Constructing Metropolis-Hastings proposals using damped BFGS updates

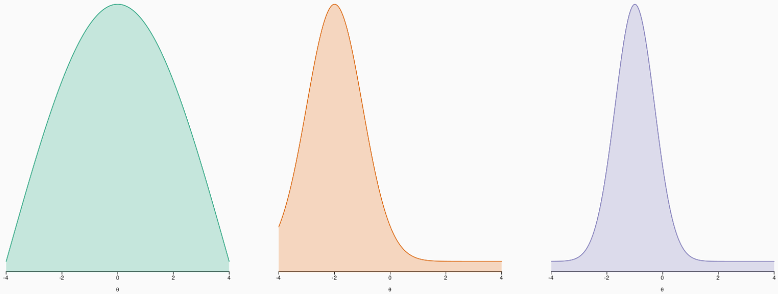
The 18th IFAC Symposium on System Identification, Stockholm, Sweden.

Johan Dahlin, University of Newcastle, Australia.

This is work in collaboration with

Dr. Adrian Wills (University of Newcastle, Australia)
Prof. Brett Ninness (University of Newcastle, Australia)

Bayesian inference in one slide



$$\pi(\theta) \triangleq p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)},$$

$$\pi[\varphi] \triangleq \mathbb{E}_{\pi}\left[\varphi(\theta)\right] = \int_{\Theta} \varphi(\theta)\pi(\theta)\,\mathsf{d}\theta.$$

What are we going to do?

- Estimate posterior distributions using Markov chains.
- Improve the standard choice to handle high-dimensional problems.

Why are we doing this?

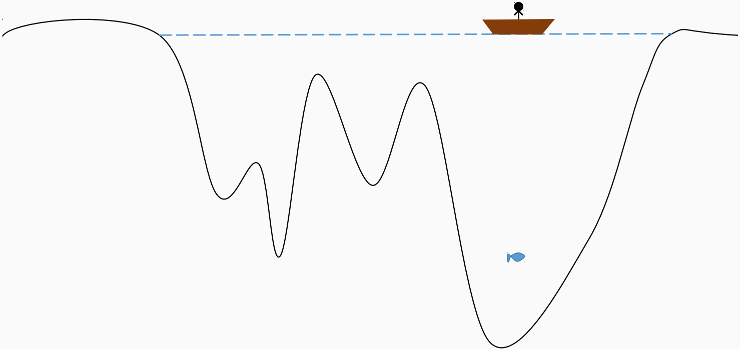
- Standard choice gives inefficient sampling: slow or not working at all.
- Bayesian methods give uncertainty and valid estimates for finite data.

How will we do this?

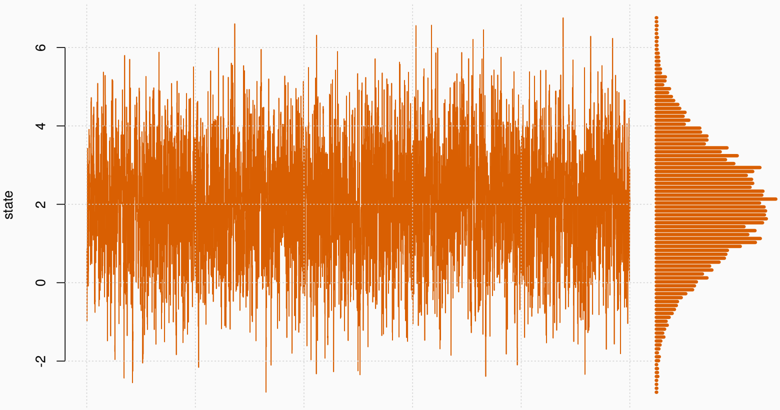
- Make use of gradient and Hessian information.
- Estimate the Hessian using quasi-Newton and least squares.

Exploring posteriors using Markov chains.

