Variational Sequential Monte Carlo

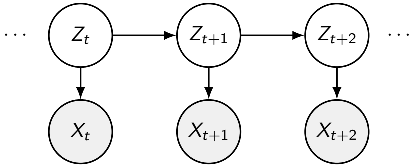
Shvechikov Pavel

Samsung AI Center
National Research University Higher School of Economics

September 22, 2018

Overview

State-Space Models



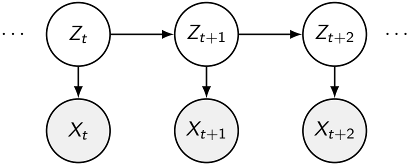
1. $\{Z_t\}_{t\geq 1}$ is a hidden Markov process

$$Z_1 \sim \mu(\cdot) \qquad Z_t \mid (Z_{t-1} = z) \sim f_\theta(\cdot|z) \tag{1}$$

2. $\{X_t\}_{t\geq 1}$ is Markov observation process

$$X_t \mid (Z_t = z) \sim g_\theta(\cdot|z) \tag{2}$$

State-Space Models: Examples



1. Hidden Markov Model: $\{Z_t\}$ is a finite Markov Chain
2. Linear Gaussian SSM:

$$
\begin{aligned}
Z_t &= A_t Z_{t-1} + B_t V_t \qquad V_t \overset{iid}{\sim} \mathcal{N}(0, I) \\
X_t &= B_t Z_t + D_t W_t \qquad W_t \overset{iid}{\sim} \mathcal{N}(0, I)
\end{aligned}
\tag{3}
$$

3. Non-linear non-Gaussian model – stochastic volatility model

$$
\begin{aligned}
Z_t &= \phi Z_{t-1} + \sigma V_t \qquad V_t \overset{iid}{\sim} \mathcal{N}(0, I) \\
X_t &= \beta \exp(Z_t/2) W_t \qquad W_t \overset{iid}{\sim} \mathcal{N}(0, I)
\end{aligned}
\tag{4}
$$

Inference in SSM

Goals:

- θ **is known**: infer $\{z_t\}_{t\geq 1}$ from $\{x_t\}_{t\geq 1}$
  - Filtering: $p(z_t|x_{1:t})$, $p(x_{1:t})$
  - Smoothing: $p(z_t|x_{1:T})$, $p(z_{1:T}|x_{1:T})$
- θ **is unknown**: identify dynamics, i.e. $\log p(x_{1:T}|\theta) \to \max_\theta$

Inference in SSM

Goals:

- θ **is known**: infer $\{z_t\}_{t\geq1}$ from $\{x_t\}_{t\geq1}$
  - Filtering: $p(z_t|x_{1:t}),\ p(x_{1:t})$
  - Smoothing: $p(z_t|x_{1:T}),\ p(z_{1:T}|x_{1:T})$
- θ **is unknown**: identify dynamics, i.e. $\log p(x_{1:T}|\theta) \to \max_\theta$

$$p(z_{1:T}|x_{1:T}) = \frac{p(x_{1:T}, z_{1:T})}{p(x_{1:T})} \tag{5}$$

$$p(x_{1:T}, z_{1:T}) = \underbrace{\mu(z_1)\prod_{t=2}^{T} f(z_t|z_{t-1})}_{p(z_{1:T})} \underbrace{\prod_{t=1}^{T} g(x_t|z_t)}_{p(x_{1:T}|z_{1:T})} \tag{6}$$

$$p(x_{1:T}) = \int p(x_{1:T}, z_{1:T})dz_{1:T} \tag{7}$$

Analytic Inference

Posterior

$$p(z_{1:t}|x_{1:t}) = \frac{p(z_{1:t}, x_{1:t})}{p(x_{1:t})} = \frac{p(z_{1:t-1}, x_{1:t-1})g(x_t|z_t)f(z_t|z_{t-1})}{p(x_{1:t})}$$
$$= p(z_{1:t-1}|x_{1:t-1})\frac{g(x_t|z_t)f(z_t|z_{t-1})}{p(x_t|x_{1:t-1})} \tag{8}$$

Denominator

$$p(x_t|x_{1:t-1}) = \int g(x_t|z_t)f(z_t|z_{t-1})p(z_{t-1}|x_{1:t-1})dz_{t-1:t} \tag{9}$$

Marginal likelihood decomposes naturally

$$p(x_{1:t}) = p(x_1)\prod_{k=2}^{t} p(x_k|x_{1:k-1}) \tag{10}$$

Non-Gaussian non-linear dynamics?

