g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) > 0$.

- This justify the use of the following Monte Carlo estimator

$$\int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{\theta} \approx \frac{1}{R}\sum_{r=1}^{R} w(\boldsymbol{\theta}^{(r)})g(\boldsymbol{\theta}^{(r)}), \qquad w(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{X})}{q(\boldsymbol{\theta})}, \qquad \boldsymbol{\theta}^{(r)} \overset{iid}{\sim} q(\boldsymbol{\theta}).$$

Accuracy of IS

- For a generic unbiased Monte Carlo estimator,

$$MSE_\pi\left(\frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}^{(r)})\right) = \frac{1}{R}\mathrm{Var}_\pi(g(\boldsymbol{\theta})).$$

- The IS estimator is unbiased, since

$$\mathrm{E}_q\left[\frac{1}{R}\sum_{r=1}^{R} w(\boldsymbol{\theta}^{(r)})g(\boldsymbol{\theta}^{(r)})\right] = \mathrm{E}_\pi[g(\boldsymbol{\theta})].$$

- The MSE committed while doing IS is equal to

$$MSE\left(\frac{1}{R}\sum_{r=1}^{R} w(\boldsymbol{\theta}^{(r)})g(\boldsymbol{\theta}^{(r)})\right) = \frac{1}{R}\mathrm{Var}_q(w(\boldsymbol{\theta})g(\boldsymbol{\theta})).$$

- The IS estimator is much useless if $\mathrm{E}_q[w(\boldsymbol{\theta})^2 g(\boldsymbol{\theta})^2] < \infty$ does not hold.
- A sufficient condition is that $w(\boldsymbol{\theta})$ is upper bounded in $\boldsymbol{\theta}$.

Optimal IS

- Among all the possible importance distribution, we can find an optimal one which maximize the accuracy of the sampler.

Theorem

The MSE of the normalized IS estimator of $\mathrm{E}_\pi[g(\boldsymbol{\theta})]$ *is minimized by*

$$\tilde{q}(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta} \mid \boldsymbol{X})|g(\boldsymbol{\theta})|}{\int_\Theta |g(\boldsymbol{\theta})|\pi(\boldsymbol{\theta} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{\theta}}.$$

- Unfortunately, the previous is quite useless given that we need to compute the denominator, which contains a similar functional to the one we try to avoid.
- In practice, we want the proposal distribution to be close as possible to $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ within a class of tractable distribution.
- Further, we can tune the proposal distribution depending on specific functional we are considering.

IS in practice

- One of the main application of IS is when we are interested on specific functional of the posterior distribution.
- For example, if we want to integrate distributional tails with extreme quantiles, associated to rare event probabilities.

Suppose we want to quantify the probability mass that a standard Gaussian distribution $\pi(\theta \mid \boldsymbol{X})$ is leaving on the right tail, after the value $\theta = 10$, i.e. $1 - \Phi(10)$. Hence, the functional of interest is given by $g(\theta) = \phi(\theta) 1_{[\theta > 10]}$. Resorting to other samplers, the accuracy on the tail is quite low since is rare to sample values on that part of the support.

With a generic IS the accuracy can also be quite low. A more smart proposal distribution concentrate the effort in the relevant part of the support. Hence, we can consider an IS with

$$q(\theta) = \lambda e^{-\lambda(\theta - c)} 1_{[\theta \geq c]},$$

where all the sampling is concentrated on relevant part of the support, and then weighted by

$$w(\theta) = \frac{\pi(\theta \mid \boldsymbol{X})}{q(\theta)},$$

the ratio between the density function of a gaussian distribution and a shifted exponential distribution.

Auto-normalised importance sampling

- In practice, the posterior distribution of interest is usually known up to a normalization constant. But usually, we can easily evaluate

$$\pi(\boldsymbol{\theta}, \boldsymbol{X}) = \mathrm{L}(\boldsymbol{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

- Similarly, let us consider $q(\boldsymbol{\theta}) = \tilde{q}(\boldsymbol{\theta})/C_q$, with C_q being the normalization constant of $\tilde{q}(\boldsymbol{\theta})$.

- The IS works also up to normalization constants of both target distribution and proposal distribution, since

$$\int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{X}\mathrm{d}\boldsymbol{\theta}) = \frac{\int_{\Theta} g(\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta}|\boldsymbol{X})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta} \frac{\pi(\boldsymbol{\theta}|\boldsymbol{X})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{\int_{\Theta} g(\boldsymbol{\theta}) \frac{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})}{\tilde{q}(\boldsymbol{\theta})} q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int_{\Theta} \frac{\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{X})}{\tilde{q}(\boldsymbol{\theta})} q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},$$

  which leads to the following Monte Carlo estimator

$$\mathrm{E}_{\pi}[g(\boldsymbol{\theta})] \approx \frac{\sum_{r=1}^{R} w(\boldsymbol{\theta}^{(r)}) g(\boldsymbol{\theta}^{(r)})}{\sum_{r=1}^{R} w(\boldsymbol{\theta}^{(r)})}, \qquad w(\boldsymbol{\theta}) = \frac{\tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{X})}{q(\boldsymbol{\theta})}, \qquad \boldsymbol{\theta}^{(r)} \overset{iid}{\sim} q.$$

A formal viewpoint of IS

- Instead of thinking as density function $\pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ and $q(\boldsymbol{\theta})$, we can view the IS in terms of probability measure $\mathbb{P}$ and $\mathbb{Q}$.
- We assume $\mathbb{P} \ll \mathbb{Q}$.
- For any test function $g(\cdot)$, we denote the expectation w.r.t. $\mathbb{P}$ as $\mathbb{P}(g) := \int_\Theta g(\boldsymbol{\theta})\mathbb{Q}(\mathrm{d}\boldsymbol{\theta})$.
- Recall that the potentially un-normalized Radon-Nikodym derivative between $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$w(\boldsymbol{\theta}) \propto \frac{\mathbb{P}(\mathrm{d}\boldsymbol{\theta})}{\mathbb{Q}(\mathrm{d}\boldsymbol{\theta})},$$

meaning that

$$\mathbb{P}(A) = \int_A w(\boldsymbol{\theta})\mathbb{Q}(\mathrm{d}\boldsymbol{\theta}), \qquad \forall A \subseteq \Theta.$$

Theorem

Let $w(\boldsymbol{\theta}) \propto \mathbb{P}(\mathrm{d}\boldsymbol{\theta})/\mathbb{Q}(\boldsymbol{\theta})$. Then, for any test function g, we have

$$\mathbb{Q}(gw) = \mathbb{P}(g)\mathbb{Q}(w).$$

A formal viewpoint of IS

- In force of the previous identity, we can rethink at the IS as a change of measure, where we are changing from $\mathbb{Q}$ to $\mathbb{P}$.

- From the previous identity,

$$\mathbb{Q}(gw) = \mathbb{P}(g)\mathbb{Q}(w),$$

  we have the following estimator of $\mathbb{P}(g)$

$$\mathbb{P}(g) \approx \sum_{r=1}^{R} W^{(r)} g(\boldsymbol{\theta}_r), \qquad W^{(r)} = \frac{w(\boldsymbol{\theta}_r)}{\sum_{r=1}^{R} w(\boldsymbol{\theta}_r)}, \qquad \boldsymbol{\theta}_r \sim \mathbb{Q}.$$

- We can interpret the estimator as the expectation of g w.r.t. the random probability measure

$$\mathbb{P}^R(\mathrm{d}\boldsymbol{\theta}) = \sum_{r=1}^{R} W^{(r)} \delta_{\boldsymbol{\theta}_r}(\mathrm{d}\boldsymbol{\theta}), \qquad \boldsymbol{\theta}_r \sim \mathbb{Q}.$$

- $\mathbb{P}^R$ is called the particle approximation of $\mathbb{P}$, while we refer to $\{\boldsymbol{\theta}_r, W^{(r)}\}_{r=1}^{R}$ as a weighted sample.

- Weak convergence hold, for which $\mathbb{P}^R \Rightarrow \mathbb{P}$.

Measuring IS performance

- Similarly to what we have done with MCMC, we can assess the behavior of IS resorting to the effective sample size.
- For IS, it can be expressed in a closed form as

$$\mathrm{ESS}(\{W^{(r)}\}_{r=1}^{R}) = \frac{1}{\sum_{r=1}^{R}(W^{(r)})^2} = \frac{\left[\sum_{r=1}^{R} w(\boldsymbol{\theta}_r)\right]^2}{\sum_{r=1}^{R} w(\boldsymbol{\theta}_r)^2},$$

  where $\mathrm{ESS}(\{W^{(r)}\}_{r=1}^{R}) \in [1, R]$.
- $\mathrm{ESS}(\{W^{(r)}\}_{r=1}^{R}) = R$ if $\mathbb{P} = \mathbb{Q}$, hence our proposal is the target density.
- The closer is getting $\mathrm{ESS}(\{W^{(r)}\}_{r=1}^{R}) = R$ to R, the better is our proposal distribution (and close to $\mathbb{P}$).
- In practice, once we collect our sample, we can assess its quality by computing empirically the $\mathrm{ESS}(W_{1:R})$ with the observed weights.

Curse of dimensionality and IS

- It is a commonplace that importance sampling suffers from the curse of dimensionality.
- Consider again distributions $\mathbb{P}$ and $\mathbb{Q}$, with $\mathbb{P} \ll \mathbb{Q}$, and let

$$\mathbb{P}_p(\mathrm{d}\theta_{1:p}) = \prod_{j=1}^{p} \mathbb{P}(\mathrm{d}\theta_j), \qquad \mathbb{Q}_p(\mathrm{d}\theta_{1:p}) = \prod_{j=1}^{p} \mathbb{Q}(\mathrm{d}\theta_j),$$

  i.e. both $\mathbb{P}_p$ and $\mathbb{Q}_p$ factorize in i.i.d. copies of $\mathbb{P}$ and $\mathbb{Q}$, respectively.

- Hence, if we consider again $w(\theta) = \mathrm{d}\mathbb{P}(\theta)/\mathrm{d}\mathbb{Q}(\theta)$ (normalized here), we have

$$w_p(\theta_{1:p}) = \frac{\mathrm{d}\mathbb{P}_p}{\mathrm{d}\mathbb{Q}_p} = \prod_{j=1}^{p} w(\theta_j).$$

  Then,

$$\mathrm{Var}_{\mathbb{Q}_p}(w(\theta_{1:p})) = \mathbb{Q}(w^2)^p - 1,$$

  which grows exponentially in p number of dimensions.

The Pima Indian dataset

- We consider again the "famous" **Pima Indian** dataset, with $n = 532$ and $p = 8$.
- The purpose of this exercise is mainly to present the implementation of the various IS algorithms and show their performance in this specific example.
- **The following results should not be generalized** to any statistical models nor even to any logistic regression model.
- We consider once again a logistic model, with

$$y_i \mid \lambda_i \overset{ind}{\sim} \mathcal{B}ern(\lambda_i), \qquad \lambda_i = g(\eta_i), \qquad \eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

  and a priori an independent component Gaussian prior, with $\boldsymbol{\beta} \sim N_p(\mathbf{0}, 100\mathrm{I}_p)$.
- We consider different proposal distributions.

Naive covariance matrix

- Let us start with a naive choice for the proposal covariance $S = 10^3 I_p$, centering the distribution in $\mathbf{0}$.
- This "random" choice of S works terribly, producing garbage results.

```
# Covariance matrix and expectation of the proposal
S <- diag(100, ncol(X))
mu <- rep(0, 8)

# Running the IS (R = 30000)
IS_not_informed <- IS(R, y, X, S, mu)

est_coef <- colSums(IS_not_informed$values * exp(IS_small$weights))
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -0.866  0.433  1.201 -0.039  0.060  0.602  0.401  0.145

ESS <- 1 / sum(exp(IS_not_informed$weights)^2)
# 1

idx <- sample(1:R, R, TRUE, exp(IS_not_informed$weights))
length(unique(idx))
# 1
```

Informed covariance matrix

- We now set the proposal parameters according to the Laplace approximation of the posterior distribution.

```
# Covariance matrix and expectation of the proposal
S <- vcov(fit_logit)
mu <- coefficients(fit_logit)

# Running the IS (R = 30000)
IS_informed <- IS(R, y, X, S, mu)

est_coef <- colSums(IS_not_informed$values * exp(IS_small$weights))
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -0.865  0.253  0.943 -0.068  0.234  0.424  0.324  0.397

ESS <- 1 / sum(exp(IS_not_informed$weights)^2)
# 2876.212

idx <- sample(1:R, R, TRUE, exp(IS_not_informed$weights))
length(unique(idx))
# 7378
```

Importance resampling

Motivation

- Resampling is the action of drawing randomly from a weighted sample, so as to obtain an unweighted sample.
- Resampling has the curious property of potentially reducing the variance at a later stage. This point is crucial for the good performance of particle algorithms.
- We have the following particle approximation of $\mathbb{P}_0(\mathrm{d}\boldsymbol{\theta})$

$$\mathbb{P}_0^R(\mathrm{d}\boldsymbol{\theta}) = \sum_{r=1}^{R} W_0^{(r)} \delta_{\boldsymbol{\theta}_0^{(r)}}, \qquad \boldsymbol{\theta}_0^{(r)} \sim \mathbb{Q}_0, \qquad W_0^{(r)} = \frac{w_0(\boldsymbol{\theta}_0^{(r)})}{\sum_{r=1}^{R} w_0(\boldsymbol{\theta}_0^{(r)})},$$

  obtained through importance sampling based on the proposal $\mathbb{Q}_0$ and the weight function w_0.
- Ideally, we want to recycle the previous to approximate the following extended probability measure

$$(\mathbb{P}_0 Q_1)(\mathrm{d}\boldsymbol{\theta}_{0:1}) = \mathbb{P}_0(\mathrm{d}\boldsymbol{\theta}_0) Q_1(\boldsymbol{\theta}_0, \mathrm{d}\boldsymbol{\theta}_1),$$

  where $Q_1(\boldsymbol{\theta}_0, \mathrm{d}\boldsymbol{\theta}_1)$ is a probability kernel function.
  - $\rightarrow$ $Q_1(\boldsymbol{\theta}_0, \cdot)$ is a probability measure
  - $\rightarrow$ $Q_1(\boldsymbol{\theta}_0, A)$ is measurable in $\boldsymbol{\theta}_0$
- There are two solutions to this problem.

Importance resampling I

- At first, we recognize that importance sampling from $\mathbb{Q}_1 = \mathbb{Q}_0 Q_1$ to $\mathbb{P}_0 Q_1$ requires
  - **i)** to sample $\{\boldsymbol{\theta}_0^{(r)}, \boldsymbol{\theta}_1^{(r)}\}_{r=1}^R$ from $\mathbb{Q}_0 Q_1$;
  - **ii)** to compute the weights, which are precisely $w_0(\boldsymbol{\theta}_0^{(r)})$, $r = 1, \ldots, R$, since $\boldsymbol{\theta}_1^{(r)}$ condittionally on $\boldsymbol{\theta}_0^{(r)}$ is sampled from the correct distribution.

- If we want to follow this strategy, the only thing is left to sample is

$$\boldsymbol{\theta}_1^{(r)} \sim Q_1(\boldsymbol{\theta}_0^{(r)}, \mathrm{d}\boldsymbol{\theta}_1), \qquad r = 1, \ldots, R.$$

- This strategy belongs to the **sequential importance sampling** framework.

- Inference relative to $(\mathbb{P}_0 Q_1)$ may be analyzed as standard importance sampling, without paying attention to the intermediate step.

- In practice, we are keeping fixed the weight values, but we resample our particles according to the probability kernel function Q_1, filtering our particles over time.

Importance resampling II

- The second strategy considers a two-step approximation. We first replace in the previous definition of the target extended probability measure $\mathbb{P}_0$ by $\mathbb{P}_0^R$, with

$$\mathbb{P}_0^R(\mathrm{d}\boldsymbol{\theta}_0)Q_1(\boldsymbol{\theta}_0, \mathrm{d}\boldsymbol{\theta}_1) = \sum_{r=1}^{R} W_0^{(r)} Q_1(\boldsymbol{\theta}_0^{(r)}, \mathrm{d}\boldsymbol{\theta}_1)\delta_{\boldsymbol{\theta}_0^{(r)}}(\mathrm{d}\boldsymbol{\theta}_0).$$

- Then, we resample R times from the previous, which is an intermediate approximation, so we sample from

$$\frac{1}{R}\sum_{r=1}^{R}\delta_{\tilde{\boldsymbol{\theta}}_{0:1}^{(r)}}, \qquad \tilde{\boldsymbol{\theta}}_{0:1}^{(r)} \sim \mathbb{P}_0^R(\mathrm{d}\boldsymbol{\theta}_0)Q_1(\boldsymbol{\theta}_0, \mathrm{d}\boldsymbol{\theta}_1).$$

- This second approach is called **importance resampling**, as it is connected to other resampling techniques, such as the bootstrap, where we sample with replacement.

- At first, it seems unappealing, as it has two approximation steps. Intuitively, it should have a larger Monte Carlo error and more computationally expensive. However, the first intuition turns out to be incorrect.

Toy example

- Let us consider the following toy example (Section 9.2, Chopin and Papaspiliopoulos, 2020). We set our sampling space $\Theta = \mathbb{R}$.
  - → Our approximating measure at time 0 is a standard Gaussian distribution, i.e. $\mathbb{Q}_0 \overset{d}{=} N(0,1)$.
  - → Our target distribution is a truncated Gaussian distribution, setting mass in $[-\tau, +\tau]$, i.e. $\mathbb{P}_0 \overset{d}{=} TN_\tau(0,1)$.
  - → $Q_1(\theta_0, \mathrm{d}\theta_1)$ is our probability kernel function, with

$$\theta_1 = \rho\theta_0 + \sqrt{1-\rho^2}U, \qquad U \sim N(0,1).$$

  - → We consider as test function $g(\theta_1) = \theta_1$, identity function, we want to obtain a sample at time 1.

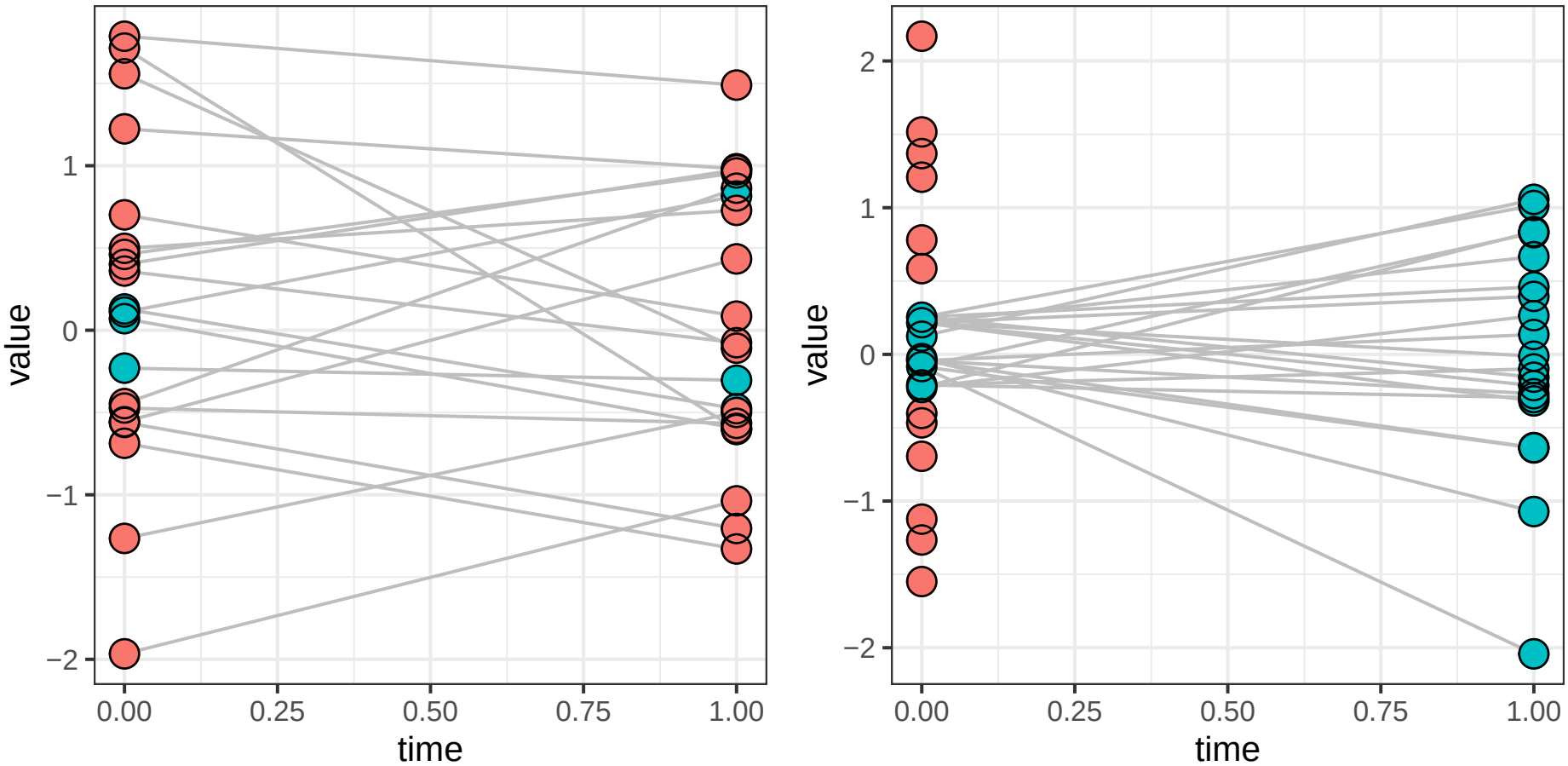
- We have two alternatives

$$\text{importance sampling:} \qquad \hat{g}_{IS}(\theta) = \sum_{r=1}^{R} W_0^{(r)} \theta_1^{(r)}, \qquad (\theta_0^{(r)}, \theta_1^{(r)}) \sim \mathbb{Q}_0 Q_1,$$

$$\text{importance resampling:} \qquad \hat{g}_{IR}(\theta) = \frac{1}{R} \sum_{r=1}^{R} \theta_1^{(r)}, \qquad \theta_1^{(r)} \sim \mathbb{P}_1^R Q_1,$$

where $\theta_1^{(r)} \sim \mathbb{P}_1^R Q_1$ means we draw from the corresponding marginal distribution.

Toy example



Effect of resampling for the toy example with $N = 20$, $\tau = 0.3$, $\rho = 0.5$. Green dots represent particles with non-zero weights.

Toy example

- For this specific case, we have that in the limit behavior

$$\sqrt{R}\hat{g}_{IS} \xrightarrow{d} N(0, \sigma^2_{IS}),$$
$$\sqrt{R}\hat{g}_{IR} \xrightarrow{d} N(0, \sigma^2_{IR}),$$

with

$$\sigma^2_{IS} = \rho^2 \frac{\gamma(\tau)}{S(\tau)} + (1-\rho^2)\frac{1}{S(\tau)},$$
$$\sigma^2_{IR} = \rho^2 \frac{\gamma(\tau)}{S(\tau)} + (1-\rho^2) + \rho^2\gamma(\tau),$$

where $S(\tau) = \mathbb{Q}_0(w_0) = \mathrm{P}(|\theta_0| \leq \tau) = 2\Phi(\tau) - 1$ and $\gamma(\tau) = \mathbb{P}_0(\theta_0^2)$.

- The first term is due to the variability of the particles θ_0 at time 0, equal for both.
- The second term is due to the variability of the innovation terms U (simulated when sampling from kernel Q_1), smaller for importance resampling since $S(\tau) < 1$.
- The third term appears only for IR, and comes from the randomness introduced by the resampling step.
- Further, $(\sigma^2_{IS} - \sigma^2_{IR}) \to +\infty$ for $\tau \to 0$ and any fixed ρ.

Summary

- We have a general structure to filter particles from time 0 to 1, which we can generalize to multiple times and combined with many possible model choices.
  - $\rightarrow$ We can define general algorithms to deal with state-space models.

- Against our intuitions, resampling does not always result in an increase of the estimation variance. Indeed, sometimes, e.g. in presence of some hard constraints such as truncation, the variance can be reduced, as shown in the toy example.

- There are many possible sampling schemes, some of them can be better in specific scenarios. In practice, systematic sampling works pretty well.

- Resampling can be viewed as random weights importance sampling.

Importance resampling with Pima dataset

- We consider again the Pima dataset, but now with the proposal distribution at time 0 being a multivariate Gaussian with parameter based on the Laplace approximation.
- Moving from time 0 to time 1, we resort to the transition kernel of the MALA algorithm.

```
# Covariance matrix and expectation of the proposal
S <- vcov(fit_logit)
mu <- coefficients(fit_logit)
sigma2 <- 0.1

# Running the IrS (R = 30000)
IrS_not_informed <- IrS(R, y, X, S, mu, sigma2)

est_coef <- colSums(IS_not_informed$values * exp(IS_small$weights))
# X       Xnpreg   Xglu    Xbp  Xskin   Xbmi   Xped   Xage
# -0.866  0.433  1.201 -0.039  0.060  0.602  0.401  0.145

ESS <- 1 / sum(exp(IS_not_informed$weights)^2)
# 12362.98

idx <- sample(1:R, R, TRUE, exp(IS_not_informed$weights))
length(unique(idx))
# 13740
```