Exploring ”the lake”



Markov chains: stationary distribution



Metropolis-Hastings: algorithm

Get samples from target $\pi(\theta) \propto p(y|\theta)p(\theta)$ by iterating over k :

(i) Propose candidate parameter θ' by

$$\theta' \sim q(\theta'|\theta_{k-1}).$$

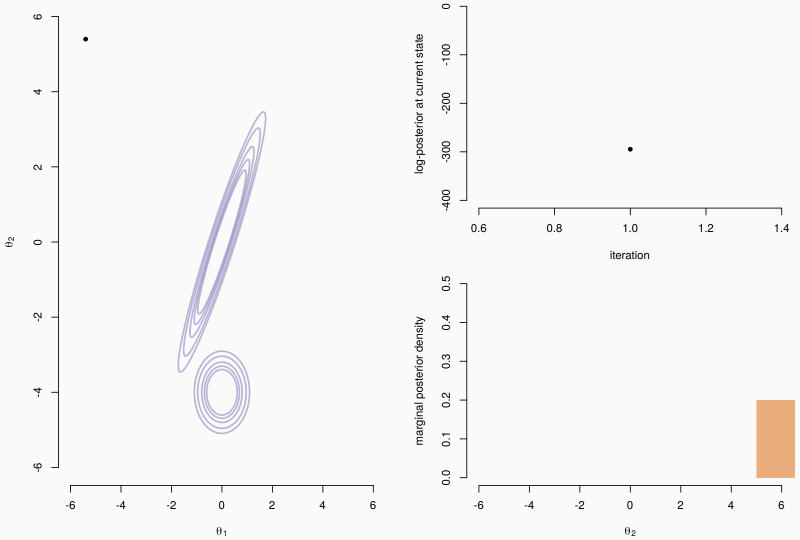
(ii) Accept θ' by setting $\theta_k \leftarrow \theta'$ with probability

$$\min\left\{1, \frac{\pi(\theta')}{\pi(\theta_{k-1})}\right\},$$

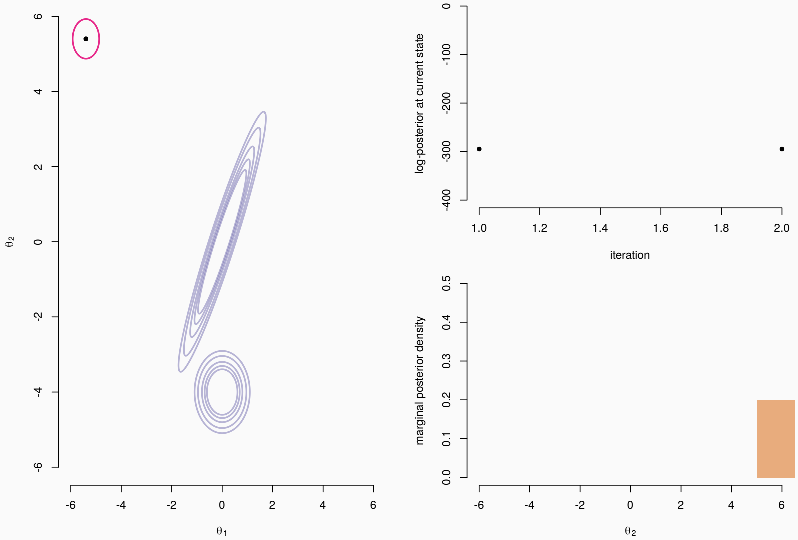
otherwise reject θ' by setting $\theta_k \leftarrow \theta_{k-1}$.

Output: Samples $\{\theta_k\}_{k=1}^{K}$ from $\pi(\theta)$.

