Scaling up MCMC for Bayesian inference through adaptive data subsampling

Ayoub El Hanchi

McGill University

April 30, 2019

Basic MCMC

- In Bayesian inference, we are interested in evaluating integrals of the form:

$$\mathop{\mathrm{E}}_{\theta\sim\pi}[\mathcal{L}(\theta)] = \int_{R^d} \mathcal{L}(\theta)\pi_N(\theta|x)\mathrm{d}\theta$$

for some loss function $\mathcal{L}(\theta)$ with respect to the posterior distribution of the parameters $\pi_N(\theta|x)$ for some observed data $x = \{x_i\}_{i=1}^N$.

- Under the usual exchangeability assumption, the posterior has the form:

$$\pi_N(\theta|x) = \frac{1}{Z}\left\{\prod_{i=1}^N \pi(x_i|\theta)\right\}\pi(\theta)$$

where Z is the normalizing constant. On the log scale this becomes:

$$\log\pi(\theta|x) = \sum_{i=1}^N \log\pi(x_i|\theta) + \log\pi(\theta) - \log Z$$

Basic MCMC

- The MCMC solution to this problem is to construct an ergodic Markov chain $\{\theta_i\}_{i\in\mathbb{N}}$ with stationary distribution $\pi(\theta|x)$ and estimate the above integral with the finite sum:

$$\mathop{\mathrm{E}}_{\theta\sim\pi}[\mathcal{L}(\theta)] \approx \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\theta_i)$$

- Under some regularity conditions, we have:

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\theta_i) - \mathop{\mathrm{E}}_{\theta\sim\pi}[\mathcal{L}(\theta)] \to \mathcal{N}\left(0, \frac{1}{n}\sigma_{\mathcal{L}}\right)$$

where:

$$\sigma_{\mathcal{L}} = \mathop{\mathrm{Var}}_{\theta_0\sim\pi}[\mathcal{L}(\theta_0)] + 2\sum_{i=1}^{\infty}\mathop{\mathrm{Cov}}_{\theta_0\sim\pi}[\mathcal{L}(\theta_0), \mathcal{L}(\theta_i)]$$

Complexity

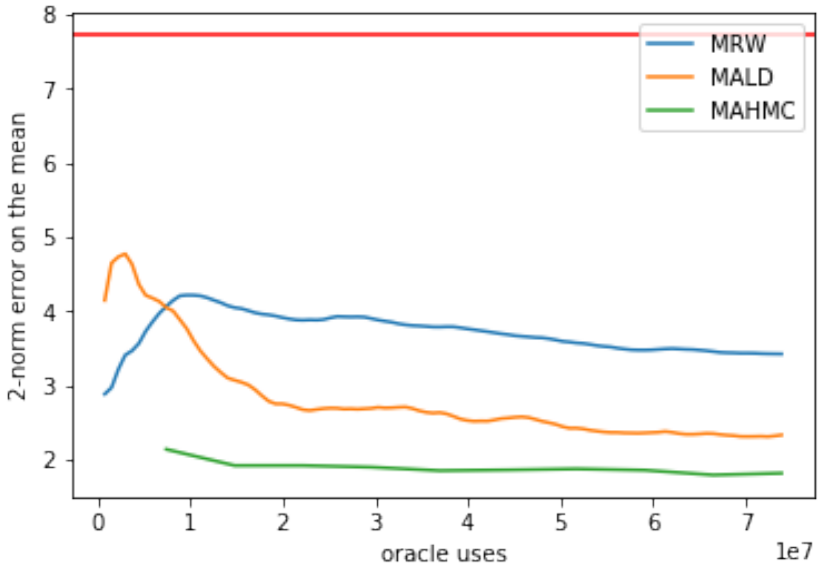
- An efficient Markov Chain is one that:
  - converges rapidly to its limiting distribution (fast CLT convergence)
  - has low autocorrelation (small asymptotic variance).
- We focus on the convergence to the limiting distribution.
- We assume we have an oracle that returns $\log \pi(x_i|\theta)$ or $\nabla_\theta \log \pi(x_i|\theta)$, and ask how many oracle uses are needed for the Markov Chain to be ϵ close to the limiting distribution.
- We are interested in the case of large data with $N \gg 1$ and $d \gg 1$, and look for a sublinear algorithm in both d and N.

