The Viterbi Process, Parallelized Optimization and Sampling

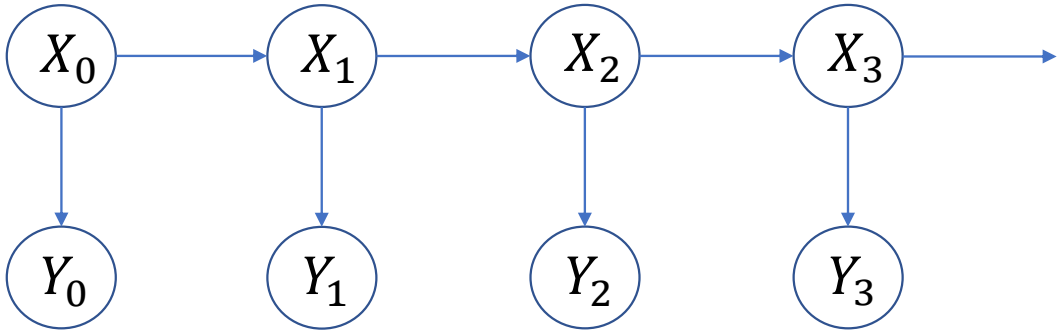
Nick Whiteley

University of Bristol

Outline

1. Hidden Markov Models and the Viterbi Process
2. Motivation: high-dimensional neural data analysis
3. Convexity, high-dimensional optimization and sampling
4. ODE approach to the Viterbi process
5. Parallelized ULA

Hidden Markov model



$$p(x_0, \ldots, x_n, y_0, \ldots, y_n) = p(x_0)p(y_0|x_0) \prod_{m=1}^{n} p(x_m|x_{m-1})p(y_m|x_m)$$

Hidden Markov model

MAP smoothing estimator

$$(\xi_0^n, \ldots, \xi_n^n) = \underset{x_0, \ldots, x_n}{\operatorname{argmax}}\, p(x_0, \ldots, x_n | y_0, \ldots, y_n)$$

Hidden Markov model

MAP smoothing estimator

$$(\xi_0^n, \ldots, \xi_n^n) = \operatorname*{argmax}_{x_0,\ldots,x_n} p(x_0, \ldots, x_n | y_0, \ldots, y_n)$$

The Viterbi process

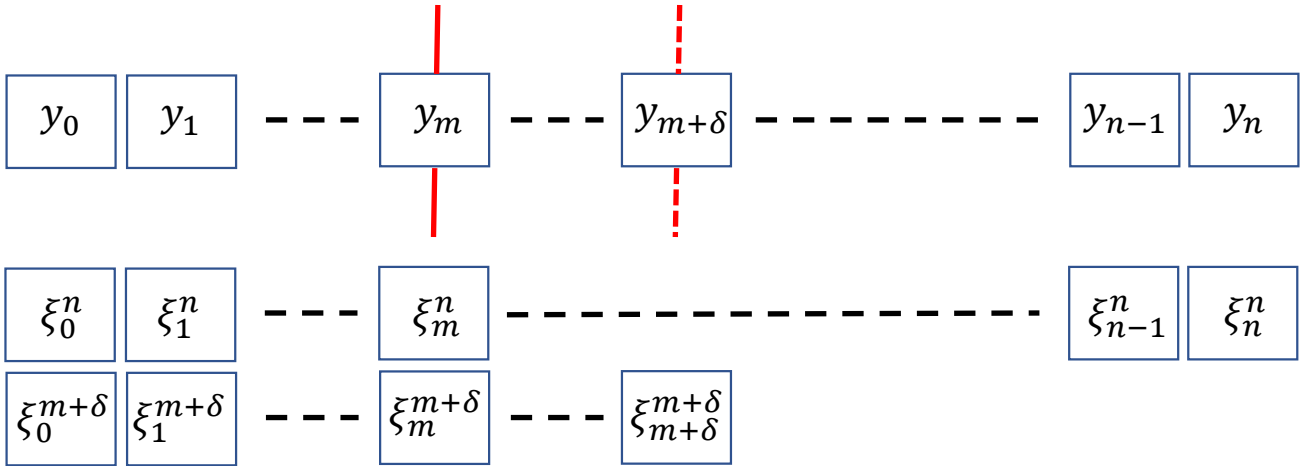
- A sequence $(\xi_0^\infty, \xi_1^\infty, \ldots)$ such that for each $m \geq 0$,

$$\lim_{n\to\infty} \|(\xi_0^n, \ldots, \xi_m^n) - (\xi_0^\infty, \ldots, \xi_m^\infty)\| = 0.$$

- Caliebe and Rosler (IEEE Inf. The. 2002), $\|\cdot\|$ is discrete norm
- Chigansky and Ritov (Bernoulli 2011), $\|\cdot\|$ is Euclidean

Viterbi Process $\rightsquigarrow$ parallelized MAP estimation?

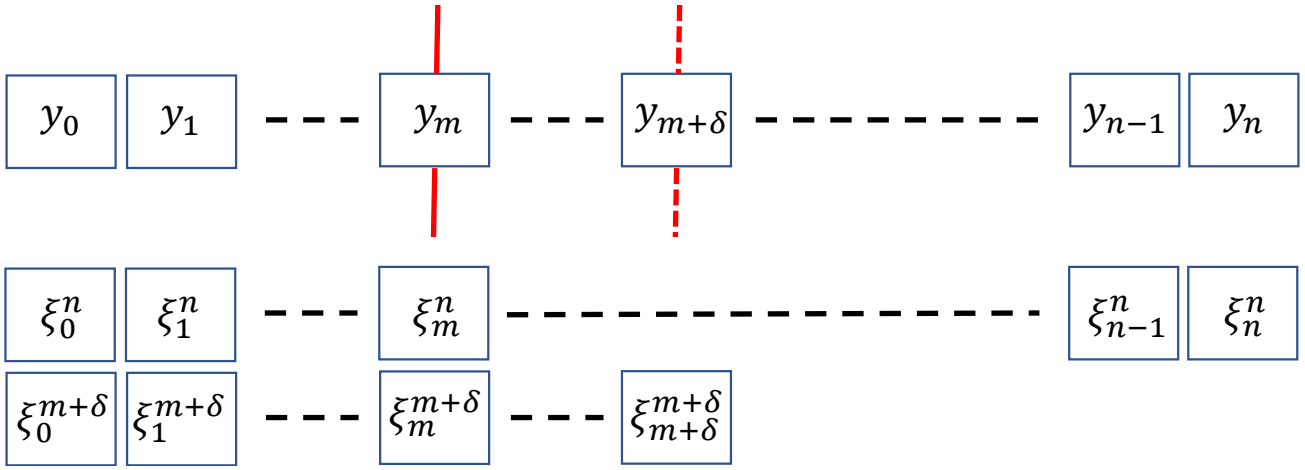
Idea: solve $\operatorname{argmax}_{x_0,\ldots,x_n} p(x_0,\ldots,x_n|y_0,\ldots,y_n)$ approximately