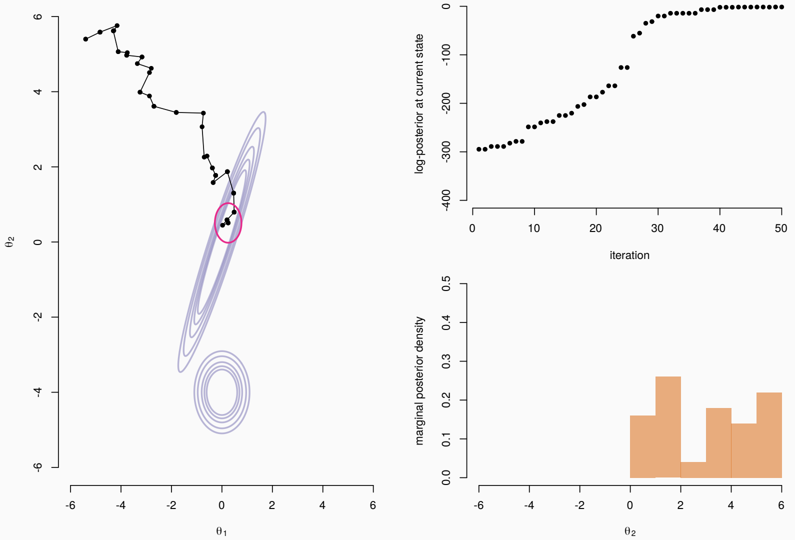
Metropolis-Hastings: toy example



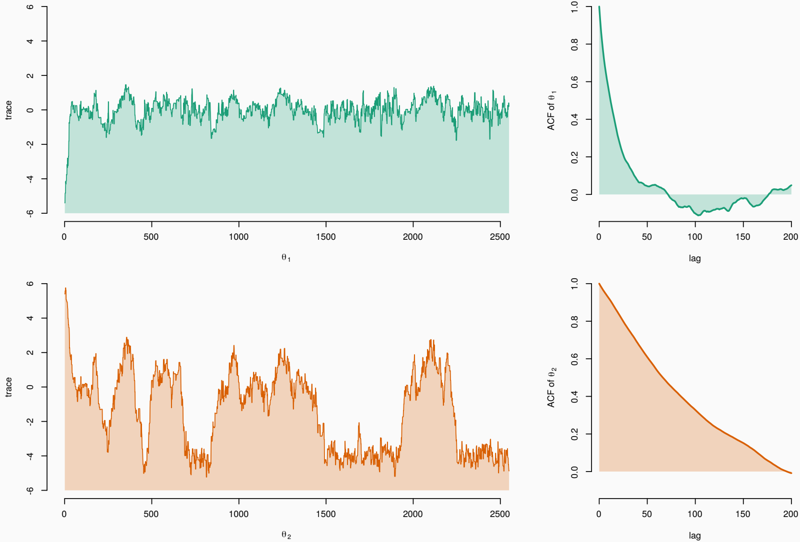
Metropolis-Hastings: toy example



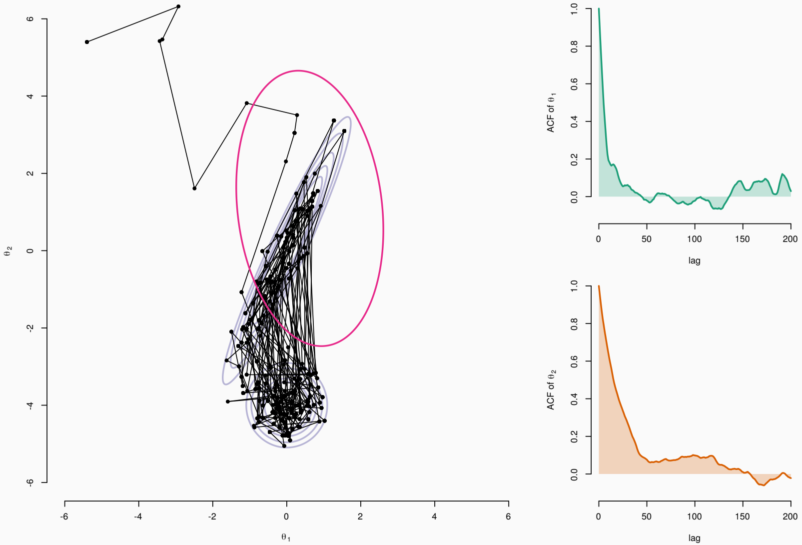
Metropolis-Hastings: toy example



Metropolis-Hastings: proposal and mixing

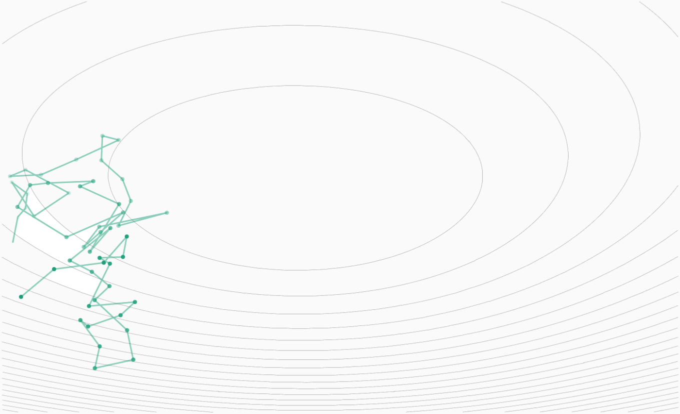


Metropolis-Hastings: proposal and mixing



Constructing efficient proposal distributions.