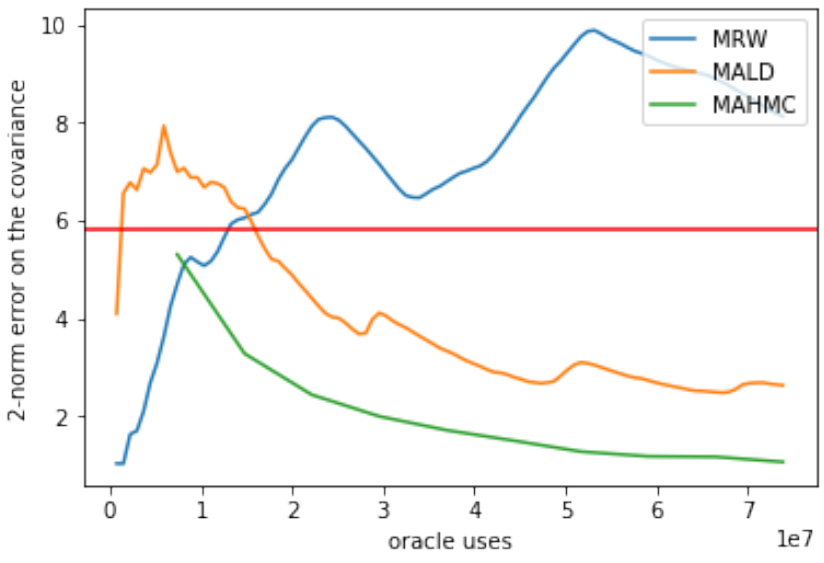
Conjectured rates for Traditional MCMC

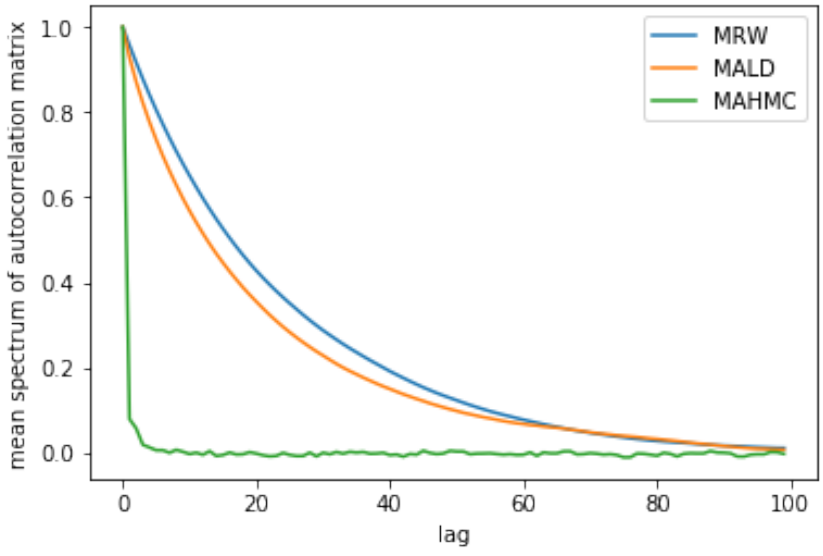
Metropolis random walk (MRW)	$\mathcal{O}\left(\frac{Nd}{\log(\epsilon)}\right)$
Metropolis adjusted Langevin dynamics (MALD)	$\mathcal{O}\left(\frac{Nd^{1/3}}{\log(\epsilon)}\right)$
Metropolis Hamiltonian Monte Carlo (MAHMC)	$\mathcal{O}\left(\frac{Nd^{1/4}}{\log(\epsilon)}\right)$

- The dependence on the dimension d comes from studying the optimal acceptance rates for each of the algorithms.
- Some recent results prove variants of these rates for strongly log-concave densities.

Experiments







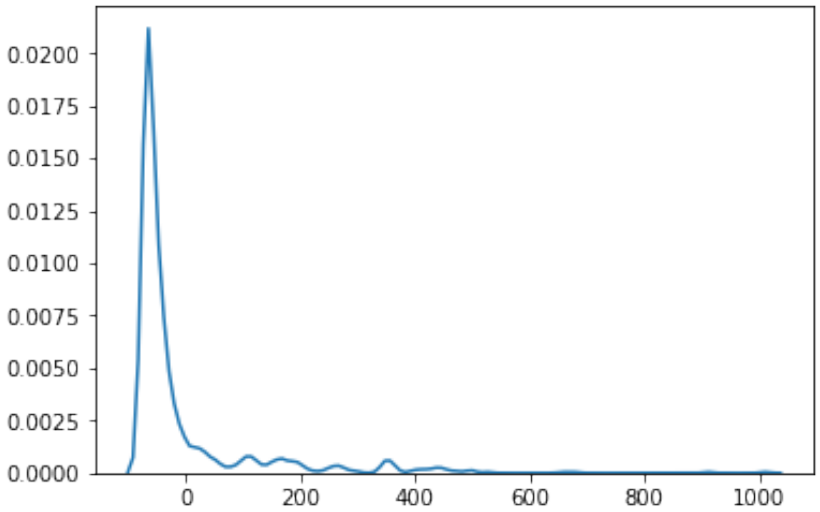
Bias-Variance Tradeoff

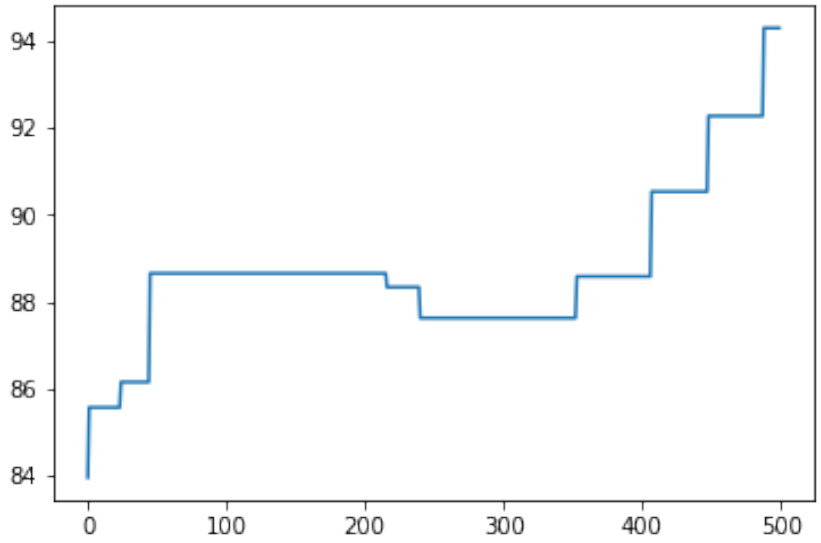
- Traditional MCMC algorithms asymptotically converge to the posterior distribution, and are therefore (asymptotically) unbiased.
- Their high computational complexity on the other hand prevents us from collecting a large number of samples, resulting in high variance.
- This suggests the development of efficient approximate MCMC algorithms with an adjustable bias-controlling parameter $\epsilon > 0$.
- Many approaches fit in this framework. We focus here on approaches based on data subsampling.