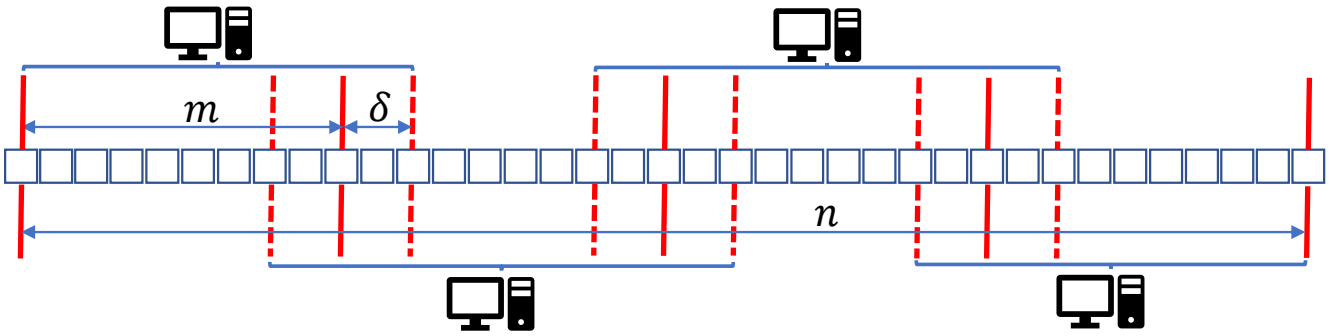
$$(\xi_0^n,\ldots,\xi_m^n) \approx (\xi_0^{m+\delta},\ldots,\xi_m^{m+\delta})$$

Viterbi Process $\rightsquigarrow$ parallelized MAP estimation?

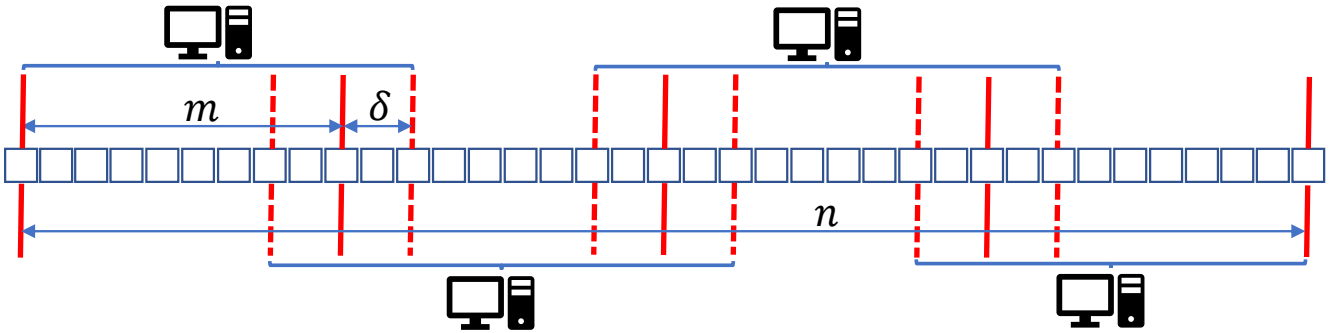
Idea: solve $\operatorname{argmax}_{x_0,\ldots,x_n} p(x_0,\ldots,x_n|y_0,\ldots,y_n)$ approximately



$$(\xi_0^n,\ldots,\xi_m^n) \approx (\xi_0^{m+\delta},\ldots,\xi_m^{m+\delta})$$



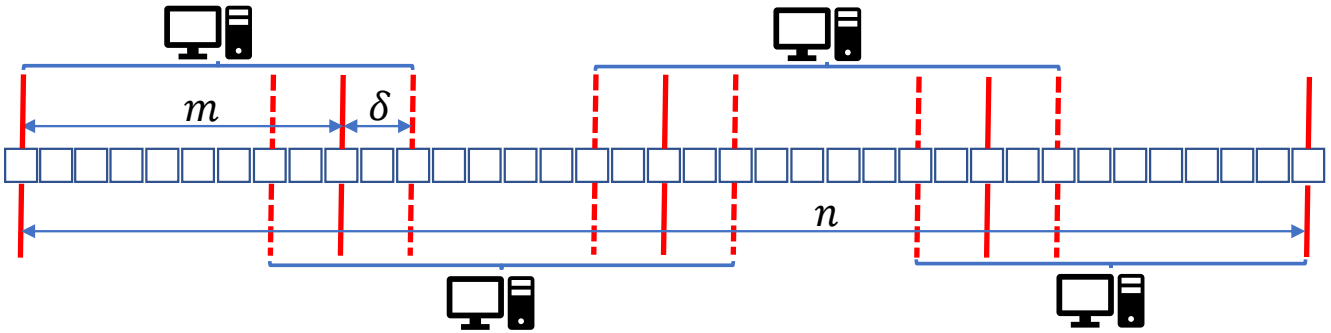
Viterbi Process $\rightsquigarrow$ parallelized MAP estimation?



Speed-up/error trade-off

- **speed-up** $\approx T(n)/T(m+2\delta)$
- **squared error** $\approx \underbrace{n/m}_{\text{number of segments}} \times \underbrace{\|(\xi_0^{m+\delta},\ldots,\xi_m^{m+\delta}) - (\xi_0^n,\ldots,\xi_m^n)\|^2}_{\text{squared error per segment}}$

Viterbi Process $\rightsquigarrow$ parallelized MAP estimation?



Speed-up/error trade-off

- **speed-up** $\approx T(n)/T(m+2\delta)$
- **squared error** $\approx \underbrace{n/m}_{\text{number of segments}} \times \underbrace{\|(\xi_0^{m+\delta},\ldots,\xi_m^{m+\delta}) - (\xi_0^n,\ldots,\xi_m^n)\|^2}_{\text{squared error per segment}}$

- rate of convergence to Viterbi process? dimension-dependence?
- MAP estimation $\rightsquigarrow$ sampling

Outline

1. Hidden Markov Models and the Viterbi Process
2. Motivation: high-dimensional neural data analysis
3. Convexity, high-dimensional optimization and sampling
4. ODE approach to the Viterbi process
5. Parallelized ULA