Gaussian random walk proposal



$$\theta' | \theta_{k-1} \sim \mathcal{N}\left(\theta' \middle| \theta_{k-1}, \Sigma\right),$$

A second-order approximation

Second-order Taylor expansion of log-target

$$\log \pi(\theta + \Delta\theta) \approx \log \pi(\theta) + G(\theta)^\top \Delta\theta - \frac{1}{2}\Delta\theta^\top H(\theta)\Delta\theta,$$

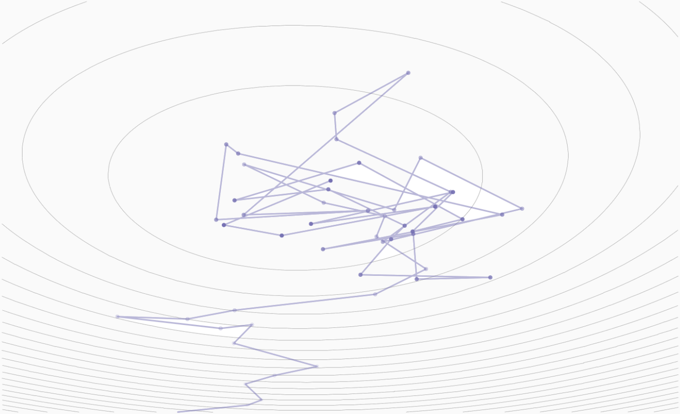
with the approximate gradient

$$\nabla \log \pi(\theta + \Delta\theta) \approx G(\theta) - H(\theta)\Delta\theta,$$

which by setting to zero gives the search direction

$$\Delta\theta = H(\theta)G(\theta).$$

Second-order Gaussian proposal



$$\theta' | \theta_{k-1} \sim \mathcal{N}\Big(\theta' \Big| \theta_{k-1} + \frac{1}{2} H(\theta_{k-1}) G(\theta_{k-1}), H(\theta_{k-1})\Big).$$

Hessian estimation using quasi-Newton methods.

Hessian estimates using quasi-Newton

BFGS $$\bar{H}_k = \bar{H}_{k-1} + \frac{s_k s_k^\top}{s_k^\top y_k} - \frac{\bar{H}_{k-1} y_k y_k^\top \bar{H}_{k-1}}{y_k^\top \bar{H}_{k-1} y_k},$$

SR1 $$\bar{H}_k = \bar{H}_{k-1} + \frac{(s_k - \bar{H}_{k-1} y_k)(s_k - \bar{H}_{k-1} y_k)^\top}{(s_k - \bar{H}_{k-1} y_k)^\top y_k},$$

$$s_k \triangleq \theta_k - \theta_{k-1}, \qquad y_k \triangleq G(\theta_k) - G(\theta_{k-1}).$$

Hessian estimate using least squares

The quasi-Newton estimate $\bar{H}(\theta)$ is assumed to fulfill

$$G(\theta + \Delta\theta) = G(\theta) + \bar{H}(\theta)\Delta\theta,$$

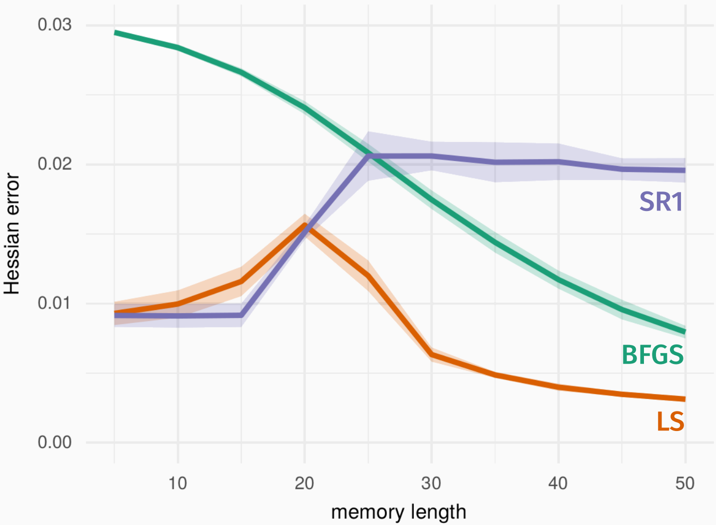
which is called the secant condition. Introduce

$$Y_k = [y_2, \ldots, y_k], \quad S_k = [s_2, \ldots, s_k],$$

and compute a least squares estimate of

$$Y_k = \bar{H}_k S_k.$$

Hessian approximation error



Numerical illustrations.

Detecting the Higgs boson

$N = 11,000,000$ events generated with binary observations (detection/background) given $p = 21$ kinematic properties.