Approximate Metropolis random walk

- Main idea: treat the Metropolis accept-reject step as a statistical decision problem.
- Relies on the assumption that a CLT holds for the subsampled data.
- Given $\epsilon > 0$, use a one-sample t-test with threshold ϵ to decide whether to accept or reject.
- if the test is inconclusive, sample more data points.
- Guaranteed to halt under sampling without replacement.
- Difficult to directly quantify the bias.

Experiments



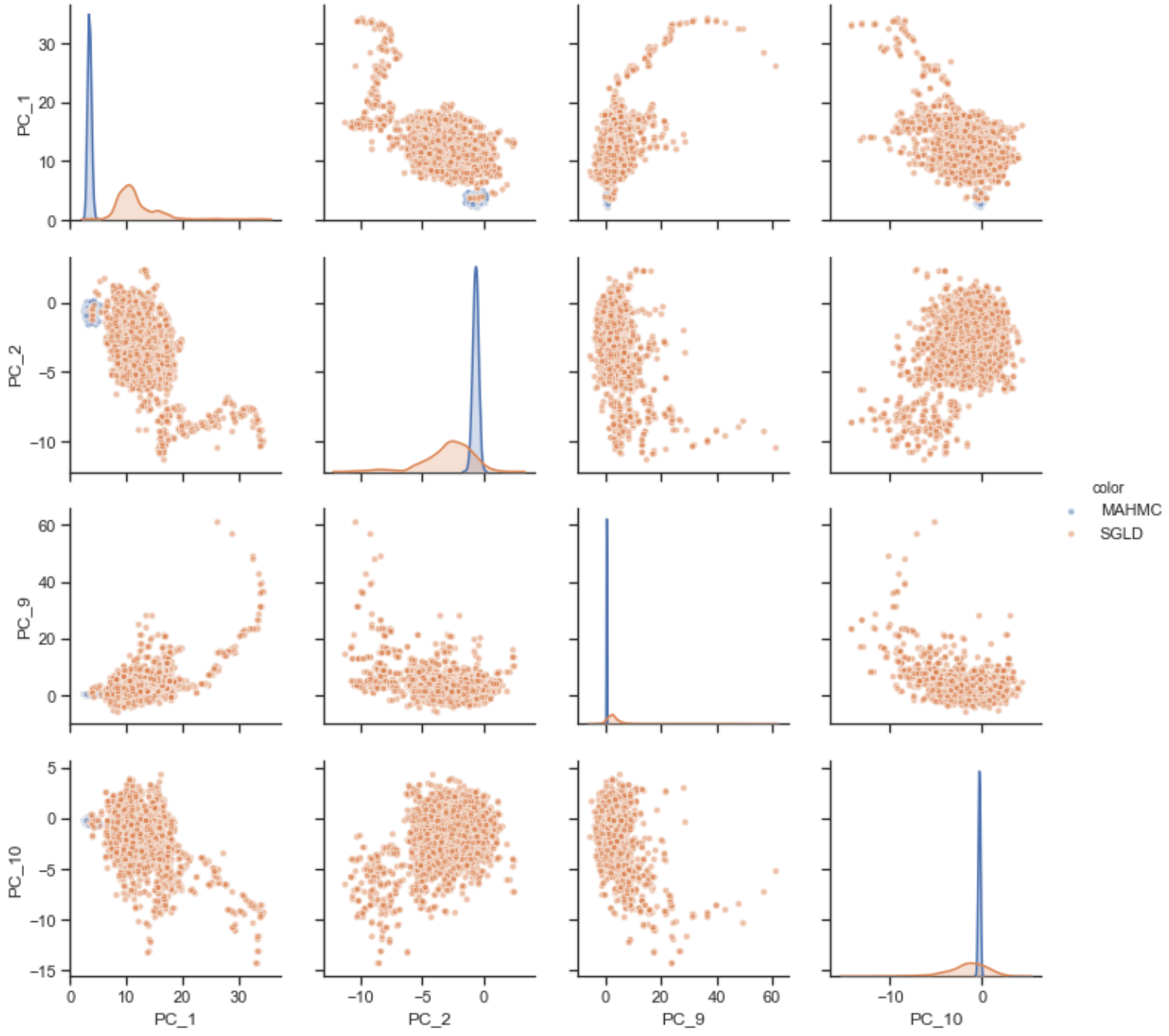


- CLT assumption does not hold.
- Bad estimates of the log posterior have a small, but non-zero chance of being accepted. The chain gets stuck at those points with no way to recover.

