Bayesian Learning via Stochastic Gradient Langevin Dynamics

Max Welling and Yee Whye Teh
Presented by:
Andriy Mnih and Levi Boyles

University of California Irvine
University College London

June 2011 / ICML

Outline

Outline

Motivation

Method

Demonstrations

Discussion

Motivation

- Large scale datasets are becoming more commonly available across many fields.
- Learning complex models from these datasets will be the future.
- Current successes in scalable learning methods are optimization-based and non-Bayesian.
- Bayesian methods are currently not scalable, e.g. each iteration of MCMC sampling requires computations over the whole datasets.
- **Aim: develop Bayesian methodologies applicable to large scale datasets.**
    - Best of both worlds: scalability, and Bayesian protection against overfitting.

Contribution

- A very simple twist to standard **stochastic gradient ascent**.
- Turns it into a Bayesian algorithm which samples from the full posterior distribution rather than converges to a MAP mode.
- Resulting algorithm is related to **Langevin dynamics**—a classical physics method for sampling from a distribution.
- Applied to Bayesian mixture models, logistic regression, and independent components analysis.

Outline

Setup

- Parameter vector θ.
- Large numbers of data items $x_1, x_2, \dots, x_N$ where $N \gg 1$.
- Model distribution is:

$$p(\theta, X) = p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$

- **Aim: obtain samples from posterior distribution $p(\theta|X)$.**

Stochastic Gradient Ascent

- Also known as: **stochastic approximation**, **Robbins-Munro**.
- At iteration $t = 1, 2, \ldots$:
  - Get a subset (minibatch) $x_{t1}, \ldots, x_{tn}$ of data items where $n \ll N$.
  - Approximate gradient of log posterior using the subset:

$$\nabla \log p(\theta_t | X) \approx \nabla \log p(\theta_t) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{ti} | \theta_t)$$

  - Take a gradient step:

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2} \left(\nabla \log p(\theta_t) + \frac{N}{n} \sum_{i=1}^{n} \nabla \log p(x_{ti} | \theta_t) \right)$$

Stochastic Gradient Ascent

- Major requirement for convergence on step-sizes[1]:

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \qquad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty$$

- Intuition:
  - Step sizes cannot decrease too fast, otherwise will not be able to traverse parameter space.
  - Step sizes must decrease to zero, otherwise parameter trajectory will not converge to a local MAP mode.

[1] In addition to other technical assumptions.

Langevin Dynamics

- Stochastic differential equation describing dynamics which converge to posterior $p(\theta|X)$:

  $$d\theta(t) = \frac{1}{2}\nabla \log p(\theta(t)|X) + db(t)$$

  where $b(t)$ is Brownian motion.
- Intuition:
  - Gradient term encourages dynamics to spend more time in high probability areas.
  - Brownian motion provides noise so that dynamics will explore the whole parameter space.

Langevin Dynamics

- First order Euler discretization:

$$\theta_{t+1} = \theta_t + \frac{\epsilon}{2} \nabla \log p(\theta_t|X) + \eta_t \qquad \eta_t = N(0, \epsilon)$$

- Amount of noise is balanced to gradient step size.
- With finite step size there will be discretization errors.
- Discretization can be fixed by Metropolis-Hastings accept/reject step.
- As $\epsilon \to 0$ acceptance rate goes to 1.

Stochastic Gradient Langevin Dynamics

- **Idea: Langevin dynamics with stochastic gradients.**

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \eta_t$$
$$\eta_t = N(0, \epsilon_t)$$

- Update is just stochastic gradient ascent plus Gaussian noise.
- Noise variance is balanced with gradient step sizes.
- ϵ_t decreases to 0 slowly (step-size requirement).

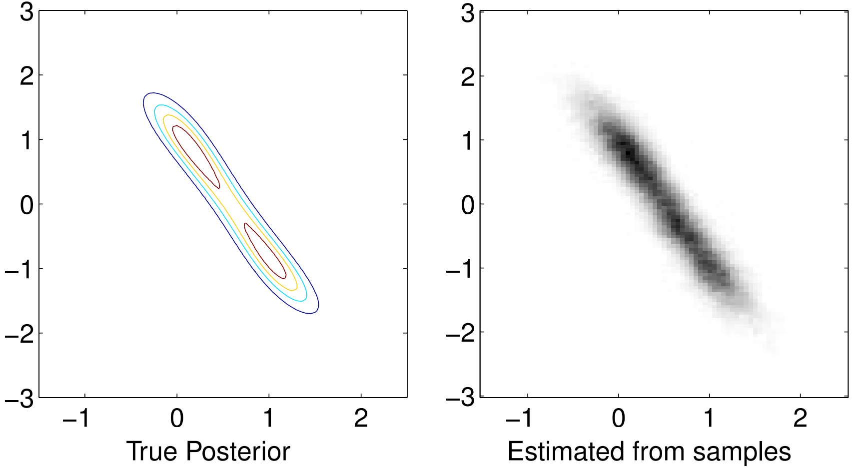
Stochastic Gradient Langevin Dynamics—Intuition

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \eta_t$$
$$\eta_t = N(0, \epsilon_t)$$

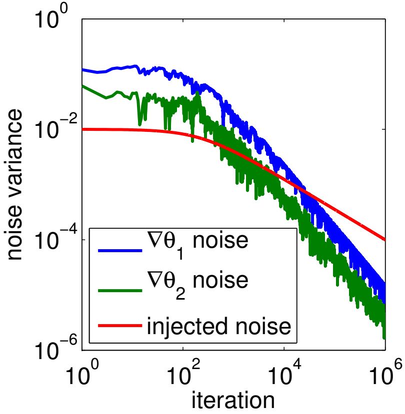
- Only computationally expensive part of Langevin dynamics is the gradient computation. If gradient can be well-approximated on small minibatches the algorithm will work well.
- As $\epsilon_t \to 0$:
  - Variance of gradient noise is $O(\epsilon_t^2)$, while variance of η_t is $\epsilon_t \gg \epsilon_t^2$. Gradient noise dominated by η_t so can be ignored.
    Result: Langevin dynamic updates with decreasing step sizes.
  - MH acceptance probability approaches 1, so we can ignore the expensive MH accept/reject step.
  - ϵ_t approaches 0 slowly enough, so discretized Langevin dynamics still able to explore whole parameter space.

Outline

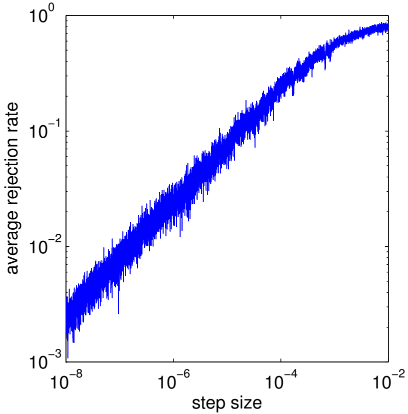
Mixture of Gaussians



Mixture of Gaussians

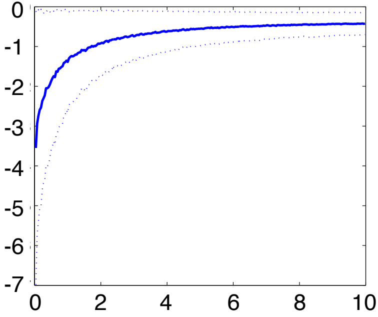


Noise amounts vs iterations

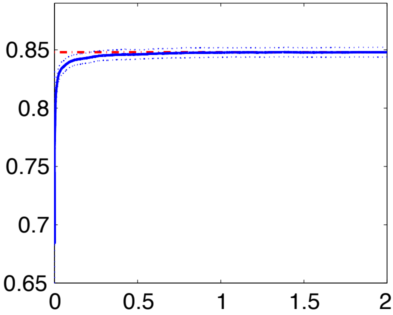


Rejection probabilities vs step sizes

Logistic Regression

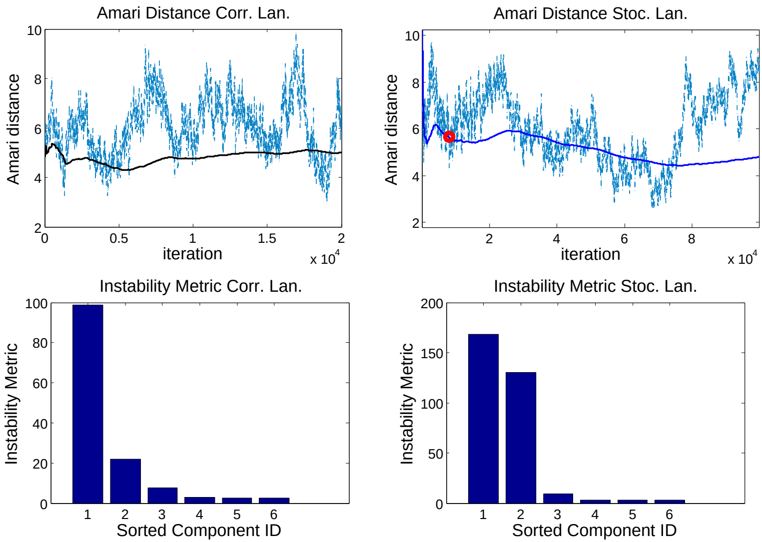


Average log joint probability vs iterations through dataset



Accuracies vs iterations through dataset

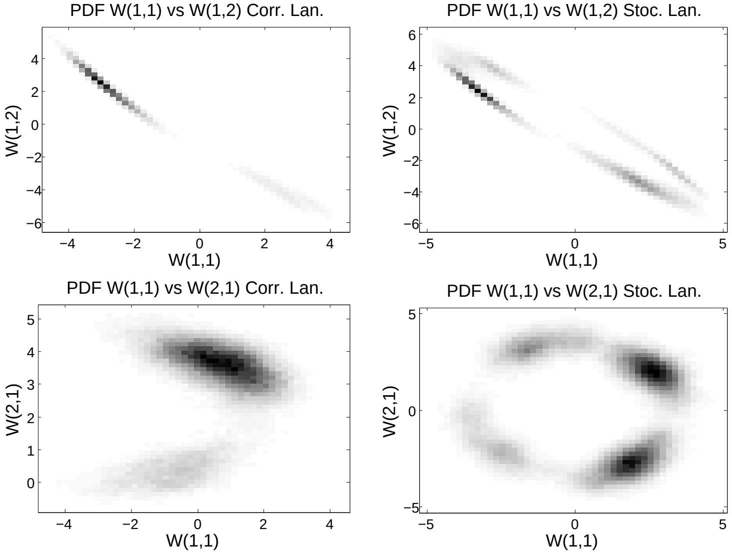
Independent Components Analysis



Corrected Langevin

Stochastic Langevin

Independent Components Analysis



Corrected Langevin

Stochastic Langevin

Outline

Discussion

- This is the first baby step towards Bayesian learning for large scale datasets.
- Future work:
  - Theoretical convergence proof.
  - Better scalable MCMC techniques.
  - Methods that do not require decreasing step sizes.