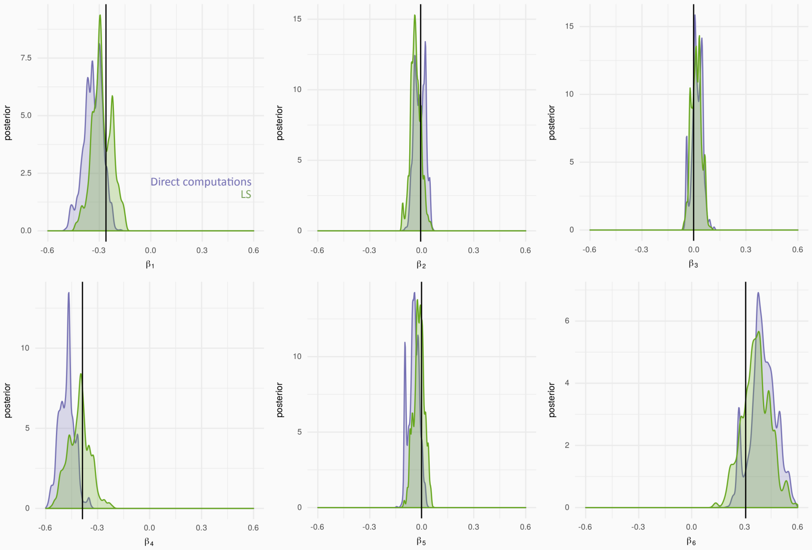
Modelling the data by a logistic regression model:

$$y_t \sim \text{Bernoulli}(p_t), \qquad p_t = \text{logit}\left(\beta_0 + \sum_{i=1}^{p} \beta_i x_{it}\right).$$

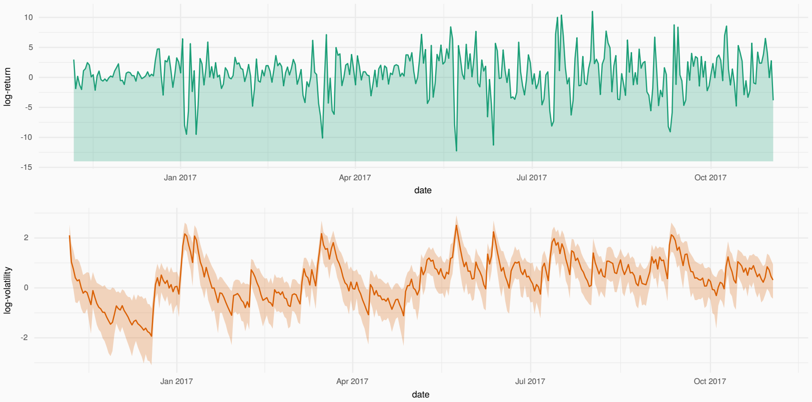
where the parameters are $\theta = \{\beta_0, \beta_1, \ldots, \beta_{21}\}$.

Big data setting: sub-sampling methods are required.

Detecting the Higgs boson, cont.



Modelling Bitcoin prices



Modelling Bitcoin prices

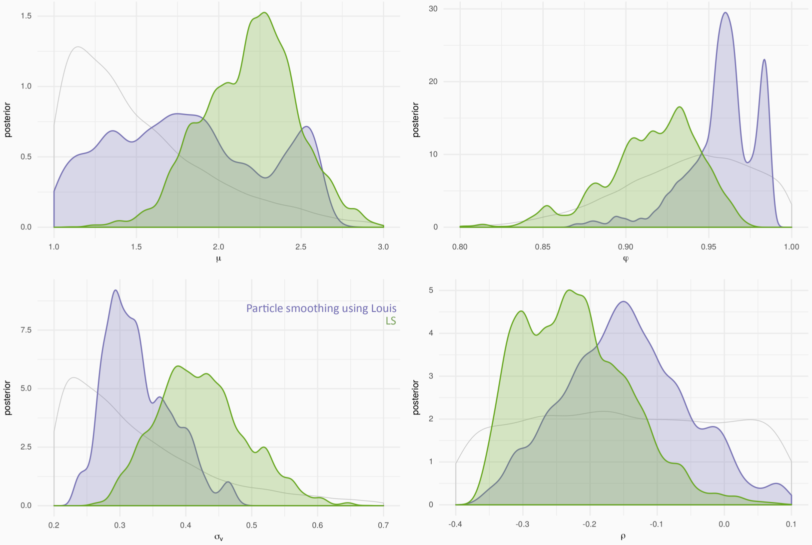
Modelling the data using a stochastic volatility model:

$$x_{t+1}|x_t \sim \mathsf{N}\Big(x_{t+1}\Big|\mu + \phi(x_t - \mu) + \rho\sigma_v y_t \exp(-x_t), \sigma_v^2\Big),$$
$$y_t|x_t \sim \mathsf{N}\Big(y_t\Big|0, \exp(x_t)\Big)$$

where the parameters are $\theta = \{\mu, \phi, \sigma_v, \rho, x_{0:T}\}$.

