Multilevel Monte Carlo and transport maps

Jérémie Houssineau

National University of Singapore

September 19, 2018

Joint work with: Ajay Jasra and Sumeetpal S. Singh

Outline

1. Multi-level Monte Carlo
2. Transport maps
3. Applications

Outline

Principle

Giles 2008

Ingredients:

- Identify in the model some levels of accuracy $0, \ldots, L$
- Write the target distribution π^L as a telescopic sum

$$\pi^L(\varphi) = \pi^0(\varphi) + \sum_{l=1}^{L} \left[\pi^l(\varphi) - \pi^{l-1}(\varphi)\right]$$

- Sample coupled pairs $(X_i^l, X_i^{l-})_{i=1}^{N_l}$ from (π^{l-1}, π^l)

$$\tilde{\pi}^L(\varphi) = \frac{1}{N_0} \sum_{i=1}^{N_0} \varphi(X_i^0) + \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[\varphi(X_i^l) - \varphi(X_i^{l-})\right]$$

- Could use the idea of Rhee and Glynn 2015

Principle

Giles 2008

Ingredients:

- Identify in the model some levels of accuracy $0, \ldots, L$
- Write the target distribution π^L as a telescopic sum

$$\pi^L(\varphi) = \pi^0(\varphi) + \sum_{l=1}^{L} \left[\pi^l(\varphi) - \pi^{l-1}(\varphi)\right]$$

- Sample coupled pairs $(X_i^l, X_i^{l-})_{i=1}^{N_l}$ from (π^{l-1}, π^l)

$$\tilde{\pi}^L(\varphi) = \frac{1}{N_0}\sum_{i=1}^{N_0} \varphi(X_i^0) + \sum_{l=1}^{L} \frac{1}{N_l}\sum_{i=1}^{N_l} \left[\varphi(X_i^l) - \varphi(X_i^{l-})\right]$$

- Could use the idea of Rhee and Glynn 2015

Principle

Giles 2008

Ingredients:

- Identify in the model some levels of accuracy $0, \dots, L$
- Write the target distribution π^L as a telescopic sum

$$\pi^L(\varphi) = \pi^0(\varphi) + \sum_{l=1}^{L} \left[\pi^l(\varphi) - \pi^{l-1}(\varphi)\right]$$

- Sample coupled pairs $(X_i^l, X_i^{l-})_{i=1}^{N_l}$ from (π^{l-1}, π^l)

$$\tilde{\pi}^L(\varphi) = \frac{1}{N_0}\sum_{i=1}^{N_0} \varphi(X_i^0) + \sum_{l=1}^{L} \frac{1}{N_l}\sum_{i=1}^{N_l} \left[\varphi(X_i^l) - \varphi(X_i^{l-})\right]$$

- Could use the idea of Rhee and Glynn 2015

Principle

Giles 2008

Ingredients:

- Identify in the model some levels of accuracy $0, \dots, L$
- Write the target distribution π^L as a telescopic sum

$$\pi^L(\varphi) = \pi^0(\varphi) + \sum_{l=1}^{L} \left[\pi^l(\varphi) - \pi^{l-1}(\varphi)\right]$$

- Sample coupled pairs $(X_i^l, X_i^{l-})_{i=1}^{N_l}$ from (π^{l-1}, π^l)

$$\tilde{\pi}^L(\varphi) = \frac{1}{N_0}\sum_{i=1}^{N_0} \varphi(X_i^0) + \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[\varphi(X_i^l) - \varphi(X_i^{l-})\right]$$

- Could use the idea of Rhee and Glynn 2015

Bayesian inference

Hidden Markov model:

- Initial distribution p_0
- Observations $y_{0:T}$ at integer times
- Markov kernel $Q(x, \cdot)$
- Likelihood $\ell(x, \cdot)$

Smoothing distribution:

$$\boldsymbol{p}(x_{0:T} \mid y_{0:T}) \propto p_0(x_0)\ell(x_0, y_0) \prod_{k=1}^{T} Q(x_{k-1}, x_k)\ell(x_k, y_k)$$

Bayesian inference

Hidden Markov model:

- Initial distribution p_0
- Observations $y_{0:T}$ at integer times
- Markov kernel $Q(x, \cdot)$
- Likelihood $\ell(x, \cdot)$

Smoothing distribution:

$$\boldsymbol{p}(x_{0:T} \mid y_{0:T}) \propto p_0(x_0)\ell(x_0, y_0) \prod_{k=1}^{T} Q(x_{k-1}, x_k)\ell(x_k, y_k)$$

Examples

Filtering and smoothing for SDEs

Diffusion process on $\mathbb{R}^d$

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t, \qquad t \in [0, T]$$

- Discretization with time step $h_l = 2^{-l}$

$$X_{t+h_l} = X_t + h_l a(X_t) + \sqrt{h_l} b(X_t) U_t$$

- Induced kernel:

$$Q^l(x, \cdot) = \underbrace{K^l \dots K^l}_{2^l \text{ times}}(x, \cdot),$$

  with $K^l(x, \cdot) = \phi\big(\cdot\,; x + h_l a(x), h_l b(x) b(x)^{\mathrm{t}}\big)$

- Consider the extended distribution $\boldsymbol{p}^l(x_0, x_{h_l}, \dots, x_T \,|\, y_{0:T})$ based on K^l

$\rightarrow$ *Can apply ML idea to different discretizations*

Examples

Filtering and smoothing for SDEs

Diffusion process on $\mathbb{R}^d$

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t, \qquad t \in [0, T]$$

- Discretization with time step $h_l = 2^{-l}$

$$X_{t+h_l} = X_t + h_l a(X_t) + \sqrt{h_l} b(X_t) U_t$$

- Induced kernel:

$$Q^l(x, \cdot) = \underbrace{K^l \dots K^l}_{2^l \text{ times}}(x, \cdot),$$

with $K^l(x, \cdot) = \phi\big(\cdot\,; x + h_l a(x), h_l b(x) b(x)^{\mathrm{t}}\big)$

- Consider the extended distribution $\boldsymbol{p}^l(x_0, x_{h_l}, \dots, x_T \mid y_{0:T})$ based on K^l

→ *Can apply ML idea to different discretizations*

Examples

Filtering and smoothing for SDEs

Diffusion process on $\mathbb{R}^d$

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t, \qquad t \in [0, T]$$

- Discretization with time step $h_l = 2^{-l}$

$$X_{t+h_l} = X_t + h_l a(X_t) + \sqrt{h_l} b(X_t) U_t$$

- Induced kernel:

$$Q^l(x, \cdot) = \underbrace{K^l \dots K^l}_{2^l \text{ times}}(x, \cdot),$$

with $K^l(x, \cdot) = \phi\big(\cdot\,; x + h_l a(x), h_l b(x) b(x)^{\mathrm{t}}\big)$

