Robust Measurement of Heavy-Tailed Risks: Theory and Application

joint with Nikolaus Schweizer

Judith C. Schneider

University of Muenster-Finance Center Muenster

Loccum, September 2013

Starting Point

- Incomplete Market: Hedging strategy has to be derived w.r.t. an objective function:
  utility-based, variance-based, VaR-based ...
- Variance minimizing hedging ▶ not a unique hedging strategy: static vs dynamic, myopic vs. directional views, precommitment vs. time consistent
- Criterium: Robustness w.r.t. model risk

Concept

Robustness Approach: Model risk is measured in a non-parametric way, i.e., in terms of the divergence from a nominal model

Related literature

- Operations Research/Optimal Control (Ben-Tal et al. 2009), Macroeconomics (Hansen and Sargent 2009)
- Finance (largely independent): Friedmann (2002), Calafiore (2007), Breuer and Csiszar (2013) , Glasserman and Xu (2013), Hu and Hong (2013), Ben-Tal et al. (2013)

Motivation

Research Questions:

1. Given a nominal model, a distance to the ”true“ data generating process and a target quantity to be optimized: What is the worst case value of that quantity?
2.

Motivation

Research Questions:

1. Given a nominal model, a distance to the ”true“ data generating process and a target quantity to be optimized: What is the worst case value of that quantity?
2. How does the answer to question (1) influence optimal decisions?

Motivation

Research Questions:

3 How does the choice of the distance measure influence the worst case analysis?

4

5

Motivation

Research Questions:

3 How does the choice of the distance measure influence the worst case analysis?

4 How can we calculate worst-case quantities in a numerically reliable way?

5

Motivation

Research Questions:

3. How does the choice of the distance measure influence the worst case analysis?
4. How can we calculate worst-case quantities in a numerically reliable way?
5. How can we meaningful determine the distance between our model and the ”true“ data generating process?

Motivation

Agenda

1. Problem definition
2. Choice of the distance measure
3. Choice of the risk measure
4. Numerical approach
5. Estimating the distance
6. Numerical application

Finance Center Münster
Juniorprofessur Finance

SCR non-life premium and reserve risk

- Standard approach: combined non-life premium and reserve risk is assumed to be log-normally distributed

$$NL_{P,R} = E[L(RPR)] \cdot V$$

- log-normal assumption for RPR, i.e., heavy-tailed distribution
- risk measure
- parameter and quantity uncertainties
- how sensitive is the SCR towards uncertainties from modeling and which changes affect the SCR most severely?

Formulation of the problem

Risk under the nominal model

$$\Lambda_{\nu}(Y) = E_{\nu}[L(Y)] < \infty,$$

Worst case risk

$$\Lambda_{\nu}^{\kappa}(Y) = \sup_{\eta:D_{\nu}(\eta)\leq\kappa} E_{\eta}[L(Y)] = \sup_{\eta:D_{\nu}(\eta)\leq\kappa} E_{\nu}\left[\frac{d\eta}{d\nu}(Y)\,L(Y)\right]$$

where the measures η are absolutely continuous w.r.t. ν.

How do we choose D?

KL-divergence and sufficient condition

$$D_\nu^{KL}(\eta) = E_\nu\left[\frac{d\eta}{d\nu}\log\left(\frac{d\eta}{d\nu}\right)\right] = E_\eta\left[\log\left(\frac{d\eta}{d\nu}\right)\right]$$

$$E_\nu[\exp(\varepsilon L(Y))] < \infty.$$

- postulates the existence of the moment generating function

α-divergence and sufficient condition

$$D_\nu^{\alpha}(\eta) = \frac{\left(E_\nu\left[\left(\frac{d\eta}{d\nu}\right)^{\alpha}\right] - 1\right)}{\alpha(\alpha-1)}$$

$$E_\nu[L(Y)^{\frac{\alpha}{\alpha-1}}] < \infty.$$

- postulates the existence of a certain moment of $L(Y)$.

How do these rather opaque conditions influence the worst case?