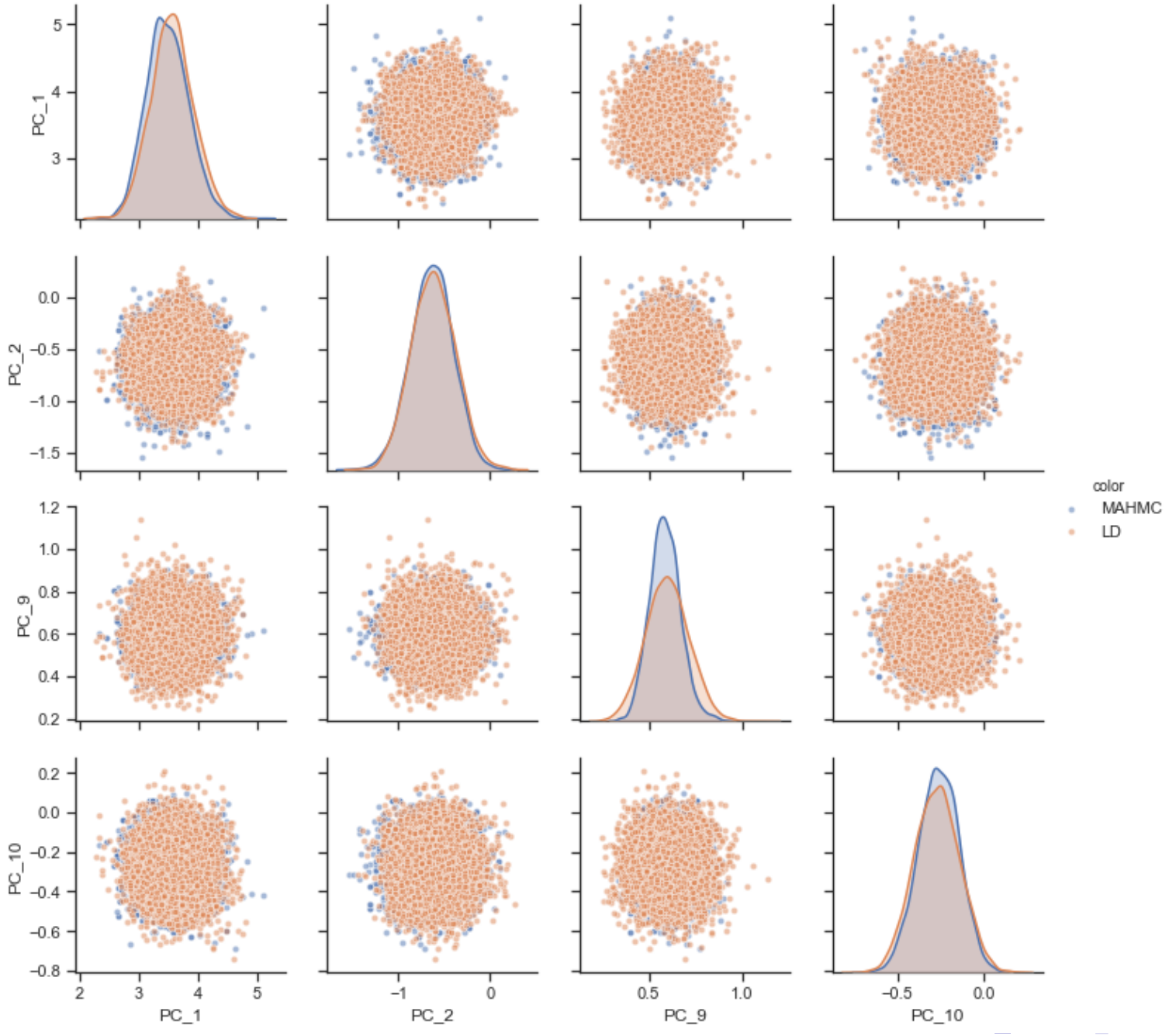
Stochastic gradient Langevin dynamics (SGLD)

- Based on the discretization of a continuous time Markov process.
- Replaces the gradient with an unbiased estimate of the gradient obtained using a subset of the data.
- Two sources of bias: the discretization error and the use of a noisy gradient.
- Bias due to noisy gradient can be removed if the CLT holds.
- Highly reminiscent of stochastic gradient descent.

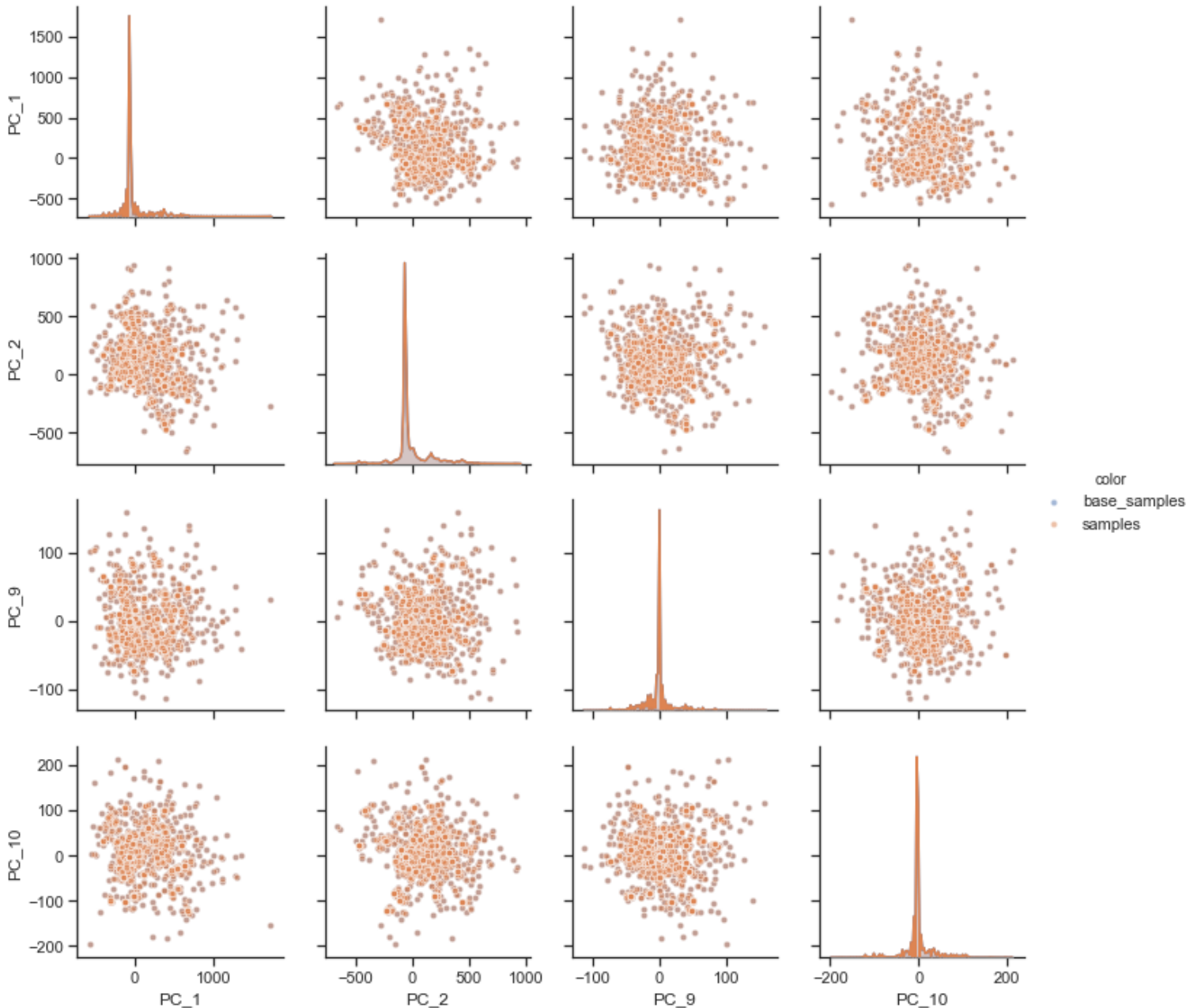
Experiments



Experiments



Experiments



Some Theory

- Recently, it was proven that every Markov process with stationary distribution with density $\pi(z)$ can be written in the following SDE form:

$$dz = g(z)dt + \sqrt{A(z) + A(z)^T}dW$$

where:

$$g(z) = A(z)\nabla \log \pi(z) + \Gamma(z)$$

$$\Gamma_i(z) = \sum_{i=1}^{d} \frac{\partial}{\partial \theta_j} A_{ij}(z)$$

$$A(z) \in \mathbb{R}^{d \times d}$$

$$z \in \mathbb{R}^d$$

Langevin dynamics (LD)

- Taking $z := \theta$ and $A(z) := A$ we recover the Langevin dynamics:
$$d\theta = A\nabla \log \pi(\theta)dt + \sqrt{2A}dW_t$$
- It's Euler discretization with step size $t > 0$ is given by:
$$\theta_{i+1} = \theta_i + tA\nabla \log \pi(\theta) + \mathcal{N}(0, 2tA)$$
- The discrete chain does not converge to $\pi(\theta)$ for any constant step size $t > 0$. Furthermore the unbiased discretization is not implementable.
- The above discretization is of order 1 (the local error decreases linearly with t). A discretization of order 2 exists but makes use of the proximal gradient operator.

Langevin dynamics (LD)

- No discretization of higher order exists, for this process, and for any other continuous-time stochastic process. (it requires running the noise backwards in time).
- While it does not converge to $\pi(\theta)$, the discretized process can get arbitrarily close to it by a suitable choice of step size.
- It was recently shown that for the discretized process to be ϵ away from $\pi(\theta)$, it needs $\mathcal{O}(\frac{d}{\epsilon}\log\frac{d}{\epsilon})$ iterations.
- The second order discretization is conjectured to require only $\mathcal{O}(\sqrt{\frac{d}{\epsilon}}\log\frac{d}{\epsilon})$ iterations to achieve the same accuracy.

Connections with optimization

- It was shown that the continuous time Langevin dynamics process can be interpreted as gradient flow on the space of probability measures with the Wasserstein metric of order 2, and the KL-divergence as the objective functional.
- One can therefore interpret the discretized version as gradient descent on the space of probability measures.
- In optimization, it is well known that no algorithm can get ϵ close to the optimum in less than $\Omega(\frac{1}{\sqrt{\epsilon}})$ iterations.
- This rate is achieved using Nesterov acceleration. Is there an analog to Nesterov acceleration in this case ?

Underdamped Langevin dynamics (ULD)

- Yes !
- Taking

$$z = \begin{bmatrix} \theta \\ v \end{bmatrix}$$

and

$$A(z) = A = \begin{bmatrix} 0 & -B \\ B & \gamma B \end{bmatrix}$$

and

$$\log \pi(z) = \log \pi(\theta) - \frac{1}{2} v^T B^{-1} v$$

yields the underdamped langevin dynamics:

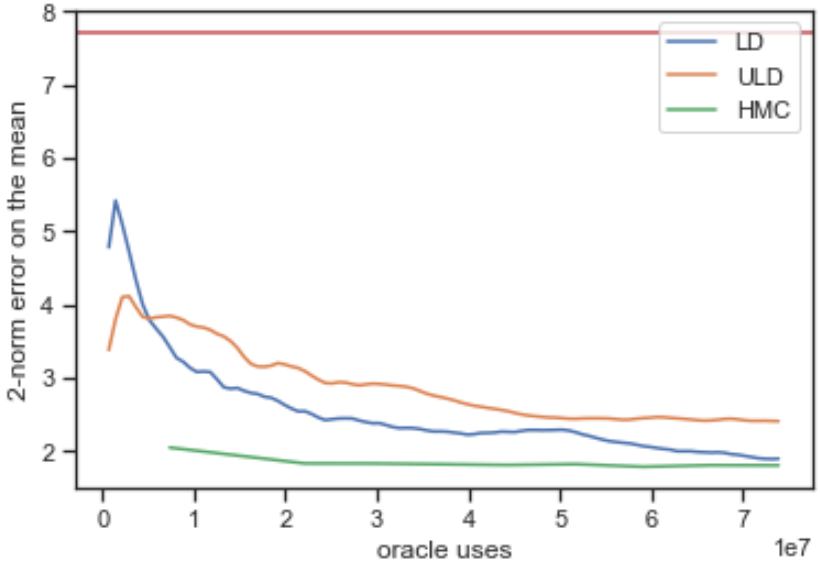
$$d\theta = v dt$$

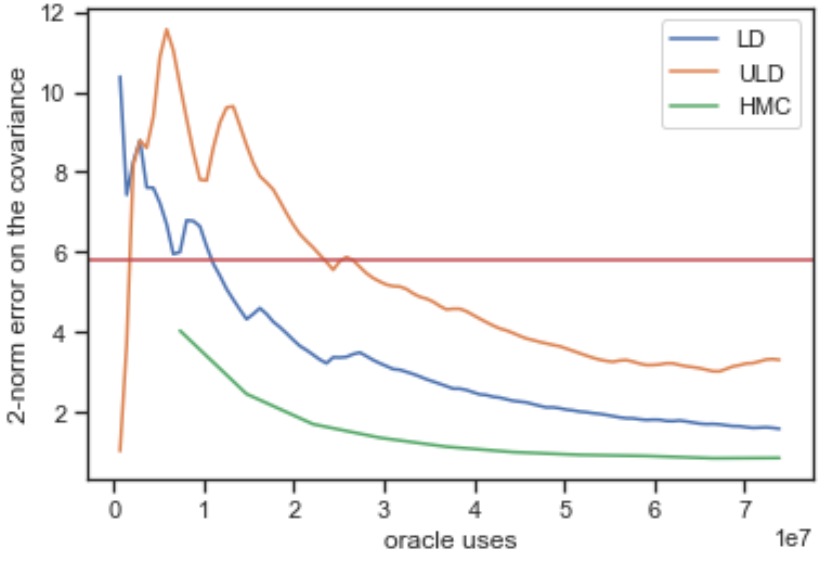
$$dv = -\gamma v dt + B \nabla_\theta \log \pi(\theta) dt + \mathcal{N}(0, 2\gamma(B + B^T))$$

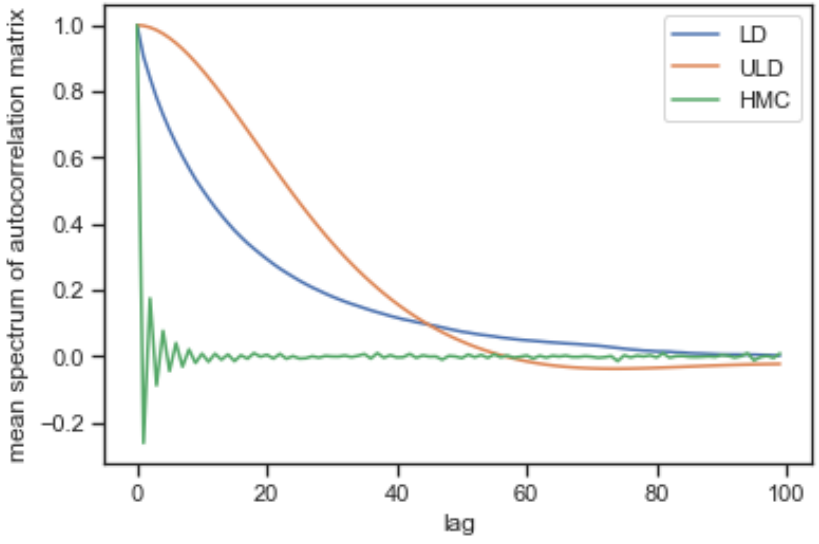
Underdamped Langevin dynamics(ULD)

- An order 2 discretization of this process is available that uses only the gradient and normal noise.
- it has been shown that ULD reaches an accuracy of ϵ in $\mathcal{O}(\sqrt{\frac{d}{\epsilon}} \log \frac{d}{\epsilon})$ iterations.
- we can recover LD from ULD by letting $\gamma \to \infty$
- At the other extreme, we can recover HMC by letting $\gamma \to 0$
- is this the best possible ?

Experiments







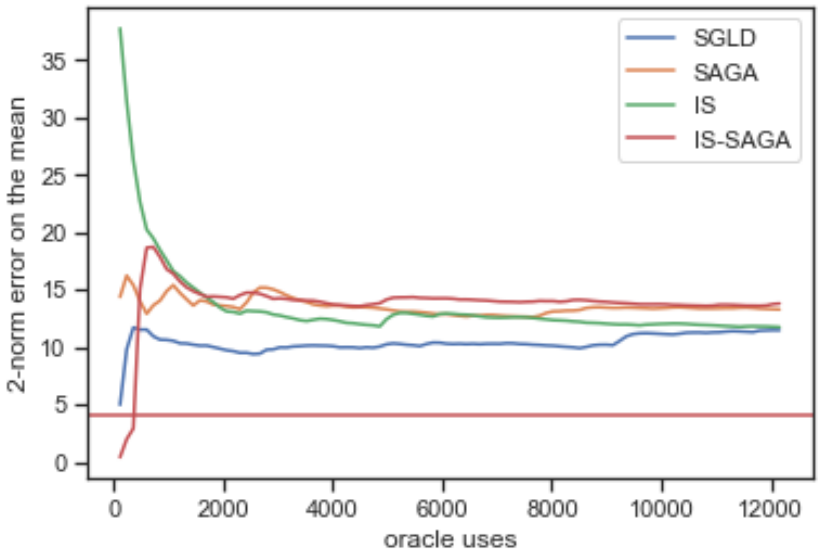
Stochastic gradient Langevin dynamics (SGLD)

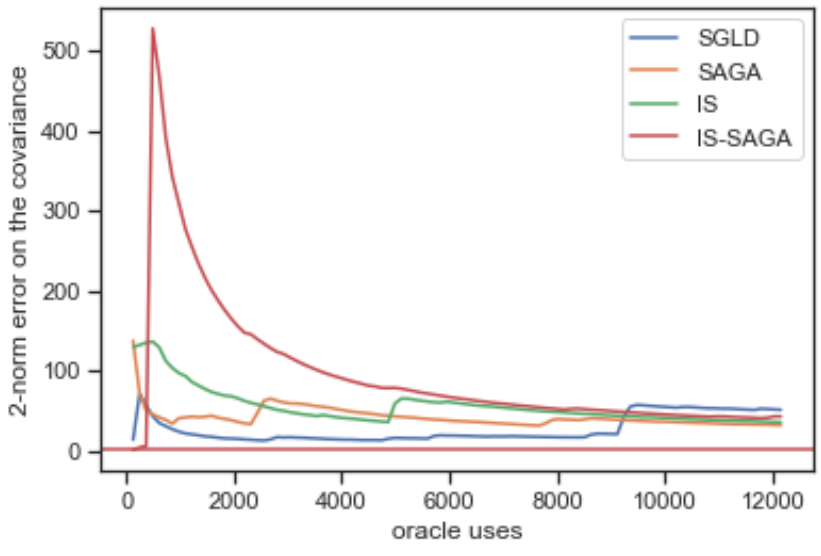
- We have seen that SGLD fails to converges to the target distribution and that it was due to the large noise in the gradient estimates.
- Assuming that SGLD has access to gradients with bounded variance $\sigma^2 d$, it was recently shown that SGLD requires $\mathcal{O}(\frac{\sigma^2}{\epsilon}\frac{d}{\epsilon}\log\frac{d}{\epsilon})$ iterations to get ϵ close to the target distribution, making it unpractical for large σ.

More connections with optimization

- In order to recover the fast linear rate of gradient descent (GD) for stochastic gradient descent (SGD), the idea of variance reduction was introduced in optimization.
- The most well known such variance reduction technique is SAGA.
- It was recently proven that SGLD with SAGA, usually called SAGA-LD, recovers the fast rate of $\mathcal{O}(\frac{d}{\epsilon}\log\frac{d}{\epsilon})$ of LD.
- In parallel, the idea of importance sampling was also used to prove sharper convergence rates in optimization.
- Currently the best known convergence rate relies on the use of SAGA, and an importance sampling scheme that assigns $\frac{L_i}{\sum_j L_j}$ probability to each sample i, where L_i is the Lipschitz constant of the i^{th} function.
- No analysis of SGLD with importance sampling is available to my knowledge.