- Consider the extended distribution $\boldsymbol{p}^l(x_0, x_{h_l}, \dots, x_T \mid y_{0:T})$ based on K^l

$\rightarrow$ *Can apply ML idea to different discretizations*

Examples

Filtering and smoothing for SDEs

Diffusion process on $\mathbb{R}^d$

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t, \qquad t \in [0, T]$$

- Discretization with time step $h_l = 2^{-l}$

$$X_{t+h_l} = X_t + h_l a(X_t) + \sqrt{h_l} b(X_t) U_t$$

- Induced kernel:

$$Q^l(x, \cdot) = \underbrace{K^l \dots K^l}_{2^l \text{ times}}(x, \cdot),$$

  with $K^l(x, \cdot) = \phi\big(\cdot\,; x + h_l a(x), h_l b(x) b(x)^{\mathrm{t}}\big)$

- Consider the extended distribution $\boldsymbol{p}^l(x_0, x_{h_l}, \dots, x_T \,|\, y_{0:T})$ based on K^l

$\rightarrow$ *Can apply ML idea to different discretizations*

Examples

Filtering and smoothing for SDEs

Diffusion process on $\mathbb{R}^d$

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t, \qquad t \in [0, T]$$

- Discretization with time step $h_l = 2^{-l}$

$$X_{t+h_l} = X_t + h_l a(X_t) + \sqrt{h_l} b(X_t) U_t$$

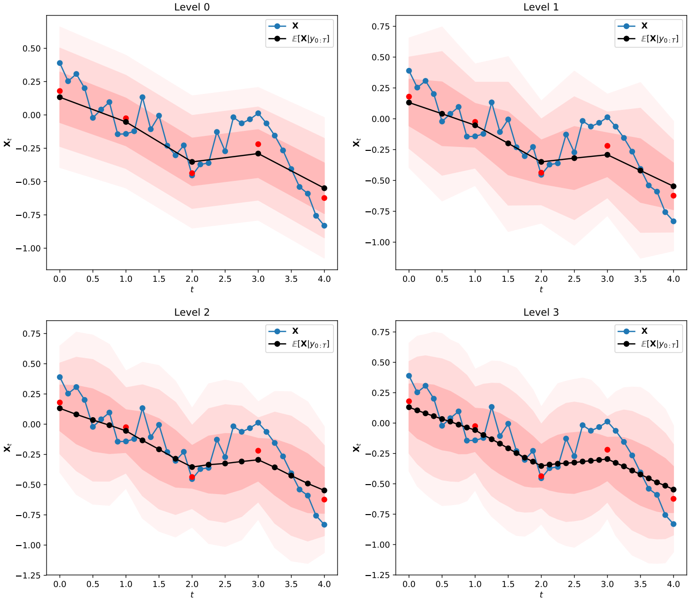
- Induced kernel:

$$Q^l(x, \cdot) = \underbrace{K^l \dots K^l}_{2^l \text{ times}}(x, \cdot),$$

  with $K^l(x, \cdot) = \phi\big(\cdot\,; x + h_l a(x), h_l b(x) b(x)^{\mathrm{t}}\big)$

- Consider the extended distribution $\boldsymbol{p}^l(x_0, x_{h_l}, \dots, x_T \mid y_{0:T})$ based on K^l

$\rightarrow$ *Can apply ML idea to different discretizations*



Mean and percentiles for a linear-Gaussian SDE at different levels

Examples

Large-lag smoothing

Only interested in the marginal posterior distribution (on $\mathbb{R}^d$)

$$p_0(x_0 \,|\, y_{0:T}) \propto p_0(x_0)\ell(x_0, y_0) \int \prod_{k=1}^{T} Q(x_{k-1}, x_k)\ell(x_k, y_k) dx_{1:T}$$

Introduce the progressive smoothing distribution:

$$p_0(x_0 \,|\, y_{0:l}) \propto p_0(x_0)\ell(x_0, y_0) \int \prod_{k=1}^{l} Q(x_{k-1}, x_k)\ell(x_k, y_k) dx_{1:T}$$

$\rightarrow$ *Can apply ML idea to different progressive smoothing distributions*

Examples

Large-lag smoothing

Only interested in the marginal posterior distribution (on $\mathbb{R}^d$)

$$p_0(x_0 \,|\, y_{0:T}) \propto p_0(x_0)\ell(x_0, y_0) \int \prod_{k=1}^{T} Q(x_{k-1}, x_k)\ell(x_k, y_k) dx_{1:T}$$

Introduce the progressive smoothing distribution:

$$p_0(x_0 \,|\, y_{0:l}) \propto p_0(x_0)\ell(x_0, y_0) \int \prod_{k=1}^{l} Q(x_{k-1}, x_k)\ell(x_k, y_k) dx_{1:T}$$

$\rightarrow$ *Can apply ML idea to different progressive smoothing distributions*

Examples

Large-lag smoothing

Only interested in the marginal posterior distribution (on $\mathbb{R}^d$)

$$p_0(x_0 \mid y_{0:T}) \propto p_0(x_0)\ell(x_0, y_0) \int \prod_{k=1}^{T} Q(x_{k-1}, x_k)\ell(x_k, y_k) dx_{1:T}$$

Introduce the progressive smoothing distribution:

$$p_0(x_0 \mid y_{0:l}) \propto p_0(x_0)\ell(x_0, y_0) \int \prod_{k=1}^{l} Q(x_{k-1}, x_k)\ell(x_k, y_k) dx_{1:T}$$

$\rightarrow$ *Can apply ML idea to different progressive smoothing distributions*

Ok, but...

Question: How to obtain correlated samples efficiently?

Giles 2013 suggests:

- Obtain samples Z_{f} and Z_{c} in a way which minimises $\mathbb{E}(|Z_{\mathrm{f}} - Z_{\mathrm{c}}|^p)$
- Corresponds to the Wasserstein metric, expressed in 1D as

$$\left(\int_0^1 \left| \Phi_{\mathrm{f}}^{-1}(u) - \Phi_{\mathrm{c}}^{-1}(u) \right|^p \mathrm{d}u \right)^{1/p}$$

- Leads to

$$Z_{\mathrm{f}} = \Phi_{\mathrm{f}}^{-1}(U) \qquad \text{and} \qquad Z_{\mathrm{c}} = \Phi_{\mathrm{c}}^{-1}(U)$$

Ok, but...

Question: How to obtain correlated samples efficiently?

Giles 2013 suggests:

- Obtain samples Z_f and Z_c in a way which minimises $\mathbb{E}(|Z_\mathrm{f} - Z_\mathrm{c}|^p)$
- Corresponds to the Wasserstein metric, expressed in 1D as

$$\left(\int_0^1 \left| \Phi_\mathrm{f}^{-1}(u) - \Phi_\mathrm{c}^{-1}(u) \right|^p \mathrm{d}u \right)^{1/p}$$

- Leads to

$$Z_\mathrm{f} = \Phi_\mathrm{f}^{-1}(U) \qquad \text{and} \qquad Z_\mathrm{c} = \Phi_\mathrm{c}^{-1}(U)$$

Outline

Transport maps and multilevel

In principle:

- Say we have two target distributions π and π'
- If we can find maps G and G' such that

$$\pi(\mathrm{d}x) = G_{\#}\eta(\mathrm{d}x) \doteq \eta(G^{-1}(\mathrm{d}x)) \qquad \text{and} \qquad \pi'(\mathrm{d}x) = G'_{\#}\eta(\mathrm{d}x)$$

  for some *base* distribution η
- Then correlated samples (x, x') can be obtained as

$$z \sim \eta, \qquad x = G(z), \qquad x' = G'(z)$$

Remarks:

- With densities:

$$G_{\#}\eta(x) = \eta\big(G^{-1}(x)\big)\big|\det \nabla G^{-1}(x)\big|$$

- On $\mathbb{R}$: increasing rearrangement

$$G = \Phi_{\pi}^{-1} \circ \Phi_{\eta}$$

Transport maps and multilevel

In principle:

- Say we have two target distributions π and π'
- If we can find maps G and G' such that

$$\pi(\mathrm{d}x) = G_{\#}\eta(\mathrm{d}x) \doteq \eta(G^{-1}(\mathrm{d}x)) \qquad \text{and} \qquad \pi'(\mathrm{d}x) = G'_{\#}\eta(\mathrm{d}x)$$

  for some *base* distribution η

- Then correlated samples (x, x') can be obtained as

$$z \sim \eta, \qquad x = G(z), \qquad x' = G'(z)$$

Remarks:

- With densities:

$$G_{\#}\eta(x) = \eta\big(G^{-1}(x)\big)\big|\det \nabla G^{-1}(x)\big|$$

- On $\mathbb{R}$: increasing rearrangement

$$G = \Phi_{\pi}^{-1} \circ \Phi_{\eta}$$

Transport maps and multilevel

In principle:

- Say we have two target distributions π and π'
- If we can find maps G and G' such that

$$\pi(\mathrm{d}x) = G_{\#}\eta(\mathrm{d}x) \doteq \eta(G^{-1}(\mathrm{d}x)) \qquad \text{and} \qquad \pi'(\mathrm{d}x) = G'_{\#}\eta(\mathrm{d}x)$$

  for some *base* distribution η

- Then correlated samples (x, x') can be obtained as

$$z \sim \eta, \qquad x = G(z), \qquad x' = G'(z)$$

Remarks:

- With densities:

$$G_{\#}\eta(x) = \eta\big(G^{-1}(x)\big)\big|\det \nabla G^{-1}(x)\big|$$

- On $\mathbb{R}$: increasing rearrangement

$$G = \Phi_{\pi}^{-1} \circ \Phi_{\eta}$$

Transport maps and multilevel

In principle:

- Say we have two target distributions π and π'
- If we can find maps G and G' such that

$$\pi(\mathrm{d}x) = G_{\#}\eta(\mathrm{d}x) \doteq \eta(G^{-1}(\mathrm{d}x)) \qquad \text{and} \qquad \pi'(\mathrm{d}x) = G'_{\#}\eta(\mathrm{d}x)$$

  for some *base* distribution η
- Then correlated samples (x, x') can be obtained as

$$z \sim \eta, \qquad x = G(z), \qquad x' = G'(z)$$

Remarks:

- With densities:

$$G_{\#}\eta(x) = \eta\big(G^{-1}(x)\big)\big|\det \nabla G^{-1}(x)\big|$$

- On $\mathbb{R}$: increasing rearrangement

$$G = \Phi_{\pi}^{-1} \circ \Phi_{\eta}$$

Representing and approximating transport maps

General approach:

- Knothe-Rosenblatt rearrangement:

$$G(z) = \begin{bmatrix} G^1(z_1) \\ G^2(z_1, z_2) \\ \vdots \\ G^n(z_1, \dots, z_n) \end{bmatrix}$$

- Parametrise each (monotone increasing) component (Ramsay 1998):

$$G^i(z_1, \dots, z_i) = a_i(z_1, \dots, z_{i-1}) + \int_0^{z_i} b_i(z_1, \dots, z_{i-1}, t)^2 \mathrm{d}t$$

- Find the map minimising the KL divergence (Moselhy and Marzouk 2012):

$$G^* = \operatorname*{argmin}_G -\mathbb{E}\left(\log \pi(G(Z)) + \sum_{i=1}^{n} \log \partial_i G^i(Z) - \log \eta(Z) \right)$$

Representing and approximating transport maps

General approach:

- Knothe-Rosenblatt rearrangement:

$$G(z) = \begin{bmatrix} G^1(z_1) \\ G^2(z_1, z_2) \\ \vdots \\ G^n(z_1, \dots, z_n) \end{bmatrix}$$

- Parametrise each (monotone increasing) component (Ramsay 1998):

$$G^i(z_1, \dots, z_i) = a_i(z_1, \dots, z_{i-1}) + \int_0^{z_i} b_i(z_1, \dots, z_{i-1}, t)^2 \mathrm{d}t$$

- Find the map minimising the KL divergence (Moselhy and Marzouk 2012):

$$G^* = \operatorname*{argmin}_G -\mathbb{E}\bigg(\log \pi(G(Z)) + \sum_{i=1}^{n} \log \partial_i G^i(Z) - \log \eta(Z) \bigg)$$

Representing and approximating transport maps

General approach:

- Knothe-Rosenblatt rearrangement:

$$G(z) = \begin{bmatrix} G^1(z_1) \\ G^2(z_1, z_2) \\ \vdots \\ G^n(z_1, \dots, z_n) \end{bmatrix}$$

- Parametrise each (monotone increasing) component (Ramsay 1998):

$$G^i(z_1, \dots, z_i) = a_i(z_1, \dots, z_{i-1}) + \int_0^{z_i} b_i(z_1, \dots, z_{i-1}, t)^2 \mathrm{d}t$$

- Find the map minimising the KL divergence (Moselhy and Marzouk 2012):

$$G^* = \operatorname*{argmin}_G -\mathbb{E}\bigg(\log \pi(G(Z)) + \sum_{i=1}^{n} \log \partial_i G^i(Z) - \log \eta(Z) \bigg)$$

Limitation

Limitation

Finding the map G^* is difficult if n is large!

Even if $d = 1$:

(SDE) $\boldsymbol{p}^l(\cdot \,|\, y_{0:T})$ is a distribution on $\mathbb{R}^{2^l T+1}$

(Smoothing) $p_0(\cdot \,|\, y_{0:l})$ is a distribution on $\mathbb{R}$ but cannot be computed directly

Idea: Use conditional independence in the HMM to break down the problem

→Spantini, Bigoni, and Marzouk 2017

Limitation

Limitation

Finding the map G^* is difficult if n is large!

Even if $d = 1$:

(SDE) $\boldsymbol{p}^l(\cdot \,|\, y_{0:T})$ is a distribution on $\mathbb{R}^{2^l T+1}$

(Smoothing) $p_0(\cdot \,|\, y_{0:l})$ is a distribution on $\mathbb{R}$ but cannot be computed directly

Idea: Use conditional independence in the HMM to break down the problem

→Spantini, Bigoni, and Marzouk 2017

Limitation

Limitation

Finding the map G^* is difficult if n is large!

Even if $d = 1$:

(SDE) $\boldsymbol{p}^l(\cdot \,|\, y_{0:T})$ is a distribution on $\mathbb{R}^{2^l T+1}$

(Smoothing) $p_0(\cdot \,|\, y_{0:l})$ is a distribution on $\mathbb{R}$ but cannot be computed directly

Idea: Use conditional independence in the HMM to break down the problem

→Spantini, Bigoni, and Marzouk 2017

Limitation

Limitation

Finding the map G^* is difficult if n is large!

Even if $d = 1$:

(SDE) $\boldsymbol{p}^l(\cdot \mid y_{0:T})$ is a distribution on $\mathbb{R}^{2^l T+1}$

(Smoothing) $p_0(\cdot \mid y_{0:l})$ is a distribution on $\mathbb{R}$ but cannot be computed directly

Idea: Use conditional independence in the HMM to break down the problem

→Spantini, Bigoni, and Marzouk 2017

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \,|\, y_{0:2}) \propto p_0(x_0) Q(x_0, x_1) \overbrace{\ell_0(x_0)}^{\doteq \ell(x_0, y_0)} \ell_1(x_1) Q(x_1, x_2) \ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \,|\, y_{0:1})$

with $G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix}$ and $(G_0^2)_{\#}\eta_1 = p(x_1 \,|\, y_{0:1})$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1) Q(G_0^2(z_1), x_2)\ell_2(x_2)$

s.t. $C = \int p(x_1 \,|\, y_{0:1}) Q(x_1, x_2)\ell_2(x_2) \mathrm{d}x_1 \mathrm{d}x_2$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \mid y_{0:2}) \propto p_0(x_0)Q(x_0, x_1)\ell_0(x_0)\ell_1(x_1)Q(x_1, x_2)\ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \mid y_{0:1})$

with $G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix}$ and $(G_0^2)_{\#}\eta_1 = p(x_1 \mid y_{0:1})$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$

s.t. $C = \int p(x_1 \mid y_{0:1})Q(x_1, x_2)\ell_2(x_2)\mathrm{d}x_1\mathrm{d}x_2$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \mid y_{0:2}) \propto p_0(x_0) Q(x_0, x_1) \ell_0(x_0) \ell_1(x_1) Q(x_1, x_2) \ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \mid y_{0:1})$

with $G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix}$ and $(G_0^2)_{\#}\eta_1 = p(x_1 \mid y_{0:1})$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1) Q(G_0^2(z_1), x_2)\ell_2(x_2)$

s.t. $C = \int p(x_1 \mid y_{0:1}) Q(x_1, x_2)\ell_2(x_2) \mathrm{d}x_1 \mathrm{d}x_2$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \stackrel{?}{\to} \boldsymbol{p}(x_{0:2} \mid y_{0:2}) \propto p_0(x_0)Q(x_0, x_1)\ell_0(x_0)\ell_1(x_1)Q(x_1, x_2)\ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \mid y_{0:1})$

$$\text{with } G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix} \qquad \text{and} \qquad (G_0^2)_{\#}\eta_1 = p(x_1 \mid y_{0:1})$$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$

$$\text{s.t. } C = \int p(x_1 \mid y_{0:1})Q(x_1, x_2)\ell_2(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \,|\, y_{0:2}) \propto p_0(x_0) Q(x_0, x_1) \ell_0(x_0) \ell_1(x_1) Q(x_1, x_2) \ell_2(x_2)$$

First step: $(G_0)_\# \eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \,|\, y_{0:1})$

$$\text{with } G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix} \qquad \text{and} \qquad (G_0^2)_\# \eta_1 = p(x_1 \,|\, y_{0:1})$$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1} \eta_1(z_1) Q(G_0^2(z_1), x_2) \ell_2(x_2)$

$$\text{s.t. } C = \int p(x_1 \,|\, y_{0:1}) Q(x_1, x_2) \ell_2(x_2) \mathrm{d}x_1 \mathrm{d}x_2$$

Third step: $(G_1)_\# \eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \mid y_{0:2}) \propto p_0(x_0)Q(x_0, x_1)\ell_0(x_0)\ell_1(x_1)Q(x_1, x_2)\ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \mid y_{0:1})$

$$\text{with } G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix} \qquad \text{and} \qquad (G_0^2)_{\#}\eta_1 = p(x_1 \mid y_{0:1})$$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$

$$\text{s.t. } C = \int p(x_1 \mid y_{0:1})Q(x_1, x_2)\ell_2(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \mid y_{0:2}) \propto p_0(x_0)Q(x_0, x_1)\ell_0(x_0)\ell_1(x_1)Q(x_1, x_2)\ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \mid y_{0:1})$

$$\text{with } G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix} \qquad \text{and} \qquad (G_0^2)_{\#}\eta_1 = p(x_1 \mid y_{0:1})$$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$

$$\text{s.t. } C = \int p(x_1 \mid y_{0:1})Q(x_1, x_2)\ell_2(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

$$\overbrace{\mathcal{N}(\mathbf{0}_3, \boldsymbol{I}_3)}^{\eta_3 \doteq} \xrightarrow{?} \boldsymbol{p}(x_{0:2} \mid y_{0:2}) \propto p_0(x_0) Q(x_0, x_1) \ell_0(x_0) \ell_1(x_1) Q(x_1, x_2) \ell_2(x_2)$$

First step: $(G_0)_{\#}\eta_2(x_{0:1}) = \boldsymbol{p}(x_{0:1} \mid y_{0:1})$

$$\text{with } G_0(z_0, z_1) = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \end{bmatrix} \qquad \text{and} \qquad (G_0^2)_{\#}\eta_1 = p(x_1 \mid y_{0:1})$$

Second step: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1) Q(G_0^2(z_1), x_2) \ell_2(x_2)$

$$\text{s.t. } C = \int p(x_1 \mid y_{0:1}) Q(x_1, x_2) \ell_2(x_2) \mathrm{d}x_1 \mathrm{d}x_2$$

Third step: $(G_1)_{\#}\eta_2(z_1, x_2) = \tilde{\pi}(z_1, x_2)$ with $G_1(z_1, z_2) = \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned} \boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\ &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\ &= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\ &= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\ &= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2}) \end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned} \boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\ &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\ &= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\ &= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\ &= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2}) \end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$
\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}
$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \mid y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \mid y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \mid y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \,|\, y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \,|\, y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \,|\, y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \,|\, y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Principle

1-lag smoothing: $(z_1, z_2) \mapsto \begin{bmatrix} G_0^2(G_1^1(z_1, z_2)) \\ G_1^2(z_2) \end{bmatrix}$ or $\begin{bmatrix} G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix}$

Full smoothing: $\boldsymbol{G} = \begin{bmatrix} G_0^1(z_0, z_1) \\ G_0^2(z_1) \\ z_2 \end{bmatrix} \circ \begin{bmatrix} z_0 \\ G_1^1(z_1, z_2) \\ G_1^2(z_2) \end{bmatrix} \doteq \bar{G}_0 \circ \bar{G}_1$

Sketch of proof: (reminder: $\tilde{\pi}(z_1, x_2) = C^{-1}\eta_1(z_1)Q(G_0^2(z_1), x_2)\ell_2(x_2)$)

$$\begin{aligned}
\boldsymbol{G}^{\#}\boldsymbol{p}(z_{0:2} \,|\, y_{0:2}) &= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}\boldsymbol{p}(z_{0:2} \,|\, y_{0:2})\big) \\
&= \bar{G}_1^{\#}\big(\bar{G}_0^{\#}(C^{-1}\boldsymbol{p}(z_{0:1} \,|\, y_{0:1})Q(z_1, z_2)\ell_2(z_2))\big) \\
&= \bar{G}_1^{\#}\big(G_0^{\#}\boldsymbol{p}(z_{0:1} \,|\, y_{0:1})C^{-1}Q(G_0^2(z_1), z_2)\ell_2(z_2)\big) \\
&= \bar{G}_1^{\#}\big(\eta_1(z_0)\tilde{\pi}(z_1, z_2)\big) \\
&= \eta_1(z_0)G_1^{\#}\tilde{\pi}(z_1, z_2) = \eta_3(z_{0:2})
\end{aligned}$$

Back to MLMC

In general:

$$\pi^L(\varphi) \approx \frac{1}{N_0} \sum_{i=1}^{N_0} \varphi(G_0(Z_i^0)) + \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \left[\varphi(G_l(Z_i^l)) - \varphi(G_{l-1}(Z_i^l))\right]$$

where $Z_i^l \sim \eta^l$, $i \in \{1, \dots, N_l\}$, $l \in \{0, \dots, L\}$

Particularities:

(SDE) Distributions at different level have different dimensions

$\rightarrow$ "thinning"

(Smoothing) Several options to obtain samples:

- Propagate the smoothing distribution
- Consider X_0 as a parameter (see Spantini, Bigoni, and Marzouk 2017)

Back to MLMC

In general:

$$\pi^L(\varphi) \approx \frac{1}{N_0}\sum_{i=1}^{N_0}\varphi(G_0(Z_i^0)) + \sum_{l=1}^{L}\frac{1}{N_l}\sum_{i=1}^{N_l}\left[\varphi(G_l(Z_i^l)) - \varphi(G_{l-1}(Z_i^l))\right]$$

where $Z_i^l \sim \eta^l$, $i \in \{1, \ldots, N_l\}$, $l \in \{0, \ldots, L\}$

Particularities:

(SDE) Distributions at different level have different dimensions

$\rightarrow$ "thinning"

(Smoothing) Several options to obtain samples:

- Propagate the smoothing distribution
- Consider X_0 as a parameter (see Spantini, Bigoni, and Marzouk 2017)

Outline

SDE

H., Jasra, and Singh 2018a

MSE as variance + bias:

$$\mathbb{E}\Big((\tilde{\pi}^L - \pi)(\varphi)^2\Big) = \sum_{l=0}^{L} \mathcal{V}_l + (\pi^L - \pi)(\varphi)^2$$

with

$$\mathcal{V}_l = \mathbb{E}\left(\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\big(\varphi(X_i^l) - \varphi(X_i^{l-})\big) - (\pi^l - \pi^{l-1})(\varphi)\right]^2\right)$$

To obtain a MSE of order ϵ^2

- If bias in $\mathcal{O}(h_L^{\alpha})$ then $L \propto -\frac{1}{\alpha}\log_2(\epsilon)$
- If $\mathcal{V}_l$ in $\mathcal{O}(h_l^{\beta})$ cost $\mathcal{C}_l$ in $\mathcal{O}(h_l^{-\zeta})$

$$N_l = N_1 2^{-(\beta+\zeta)(l-1)/2},$$

with $N_1 \propto \epsilon^{-2}\sum_{l=1}^{L} 2^{(\zeta-\beta)l/2}$

SDE

H., Jasra, and Singh 2018a

MSE as variance + bias:

$$\mathbb{E}\big((\tilde{\pi}^L - \pi)(\varphi)^2\big) = \sum_{l=0}^{L} \mathcal{V}_l + (\pi^L - \pi)(\varphi)^2$$

with

$$\mathcal{V}_l = \mathbb{E}\left(\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\big(\varphi(X_i^l) - \varphi(X_i^{l-})\big) - (\pi^l - \pi^{l-1})(\varphi)\right]^2\right)$$