α-divergence implies that the tails of the ”true“ risk are not much heavier than those of the risk under the nominal model

Consider a log-normal distribution for the nominal model ν with volatility σ

- For KL divergence the ball around ν contains basically all log-normal and power decay distributions

Implications of divergence measures

- For $\alpha-$ divergence the ball only contains lognormal distribution with a not much higher σ
- Similar results apply if the nominal model follows a power law
- Consider an arbitrarily small KL-divergence ball *B* around the nominal model ν and some measure η which has finite KL-divergence w.r.t. ν. Then *B* contains models which have the same tail behavior as η.

Suppose that the nominal model is a lognormal distribution with volatility $\sigma = 0.2$:

- In the sense of 2-divergence, all lognormal distributions with volatility $\sigma > 0.283$ are infinitely different from the nominal model
- In the sense of 4-divergence, all lognormal distributions with volatility $\sigma > 0.231$ are infinitely different from the nominal model
- In the sense of KL-divergence, all these models are comparable

Risk measures of interest

- We consider risk measures of the form $E_\nu[L(Y)]$
- In particular, we are interested in partial moments of the form

$$PM_n(Y) = \int_q^\infty (y - q)^n \, d\nu(y)$$

- This includes: shortfall probability, expected shortfall, expected shortfall given default
- Important: In the heavy-tailed case, KL-divergence only works well with bounded risk measures or bounded nominal distributions

Probability

Given some event with nominal probability β, e.g. a shortfall probability, we can infer the event's worst case probability without further distributional assumptions:

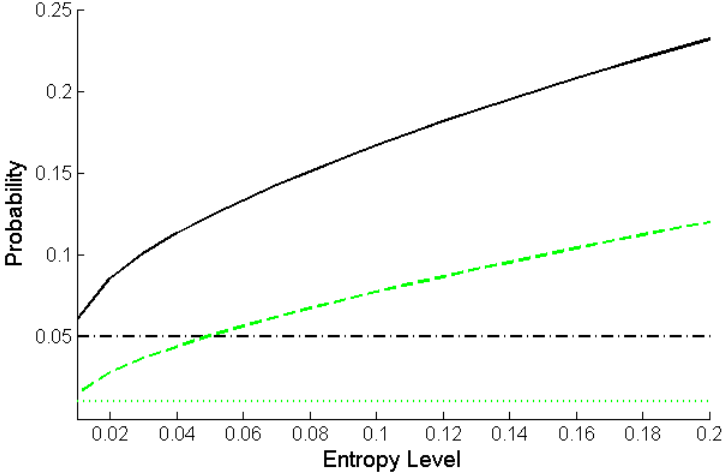
$$\beta^{wc} = \frac{\beta \exp\{\theta\}}{1 - \beta + \beta \exp\{\theta\}}$$

where θ can be inferred from

$$\kappa = \theta / Z(\theta) - log(Z(\theta))$$
$$Z(\theta) = 1 - \beta + \beta * exp(\theta)$$

and κ denotes the relative entropy

Natural bounded risk measures -Example



Cutting off

Definition of the *bounded* risk measure

$$L_i^{\mathrm{b}}(Y) = L_i(Y)1_{\{Y \leq \tau\}} + c1_{\{Y > \tau\}}.$$

Risk under the nominal model

$$E_\nu[L_i{}^{\mathrm{b}}(Y)] = \int_{-\infty}^{\tau} L_i(y)d\nu(y) + \beta c$$

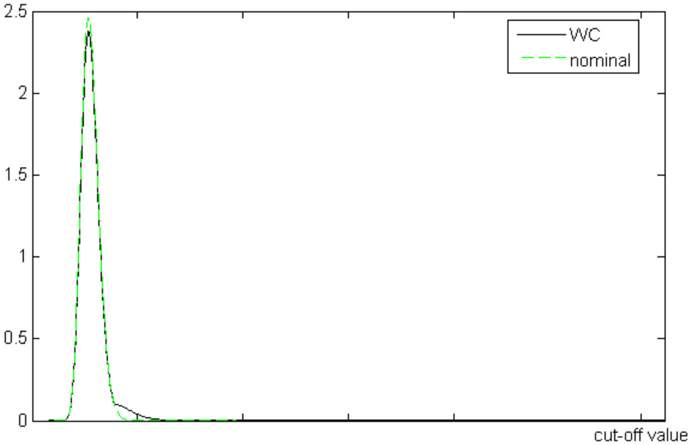
Worst case risk

$$E_\eta[L_i^{\mathrm{b}}(Y)] = \frac{1}{Z(\theta)}\left(\int_{-\infty}^{\tau} \exp\{\theta L_i(y)\}L_i(y)d\nu(y) + \beta c \exp\{\theta c\}\right)$$