"Moore's law" of Neural Data Science

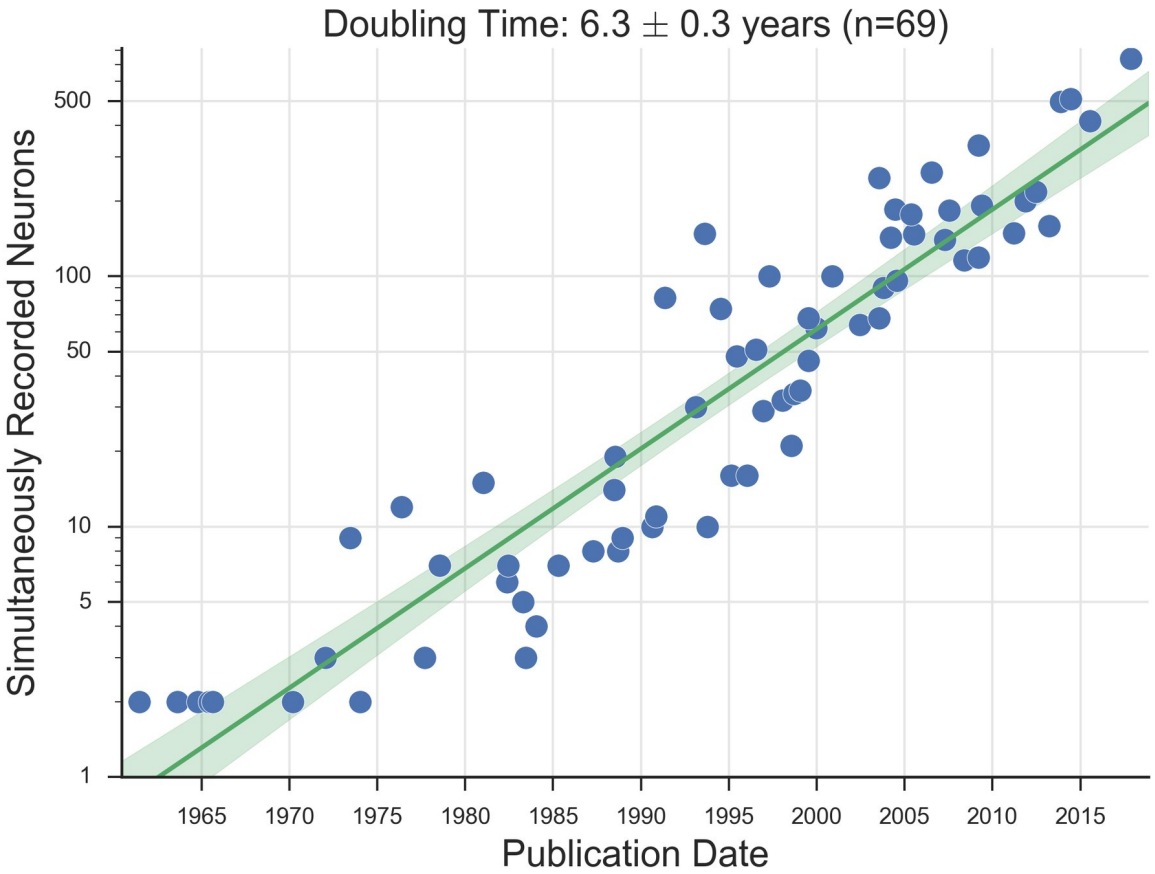


Figure: Stevenson and Kording. *How advances in neural recording affect data analysis*. Nature Neuroscience.

State of the art: Neuropixels (Jun et al., Nature 2017)

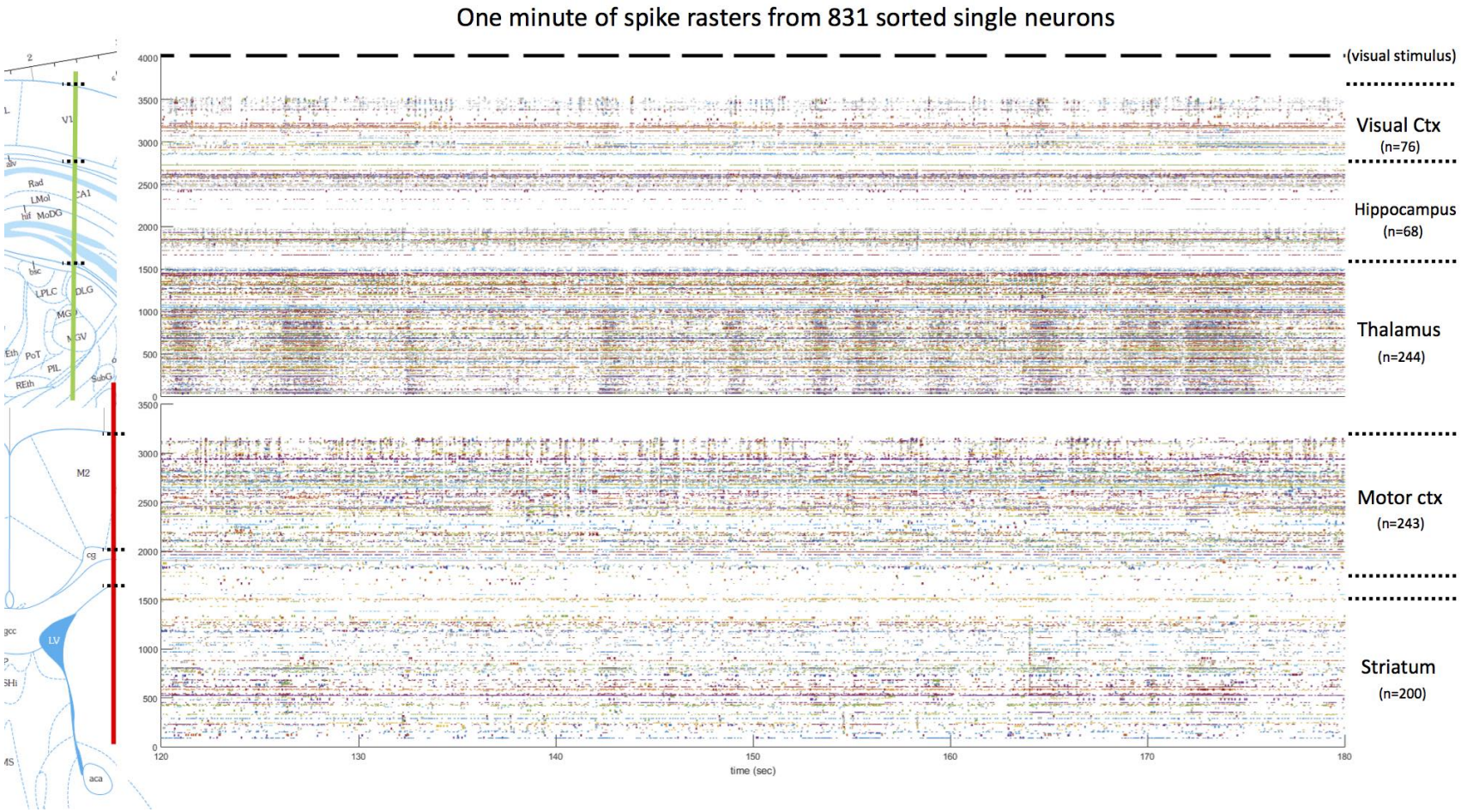


Figure: data.coretexlab.net. Simultaneous recording with two Neuropixels electrode arrays in a mouse.

Data and model sketch

Data characteristics

- electrode $\rightarrow$ "spike" vs. "no spike" each neuron $\rightarrow$ m.v. binary time series $(Y_n)_{n\geq 0}$
- covariates: stimulus $(z_n)_{n\geq 0}$, anatomical classification of neurons $\in \mathcal{P}, \mathcal{Q}, \ldots$

Data and model sketch

Data characteristics

- electrode $\to$ "spike" vs. "no spike" each neuron $\to$ m.v. binary time series $(Y_n)_{n\geq 0}$
- covariates: stimulus $(z_n)_{n\geq 0}$, anatomical classification of neurons $\in \mathcal{P}, \mathcal{Q}, \ldots$

Cartoon of model

$$X_n = AX_{n-1} + W_n, \quad W_n \sim \mathcal{N}(0, \Sigma), \qquad \{0,1\}^N \ni Y_n \sim p(\cdot | X_n; z_n)$$

$$p(y|x;z) \propto \exp\left(\sum_{i=1}^{N} y^i x^i + x^{\mathcal{P}\mathcal{Q}} \left\{\sum_{i\in\mathcal{P}} y^i\right\}\left\{\sum_{i\in\mathcal{Q}} y^i\right\} + x^z z\right)$$

- $p(\cdot|x;z)$ exp. fam. $\rightsquigarrow$ $x \mapsto p(y|x;z)$ log-concave
- $\rightsquigarrow$ $p(x_0, \ldots, x_n | y_0, \ldots, y_n)$ log-concave
- typically: $n = O(10^6)$, $d = O(10^3)$