To obtain a MSE of order ϵ^2

- If bias in $\mathcal{O}(h_L^\alpha)$ then $L \propto -\frac{1}{\alpha}\log_2(\epsilon)$
- If $\mathcal{V}_l$ in $\mathcal{O}(h_l^\beta)$ cost $\mathcal{C}_l$ in $\mathcal{O}(h_l^{-\zeta})$

$$N_l = N_1 2^{-(\beta+\zeta)(l-1)/2},$$

with $N_1 \propto \epsilon^{-2}\sum_{l=1}^{L} 2^{(\zeta-\beta)l/2}$

SDE

H., Jasra, and Singh 2018a

MSE as variance + bias:

$$\mathbb{E}\big((\tilde{\pi}^L - \pi)(\varphi)^2\big) = \sum_{l=0}^{L} \mathcal{V}_l + (\pi^L - \pi)(\varphi)^2$$

with

$$\mathcal{V}_l = \mathbb{E}\left(\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\big(\varphi(X_i^l) - \varphi(X_i^{l-})\big) - (\pi^l - \pi^{l-1})(\varphi)\right]^2\right)$$

To obtain a MSE of order ϵ^2

- If bias in $\mathcal{O}(h_L^{\alpha})$ then $L \propto -\frac{1}{\alpha}\log_2(\epsilon)$
- If $\mathcal{V}_l$ in $\mathcal{O}(h_l^{\beta})$ cost $\mathcal{C}_l$ in $\mathcal{O}(h_l^{-\zeta})$

$$N_l = N_1 2^{-(\beta+\zeta)(l-1)/2},$$

with $N_1 \propto \epsilon^{-2}\sum_{l=1}^{L} 2^{(\zeta-\beta)l/2}$

SDE

H., Jasra, and Singh 2018a

MSE as variance + bias:

$$\mathbb{E}\big((\tilde{\pi}^L - \pi)(\varphi)^2\big) = \sum_{l=0}^{L} \mathcal{V}_l + (\pi^L - \pi)(\varphi)^2$$

with

$$\mathcal{V}_l = \mathbb{E}\left(\left[\frac{1}{N_l}\sum_{i=1}^{N_l}\big(\varphi(X_i^l) - \varphi(X_i^{l-})\big) - (\pi^l - \pi^{l-1})(\varphi)\right]^2\right)$$

To obtain a MSE of order ϵ^2

- If bias in $\mathcal{O}(h_L^{\alpha})$ then $L \propto -\frac{1}{\alpha}\log_2(\epsilon)$
- If $\mathcal{V}_l$ in $\mathcal{O}(h_l^{\beta})$ cost $\mathcal{C}_l$ in $\mathcal{O}(h_l^{-\zeta})$

$$N_l = N_1 2^{-(\beta+\zeta)(l-1)/2},$$

with $N_1 \propto \epsilon^{-2}\sum_{l=1}^{L} 2^{(\zeta-\beta)l/2}$

SDE

Verification in the linear-Gaussian case

Variance at time $k \in \{1, \ldots, T\}$ for the filtering problem:

$$\mathcal{V}_{l,k} = \operatorname{Var}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\Big(\varphi\big(G_k^l(Z_i)\big) - \varphi\big(G_k^{l-1}(Z_i)\big)\Big)\right)$$

Theorem

In the case of a 1*-dimensional diffusion process with linear drift and constant diffusion coefficient and with a linear-Gaussian likelihood, the variance* $\mathcal{V}_{l,k}$ *obtained at level* l *for Euler's method verifies*

$$\mathcal{V}_{l,k} = \mathcal{O}(h_l^2)$$

for any $k \in \{1, \ldots, T\}$.

SDE

Models:

- Linear-Gaussian
- Langevin

SDE

Models:

- **Linear-Gaussian**:

$$\mathrm{d}X_t = aX_t\mathrm{d}t + b\mathrm{d}W_t, \qquad t \in [0, T]$$

with $a = -0.1$ and $b = 1$, observed via

$$Y_k \mid X_k \sim \mathcal{N}\big(X_k, \tau^2\big)$$

with $\tau = 0.25$

- Langevin

SDE

Models:

- Linear-Gaussian
- **Langevin**:

$$\mathrm{d}X_t = \frac{1}{2}\nabla \log \mathcal{S}_\nu(X_t)\mathrm{d}t + b\,\mathrm{d}W_t, \qquad t \in [0, T]$$

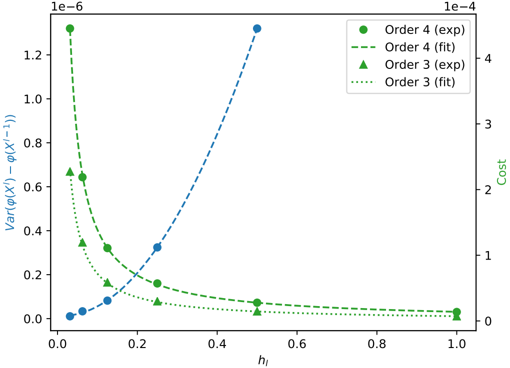
with $\mathcal{S}_\nu$ the Student's t distribution with $\nu = 10$ and with $b = 1$, observed via

$$Y_k \mid X_k \sim \mathcal{N}\big(0, \tau^2 \exp(X_k)\big)$$

with $\tau = 1$

SDE

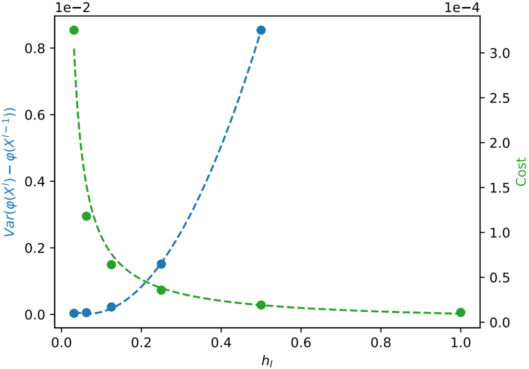
Orders



Linear-Gaussian

SDE

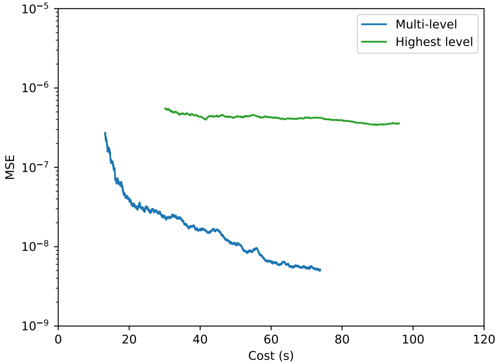
Orders



Langevin

SDE

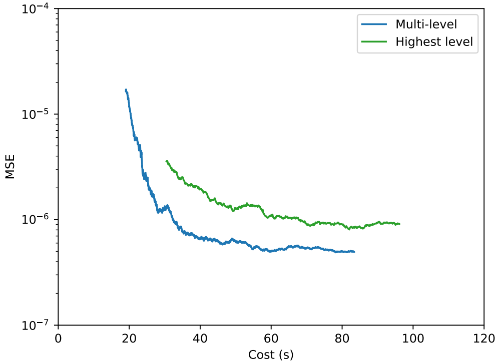
Performance



Linear-Gaussian $\varphi(x_{0:T}) = x_T$

SDE

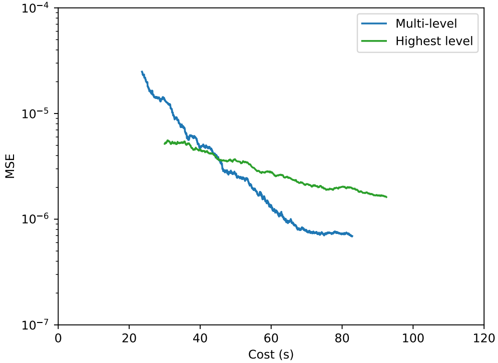
Performance



Langevin $\varphi(x_{0:T}) = x_T$

SDE

Performance



Langevin $\varphi(x_{0:T}) = \sum_{t=0}^{T} e^{-\kappa(T-t)} x_t$

Smoothing

H., Jasra, and Singh 2018b

(A1) There exists $0 < \underline{C} < \overline{C} < +\infty$ such that

$$\inf_x \ell(x, y_0) p_0(x) \wedge \inf_{k \geq 1} \inf_{x, x'} \ell(x', y_k) Q(x, x') \geq \underline{C}$$

$$\sup_x \ell(x, y_0) p_0(x) \vee \sup_{k \geq 1} \sup_{x, x'} \ell(x', y_k) Q(x, x') \leq \overline{C}.$$

Theorem (on $\mathbb{R}$, can be generalised to $\mathbb{R}^d$)

Under (A1), there exists $\rho \in (0, 1)$, $C < +\infty$ such that for any bounded measurable Lipschitz φ with Lipschitz constant K, any $N_l \geq 1$ and any $l \geq 1$ we have

$$\mathrm{Var}\Big(\frac{1}{N_l} \sum_{i=1}^{N_l} [\varphi(\Phi_l^{-1}(U^i)) - \varphi(\Phi_{l-1}^{-1}(U^i))] \Big) \leq \frac{C K^2 \rho^{l-1}}{N_l}$$

with Φ_l is the cdf of $p_0(\cdot \,|\, y_{0:l})$ and $\{U^i\}_i \overset{i.i.d.}{\sim} \mathcal{U}([0, 1])$.

Smoothing

H., Jasra, and Singh 2018b

(A1) There exists $0 < \underline{C} < \overline{C} < +\infty$ such that

$$\inf_x \ell(x, y_0)p_0(x) \wedge \inf_{k\geq 1} \inf_{x,x'} \ell(x', y_k)Q(x, x') \geq \underline{C}$$

$$\sup_x \ell(x, y_0)p_0(x) \vee \sup_{k\geq 1} \sup_{x,x'} \ell(x', y_k)Q(x, x') \leq \overline{C}.$$

Theorem (on $\mathbb{R}$, can be generalised to $\mathbb{R}^d$)

Under (A1), there exists $\rho \in (0, 1)$, $C < +\infty$ such that for any bounded measurable Lipschitz φ with Lipschitz constant K, any $N_l \geq 1$ and any $l \geq 1$ we have

$$\mathrm{Var}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}[\varphi(\Phi_l^{-1}(U^i)) - \varphi(\Phi_{l-1}^{-1}(U^i))]\right) \leq \frac{CK^2\rho^{l-1}}{N_l}$$

with Φ_l is the cdf of $p_0(\cdot \mid y_{0:l})$ and $\{U^i\}_i \overset{i.i.d.}{\sim} \mathcal{U}([0, 1])$.

Smoothing

Consequence of the theorem

To achieve an MSE of $\mathcal{O}(\epsilon^2)$

- Change the final time: $T^* = \lceil |\log(\epsilon)/\log(\rho)| \rceil$
- Number of samples: $N_l = \epsilon^{-2}(l+1)^{-1-\delta}$ for any $\delta > 0$
- Cost: $\mathcal{O}(\epsilon^{-2})$

Without ML: cost in $\mathcal{O}(T\epsilon^{-2})$

Smoothing

Consequence of the theorem

To achieve an MSE of $\mathcal{O}(\epsilon^2)$

- Change the final time: $T^* = \lceil |\log(\epsilon)/\log(\rho)| \rceil$
- Number of samples: $N_l = \epsilon^{-2}(l+1)^{-1-\delta}$ for any $\delta > 0$
- Cost: $\mathcal{O}(\epsilon^{-2})$

Without ML: cost in $\mathcal{O}(T\epsilon^{-2})$

Smoothing

Consequence of the theorem

To achieve an MSE of $\mathcal{O}(\epsilon^2)$

- Change the final time: $T^* = \lceil |\log(\epsilon)/\log(\rho)| \rceil$
- Number of samples: $N_l = \epsilon^{-2}(l+1)^{-1-\delta}$ for any $\delta > 0$
- Cost: $\mathcal{O}(\epsilon^{-2})$

Without ML: cost in $\mathcal{O}(T\epsilon^{-2})$

Smoothing

Consequence of the theorem

To achieve an MSE of $\mathcal{O}(\epsilon^2)$

- Change the final time: $T^* = \lceil |\log(\epsilon)/\log(\rho)| \rceil$
- Number of samples: $N_l = \epsilon^{-2}(l+1)^{-1-\delta}$ for any $\delta > 0$
- Cost: $\mathcal{O}(\epsilon^{-2})$

Without ML: cost in $\mathcal{O}(T\epsilon^{-2})$

Smoothing

Models

Models:

- Linear-Gaussian
- Stochastic volatility

Smoothing

Models

Models:

- **Linear-Gaussian**:
$$X_k|X_{k-1} \sim \mathcal{N}(aX_{k-1}, b^2)$$
with $X_0 \sim \mathcal{N}(x_0, \sigma^2)$, where $x_0 = 1$, $\sigma = 2$ and with $a = b = 1$, observed via
$$Y_k \mid X_k \sim \mathcal{N}\big(X_k, \tau^2\big)$$
with $\tau = 1$
- Stochastic volatility

Smoothing

Models

Models:

- Linear-Gaussian
- **Stochastic volatility**:

$$X_k = \mu + \phi(X_{k-1} - \mu) + V_k, \qquad X_0 \sim \mathcal{N}\left(\mu, \frac{1}{1-\phi^2}\right)$$

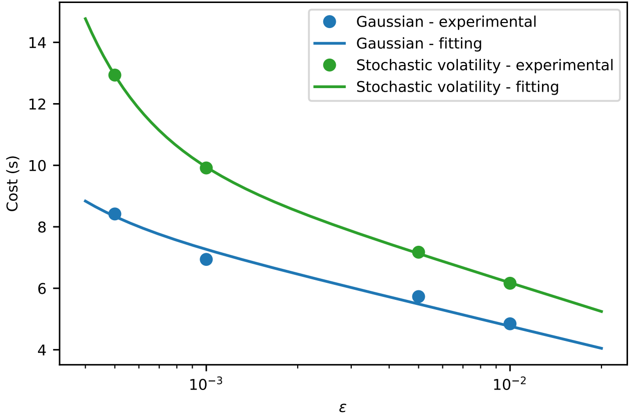
with $V_k \sim \mathcal{N}(0, \beta^2)$, $\mu = -0.5$, $\phi = 0.95$ and $\beta = 0.25$, observed via

$$Y_k = W_k \exp\left(\frac{1}{2} X_n\right)$$

with $W_n \sim \mathcal{N}(0, 1)$

Smoothing

Cost



The fitted curves are based on a function of the form $\epsilon \mapsto -\alpha\epsilon^{-2} - \beta \log(\epsilon)$

Smoothing

Result in the linear-Gaussian case

Theorem

Assuming that $\mathrm{Var}(X_k \,|\, y_{0:k}) \approx \gamma^2$ *for all* k *large enough, it holds that*

$$\mathrm{Var}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}[\Phi_l^{-1}(U^i) - \Phi_{l-1}^{-1}(U^i)]\right) = \mathcal{O}\left(\frac{1}{N_l}\left(a + \frac{b^2}{a\gamma^2}\right)^{-2l}\right).$$

- Assumption justified in terms of reachability and observability
- Beneficial if b is large and γ is small

Smoothing

Result in the linear-Gaussian case

Theorem

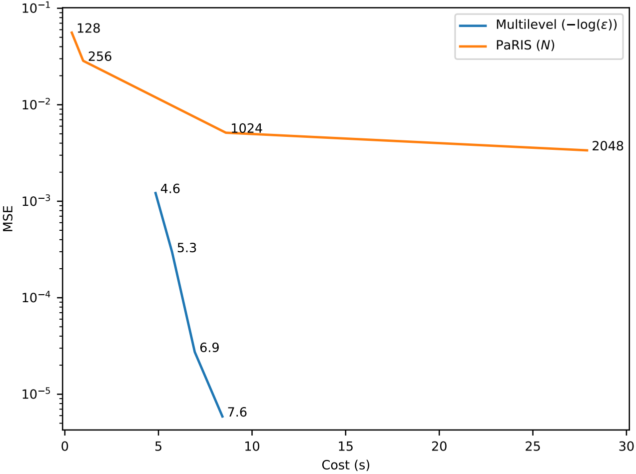
Assuming that $\mathrm{Var}(X_k \,|\, y_{0:k}) \approx \gamma^2$ *for all* k *large enough, it holds that*

$$\mathrm{Var}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}[\Phi_l^{-1}(U^i) - \Phi_{l-1}^{-1}(U^i)]\right) = \mathcal{O}\left(\frac{1}{N_l}\Big(a + \frac{b^2}{a\gamma^2}\Big)^{-2l}\right).$$

- Assumption justified in terms of reachability and observability
- Beneficial if b is large and γ is small

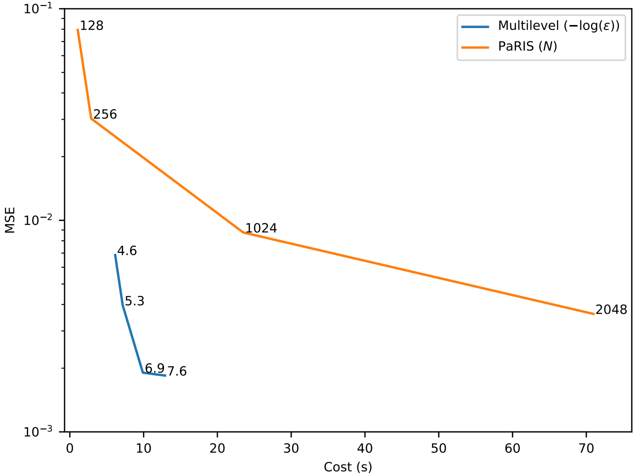
Smoothing

Performance in the linear-Gaussian case



Smoothing

Performance in the stochastic volatility model



Summary

- Transport maps are well-suited to the multilevel idea
- Full smoothing distributions can be approximated
- Quantifying the approximation error for finite order would help

Summary

- Transport maps are well-suited to the multilevel idea
- Full smoothing distributions can be approximated
- Quantifying the approximation error for finite order would help

Summary

- Transport maps are well-suited to the multilevel idea
- Full smoothing distributions can be approximated
- Quantifying the approximation error for finite order would help

References

M.B. Giles. "Multilevel monte carlo methods". In: *Monte Carlo and Quasi-Monte Carlo Methods 2012*. Springer, 2013, pp. 83–103.

M.B. Giles. "Multilevel Monte Carlo path simulation". In: *Operations Research* 56.3 (2008), pp. 607–617.

J. H., A. Jasra, and S.S. Singh. "Multilevel Monte Carlo for Smoothing via Transport Methods". In: *SIAM Journal on Scientific Computing* 40.4 (2018), A2315–A2335.

J. H., A. Jasra, and S.S. Singh. "On Large Lag Smoothing for Hidden Markov Models". In: *arXiv preprint arXiv:1804.07117* (2018).

T.A. Moselhy and Y.M. Marzouk. "Bayesian inference with optimal maps". In: *Journal of Computational Physics* 231.23 (2012), pp. 7815–7850.

J.O. Ramsay. "Estimating smooth monotone functions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2 (1998), pp. 365–375.

C. Rhee and P.W. Glynn. "Unbiased estimation with square root convergence for SDE models". In: *Operations Research* 63.5 (2015), pp. 1026–1043.

A. Spantini, D. Bigoni, and Y. Marzouk. "Inference via low-dimensional couplings". In: *arXiv preprint arXiv:1703.06131* (2017).

Thank you!

houssineau.j@gmail.com
jeremiehoussineau.com