Analytic Inference

Posterior

$$p(z_{1:t}|x_{1:t}) = \frac{p(z_{1:t}, x_{1:t})}{p(x_{1:t})} = \frac{p(z_{1:t-1}, x_{1:t-1})g(x_t|z_t)f(z_t|z_{t-1})}{p(x_{1:t})}$$
$$= p(z_{1:t-1}|x_{1:t-1})\frac{g(x_t|z_t)f(z_t|z_{t-1})}{p(x_t|x_{1:t-1})} \tag{8}$$

Denominator

$$p(x_t|x_{1:t-1}) = \int g(x_t|z_t)f(z_t|z_{t-1})p(z_{t-1}|x_{1:t-1})dz_{t-1:t} \tag{9}$$

Marginal likelihood decomposes naturally

$$p(x_{1:t}) = p(x_1)\prod_{k=2}^{t} p(x_k|x_{1:k-1}) \tag{10}$$

Non-Gaussian non-linear dynamics? No way!

Monte Carlo Integration

$$p(x_{1:T}) = \int p(x_{1:T}, z_{1:T}) dz_{1:T} \tag{11}$$

Monte Carlo Integration

$$p(x_{1:T}) = \int p(x_{1:T}, z_{1:T}) dz_{1:T} \tag{11}$$

Trotter and Tukey, 1954:

The only good Monte Carlos are dead Monte Carlos

Monte Carlo Integration

$$p(x_{1:T}) = \int p(x_{1:T}, z_{1:T}) dz_{1:T} \tag{11}$$

Trotter and Tukey, 1954:

~~*The only good Monte Carlos are dead Monte Carlos*~~

Let us for the moment review the basic Monte Carlo methods.

$$\begin{aligned} \gamma_t(z_{1:t}) &\triangleq p(z_{1:t}, x_{1:t}) \\ C_t &\triangleq p(x_{1:t}) \\ \pi_t(z_{1:t}) &\triangleq \frac{\gamma_t(z_{1:t})}{C_t} \end{aligned} \tag{12}$$

Basic Monte Carlo

Assuming we can sample $z^i_{1:t} \sim \pi_t(z_{1:t})$

$$\pi_t(z_{1:t}) = \frac{\gamma(z_{1:t})}{C_t} \approx \frac{1}{N}\sum_{i=1}^{N}\delta(z_{1:t} - Z^i_{1:t}) \triangleq \widehat{\pi}_t(z_{1:t})$$

$$I_t(\varphi) = \int \varphi(z_{1:t})\pi(z_{1:t})dz_{1:t} \approx \frac{1}{N}\sum_{i=1}^{N}\varphi(Z^i_{1:t}) \triangleq I^{MC}_t(\varphi)$$

This estimate is unbiased and have a variance of

$$\mathbb{V}\mathrm{ar}\left[I^{MC}_t(\varphi)\right] = \frac{1}{N}\left(\int \varphi^2(z_{1:t})\pi(z_{1:t})dz_{1:t} - I_t(\varphi)^2\right) \tag{13}$$

Variance of MC estimate

Problems of basic Monte Carlo

1. We cannot sample from high dimensional complex $\pi(z_{1:t})$
2. We dont want to resample $z_{1:t}$ on increment of t

Variance of MC estimate

Problems of basic Monte Carlo

1. We cannot sample from high dimensional complex $\pi(z_{1:t})$
2. We dont want to resample $z_{1:t}$ on increment of t

We are going to address

1. the first problem with Importance Sampling (IS)
2. the second problem with Sequential IS

Importance Sampling

1. choose a proposal distribution q: $\pi(z_{1:t}) > 0 \;\Rightarrow\; q(z_{1:t}) > 0$
2. sample from q, i.e. $z^i_{1:t} \sim q(z_{1:t})$,
3. reweight samples with importance weights $w(z_{1:t}) = \frac{\gamma(z_{1:t})}{q(z_{1:t})}$

Then we can

- renormalize the weights $W^i_t = \frac{w(z^i_{1:t})}{\sum_j w(z^j_{1:t})}$
- and approximate π with

$$\widehat{\pi}(z_{1:t}) = \sum_{i=1}^{N} W^i_t \delta(z_{1:t} - z^i_{1:t})$$

- estimate of the normalizing constant

$$\widehat{C}_t = \frac{1}{N} \sum_{i=1}^{N} w(z^i_{1:t})$$

Properties of IS estimation

Estimate of C_t

- is unbiased

$$\mathbb{E}\left[\widehat{C}_t\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{z_{1:t}\sim q}\left[\frac{\gamma_t(z_{1:t})}{q(z_{1:t})}\right] = \frac{N}{N}C_t$$

- has relative variance of $\mathcal{O}(\frac{1}{N})$

$$\frac{\mathbb{V}\mathrm{ar}\left[\widehat{C}_t\right]}{C_t^2} = \frac{1}{N}\left(\int \frac{\pi^2(z_{1:t})}{q(z_{1:t})}dz_{1:t} - 1\right)$$

To address the second problem we introduce **Sequential IS**

Sequential Importance Sampling

1. choose a proposal of the form $q(z_{1:t}) = q(z_t | z_{1:t-1}) q(z_{1:t-1})$
2. on increment of t, sample $z_t^i \sim q(z_t | z_{1:t-1}^i)$
3. recompute IS weights according to the recurrence

$$w(z_{1:t}) \triangleq \frac{\gamma(z_{1:t})}{q(z_{1:t})} = \frac{\gamma(z_{1:t})}{q(z_t | z_{1:t-1}) \gamma(z_{1:t-1})} \cdot \frac{\gamma(z_{1:t-1})}{q(z_{1:t-1})}$$

$$\triangleq \alpha(z_{1:t}) \cdot w(z_{1:t-1}) = w_1(z_1) \prod_{k=2}^{t} \alpha(z_{1:k})$$

We have mitigated both problems.

So what could go wrong?

Enormous variance

Consider the simplest example possible

$$\pi_t(z_{1:t}) = \prod_{k=1}^{t} \mathcal{N}(z_k|0,1) = \frac{\gamma_t(z_{1:t})}{C_t} = \frac{\prod_{k=1}^{t} \exp\left(-\frac{z_k^2}{2}\right)}{(2\pi)^{t/2}}$$

$$q_t(z_{1:t}) = \prod_{k=1}^{t} \mathcal{N}(z_k|0,\sigma^2)$$

Then

$$\frac{\mathbb{V}\mathrm{ar}\left[\widehat{C}_t\right]}{C_t^2} = \frac{1}{N}\left(\int \frac{\pi^2(z_{1:t})}{q(z_{1:t})} dz_{1:t} - 1\right) = \left[\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{t/2} - 1\right]$$

For almost perfect q with $\sigma = 1.2$ to obtain relative variance of 0.01 for $t = 1000$ we would need $N \approx 2 \times 10^{23}$ particles.

Enormous variance

Consider the simplest example possible

$$\pi_t(z_{1:t}) = \prod_{k=1}^{t} \mathcal{N}(z_k|0,1) = \frac{\gamma_t(z_{1:t})}{C_t} = \frac{\prod_{k=1}^{t} \exp\left(-\frac{z_k^2}{2}\right)}{(2\pi)^{t/2}}$$

$$q_t(z_{1:t}) = \prod_{k=1}^{t} \mathcal{N}(z_k|0,\sigma^2)$$

Then

$$\frac{\mathbb{V}\mathrm{ar}\left[\widehat{C}_t\right]}{C_t^2} = \frac{1}{N}\left(\int \frac{\pi^2(z_{1:t})}{q(z_{1:t})} dz_{1:t} - 1\right) = \left[\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{t/2} - 1\right]$$

For almost perfect q with $\sigma = 1.2$ to obtain relative variance of 0.01 for $t = 1000$ we would need $N \approx 2 \times 10^{23}$ particles.
SMC in the same setting will require only $N \approx 10^4$

Sequential Monte Carlo

Definition of SMC

SMC = Sequential IS + Resampling

SMC is a family of methods for sampling from a sequence of distributions $\{\pi_t(z_{1:t})\}$ of increasing dimension t.
Note: π_t may not be nested, i.e. $\pi_t(z_{1:t-1}) \neq \pi_{t-1}(z_{1:t-1})$

At each time step SMC provides

1. approximation $\widehat{\pi}_t$ of π_t
2. estimates normalization constant C_t

Definition of SMC

SMC = Sequential IS + Resampling

SMC is a family of methods for sampling from a sequence of distributions $\{\pi_t(z_{1:t})\}$ of increasing dimension t.
Note: π_t may not be nested, i.e. $\pi_t(z_{1:t-1}) \neq \pi_{t-1}(z_{1:t-1})$

At each time step SMC provides

1. approximation $\widehat{\pi}_t$ of π_t
2. estimates normalization constant C_t

- simple technique, hard to analyze due to resampling
- very strong theoretical guarantees
- well explored field (over 20 years of thorough investigation)
- very good in practice

SMC procedure: Bootstrap filter (Gordon, 1993)

At $t = 1$

1. Sample N particles $z_1^i \sim q(z_1)$
2. Compute weights
$$w_1(z_1^i) = \frac{\gamma(z_1^i)}{q(z_1^i)} \qquad W_1^i = \frac{w_1(z_1^i)}{\sum_i w_1(z_1^i)}$$

SMC procedure: Bootstrap filter (Gordon, 1993)

At $t = 1$

1. Sample N particles $z_1^i \sim q(z_1)$
2. Compute weights

$$w_1(z_1^i) = \frac{\gamma(z_1^i)}{q(z_1^i)} \qquad W_1^i = \frac{w_1(z_1^i)}{\sum_i w_1(z_1^i)}$$

At $t \geq 2$

1. Sample ancestor indices $a_{t-1}^i \sim \mathrm{Cat}(W_{t-1}^1, \dots, W_{t-1}^N)$
2. Sample N particles $z_t^i \sim q(z_t | z_{1:t-1}^{a_{t-1}^i})$
3. Compute weights

$$w_t(z_{1:t}^i) = \frac{\gamma(z_{1:t-1}^i)}{q(z_{1:t}^i)} \qquad W_t^i = \frac{w_t(z_{1:t}^i)}{\sum_i w_t(z_{1:t}^i)}$$

SMC procedure: Bootstrap filter (Gordon, 1993)

At $t = 1$

1. Sample N particles $z_1^i \sim q(z_1)$
2. Compute weights
$$w_1(z_1^i) = \frac{\gamma(z_1^i)}{q(z_1^i)} \qquad W_1^i = \frac{w_1(z_1^i)}{\sum_i w_1(z_1^i)}$$

At $t \geq 2$

1. Sample ancestor indices $a_{t-1}^i \sim \mathrm{Cat}(W_{t-1}^1, \dots, W_{t-1}^N)$
2. Sample N particles $z_t^i \sim q(z_t | z_{1:t-1}^{a_{t-1}^i})$
3. Compute weights

$$w_t(z_{1:t}^i) = \frac{\gamma(z_{1:t-1}^i)}{q(z_{1:t}^i)} \qquad W_t^i = \frac{w_t(z_{1:t}^i)}{\sum_i w_t(z_{1:t}^i)}$$

Estimate normalization constant and target distribution

$$\widehat{C}_t = \frac{1}{N} \sum_{i=1}^{t} w_t(z_{1:t}^i) \qquad \widehat{\pi}_t(z_{1:t}) = \sum_{i=1}^{N} W_t^i \delta(z_{1:t} - z_{1:t}^i)$$

Resampling reduces variance of final estimates

$$N\frac{\mathbb{V}\mathrm{ar}\left[\widehat{C}_t^{SIS}\right]}{C_t^2} = \int \frac{\pi_t^2(z_{1:t})}{q(z_{1:t})} dz_{1:t} - 1$$

$$N\frac{\mathbb{V}\mathrm{ar}\left[\widehat{C}_t^{SMC}\right]}{C_t^2} \approx \int \frac{\pi_{t}^2(z_1)}{q_1(z_1) dz_1} - 1 + \sum_{k=2}^{t} \int \frac{\pi_{t}^2(z_{1:k})}{\pi_{k-1}(z_{1:k-1}) q_k(z_k | z_{1:k})} dz_{k-1:k} - 1$$

Resampling "resets" the system – splits the integral into parts.

Particle impoverishment

No free lunch: (Doucet, 2011)

> *It is impossible to accurately represent a distribution on a space of arbitrarily high dimension with a sample of fixed, finite size.*

At each step we can only reduce the particle set!

Particle impoverishment

No free lunch: (Doucet, 2011)

> *It is impossible to accurately represent a distribution on a space of arbitrarily high dimension with a sample of fixed, finite size.*

At each step we can only reduce the particle set!

Many techniques to partially mitigate impoverishment

- Controlled resampling: the variance of weights (ESS, ent.)
- Advanced resampling: Systematic / Residual resampling, etc.
- Look-aheads: Block Sampling, Auxiliary Particle Filter
- Resample-Move: MCMC / Gibbs steps to "jitter" particles

SMC for Filtering – Particle Filter

Recall

$$\pi_t(z_{1:t}) \triangleq \frac{\gamma_t(z_{1:t})}{C_t} = \frac{p(z_{1:t}, x_{1:t})}{p(x_{1:t})} = p(z_{1:t}|x_{1:t})$$

$$p(z_{1:t}|x_{1:t}) = p(z_{1:t-1}|x_{1:t-1})\frac{g(x_t|z_t)f(z_t|z_{t-1})}{p(x_t|x_{1:t-1})}$$

- We have $\widehat{p}(z_{1:t-1}|x_{1:t-1}) = \sum_i W_i^{t-1}\delta(z_{1:t-1} - z_{1:t-1}^i)$
- Can marginalize $\widehat{p}(z_{t-1}|x_{1:t-1}) = \sum_i W_i^{t-1}\delta(z_{t-1} - z_{t-1}^i)$
- Resample, i.e. sample from $\widehat{p}(z_{t-1}|x_{1:t-1})$:

$$\overline{p}(z_{t-1}|x_{1:t-1}) \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta(z_{t-1} - z_{t-1}^i)$$

The marginal likelihood estimate

Sampling z_t^i from proposal $q(z_t|z_{t-1}^i)$ we obtain

$$p(x_t|x_{1:t-1}) \approx \int \frac{g(x_t|z_t)f(z_t|z_{t-1})}{q(z_t|z_{t-1}^i)} q(z_t|z_{t-1}^i)\overline{p}(z_{t-1}|x_{1:t-1})dz_{t-1:t}$$

$$= \frac{1}{N}\sum_i^N \frac{g(x_t|z_t^i)f(z_t^i|z_{t-1}^i)}{q(z_t^i|z_{t-1}^i)}$$

The marginal likelihood estimate

Sampling z_t^i from proposal $q(z_t|z_{t-1}^i)$ we obtain

$$p(x_t|x_{1:t-1}) \approx \int \frac{g(x_t|z_t)f(z_t|z_{t-1})}{q(z_t|z_{t-1}^i)} q(z_t|z_{t-1}^i)\overline{p}(z_{t-1}|x_{1:t-1})dz_{t-1:t}$$

$$= \frac{1}{N}\sum_i^N \frac{g(x_t|z_t^i)f(z_t^i|z_{t-1}^i)}{q(z_t^i|z_{t-1}^i)}$$

We can model f, g, q with complex models:

$$w_t^i = \frac{f(z_t|z_{1:t-1}^{a_{t-1}^i})g(x_t|z_{1:t}^k)}{q(z_t^k|x_{1:t}, z_{1:t-1}^{a_{t-1}^i})}$$

And still easily estimate marginal likelihood (unbiasedly)

$$\widehat{p}(x_{1:t}) \triangleq \prod_{t=1}^T \frac{1}{N}\sum_{i=1}^N w_t^i,$$

Some of the theoretical results

Assumption: exponential stability – $\forall z_1, z_1'$

$$\int \left| p(z_t | x_{2:t}, z_1) - p(z_t | x_{2:t}, z_1') \right| \, dx_t \;\leq\; \alpha^t, \qquad 0 \leq \alpha < 1$$

- **L1 distance.** Bias increases linearly with t: $\exists B_1 < \infty$

$$\int \left| \mathbb{E}\left[\widehat{p}(z_{1:t}|x_{1:t})\right] - p(z_{1:t}|x_{1:t}) \right| \leq \frac{B_1 \cdot t}{N}$$

- **Central Limit Theorem.** Approximate Normality: $\exists B_2 < \infty$

$$\lim_{N\to\infty} \sqrt{N}(\log \widehat{p}(x_{1:t}) - \log p(x_{1:t})) \to \mathcal{N}(0, \sigma_t^2), \quad \sigma_t^2 \leq B_2 t$$

- **Relative Variance** increases linearly with t: $\exists B_3 < \infty$

$$\mathbb{E}\left[\left(\frac{\widehat{p}(x_{1:t})}{p(x_{1:t})} - 1\right)^2\right] \leq \frac{B_3 t}{N}$$

Improvements over standard SMC

Proposal improvements:

- Estimating the mode of a true posterior $p(z_t|x_{1:t})$
- Local approximations: local linearization of system dynamics (EKF), Unscented KF, etc.
- Implicit proposals (Chorin, 2012)

Improvements over standard SMC

Proposal improvements:

- Estimating the mode of a true posterior $p(z_t|x_{1:t})$
- Local approximations: local linearization of system dynamics (EKF), Unscented KF, etc.
- Implicit proposals (Chorin, 2012)

Can we improve upon the fixed proposal?

Variational Sequential Monte Carlo

High level overview

1. We can parametrise our proposal distribution q
2. And optimize KL between q and true posterior $p(z_{1:t}|x_{1:t})$
3. To sample from variational posterior
   - Run SMC and pick one of the particles
4. Applicable to any sequence of probabilistic models
5. VSMC allows for model learning, proposal adaptation and inference amortization

Unifying view on ELBO

For any unnormalized target density $\gamma(z)$ with normalizing constant C, $\pi(z) = \frac{\gamma(z)}{C}$ and a proposal density q

$$\mathrm{ELBO} = \int Q(z) \log \frac{\gamma(z)}{Q(z)} dz = \log C - \mathrm{KL}(Q||\pi)$$

Unifying view on ELBO

For any unnormalized target density $\gamma(z)$ with normalizing constant C, $\pi(z) = \frac{\gamma(z)}{C}$ and a proposal density q

$$\mathrm{ELBO} = \int Q(z) \log \frac{\gamma(z)}{Q(z)} dz = \log C - \mathrm{KL}(Q||\pi)$$

- Assume $\widehat{C}(z)$ is nonnegative and $\int Q(z)\widehat{C}(z) = C$
- Then we can plug $\gamma(z) = Q(z)\widehat{C}(z)$ into ELBO

$$\mathrm{ELBO} = \int Q(z) \log \frac{Q(z)\widehat{C}(z)}{Q(z)} dz = \int Q(z) \log \widehat{C}(z) dz$$

Unifying view on ELBO

For any unnormalized target density $\gamma(z)$ with normalizing constant C, $\pi(z) = \frac{\gamma(z)}{C}$ and a proposal density q

$$\mathrm{ELBO} = \int Q(z) \log \frac{\gamma(z)}{Q(z)} dz = \log C - \mathrm{KL}(Q||\pi)$$

- Assume $\widehat{C}(z)$ is nonnegative and $\int Q(z)\widehat{C}(z) = C$
- Then we can plug $\gamma(z) = Q(z)\widehat{C}(z)$ into ELBO

$$\mathrm{ELBO} = \int Q(z) \log \frac{Q(z)\widehat{C}(z)}{Q(z)} dz = \int Q(z) \log \widehat{C}(z) dz$$

- For example $\widehat{C}(z)$ may be one of these

$$\widehat{C}(z)^{VAE} = \frac{p(x,z)}{q(z|x)}, \qquad \widehat{C}(z^{1:K})^{IWAE} = \frac{1}{K}\sum_{k=1}^{K} \frac{p(x,z^k)}{q(z^k|x)}$$

VSMC

1. Based on sampling distribution of SMC

$$Q_{SMC}(z_{1:T}^{1:K}, a_{1:T-1}^{1:K}) =$$
$$\left(\prod_{k=1}^{K} q_\phi(z_1^k)\right)\left(\prod_{t=2}^{T}\prod_{k=1}^{K} q_\phi(z_t^k|z_{1:t-1}^{a_{t-1}^k})\mathrm{Cat}(a_{t-1}^k|W_{t-1}^{1:K})\right)$$

2. and unbiased estimator of marginal likelihood

$$\widehat{C}_{SMC}(z_{1:T}^{1:K}, a_{1:T-1}^{1:K}) = \prod_{t=1}^{T}\left[\frac{1}{N}\sum_{i=1}^{N} w_t^i\right] \quad w_t^i = \frac{f_\theta(z_t|z_{1:t-1}^{a_{t-1}^i})g_\theta(x_t|z_{1:t}^k)}{q_\phi(z_t^k|x_{1:t}, z_{1:t-1}^{a_{t-1}^i})}$$

3. we can form and optimize ELBO on $\log p(x_{1:T})$

$$\mathrm{ELBO}_{SMC}(\theta, \phi, x_{1:T}) =$$
$$\int Q_{SMC}(z_{1:T}^{1:K}, a_{1:T-1}^{1:K}) \log C_{SMC}(z_{1:T}^{1:K}, a_{1:T-1}^{1:K})\ dz_{1:T}^{1:K}\ da_{1:T-1}^{1:K}$$

Optimization

$$\mathrm{ELBO}_{SMC}(\theta, \phi, x_{1:T}) \quad \rightarrow \quad \max_{\phi, \theta}$$

Optimization

$$\mathrm{ELBO}_{SMC}(\theta, \phi, x_{1:T}) \quad \to \quad \max_{\phi,\theta}$$

- make proposal $q(z_t | z^k_{1:t-1})$ **reparametrizable**
- **ignore** gradient with respect to categorical sampling

Theoretical benefits

1. We can bound the KL in N

$$\mathrm{KL}(q_\phi(z_{1:t})||p(z_{1:T}|x_{1:T})) \leq \frac{c(\phi)}{N}$$

2. We can bound the KL in T if $N = bT$

$$\mathrm{KL}(q_\phi(z_{1:t})||p(z_{1:T}|x_{1:T})) \leq -\mathbb{E}\left[\log \frac{\widehat{p}(x_{1:T})}{p(x_{1:T})}\right] \xrightarrow{T\to\infty} \frac{\sigma^2(\phi)}{2b} < \infty$$

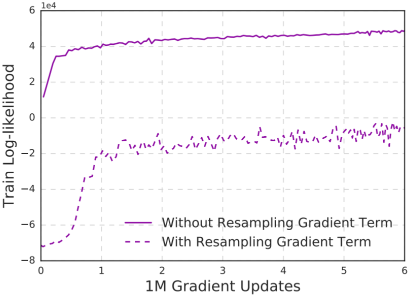
3. In general cannot achieve the marginal likelihood on optimal proposal q^*. Though, it is possible if p admits independence structure, i.e. if

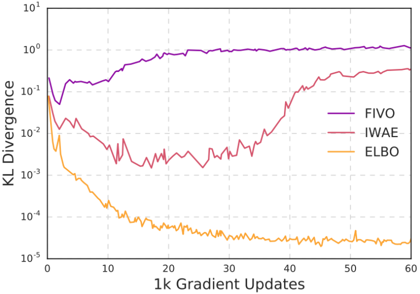
$$p(z_{1:t-1}|x_{1:t}) = p(z_{1:t-1}|x_{1:t-1})$$

Experiments

		TIMIT	
N	Bound	64 units	256 units
4	ELBO	0	10,438
	IWAE	-160	11,054
	FIVO	**5,691**	**17,822**
8	ELBO	2,771	9,819
	IWAE	3,977	11,623
	FIVO	**6,023**	**21,449**
16	ELBO	1,676	9,918
	IWAE	3,236	13,069
	FIVO	**8,630**	**21,536**

Experiments





Thank you!

References I

- Presentation of Arnaud Doucet MLSS 2012
- Doucet, Arnaud, and Adam M. Johansen. "A tutorial on particle filtering and smoothing: Fifteen years later." Handbook of nonlinear filtering 12.656-704 (2009): 3.
- Maddison, Chris J., et al. "Filtering variational objectives." Advances in Neural Information Processing Systems. 2017.
- Naesseth, Christian A., et al. "Variational Sequential Monte Carlo." arXiv preprint arXiv:1705.11140 (2017).
- Le, Tuan Anh, et al. "Auto-Encoding Sequential Monte Carlo." arXiv preprint arXiv:1705.10306 (2017).