Outline

1. Hidden Markov Models and the Viterbi Process
2. Motivation: high-dimensional neural data analysis
3. Convexity, high-dimensional optimization and sampling
4. ODE approach to the Viterbi process
5. Parallelized ULA

Convexity basics

Definition (Strong convexity)

With $\lambda \geq 0$, $U : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex if for all $\alpha \in [0, 1]$,

$$\alpha U(x) + (1 - \alpha)U(x') \geq U(\alpha x + (1 - \alpha)x') + \frac{\lambda}{2}\|x - x'\|^2 \alpha(1 - \alpha).$$

Convexity basics

Definition (Strong convexity)

With $\lambda \geq 0$, $U : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex if for all $\alpha \in [0, 1]$,

$$\alpha U(x) + (1-\alpha)U(x') \geq U(\alpha x + (1-\alpha)x') + \frac{\lambda}{2}\|x - x'\|^2 \alpha(1-\alpha).$$

Facts

- If U is C^1, λ-strong convexity is equivalent to each of 1) and 2),

  1) $U(x) \geq U(x') + \langle \nabla U(x'), x - x' \rangle + \frac{\lambda}{2}\|x - x'\|^2$,
  2) $\langle \nabla U(x) - \nabla U(x'), x - x' \rangle \geq \lambda \|x - x'\|^2$.

- If U is C^2, λ-strong convexity is equivalent to 3),

  3) $\mathrm{Hess}_U(x) \succeq \lambda I_d$.

ODE interpretation of strong convexity

Gradient ODE

- If $U : \mathbb{R}^d \to \mathbb{R}$ is C^1 and λ-strongly convex, the soln. of:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) = -\nabla U(\Phi_t(x)), \qquad \Phi_0(x) = x,$$

satisfies

$$\|\Phi_t(x) - \Phi_t(x')\| \leq e^{-\lambda t}\|x - x'\|. \tag{1}$$

ODE interpretation of strong convexity

Gradient ODE

- If $U : \mathbb{R}^d \to \mathbb{R}$ is C^1 and λ-strongly convex, the soln. of:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) = -\nabla U(\Phi_t(x)), \qquad \Phi_0(x) = x,$$

satisfies

$$\|\Phi_t(x) - \Phi_t(x')\| \leq e^{-\lambda t}\|x - x'\|. \tag{1}$$

Key facts about gradient flow

- by simple differentiation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi_t(x) - \Phi_t(x')\|^2 = -2\left\langle \nabla U(\Phi_t(x)) - \nabla U(\Phi_t(x')), \Phi_t(x) - \Phi_t(x')\right\rangle,$$

- hence

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\Phi_t(x) - \Phi_t(x')\|^2 \leq -2\lambda\|\Phi_t(x) - \Phi_t(x')\|^2,$$

and in turn (1).

Convexity and high-dimensional optimization

Gradient ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t = -\nabla U(\Phi_t)$$

Convexity and high-dimensional optimization

Gradient ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t = -\nabla U(\Phi_t)$$

Gradient descent algorithm

If $U : \mathbb{R}^d \to \mathbb{R}$ is C^1, λ-strongly convex and $\|\nabla U(x) - \nabla U(x')\| \leq \mu\|x - x'\|$,

$$x[0] = x, \qquad x[k+1] = x[k] - h\nabla U(x[k]),$$

with $0 < h < \frac{2}{\mu+\lambda}$, satisfies

$$\|x[k] - \xi\| \leq \left(1 - \frac{2h\lambda\mu}{\mu+\lambda}\right)^k \|x - \xi\|,$$

where ξ is the minimizer of U.

Convexity and high-dimensional sampling

Overdamped Langevin SDE

$$\mathrm{d}\Phi_t = -\nabla U(\Phi_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad \text{targeting} \quad \pi(x) \propto e^{-U(x)}$$

Convexity and high-dimensional sampling

Overdamped Langevin SDE

$$\mathrm{d}\Phi_t = -\nabla U(\Phi_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \qquad \text{targeting} \quad \pi(x) \propto e^{-U(x)}$$

Unadjusted Langevin alg. (Dalalyan, 2014; Durmus and Moulines, 2017; Dalayan, 2018)

If $U : \mathbb{R}^d \to \mathbb{R}$ is C^1 , λ-strongly convex and $\|\nabla U(x) - \nabla U(x')\| \leq \mu\|x - x'\|$,

$$X[0] = x, \qquad X[k+1] = X[k] - h\nabla U(X[k]) + \sqrt{2h}Z[k], \quad Z[k] \sim \mathcal{N}(0, I_d)$$

with $0 < h < \frac{2}{\mu+\lambda}$, satisfies

$$W_2(\mathrm{Law}(X[k]), \pi) \leq (1 - \lambda h)^k \, W_2(\delta_x, \pi) + 1.82\frac{\mu}{\lambda}\sqrt{hd},$$

where $Z_n \sim \mathcal{N}(0, I)$, $\pi(x) \propto e^{-U(x)}$ and W_2 is order-2 Wasserstein distance.

Outline

1. Hidden Markov Models and the Viterbi Process
2. Motivation: high-dimensional neural data analysis
3. Convexity, high-dimensional optimization and sampling
4. ODE approach to the Viterbi process
5. Parallelized ULA

How to approach existence of Viterbi process?

$$(\xi_0^n, \ldots, \xi_n^n) := \underset{x_0,\ldots,x_n}{\operatorname{argmax}}\, p(x_0, \ldots, x_n | y_0, \ldots, y_n)$$

How to approach existence of Viterbi process?

$$(\xi_0^n, \dots, \xi_n^n) := \underset{x_0,\dots,x_n}{\operatorname{argmax}}\, p(x_0, \dots, x_n | y_0, \dots, y_n)$$

Prior-posterior singularity on infinite time-horizon

may NOT characterize Viterbi process as:

$$(\xi_0^\infty, \xi_1^\infty, \dots) = \underset{x_0,x_1,\dots}{\operatorname{argmax}}\, p(x_0, x_1, \dots | y_0, y_1, \dots)$$

because density "$p(x_0, x_1, \dots | y_0, y_1, \dots)$" does not exist, typically:

$$\mathbb{P}(X_0 \in \cdot, X_1 \in \cdot, \dots) \perp \mathbb{P}(X_0 \in \cdot, X_1 \in \cdot, \dots | y_0, y_1, \dots)$$

How to approach existence of Viterbi process?

$$p(x_0, \ldots, x_n | y_0, \ldots, y_n) \propto \exp\{-U^n(x_0, \ldots, x_n)\}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n), \qquad \Phi_t^n(\xi_0^n, \ldots, \xi_n^n) = (\xi_0^n, \ldots, \xi_n^n)$$

How to approach existence of Viterbi process?

$$p(x_0, \ldots, x_n | y_0, \ldots, y_n) \propto \exp\{-U^n(x_0, \ldots, x_n)\}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n), \qquad \Phi_t^n(\xi_0^n, \ldots, \xi_n^n) = (\xi_0^n, \ldots, \xi_n^n)$$

Key observation about HMM

by definition,

$$U^n(x_0, \ldots, x_n) = -\log p(x_0) - \log p(y_0|x_0) - \sum_{m=1}^{n} \log p(x_m|x_{m-1}) - \sum_{m=1}^{n} \log p(y_m|x_m),$$

hence for any $n > m > 0$,

$$\frac{\partial U^n}{\partial x_m}(x_0, \ldots, x_n) = -\frac{\partial}{\partial x_m}\left\{\log p(x_m|x_{m-1}) + \log p(x_{m+1}|x_m) + \log p(y_m|x_m)\right\}$$

How to approach existence of Viterbi process?

$$p(x_0, \ldots, x_n | y_0, \ldots, y_n) \propto \exp\{-U^n(x_0, \ldots, x_n)\}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n), \qquad \Phi_t^n(\xi_0^n, \ldots, \xi_n^n) = (\xi_0^n, \ldots, \xi_n^n)$$

Key observation about HMM

by definition,

$$U^n(x_0, \ldots, x_n) = -\log p(x_0) - \log p(y_0|x_0) - \sum_{m=1}^{n} \log p(x_m|x_{m-1}) - \sum_{m=1}^{n} \log p(y_m|x_m),$$

hence for any $n > m > 0$,

$$\frac{\partial U^n}{\partial x_m}(x_0, \ldots, x_n) = -\frac{\partial}{\partial x_m}\left\{\log p(x_m|x_{m-1}) + \log p(x_{m+1}|x_m) + \log p(y_m|x_m)\right\}$$

Idea to establish existence of Viterbi process:

$$\nabla U^n \to \text{"}\nabla U^\infty\text{"}, \quad \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty), \quad \Phi_t^\infty(\xi_0^\infty, \xi_1^\infty, \ldots) = (\xi_0^\infty, \xi_1^\infty, \ldots)$$

Viterbi process in the Hilbert space $l_2(\gamma)$

Definition: $l_2(\gamma)$

- $x = [x_0^T \; x_1^T \; \cdots]^T \in \mathbb{R}^{\mathbb{N}}$, where each $x_n \in \mathbb{R}^d$
- for $\gamma \in (0,1)$, $\langle x, x' \rangle_\gamma := \sum_{n=0}^{\infty} \gamma^n \langle x_n, x_n' \rangle$, $\|x\|_\gamma := \langle x, x \rangle_\gamma^{1/2} = \left(\sum_{n=0}^{\infty} \gamma^n \|x_n\|^2\right)^{1/2}$,
- $l_2(\gamma)$ is $\{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_\gamma < +\infty\}$ equipped with $\langle \cdot, \cdot \rangle_\gamma$.

Viterbi process in the Hilbert space $l_2(\gamma)$

Definition: $l_2(\gamma)$

- $x = [x_0^T \; x_1^T \; \cdots]^T \in \mathbb{R}^{\mathbb{N}}$, where each $x_n \in \mathbb{R}^d$
- for $\gamma \in (0,1)$, $\langle x, x' \rangle_\gamma := \sum_{n=0}^{\infty} \gamma^n \langle x_n, x_n' \rangle$, $\|x\|_\gamma := \langle x, x \rangle_\gamma^{1/2} = \left(\sum_{n=0}^{\infty} \gamma^n \|x_n\|^2\right)^{1/2}$,
- $l_2(\gamma)$ is $\{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_\gamma < +\infty\}$ equipped with $\langle \cdot, \cdot \rangle_\gamma$.

Theorem part 1 - ODE and fixed point in $l_2(\gamma)$

Under certain conditions, $\exists \gamma \in (0,1)$ and $\nabla U^\infty : l_2(\gamma) \to l_2(\gamma)$ such that:

- $\lim_{n\to\infty} [\nabla U^n(x_0, \ldots, x_n)^T \, 0 \, 0 \, \cdots]^T = \nabla U^\infty(x)$ pointwise in $l_2(\gamma)$,

Viterbi process in the Hilbert space $l_2(\gamma)$

Definition: $l_2(\gamma)$

- $x = [x_0^T \, x_1^T \, \cdots]^T \in \mathbb{R}^{\mathbb{N}}$, where each $x_n \in \mathbb{R}^d$
- for $\gamma \in (0,1)$, $\langle x, x' \rangle_\gamma := \sum_{n=0}^{\infty} \gamma^n \langle x_n, x_n' \rangle$, $\|x\|_\gamma := \langle x, x \rangle_\gamma^{1/2} = \left(\sum_{n=0}^{\infty} \gamma^n \|x_n\|^2\right)^{1/2}$,
- $l_2(\gamma)$ is $\{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_\gamma < +\infty\}$ equipped with $\langle \cdot, \cdot \rangle_\gamma$.

Theorem part 1 - ODE and fixed point in $l_2(\gamma)$

Under certain conditions, $\exists \gamma \in (0,1)$ and $\nabla U^\infty : l_2(\gamma) \to l_2(\gamma)$ such that:

- $\lim_{n\to\infty} [\nabla U^n(x_0, \ldots, x_n)^T \, 0 \, 0 \, \cdots]^T = \nabla U^\infty(x)$ pointwise in $l_2(\gamma)$,
- $\exists \lambda > 0$ such that

$$\left\langle \nabla U^\infty(x) - \nabla U^\infty(x'), x - x' \right\rangle_\gamma \geq \lambda \|x - x'\|_\gamma^2, \quad \forall x, x' \in l_2(\gamma)$$

Viterbi process in the Hilbert space $l_2(\gamma)$

Definition: $l_2(\gamma)$

- $x = [x_0^T \, x_1^T \, \cdots]^T \in \mathbb{R}^{\mathbb{N}}$, where each $x_n \in \mathbb{R}^d$
- for $\gamma \in (0,1)$, $\langle x, x' \rangle_\gamma := \sum_{n=0}^{\infty} \gamma^n \langle x_n, x_n' \rangle$, $\|x\|_\gamma := \langle x, x \rangle_\gamma^{1/2} = \left(\sum_{n=0}^{\infty} \gamma^n \|x_n\|^2\right)^{1/2}$,
- $l_2(\gamma)$ is $\{x \in \mathbb{R}^{\mathbb{N}} : \|x\|_\gamma < +\infty\}$ equipped with $\langle \cdot, \cdot \rangle_\gamma$.

Theorem part 1 - ODE and fixed point in $l_2(\gamma)$

Under certain conditions, $\exists \gamma \in (0,1)$ and $\nabla U^\infty : l_2(\gamma) \to l_2(\gamma)$ such that:

- $\lim_{n\to\infty} [\nabla U^n(x_0, \ldots, x_n)^T \, 0 \, 0 \, \cdots]^T = \nabla U^\infty(x)$ pointwise in $l_2(\gamma)$,
- $\exists \lambda > 0$ such that

$$\left\langle \nabla U^\infty(x) - \nabla U^\infty(x'), x - x' \right\rangle_\gamma \geq \lambda \|x - x'\|_\gamma^2, \quad \forall x, x' \in l_2(\gamma)$$

- $\exists$ unique soln. to the Fréchet ODE on $l_2(\gamma)$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty), \quad \Phi_0^\infty = Id,$$

$\|\Phi_t^\infty(x) - \Phi_t^\infty(x')\|_\gamma \leq e^{-\lambda t} \|x - x'\|_\gamma$, hence $\exists$ unique fixed point $\Phi_t^\infty(\xi^\infty) = \xi^\infty$.

Viterbi process in the Hilbert space $l_2(\gamma)$

Theorem part 2 - convergence to Viterbi process

Denoting $\xi^n := [(\xi_0^n)^T \cdots (\xi_n^n)^T \, 0 \, 0 \cdots]^T$, $\xi^\infty \in l_2(\gamma)$ is the Viterbi process in the sense that

$$\|\xi^n - \xi^\infty\|_\gamma^2 \leq \frac{1}{\lambda^2}\left(\frac{\gamma^n \alpha_n}{\lambda^2} + \sum_{m=n}^{\infty} \gamma^m \beta_m\right),$$

where $\alpha_n := \sum_{m=0}^{n} \gamma^m \beta_{n-m}$,

$$\beta_n := \left\|\frac{\partial}{\partial x_n} \log p(x_n|x_{n-1})\right\|^2 + \left\|\frac{\partial}{\partial x_n} \log p(x_{n+1}|x_n)\right\|^2 + \left\|\frac{\partial}{\partial x_n} \log p(y_n|x_n)\right\|^2,$$

evaluated at $x_{n-1} = x_n = x_{n+1} = 0$.

Proof sketch for bound on $\|\xi^n - \xi^\infty\|_\gamma$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty) \qquad \text{in} \quad l_2(\gamma)$$

(with ∇U^n padded with zeros)

main steps

- $\|\xi^n - \xi^\infty\|_\gamma = \lim_{t\to\infty} \|\xi^n - \Phi_t^\infty(\xi^n)\|_\gamma = \lim_{t\to\infty} \|\Phi_t^n(\xi^n) - \Phi_t^\infty(\xi^n)\|_\gamma$

Proof sketch for bound on $\|\xi^n - \xi^\infty\|_\gamma$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty) \qquad \text{in} \quad l_2(\gamma)$$

(with ∇U^n padded with zeros)

main steps

- $\|\xi^n - \xi^\infty\|_\gamma = \lim_{t\to\infty} \|\xi^n - \Phi_t^\infty(\xi^n)\|_\gamma = \lim_{t\to\infty} \|\Phi_t^n(\xi^n) - \Phi_t^\infty(\xi^n)\|_\gamma$
- Compare ODE flows

$$\begin{aligned}\Phi_t^n(x) - \Phi_t^\infty(x) &= \int_0^t \frac{\partial}{\partial s}\Phi_{t-s}^\infty \circ \Phi_s^n(x)\mathrm{d}s \\ &= \int_0^t \left\{\nabla U^\infty(\Phi_s^n(x)) - \nabla U^n(\Phi_s^n(x))\right\} \nabla\Phi_{t-s}^\infty(\Phi_s^n(x))\mathrm{d}s\end{aligned}$$

where

$$\nabla\Phi_t^\infty(x) = I - \int_0^t \mathrm{Hess}_U(\Phi_s^\infty(x))\nabla\Phi_s^\infty(x)\mathrm{d}s$$

Proof sketch for bound on $\|\xi^n - \xi^\infty\|_\gamma$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty) \qquad \text{in} \quad l_2(\gamma)$$

(with ∇U^n padded with zeros)

main steps

- $\|\xi^n - \xi^\infty\|_\gamma = \lim_{t\to\infty} \|\xi^n - \Phi_t^\infty(\xi^n)\|_\gamma = \lim_{t\to\infty} \|\Phi_t^n(\xi^n) - \Phi_t^\infty(\xi^n)\|_\gamma$
- Compare ODE flows

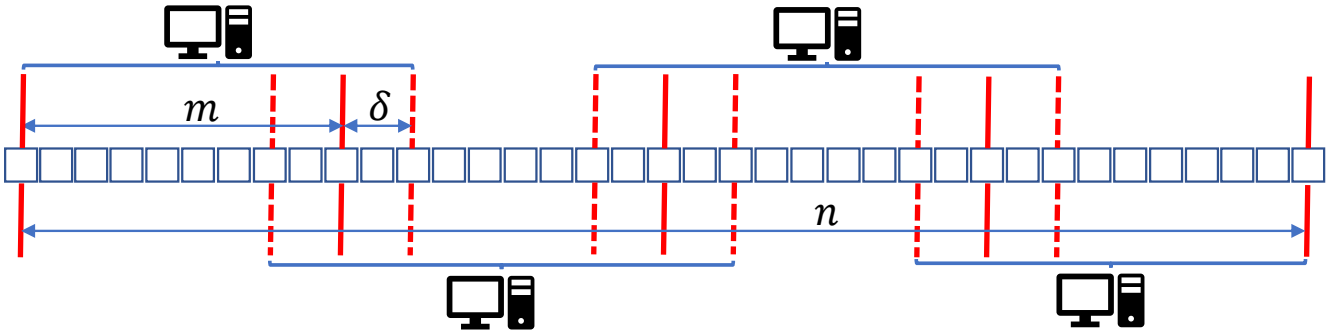
$$\begin{aligned}\Phi_t^n(x) - \Phi_t^\infty(x) &= \int_0^t \frac{\partial}{\partial s}\Phi_{t-s}^\infty \circ \Phi_s^n(x)\mathrm{d}s \\ &= \int_0^t \left\{\nabla U^\infty(\Phi_s^n(x)) - \nabla U^n(\Phi_s^n(x))\right\} \nabla\Phi_{t-s}^\infty(\Phi_s^n(x))\mathrm{d}s\end{aligned}$$

where

$$\nabla\Phi_t^\infty(x) = I - \int_0^t \mathrm{Hess}_U(\Phi_s^\infty(x))\nabla\Phi_s^\infty(x)\mathrm{d}s$$

- So $\|\xi^n - \xi^\infty\|_\gamma \le \|\nabla U^\infty(\xi^n) - \nabla U^n(\xi^n)\|_\gamma c(n) \int_0^t e^{-\lambda(t-s)}\mathrm{d}s$

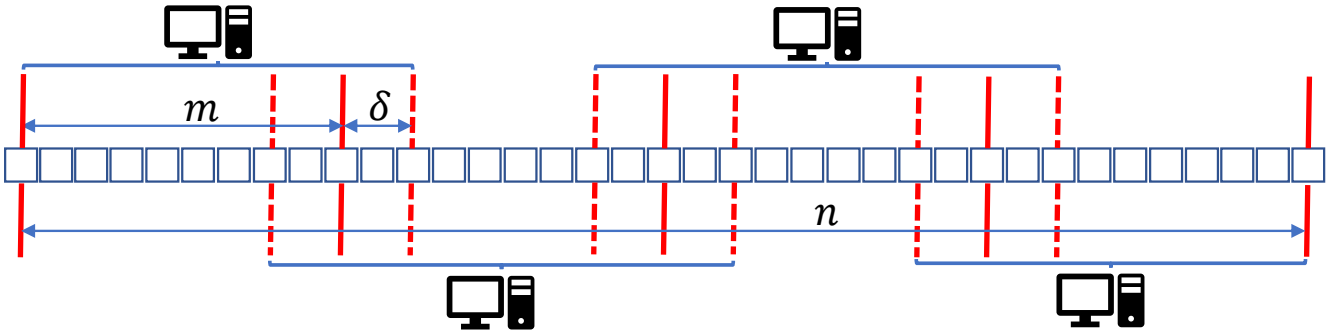
Viterbi process in the Hilbert space $l_2(\gamma)$



Speed-up/error trade-off

- **speed-up** $\approx T(n)/T(m+2\delta)$
- **squared Euclidean error** $\approx \underbrace{n/m}_{\textbf{number of segments}} \times \underbrace{\|(\xi_0^{m+\delta},\ldots,\xi_m^{m+\delta}) - (\xi_0^n,\ldots,\xi_m^n)\|^2}_{=:\mathcal{E}^2(n,m,\delta)\ \textbf{squared error per segment}}$

Viterbi process in the Hilbert space $l_2(\gamma)$



Speed-up/error trade-off

- **speed-up** $\approx T(n)/T(m+2\delta)$
- **squared Euclidean error** $\approx \underbrace{n/m}_{\textbf{number of segments}} \times \underbrace{\|(\xi_0^{m+\delta},\ldots,\xi_m^{m+\delta}) - (\xi_0^n,\ldots,\xi_m^n)\|^2}_{=:\mathcal{E}^2(n,m,\delta)\ \textbf{squared error per segment}}$

Theorem part 3 - error bound

$$\sup_{n\geq m+\delta} \mathcal{E}^2(n,m,\delta) \leq \frac{1}{\lambda^2}\gamma^\delta \left(\frac{\alpha_m}{\lambda^2} + \sum_{\ell=0}^{\infty} \gamma^\ell \beta_{m+\ell}\right)$$

Conditions on the HMM

Consider:

$$X_n = A(X_{n-1}) + W_n, \qquad Y_n \sim p(\cdot|X_n),$$

- $(W_n)_{n\geq 1}$ iid $\in \mathbb{R}^d$ with pdf $\propto \exp\{-\psi(x)\}$, X_0 with pdf $\propto \exp\{-\psi_0(x)\}$,
- where $x \mapsto \psi(x), x \mapsto \psi_0(x), x \mapsto p(y|x)$ are strictly positive and C^1

Three sets of sufficient conditions for previous theorems to hold

- A is linear, X_0 and W_n are Gaussian, and $x \mapsto p(y|x)$ is sufficiently strongly log-concave, i.e. for $\lambda > 0$ sufficiently large

$$\left\langle \frac{\partial}{\partial x_n} \log p(y_n|x_n) - \frac{\partial}{\partial x_n} \log p(y_n|x_n'), x_n - x_n' \right\rangle \leq -\lambda \|x_n - x_n'\|^2, \quad \forall n$$

- A, ψ, ψ_0 are Lipschitz, $x \mapsto p(y|x)$ is sufficiently strongly log-concave

Conditions on the HMM

Consider:

$$X_n = A(X_{n-1}) + W_n, \qquad Y_n \sim p(\cdot|X_n),$$

- $(W_n)_{n\geq 1}$ iid $\in \mathbb{R}^d$ with pdf $\propto \exp\{-\psi(x)\}$, X_0 with pdf $\propto \exp\{-\psi_0(x)\}$,
- where $x \mapsto \psi(x), x \mapsto \psi_0(x), x \mapsto p(y|x)$ are strictly positive and C^1

Three sets of sufficient conditions for previous theorems to hold

- A is linear, X_0 and W_n are Gaussian, and $x \mapsto p(y|x)$ is sufficiently strongly log-concave, i.e. for $\lambda > 0$ sufficiently large

$$\left\langle \frac{\partial}{\partial x_n} \log p(y_n|x_n) - \frac{\partial}{\partial x_n} \log p(y_n|x_n'), x_n - x_n' \right\rangle \leq -\lambda \|x_n - x_n'\|^2, \quad \forall n$$

- $x \mapsto p(y|x)$ is semi log-concave, i.e. $\exists \lambda \in \mathbb{R}$ s.t.

$$\left\langle \frac{\partial}{\partial x_n} \log p(y_n|x_n) - \frac{\partial}{\partial x_n} \log p(y_n|x_n'), x_n - x_n' \right\rangle \leq -\lambda \|x_n - x_n'\|^2, \quad \forall n,$$

  A is linear, X_0 and W_n are Gaussian, with $\|A\|_{\mathrm{op}} < 1$ sufficiently small
- A, ψ, ψ_0 are Lipschitz, $x \mapsto p(y|x)$ is sufficiently strongly log-concave

Conditions on the HMM

Consider:

$$X_n = A(X_{n-1}) + W_n, \qquad Y_n \sim p(\cdot|X_n),$$

- $(W_n)_{n\geq 1}$ iid $\in \mathbb{R}^d$ with pdf $\propto \exp\{-\psi(x)\}$, X_0 with pdf $\propto \exp\{-\psi_0(x)\}$,
- where $x \mapsto \psi(x), x \mapsto \psi_0(x), x \mapsto p(y|x)$ are strictly positive and C^2

A simple case to illustrate dimension independence of γ

- $A(x) = ax$ with $|a| < 1$, ψ is $\mathcal{N}(0, \sigma^2 I_d)$, ψ_0 is stationary density, and
- $x \mapsto p(y|x)$ is C^1, log-concave (not necessarily strongly)

then

$$\sup_{n\geq m+\delta} \mathcal{E}^2(n, m, \delta) \leq \frac{1}{\lambda^2}\gamma^\delta \left(\frac{\alpha_m}{\lambda^2} + \sum_{\ell=0}^{\infty} \gamma^\ell \beta_{m+\ell} \right)$$

holds with

$$\gamma = |a|, \quad \lambda = \frac{1 - |a|^2}{2\sigma^2}.$$

Outline

1. Hidden Markov Models and the Viterbi Process
2. Motivation: high-dimensional neural data analysis
3. Convexity, high-dimensional optimization and sampling
4. ODE approach to the Viterbi process
5. Parallelized ULA

From gradient descent to ULA

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n) \quad \rightsquigarrow \quad \mathrm{d}\Phi_t^n = -\nabla U^n(\Phi_t^n)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

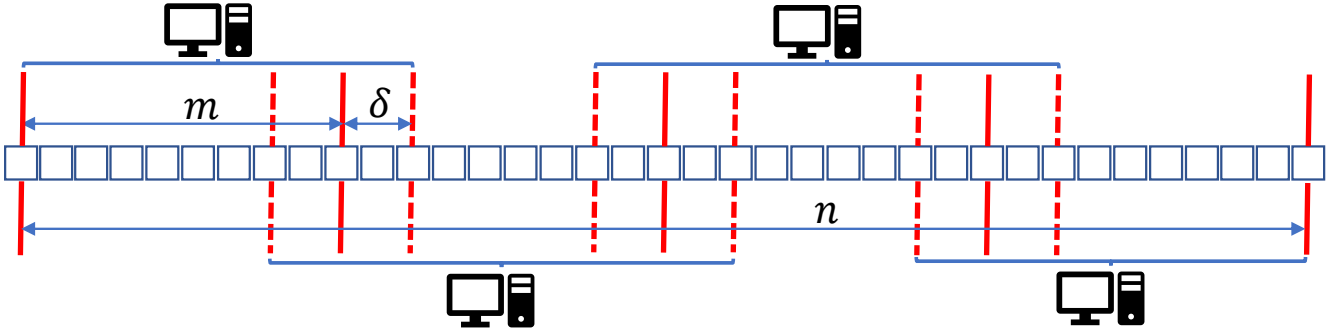
$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty) \quad \rightsquigarrow \quad \mathrm{d}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

From gradient descent to ULA

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^n = -\nabla U^n(\Phi_t^n) \quad \rightsquigarrow \quad \mathrm{d}\Phi_t^n = -\nabla U^n(\Phi_t^n)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

$$\downarrow \qquad\qquad\qquad\qquad \downarrow$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty) \quad \rightsquigarrow \quad \mathrm{d}\Phi_t^\infty = -\nabla U^\infty(\Phi_t^\infty)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$



parallelized gradient descent $\rightsquigarrow$ parallelized ULA

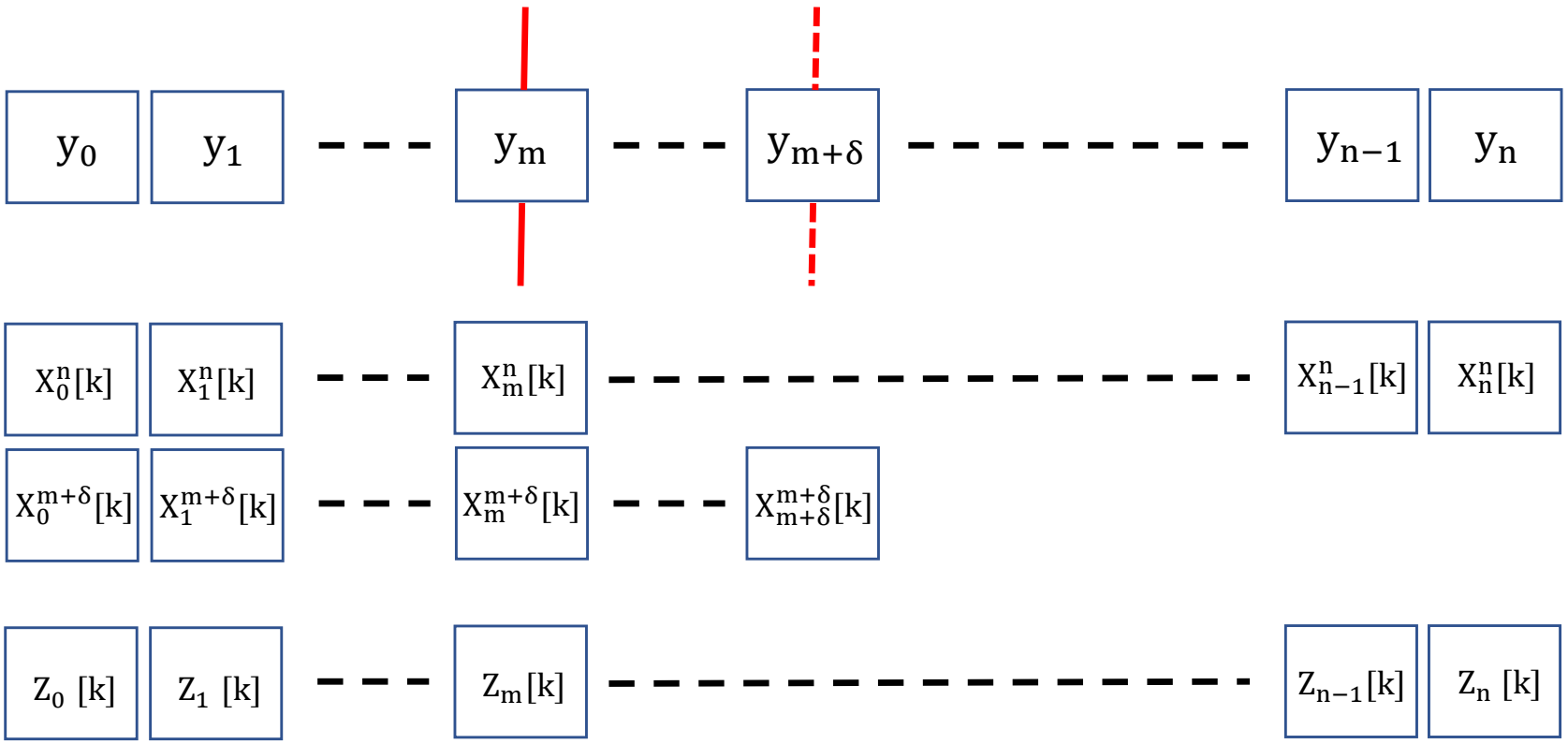
MAP estimation posterior sampling

Parallelized ULA

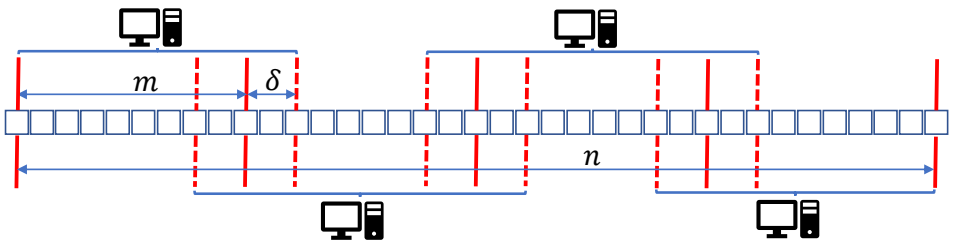
- Synchronous coupling of ULAs:

$$X^n[k] = X^n[k-1] - h\nabla U^n(X^n[k-1]) + \sqrt{2h}Z[k], \qquad X^n[0] = X[0]$$

where $Z[k] = (Z_0[k]^T \; Z_1[k]^T \; \cdots)^T \in \mathbb{R}^{\mathbb{N}}$, $Z_m[k] \overset{iid}{\sim} \mathcal{N}(0, I_d)$ are common across n



Parallelized ULA



$$X^n[k] = X^n[k-1] - h\nabla U^n(X^n[k-1]) + \sqrt{2h}Z[k], \qquad X^n[0] = X[0]$$

Theorem

Assuming additionally that $\sup_n \|\nabla U^n(x) - \nabla U^n(x')\|_\gamma \leq \mu\|x - x'\|_\gamma$,

$$\begin{aligned}
&\|(X_0^{m+\delta}[k] \cdots X_m^{m+\delta}[k]) - (X_0^n[k] \cdots X_m^n[k])\|^2 \\
&\leq 12\gamma^\delta \frac{\mu^2 h^2}{1-\rho^2}\left(V^n[k] + \frac{8}{3}\frac{\alpha_m}{\lambda^2} + \sum_{\ell=0}^{\infty}\gamma^\ell \beta_{m+\ell}\right), \quad a.s.,
\end{aligned}$$

where $\rho < 1$ depends on h, λ, μ and $V^n[k]$ is a nonnegative r.v. such that:

$$\sup_{n \geq m+\delta} \sup_k \mathbb{E}[V^n[k]] \leq \frac{2hd}{1-\gamma}\frac{1}{1-\rho} + 2\|x[0]\|_\gamma^2 + \frac{2\alpha_m}{\lambda^2}$$

Emprirical speed-up for parallelized gradient descent

Implementation

- One node of UoB's BlueChrystal Phase 3 cluster: 16 x 2.6 GHz SandyBridge cores, 4GB/core
- Matlab Parallel Computing Toolbox

Timings

- set number of segments n/m equal to number of cores (1 core $\equiv$ no approximation)
- For given ϵ and each number of cores choose δ s.t .$|rel.error| < \epsilon$, where error is in 5000 iterations of gradient descent.
- $n = O(10^5)$,$d = O(10^2)$,$\delta = O(10^1)$

cores	1	2	4	8	16
speed-up	1	1.60	2.59	5.57	11.3