Latent variables: particle smoothing is required.

Modelling Bitcoin prices, cont.



What did we do?

- Employed a second-order approximation as a proposal.
- Estimated the Hessian using quasi-Newton and least squares.
- Applied the approach to high-dimensional problems.

Why did we do this?

- Standard proposals does not scale well.
- Hessian estimates are difficult to compute directly.
- Bayesian methods work on finite data and gives uncertainty bounds.

What are you going to do now?

- Remember that high-dimensional Bayesian inference can be possible.
- Read the paper and look at the code on GitHub.

New pre-print extending the idea

Correlated pseudo-marginal Metropolis-Hastings using quasi-Newton proposals

Johan Dahlin, Adrian Wills and Brett Ninness*

June 26, 2018

Abstract

Pseudo-marginal Metropolis-Hastings (pmMH) is a versatile algorithm for sampling from target distributions which are not easy to evaluate point-wise. However, pmMH requires good proposal distributions to sample efficiently from the target, which can be problematic to construct in practice. This is especially a problem for high-dimensional targets when the standard random-walk proposal is inefficient.

arXiv pre-print: `1806.09780`.

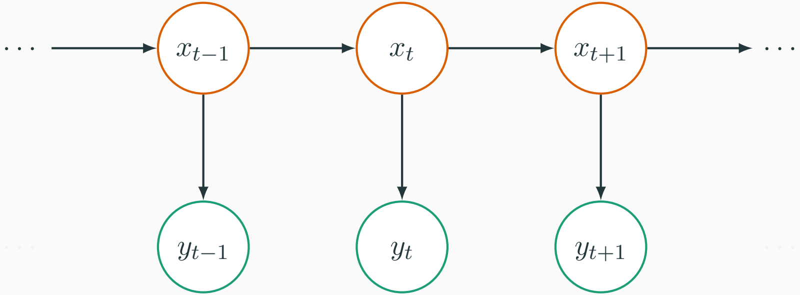
Thank you for listening

Comments, suggestions and/or questions?

Johan Dahlin

johan.dahlin@newcastle.edu.au
research.johandahlin.com

State-space models

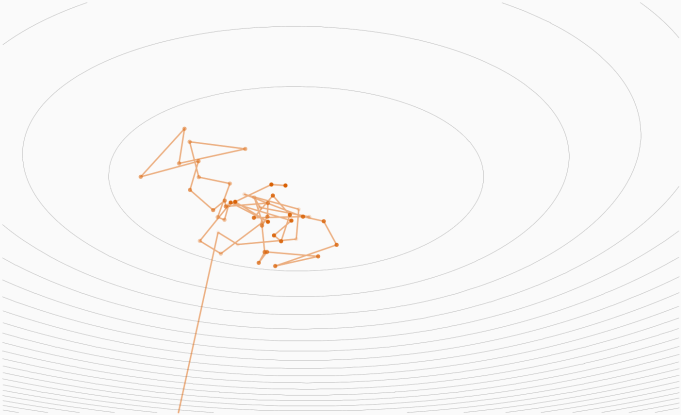


$$x_0 \sim \mu_\theta(x_0) \qquad x_{t+1}|x_t \sim f_\theta(x_{t+1}|x_t), \qquad y_t|x_t \sim g_\theta(y_t|x_t).$$

Linear Gaussian state-space model:

$$x_{t+1}|x_t \sim \mathcal{N}\Big(x_{t+1}; \mu + \phi(x_t - \mu), \sigma_v^2\Big), \qquad y_t|x_t \sim \mathcal{N}\Big(y_t; x_t, \sigma_e^2\Big).$$

Noisy gradient ascent update



$$\theta_k|\theta_{k-1} \sim \mathcal{N}\left(\theta_k; \theta_{k-1} + \frac{1}{2}\Sigma G(\theta_{k-1}), \Sigma\right).$$

Logit function

The logit function is given by

$$\text{logit}(f(x)) = \frac{1}{1 + \exp\left(-f(x)\right)}$$

and squeezes a real-valued number into the unit interval: