UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2021
Chapter 6

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Student representatives

- Give feedback to me about class on behalf of all
  - you can still email me directly if you want
- So far one volunteer: (thanks)
  - Anders Bredesen Hatlelid andebh@ifi.uio.no
- In a big class it is good with more than one
  - So still possible

Last time

- Stochastic gradient decent
  - Neural nets, back propagation
  - Spatial model $\widehat{KL}(f_\theta, g) = C - \frac{1}{\binom{n}{m}} \sum_{k=1}^{\binom{n}{m}} \log(f_\theta(y_k|s_k))$

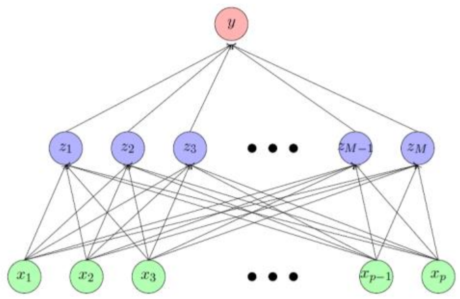


Figure 2: Visualisation of neural network with one hidden layer.

(https://www.researchgate.net/publication/259527954_A_Resampling-Based_Stochastic_Approximation_Method_for_Analysis_of_Large_Geostatistical_Data)

- 1D methods for integration $O(n^{-r})$
- Monte Carlo method in higher dimensions (R^d)
  - MC: $O\left(n^{-1/2}\right)$ vs Fubini $O\left(n^{-r/d}\right)$
  - Provided: $\mathrm{var}(h(X)) < \infty$
- Sampling by inversion and transformation
- Random number generator (RNG)
  - Reproducible randomness = assign seed in a PRNG

Last time (and some more today)

- Common set up in many cases.
  - Want to sample from $f(x)$, but get sample from $g(x)$
- Rejection sampling: Correct the error by an acceptance step
  - Need a bound of $f(x)/g(x)$ or ($q(x)/g(x)$ where $q(x) \propto f(x)$)
  - Each sample is a random sample from $f(x)$
  - Need to sample many times to get one sample

Simulation techniques

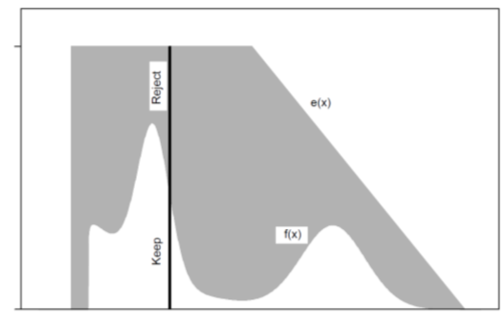
- Exact methods
  - Inversion/transformation methods
  - Rejection sampling
- Approximate methods
  - Sampling importance resampling
  - Sequential Monte Carlo
  - Markov chain Monte Carlo (Chapter 7 and 8)
- Variance reduction methods
  - Importance sampling
  - Antithetic sampling
  - Control variates
  - Rao-blackwellization
  - Common random numbers

Today

- Importance sampling
- Sampling importance Resampling (SIR)
- Sequential Monte Carlo

Rejection sampling

- Assume $\exists \alpha \leq 1$ such that for all x: $f(x) \leq g(x)/\alpha \equiv e(x)$ (the envelope)
- Algorithm:
  1. Sample $Y \sim g(\cdot)$.
  2. Sample $U \sim \text{Unif}(0,1)$.
  3. If $U \leq f(Y)/e(Y)$, put $X = Y$, otherwise return to step 1



- Distribution of X:

$$\Pr(X \leq x) = \Pr(Y \leq x | U \leq \tfrac{f(Y)}{e(Y)}) = \frac{\Pr(Y \leq x, U \leq \frac{f(Y)}{e(Y)})}{\Pr(U \leq \frac{f(Y)}{e(Y)})}$$

$$= \frac{\int_{-\infty}^{x} \int_0^{f(y)/e(y)} du g(y) dy}{\int_{-\infty}^{\infty} \int_0^{f(y)/e(y)} du g(y) dy} = \frac{\int_{-\infty}^{x} \frac{f(y)}{e(y)} g(y) dy}{\int_{-\infty}^{\infty} \frac{f(y)}{e(y)} g(y) dy}$$

$$= \int_{-\infty}^{x} f(y) dy = F(x)$$

$$\alpha = \frac{\text{Area "inner white"}}{\text{Area "inner white" \& gray}}$$

- $\alpha = \Pr(U \leq \frac{f(Y)}{e(Y)})$ is the probability for acceptance
- α^{-1} is the expected number of iterations.