Cutting off unbounded risk measures-Example

Parameter set

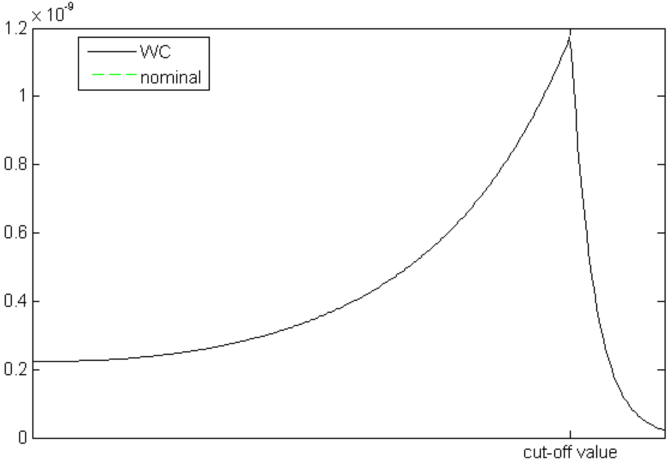
$Y \sim LN(0.1, 0.15), T = 1, c = 10, \kappa = 0.1$



Cutting off unbounded risk measures-Example

Parameter set

$Y \sim LN(0.1, 0.15), T = 1, c = 10, \kappa = 0.1$



Cutting off unbounded risk measures-Example

Parameter set

$Y \sim LN(0.1, 0.15), T = 1, c = 10, \kappa = 0.1$

cut–off level c	magnitude nominal	magnitude wc	$\frac{\int_{T-\epsilon}^{\infty} \eta(y)dy}{E_\eta[L^{\text{bnd}}(Y)]}$
5	10^{-35}	10^{-13}	$3.1 \cdot 10^{-10}$
10	10^{-49}	10^{-9}	$8.9 \cdot 10^{-7}$
15	10^{-65}	10^{-5}	0.03
25	10^{-94}	10^{-4}	0.53
30	10^{-105}	10^{-4}	0.64

Two steps

1. Determine θ (recall we estimate κ) such that the worst case distribution is defined through

$$\frac{d\eta_\theta}{d\nu}(y) \propto \exp(\theta L(y)).$$

2. Estimate $E_{\eta_\theta}[L(Y)]$ and related quantities (e.g. distribution under the worst case)

Important Sampling

So far the literature proposes to reweight samples from the nominal model using IS

- Unless θ is small, this leads to a degeneration of importance weights – very few samples carry most of the mass
- What happens in the tails of the nominal model is easily missed
- ISRS likely particles are duplicated and unlikely deleted, however does not prevent concentration of probability mass
- In many examples, the IS estimator appears to converge to a finite value, although the actual target quantities are much larger or even infinite

Markov Chain Monte Carlo

- IS: location of samples depends too much on what is likely under the nominal model
- Idea: Use MCMC to simulate an ergodic Markov chain which has the target distribution as its stationary distribution
- Works if target distribution is only known up to a constant factor
- MCMC tends to break down in multimodal applications

Markov Chain Monte Carlo

- IS: location of samples depends too much on what is likely under the nominal model
- Idea: Use MCMC to simulate an ergodic Markov chain which has the target distribution as its stationary distribution
- Works if target distribution is only known up to a constant factor
- MCMC tends to break down in multimodal applications

Sequential Monte Carlo

- Start with a sequence of mutually absolutely-continuous distribution $\mu_0, ..., \mu_n$ with $\mu_n = \eta_\theta$
- Create m independent samples $\xi_1, ..., \xi_m$ from μ_0 using e.g. MCMC
- For $i = 1, ..., n$ transform $\xi_1, ..., \xi_m$ into a sample approximating μ_i instead of μ_{i-1} using ISRS
- Afterwards, independently apply k steps of MCMC with target distribution μ_i to each ξ_i
- Thus, we approximate

$$E_{\eta_\theta}[\phi(X)] \approx \frac{1}{m} \sum_j \phi(\xi_j).$$

Sequential Monte Carlo

- Start with a sequence of mutually absolutely-continuous distribution $\mu_0, ..., \mu_n$ with $\mu_n = \eta_\theta$
- Create m independent samples $\xi_1, ..., \xi_m$ from μ_0 using e.g. MCMC
- For $i = 1, ..., n$ transform $\xi_1, ..., \xi_m$ into a sample approximating μ_i instead of μ_{i-1} using ISRS
- Afterwards, independently apply k steps of MCMC with target distribution μ_i to each ξ_i
- Thus, we approximate

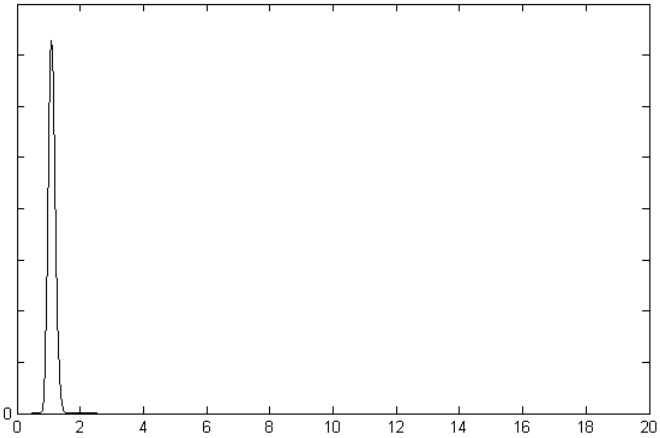
$$E_{\eta_\theta}[\phi(X)] \approx \frac{1}{m} \sum_j \phi(\xi_j).$$

Illustration

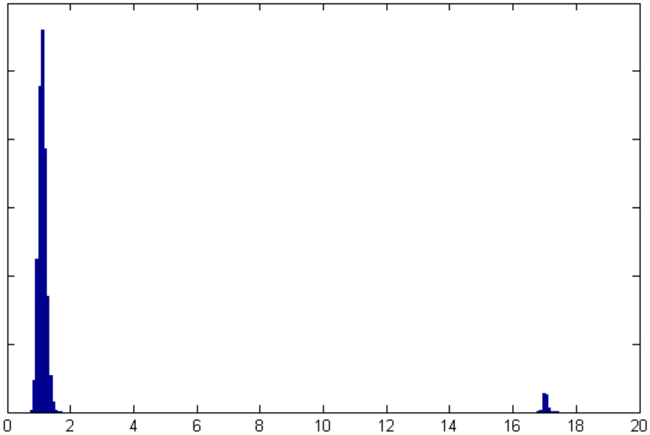
Parameter Set

$Y \sim LN(0.1, 0.1)$, $L = min(c^2, (y-m)^2)$, $c = 16$;
$m = 1.105; \kappa = 0.05; \theta = 1.4631$;
$\sqrt{(E[L])}$: nominal = 0.110;
worst case analytical = 0.2163;
IS = 0.1130;
MCMC with start in 1 = 0.1124;
MCMC with start in 15 = 15.76;
SMC = 0.2153;

Illustration



Illustration



KL-divergence estimator

Let $Y_{1:n}$ be an i.i.d. sample where Y is distributed according to ν and $X_{1:m}$ be an i.i.d. sample where X is distributed according to η. Let $y_k(i)$ denote the Euclidean distance of the kth nearest neighbor of Y_i in the sample $Y_{1:n}$ and let $x_k(i)$ be the distance of the kth nearest neighbor of Y_i in the sample $X_{1:m}$.

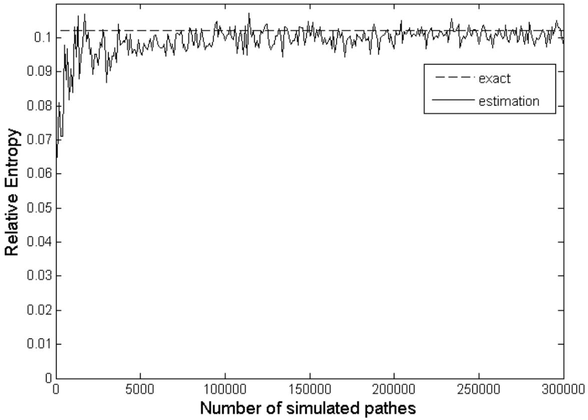
$$D_{\nu}^{\mathrm{KL}}(\eta) \approx \hat{\mathcal{D}}(Y_{1:n}||X_{1:m}) = \frac{d}{n}\sum_{i=1}^{n}\log\left(\frac{m x_k(i)}{(n-1)y_k(i)}\right) \qquad (1)$$

where d is the dimension of the state space.

KL-divergence estimator-Illustration

Parameter set

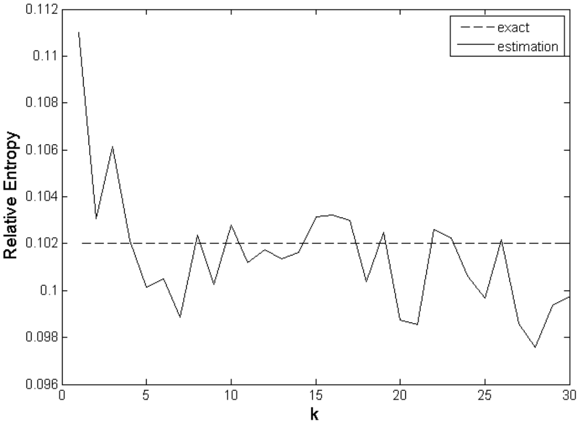
$Y \sim LN(0.1, 0.15)$, $X \sim LN(0.1, 0.2)$



KL-divergence estimator-Illustration

Parameter set

$Y \sim LN(0.1, 0.15)$, $X \sim LN(0.1, 0.2)$



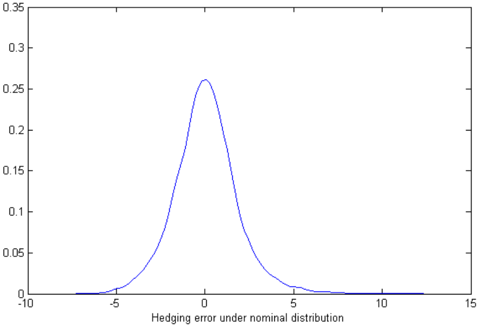
Discrete Hedging example

- We want to hedge a simple Call option, i.e., $[S_T - K]^+$
- We consider the bounded quadratic hedge error as the risk measure
- As a nominal model we stick to a simple Black-Scholes model where hedging takes place at discrete points in time
- The distance of choice is estimated w.r.t. to a sophisticated Garch model with filtered empirical innovations
- Contract parameters: $\mu = 0.097$; $\sigma = 0.189$; $T = 1$; $C_0 = 100$; $K = 100$; $r = 0.03$;

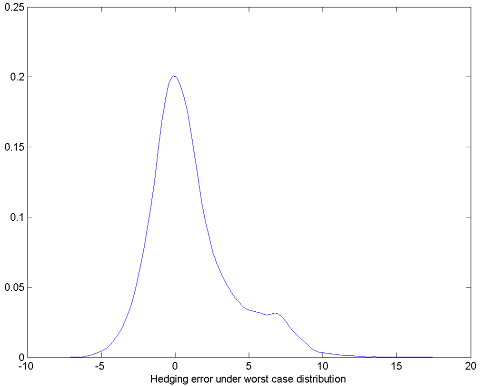
Estimation of the distance

	EGARCH e.i.	t-EGARCH	t-GARCH	t-GJR	BS($\tilde{\sigma} = 0.22$)
κ	0.20	0.23	0.27	0.54	0.20

The worst case distribution



The worst case distribution



Conclusion

- Every model presents an approximation of reality and thus modeling inevitably implies model risk
- Models risk is considered in a non-parametric way by means of a distance from a nominal model
- Take into account the special nature of financial risk and solve the following problems
  - How to choice the distance measure
  - How to estimate a plausible distance
  - A tractable numerical approach

Power Law

Suppose that the nominal model is a t-distribution with 5 degrees of freedom (equivalently: power law of order 6)
- In the sense of 2-divergence, all t-distributions with less than 2.5 degrees of freedom (equivalently: power law of order 3.5) are infinitely different from the nominal model. - In the sense of 4-divergence, all t-distributions with less than 3.75 degrees of freedom (equivalently: power law of order 4.75) are infinitely different from the nominal model. - In the sense of KL-divergence, all these models are comparable.

The Problem

- We want to calculate the expectation $\mu(f)$ of a function f with respect to a probability distribution μ on a state space E,

$$\mu(f) = \int_E f(x)\mu(dx).$$

- Assume E is "large" so that direct numerical integration is not feasible, e.g. $E = \mathbb{R}^d$, $d > 5$.
- Idea: Monte Carlo. Draw independent samples $X^1, \ldots, X^N$ from μ and approximate $\mu(f)$ by

$$\hat{\mu}(f) = \frac{1}{N}\sum_i f(X^i).$$

- Drawback: Direct random generation not feasible when μ is only known up to a constant factor.

The Problem

- We want to calculate the expectation $\mu(f)$ of a function f with respect to a probability distribution μ on a state space E,
$$\mu(f) = \int_E f(x)\mu(dx).$$
- Assume E is "large" so that direct numerical integration is not feasible, e.g. $E = \mathbb{R}^d$, $d > 5$.
- Idea: Monte Carlo. Draw independent samples $X^1, \ldots, X^N$ from μ and approximate $\mu(f)$ by
$$\hat{\mu}(f) = \frac{1}{N}\sum_i f(X^i).$$
- Drawback: Direct random generation not feasible when μ is only known up to a constant factor.

- MCMC algorithms provide a method for approximate random generation which works when μ is only known up to a constant factor.
- Idea: Construct a Markov chain $X_1, X_2, \ldots$ with stationary distribution μ.
- For sufficiently large T, X_T is approximately distributed according to μ and can be used in a Monte Carlo-type estimate.

Example: Metropolis Markov chain

- Choose initial value X_0.
- For $k = 0, \ldots, T - 1$:
  - If $X_k = x$:
  - Propose randomly to move to a new state y.
  - If y is more likely under μ than x, set $X_{k+1} = y$.
  - If y is less likely under μ than x, decide randomly whether $X_{k+1} = y$ or $X_{k+1} = y$ with probability chosen such that μ becomes stationary distribution.

Transition probabilities independent of normalizing constant.

Drawbacks of Simple MCMC

Two problems with “sufficiently long” time needed until process approaches equilibrium:

- Difficult to determine how long is “sufficiently long”.
- “Sufficiently long” may be prohibitively long.

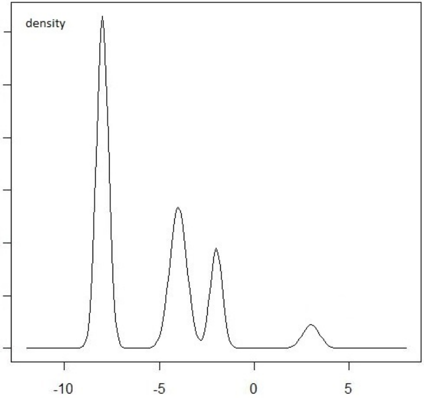
Drawbacks of Simple MCMC

MCMC works well in unimodal settings but runs into problems with multimodal distributions.

- In high dimensions, *local* dynamics are needed to guarantee reasonable acceptance probabilities.
- Thus MCMC is slow in crossing regions of small probability that separate important regions.

Example

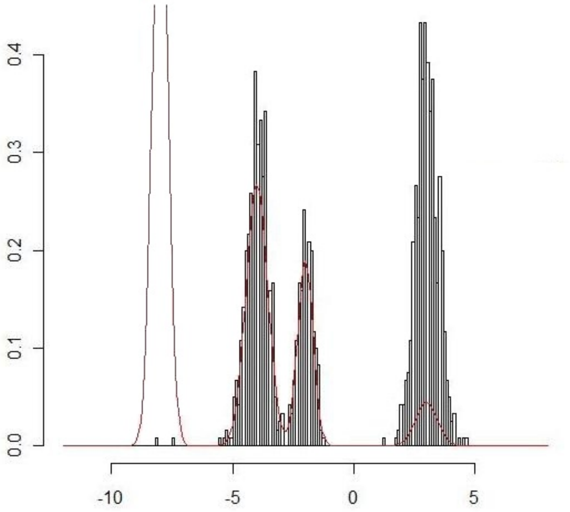
One dimensional Gaussian mixture model:



Initial distribution of Metropolis: $X_0 \sim N(0, 1)$.

Performance of Metropolis

1200 independent copies of Metropolis, 4000 steps each:



Sequential MCMC Methods

- Consider a sequence of distributions $\mu_0, \ldots, \mu_n$ on E with
  - μ_0 easy to sample (e.g. uniform),
  - $\mu_n = \mu$.
  - For all $f \in B(E)$ and all k

    $$\mu_k(f) = \mu_{k-1}(f\,\overline{g}_{k-1,k})$$

    for relative densities $\overline{g}_{k-1,k} \in B(E)$ where $B(E)$ are the bounded measurable functions from E to $\mathbb{R}$.
  - For all $x \in E$ and some γ, $\overline{g}_{k-1,k}(x) \leq \gamma$.
- Idea: Try to carry good mixing from μ_0 over to μ_n.

Sequential MCMC Methods

- Consider a sequence of distributions $\mu_0, \ldots, \mu_n$ on E with
  - μ_0 easy to sample (e.g. uniform),
  - $\mu_n = \mu$.
  - For all $f \in B(E)$ and all k

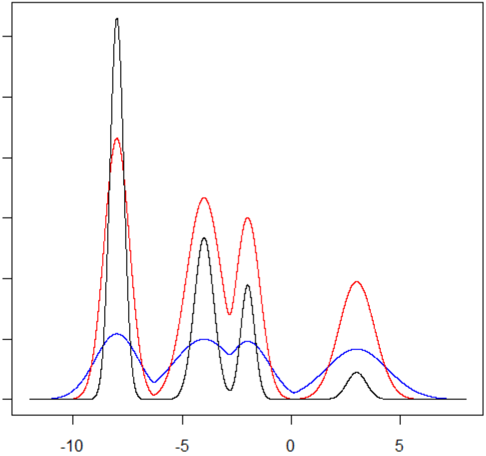
    $$\mu_k(f) = \mu_{k-1}(f\, \overline{g}_{k-1,k})$$

    for relative densities $\overline{g}_{k-1,k} \in B(E)$ where $B(E)$ are the bounded measurable functions from E to $\mathbb{R}$.
  - For all $x \in E$ and some γ, $\overline{g}_{k-1,k}(x) \leq \gamma$.
- Idea: Try to carry good mixing from μ_0 over to μ_n.

(One Choice of) Interpolating Distributions

- Write
$$\mu(dx) = \frac{1}{Z}\exp(-H(x))\pi(dx),$$
for some reference measure π.
- Introduce a sequence of inverse temperatures $0 \leq \beta_0 < ... < \beta_n = 1$.
- Tempering: Set
$$\mu_i(x) = \frac{1}{Z_i}\exp(-\beta_i H(x))\pi(dx).$$

Example



Tempering removes barriers. $\beta = 0.1, 0.3, 1.$

Importance Sampling

- How do we move from one distribution to the next?
- Importance Sampling
- Idea: Assume we have independent samples $X^1_{k-1}, \ldots, X^N_{k-1}$ from μ_{k-1}. Then

$$\mu_k(f) = \mu_{k-1}(f\,\overline{g}_{k-1,k}) \approx \frac{1}{N}\sum_{j=1}^{N} f(X^j_{k-1})\overline{g}_{k-1,k}(X^j_{k-1}).$$

Importance Sampling

If we only know $g_{k-1,k} = Z\,\overline{g}_{k-1,k}$, we still have

$$\mu_k(f) \approx \frac{\frac{1}{N}\sum_{j=1}^{N} f(X_{k-1}^j) g_{k-1,k}(X_{k-1}^j)}{\frac{1}{N}\sum_{j=1}^{N} g_{k-1,k}(X_{k-1}^j)}$$

since

$$Z = \mu_{k-1}(g_{k-1,k}) \approx \frac{1}{N}\sum_{j=1}^{N} g_{k-1,k}(X_{k-1}^j).$$