Importance sampling

- If we are unable to sample from $f(x)$ can we still use Monte Carlo methods to compute integrals?

$$\mu = \int h(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

- Lets say we can sample form another distribution $g(x)$ which is quite similar to $f(x)$

$$g(x) \approx f(x)$$

Importance sampling

- Rewriting

$$\mu = \int h(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x} = \frac{\int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x}}{\int \frac{f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x}}$$

- Assume $X_1, \dots, X_n$ iid from $g(\mathbf{x})$. (We know how to sample from g(**x**))

- Two alternative estimates

$$\hat{\mu}_{IS}^* = \frac{1}{n}\sum_{i=1}^n h(\mathbf{X}_i)w^*(\mathbf{X}_i), \quad w^*(X_i) = \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}$$

$$\hat{\mu}_{IS} = \sum_{i=1}^n h(\mathbf{X}_i)w(\mathbf{X}_i), \quad w(\mathbf{X}_i) = \frac{w^*(\mathbf{X}_i)}{\sum_{j=1}^n w^*(\mathbf{X}_j)}$$

- $w^*(\mathbf{X}_i)$ called importance weights
- $w(\mathbf{X}_i)$ called the normalized importance weights

Importance sampling version 1

$$\hat{\mu}^*_{IS} = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i), \quad w^*(\mathbf{X}_i) = \frac{f(\mathbf{X}_i)}{g(\mathbf{X}_i)}$$

$$E[w^*(\boldsymbol{X}_i)] = \int \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{x})d\boldsymbol{x} = 1$$

$$E[\hat{\mu}^*_{IS}] = \int h(\boldsymbol{x})\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x})d\boldsymbol{x} = \int h(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x} = \mu$$

$$\text{Var}[\hat{\mu}^*_{IS}] = \frac{1}{n}\text{Var}^g[h(\mathbf{X})w^*(\mathbf{X})] = \frac{1}{n}\text{Var}^g[t(\boldsymbol{X})]$$

$$\boxed{t(\boldsymbol{X}) = h(\boldsymbol{X})w^*(\boldsymbol{X})}$$

- Can be unstable if $g(\mathbf{x})$ small when $f(\mathbf{x})$ large
- $g(\mathbf{x})$ should have heavier tails than $f(\mathbf{x})$.
- If only one $h(\mathbf{X})$ of interest, should choose

$$g(\mathbf{x}) \propto |h(\mathbf{x})|f(\mathbf{x})$$

- Often interested in many functions, focus on making variability of $w^*(\mathbf{X})$ small

Importance sampling version 2

$$\hat{\mu}_{IS} = \sum_{i=1}^{n} h(\mathbf{X}_i)w(\mathbf{X}_i), \quad w(\mathbf{X}_i) = \frac{w^*(\mathbf{X}_i)}{\sum_{j=1}^{n} w^*(\mathbf{X}_j)}$$

Normalized weigths

- Based on

$$\mu = \frac{\int \frac{h(\mathbf{x})f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})d\mathbf{x}}{\int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})d\mathbf{x}} = \frac{\mu}{1} \approx \frac{\hat{\mu}^*_{IS}}{\hat{1}^*_{IS}} = \hat{\mu}_{IS}$$

$$\hat{\mu}^*_{IS} = \bar{t}, \quad t_i = t(\mathbf{X}_i) = h(\mathbf{X}_i)w^*(\mathbf{X}_i)$$

$$\hat{1}^*_{IS} = \bar{w}^*$$

- Why also estimate denominator?
  - What would be best if h(x) = c (constant)?
  - Correlations between nominator and denominator

Impact of normalization

$$\hat{\mu}^*_{IS} = \bar{t}, \quad \hat{1}^*_{IS} = \bar{w}^*$$

$$\frac{\hat{\mu}^*_{IS}}{\hat{1}^*_{IS}} = \hat{\mu}_{IS}$$

- Taylor approximation of $1/\bar{w}^*$ around 1:

$$\frac{1}{\bar{w}^*} \approx 1 - (\bar{w}^* - 1) + (\bar{w}^* - 1)^2$$

giving

$$\hat{\mu}_{IS} \approx \bar{t}\,[1 - (\bar{w}^* - 1) + (\bar{w}^* - 1)^2]$$
$$= \bar{t} - (\bar{t} - \mu)(\bar{w}^* - 1) - \mu(\bar{w}^* - 1) + \bar{t}\,(\bar{w}^* - 1)^2$$

$$E[\hat{\mu}_{IS}] = E\{\bar{t} - (\bar{t} - \mu)(\bar{w}^* - 1) - \mu(\bar{w}^* - 1) + \bar{t}\,(\bar{w}^* - 1)^2\} + \mathcal{O}(n^{-2})$$
$$= \mu - \frac{1}{n}\text{cov}[t(\boldsymbol{X}), w(\boldsymbol{X})] - 0 + \frac{\mu}{n}\text{var}(w(\boldsymbol{X})) + \mathcal{O}(n^{-2})$$

$$\text{var}[\hat{\mu}_{IS}] = E\left\{\left((\bar{t} - \mu) - \mu(\bar{w}^* - 1)\right)^2\right\} + \mathcal{O}(n^{-2})$$
$$= \frac{1}{n}\left[\text{var}\big(t(\boldsymbol{X})\big) + \mu^2\text{var}\big(w^*(\boldsymbol{X})\big) - 2\mu\cdot\text{cov}[t(X), w^*(X)]\right] + \mathcal{O}(n^{-2})$$

$$\text{MSE}[\hat{\mu}_{IS}] - \text{MSE}[\hat{\mu}^*_{IS}] = \frac{1}{n}\left(\mu^2\text{var}[w^*(\mathbf{X})] - 2\mu\text{cov}[t(\mathbf{X}), w^*(\mathbf{X})]\right) + \mathcal{O}(n^{-2})$$

When is normalization better?

$$\text{MSE}[\hat{\mu}_{IS}] - \text{MSE}[\hat{\mu}^*_{IS}] = \frac{1}{n}\left(\mu^2 \text{var}[w^*(\mathbf{X})] - 2\mu\text{cov}[t(\mathbf{X}), w^*(\mathbf{X})]\right) + \mathcal{O}(n^{-2})$$

- Gain if

$$\text{cov}[t(\mathbf{X}), w^*(\mathbf{X})] > \frac{\mu\text{var}[w^*(\mathbf{X})]}{2}$$

$$\Updownarrow$$

$$\text{cor}[t(\mathbf{X}), w^*(\mathbf{X})] > \frac{\sqrt{\text{var}[w^*(\mathbf{X})]}}{2\sqrt{\text{var}[t(\mathbf{X})]}/\mu} = \frac{\text{cv}[w^*(\mathbf{X})]}{2\text{cv}[t(\mathbf{X})]}$$

- Example: `imp_samp_beta.R`

Coefficient of variation:
cv(X)=std(X)/E(X)

What can go wrong???

- Monte Carlo integration:
  - $E_f\big(h(X)\big) < \infty$ (this is the number we want)
  - $E_f(h(X)^2) < \infty$ (this is additional requirement)
- Importance integration:
  - $E_g\big(h(X)w^*(X)\big) = E_f\,(h(X)) < \infty$ (ok ☺)
  - $E_g\Big(\big(h(X)w^*(X)\big)^2\Big) = E_f\big(h(X)^2 w^*(X)\big) < \infty$ (??)

$w^*(X) = \frac{f(X)}{g(X)}$ in rejection sampling this is bounded by α^{-1}

Effective sample size

- Assume $w_i = w(\mathbf{X}_i), i = 1, \ldots, n$ are normalized weights
- Define effective sample size by

$$\widehat{N}_{eff} = \frac{1}{\sum_{i=1}^{n} w_i^2}$$

Ex 1:	if $w_i = \frac{1}{n}$ for all i	$\widehat{N}_{eff} = n$
Ex 2:	if $w_i = 0, i \leq z, w_i = \frac{1}{n-z}, i > z$	$\widehat{N}_{eff} = n - z$
Ex 3:	if $w_i = 0, i \neq j, w_j = 1$	$\widehat{N}_{eff} = 1$

Sampling importance resampling

- Assume now we want to sample from $f(\mathbf{x})$, difficult
- Easy to sample from $g(\mathbf{x})$.
- Sampling importance resampling
  1. Sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ iid from g
  2. Calculate standardized importance weights

$$w(\mathbf{Y}_i) = \frac{f(\mathbf{Y}_i)/g(\mathbf{Y}_i)}{\sum_{j=1}^m f(\mathbf{Y}_j)/g(\mathbf{Y}_j)}, i = 1, \ldots, m$$

  3. Resample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ with probabilities $w(\mathbf{Y}_1), \ldots, w(\mathbf{Y}_m)$
- Properties: As $m \to \infty$
  - X_i converges in distribution to $f(\mathbf{x})$
  - Correlations between X_i's decreases to zero
- For finite m: Correlation between samples

Sampling importance resampling

- Assume
  - $Y_1, \ldots, Y_m$ iid from g
  - $X_1, \ldots, X_n$ resampled from $\{Y_1, \ldots, Y_m\}$, $w(Y_i) = \frac{f(Y_i}{g(Y_i)}$
- Two possible estimates of $\mu = E^f[X]$:

$$\hat{\mu}_{SIR} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

$$\hat{\mu}_{IS} = \sum_{i=1}^{m} w(Y_i)Y_i$$

Can show

$$E[(\hat{\mu}_{IS} - \mu)^2] \leq E[(\hat{\mu}_{SIR} - \mu)^2]$$

Why consider SIR?

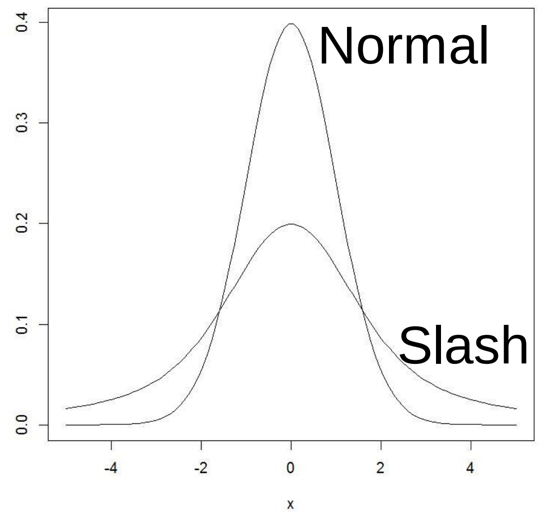
- Sometimes beneficial to have **equally weighted** samples
- May be beneficial at a later stage of analysis process
- If we want to evaluate E(*h(x)*) where *h(x)* is hard to evaluate
- Usually n < m

Example: slash distribution

- Controlled example (we know the truth)
- Y has slash distribution when $Y = \frac{X}{U}$
  X$\sim N(0,1), U \sim \text{Unif}(0,1)$

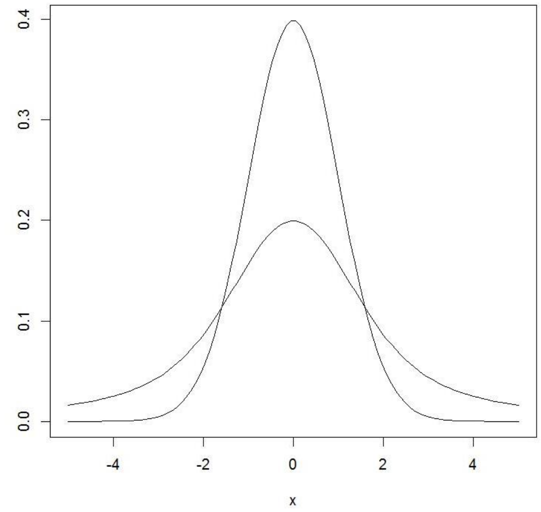
$$f(y) = \begin{cases} \frac{1-\exp\{-y^2/2\}}{y^2\sqrt{2\pi}}, & y \neq 0, \\ \frac{1}{2\sqrt{2\pi}}, & y = 0. \end{cases}$$

- Sampling Experiments
  1. X from Y
  2. Y from X
- Methods
  1. Rejection sampling
  2. Importance sampling
  3. Sampling importance resampling



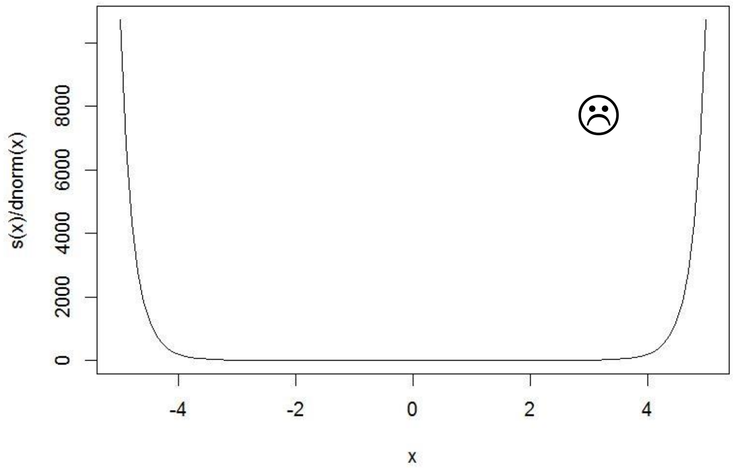
Two test functions

- Ex1: x
- Ex2: h(x)= sin(x)+0.2cos(2*pi*x)

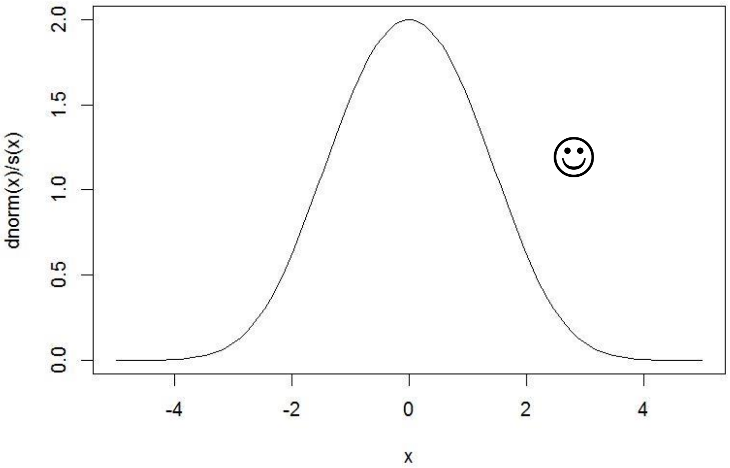


- Ratios:

Slash /normal

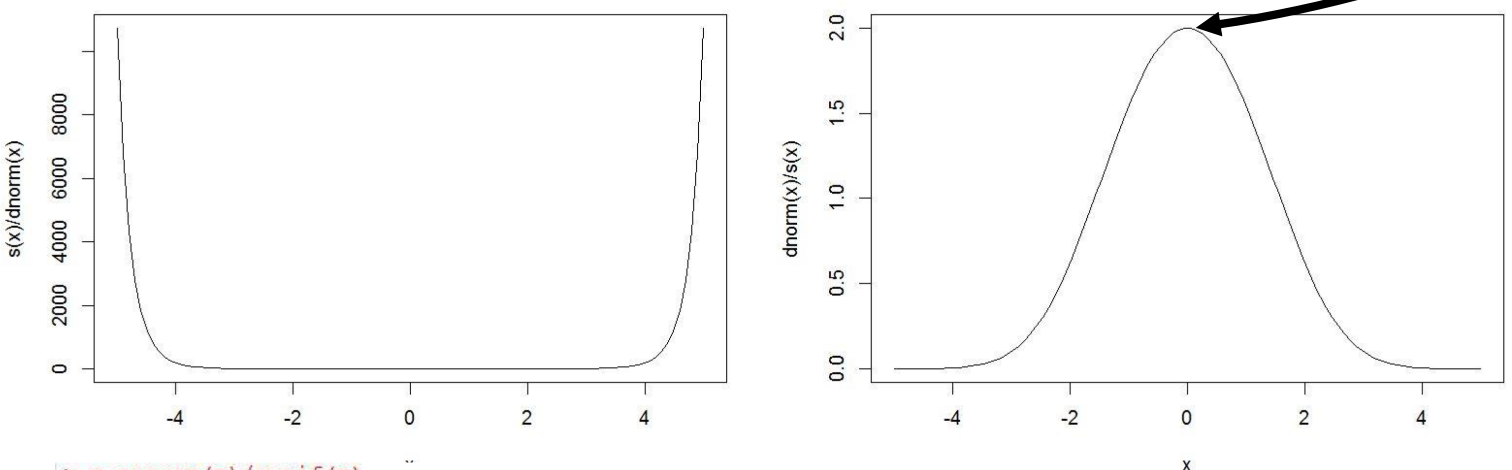


Normal/slash



Rejection sampling

- Normal from slash bounded by 2
- Slash from Normal unbounded
  (no rejection sampling possibel)




```
> y =rnorm(m)/runif(m)
> U=runif(m)
> accept = dnorm(y)/(s(y)*2)
> sample = y[U<accept]
>
> length(sample)
[1] 49979
> max(accept)
[1] 1
> min(accept)
[1] 0
```

Observed acceptance rate: 0.49979
Theoretical acceptance rate: 0.50000

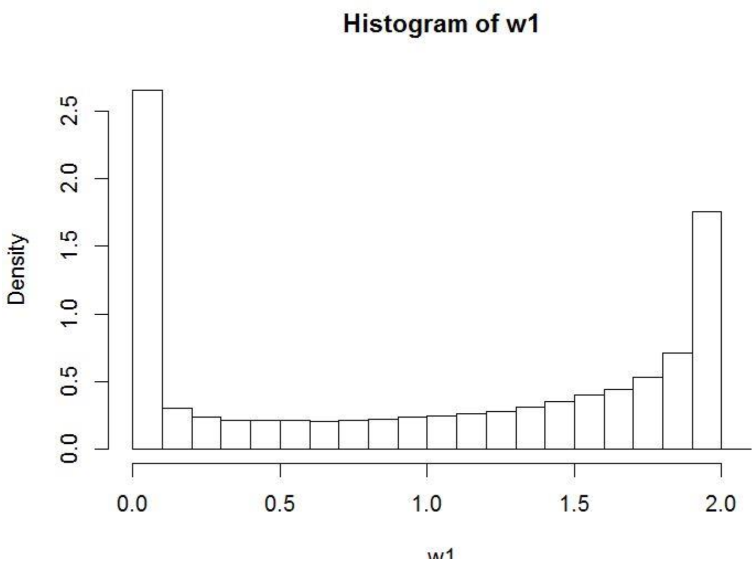
- Ex1: x
- Ex2: h(x)= sin(x)+0.2cos(2*pi*x)

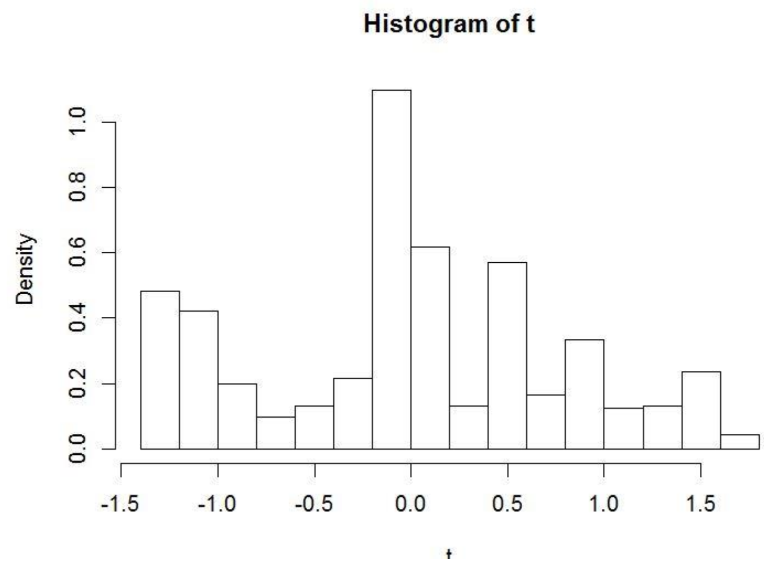
```
> m = 1000
> x=rnorm(m)
> y =rnorm(m)/runif(m)
> show(c(mean(x), mean(h(x)), mean(y), mean(h(y))))
[1] -0.04743281 -0.02314847 -0.71528509  0.01710263
>
>
>
> m = 100000
> x=rnorm(m)
> y =rnorm(m)/runif(m)
> show(c(mean(x), mean(h(x)), mean(y), mean(h(y))))
[1]  0.001369078  0.001019647  0.542154961 -0.004230678
>
>
>
> m = 10000000
> x=rnorm(m)
> y =rnorm(m)/runif(m)
> show(c(mean(x), mean(h(x)), mean(y), mean(h(y))))
[1]  3.115531e-05  4.417861e-05 -8.361942e-01 -2.532640e-05
```

```
> m = 1000
> x=rnorm(m)
> y =rnorm(m)/runif(m)
> show(c(sd(x), sd(h(x)), sd(y), sd(h(y))))
[1]  0.9951861  0.6712401 65.3199546  0.7001361
>
>
> m = 100000
> x=rnorm(m)
> y =rnorm(m)/runif(m)
> show(c(sd(x), sd(h(x)), sd(y), sd(h(y))))
[1]    0.9956829    0.6726304 1328.0501683    0.7122169
>
>
>
> m = 10000000
> x=rnorm(m)
> y =rnorm(m)/runif(m)
> show(c(sd(x), sd(h(x)), sd(y), sd(h(y))))
[1] 9.997296e-01 6.724661e-01 1.110357e+04 7.138005e-01
>
```

Slash distribution does not have a mean
=> The average does not converge

Sample from slash, estimate propeties of normal distribution

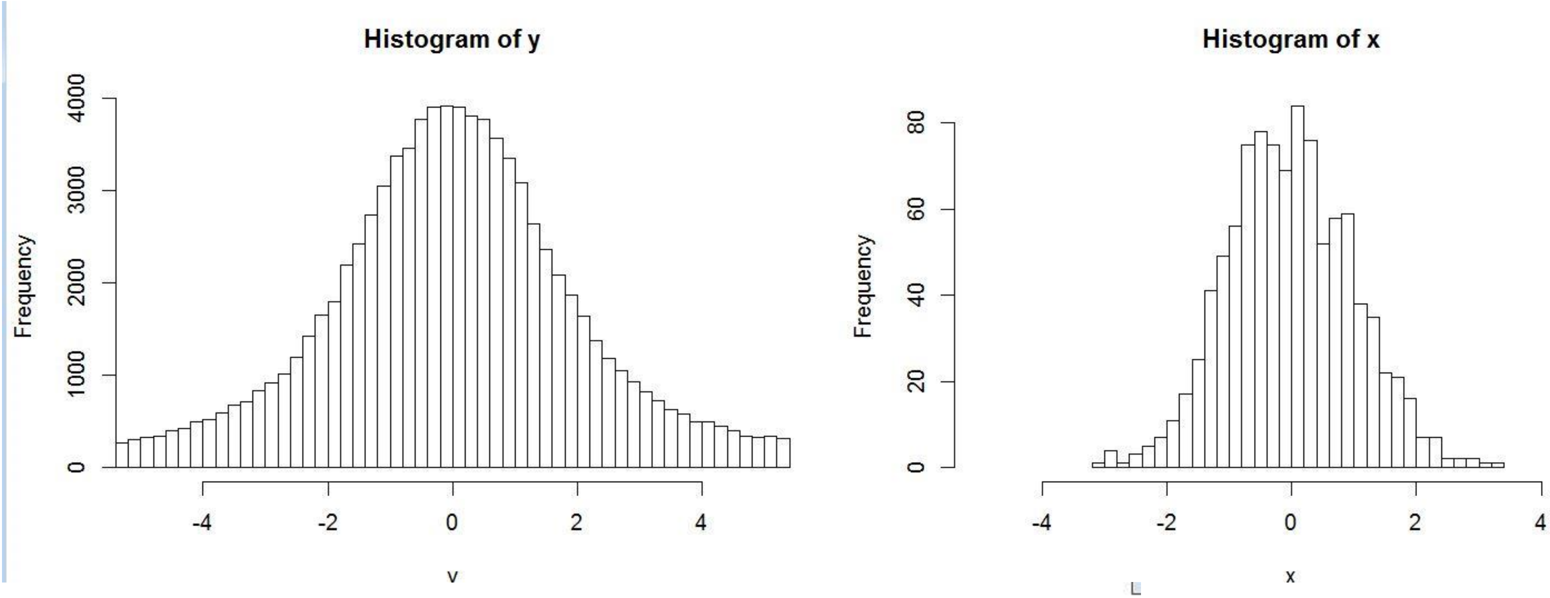





```
> m=100000
> ## importance sampling
> y =rnorm(m)/runif(m)
> w1 = dnorm(y)/s(y)
> wn1 = w1/sum(w1)
> mean(w1)
[1] 1.002287
>
>
> neff = 1/sum(wn1^2)
> show(neff/m)
[1] 0.6213973
```

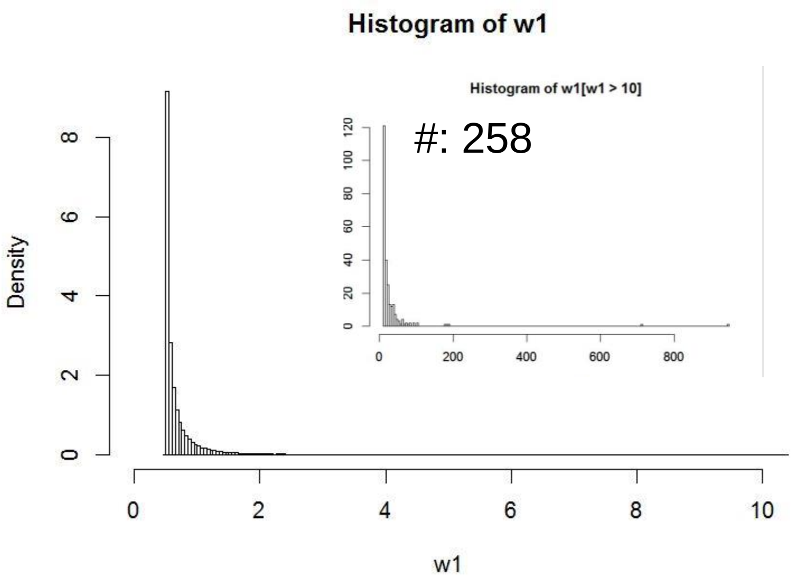
```
>
> t=h(y)*w1
> mean(t)
[1] -0.0007471106
```

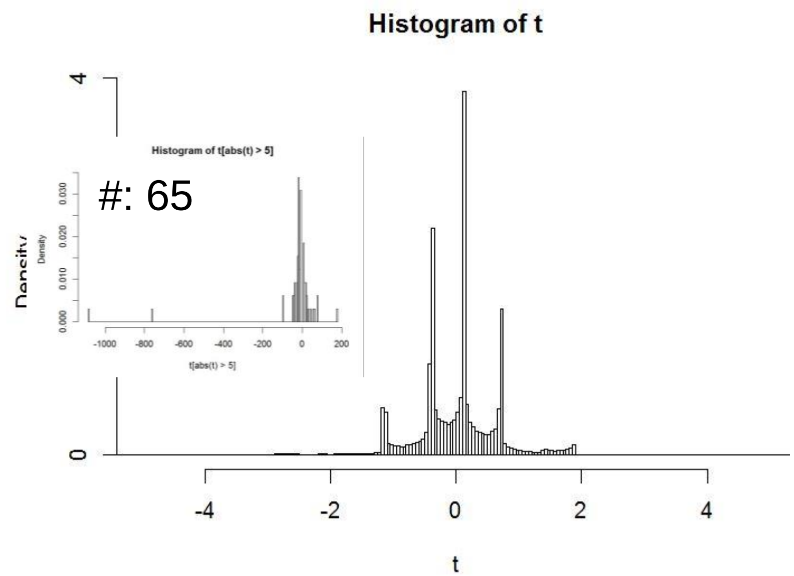
SIR normal from slash




```
n = 1000
x = sample(y,n,replace=T,prob=wn1)
par(mfrow=c(1,2))
hist(y,1000000,xlim=c(-5 ,5))
hist(x,40,xlim=c(-5 ,5))
```

Sample from normal, estimate in slash

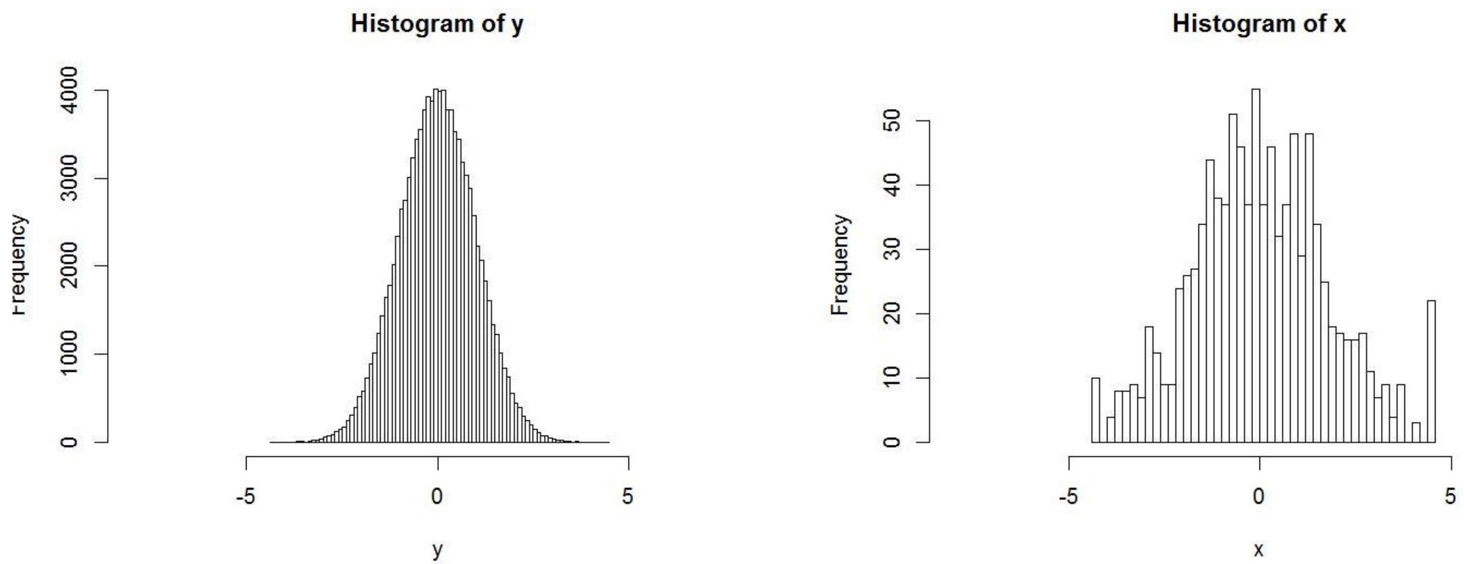





```
>
> m=100000
> ## importance sampling
> y =rnorm(m)
> w1 = s(y)/dnorm(y)
> wn1 = w1/sum(w1)
> mean(w1)
[1] 0.8170563
>
>
> neff = 1/sum(wn1^2)
> show(neff/m)
[1] 0.03708686
```

Some weigths are very large!
Gives low «effective number samples»

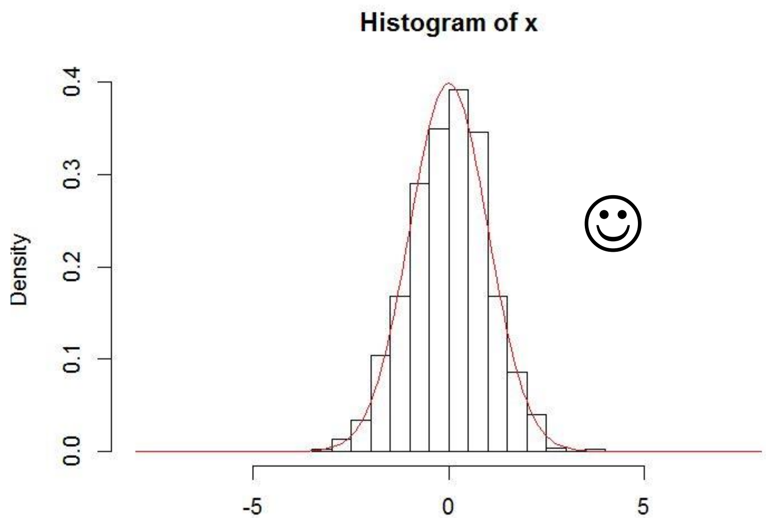
SIR slash from normal



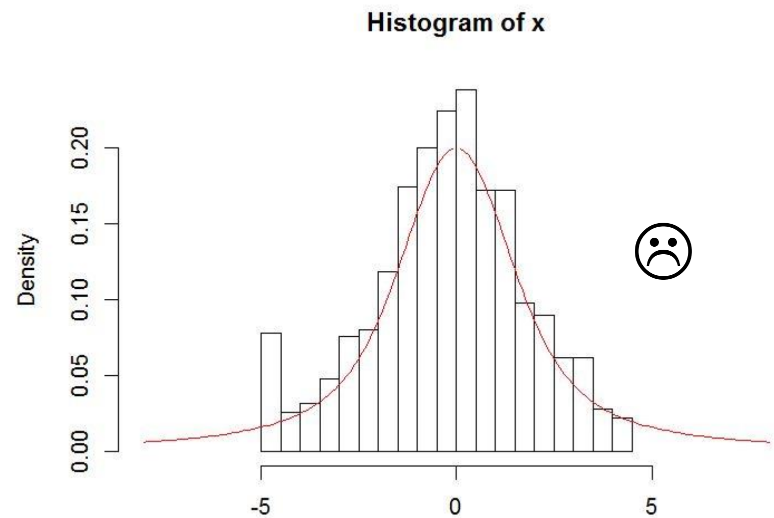

```
> ## SIR resample
> y =rnorm(m)
> w1 = s(y)/dnorm(y)
> wn1 = w1/sum(w1)
> n = 1000
> x = sample(y,n,replace=T,prob=wn1)
> par(mfrow=c(1,2))
> hist(y,1000,xlim=c(-8 ,8))
> hist(x,40,xlim=c(-8 ,8))
```

☹

SIR:
normal from slash and slash from normal




```
> y =rnorm(m)/runif(m)
> w = dnorm(y)/s(y)
> x = sample(y,n,replace=T,prob=w)
> x = sort(x)
> hist(x,20,freq=F,xlim=c(-8,8))
> xp=seq(-8,8,by=0.1)
> lines(xp,dnorm(xp),col=2)
```




```
> y = rnorm(m)
> w = s(y)/dnorm(y)
> x = sample(y,n,replace=T,prob=w)
> x = sort(x)
> hist(x,20,freq=F,xlim=c(-8,8))
> lines(xp,s(xp),col=2)
```