Importance Sampling

- Thus samples from μ_{k-1} can approximate $\mu_k(f)$.
- But: Sequential application of IS typically leads to degeneration of weights.
- Modification: Importance Sampling Resampling (ISRS)

Importance Sampling Resampling

- Let X^1_{k-1},..., X^N_{k-1} be independent samples from μ_{k-1}.
- We generate a sample $\hat{X}^1_k, \ldots, \hat{X}^N_k$ approximating μ_k by drawing N times independently from $(X^i_{k-1})_i$ with probabilities

$$w^i_{k-1} = \frac{g_{k-1,k}(X^i_{k-1})}{\sum_{j=1}^N g_{k-1,k}(X^j_{k-1})}.$$

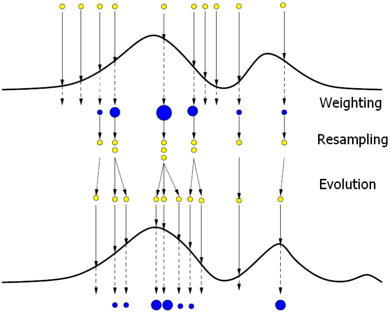
Importance Sampling Resampling

- In Importance Sampling Resampling, likely particles are duplicated and unlikely particles are deleted.
- But: Resampling alone does not prevent concentration of probability mass.
- Moreover, Resampling leads to degeneration of the sample.
- Idea: Use MCMC steps to decrease dependence between particles.

Importance Sampling Resampling

- In Importance Sampling Resampling, likely particles are duplicated and unlikely particles are deleted.
- But: Resampling alone does not prevent concentration of probability mass.
- Moreover, Resampling leads to degeneration of the sample.
- Idea: Use MCMC steps to decrease dependence between particles.

Sequential MCMC



- A typical choice of parameters could be $n = 10$ and $N = 1000$.
- If we impose $\overline{g}_{k,k+1} < 2$, then μ_0 and μ_{10} may differ locally by 2^{10}.

Sequential MCMC

Let $K_k(f)$ be an integral operator which is reversible wrt μ_k. K_k could e.g. come from T steps of Metropolis wrt μ_k:

$$“K_k(f)(x) = E[f(X_T^k)|X_0^k = x]”$$

Algorithm (e.g. Del Moral, Doucet, Jasra (2008)):

- Draw X_0^1,..., X_0^N independently from μ_0.
- For $k = 1...n$:
  - ISRS: Generate a sample $\hat{X}_k^1$,..., $\hat{X}_k^N$ approximating μ_k from X_{k-1}^1,..., X_{k-1}^N using ISRS.
  - MCMC: Generate X_k^1,..., X_k^N by moving particles $\hat{X}_k^1$,..., $\hat{X}_k^N$ independently with K_k

Estimate $\mu_n(f)$ with $\eta_n^N(f) = \frac{1}{N}\sum_{i=1}^N f(X_n^i)$

Sequential MCMC

Let $K_k(f)$ be an integral operator which is reversible wrt μ_k. K_k could e.g. come from T steps of Metropolis wrt μ_k:

$$\text{“}K_k(f)(x) = E[f(X_T^k)|X_0^k = x]\text{”}$$

Algorithm (e.g. Del Moral, Doucet, Jasra (2008)):

- Draw $X_0^1, \ldots, X_0^N$ independently from μ_0.
- For $k = 1 \ldots n$:
  - ISRS: Generate a sample $\hat{X}_k^1, \ldots, \hat{X}_k^N$ approximating μ_k from $X_{k-1}^1, \ldots, X_{k-1}^N$ using ISRS.
  - MCMC: Generate $X_k^1, \ldots, X_k^N$ by moving particles $\hat{X}_k^1, \ldots, \hat{X}_k^N$ independently with K_k

Estimate $\mu_n(f)$ with $\eta_n^N(f) = \frac{1}{N}\sum_{i=1}^N f(X_n^i)$