Experiments





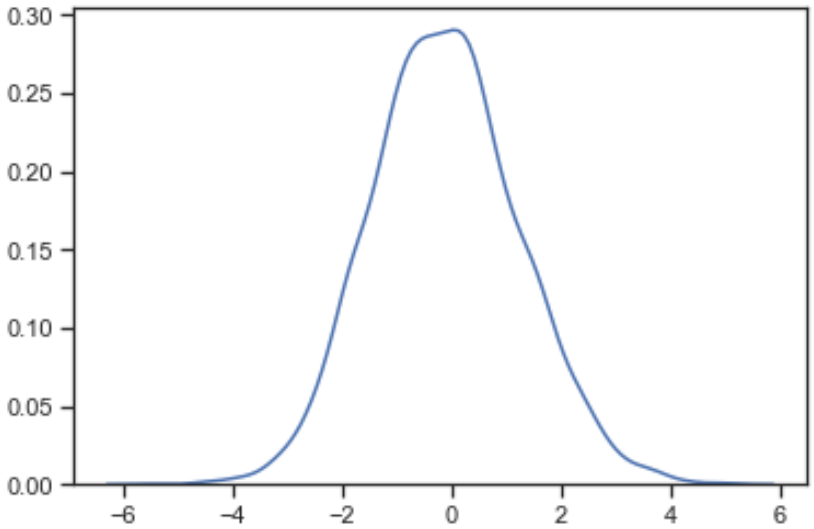
Proposed subsampling strategy

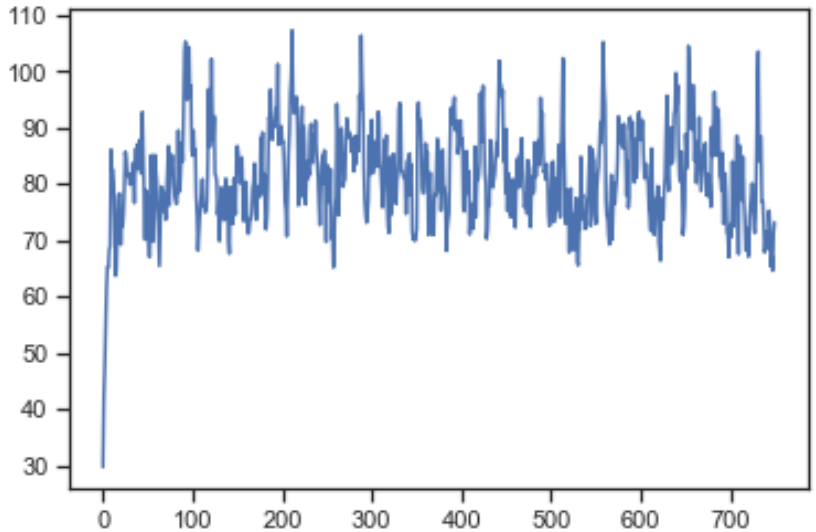
- We propose a simple new importance sampling strategy inspired by SAGA.
- Our approach is based on minimizing a simple upper bound on the variance of the gradient estimator (or log-posterior estimator).

Approximate Metropolis random walk

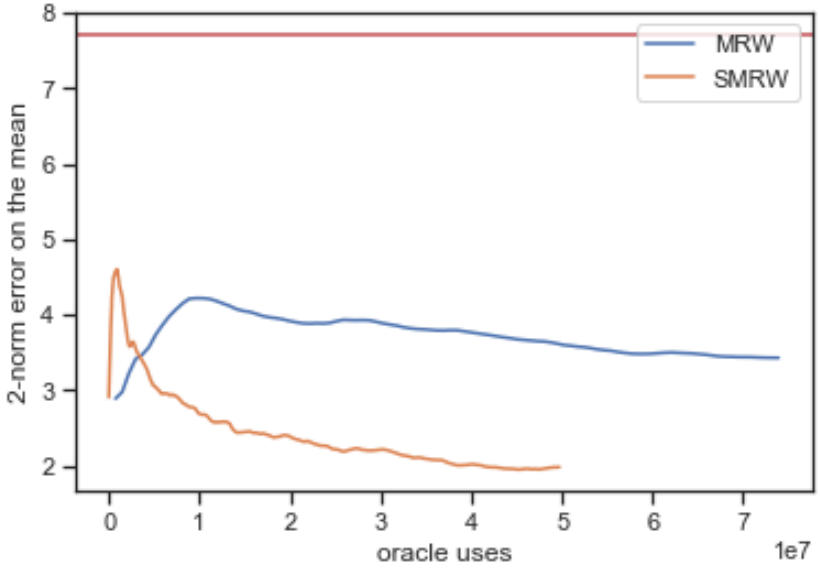
- We use our proposed subsampling strategy on the approximate Metropolis random walk algorithm.
- We also propose to replace the one-sample t-test with a two-sample t-test to allow the algorithm to revise its estimate if necessary.
- The resulting algorithm uses $\mathcal{O}(kS)$ oracle calls per iteration where k is the number of times the t-test is inconclusive and S is the batch size.

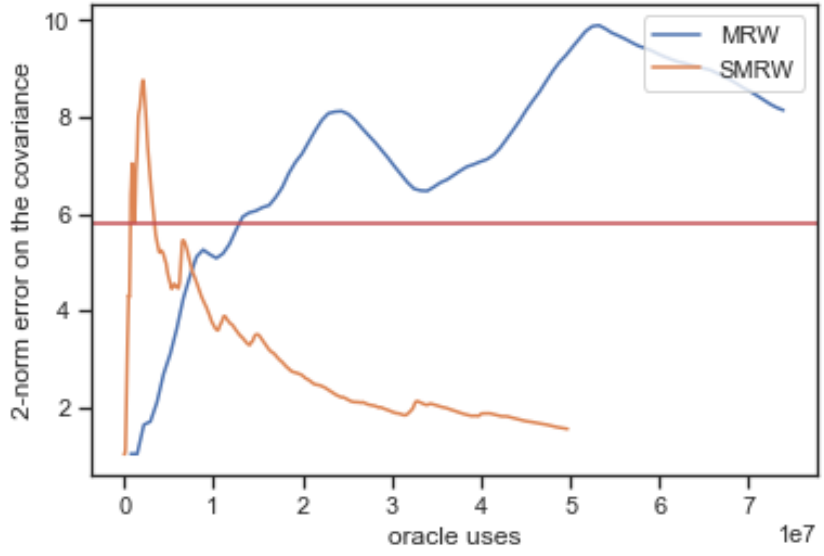
Approximate Metropolis random walk



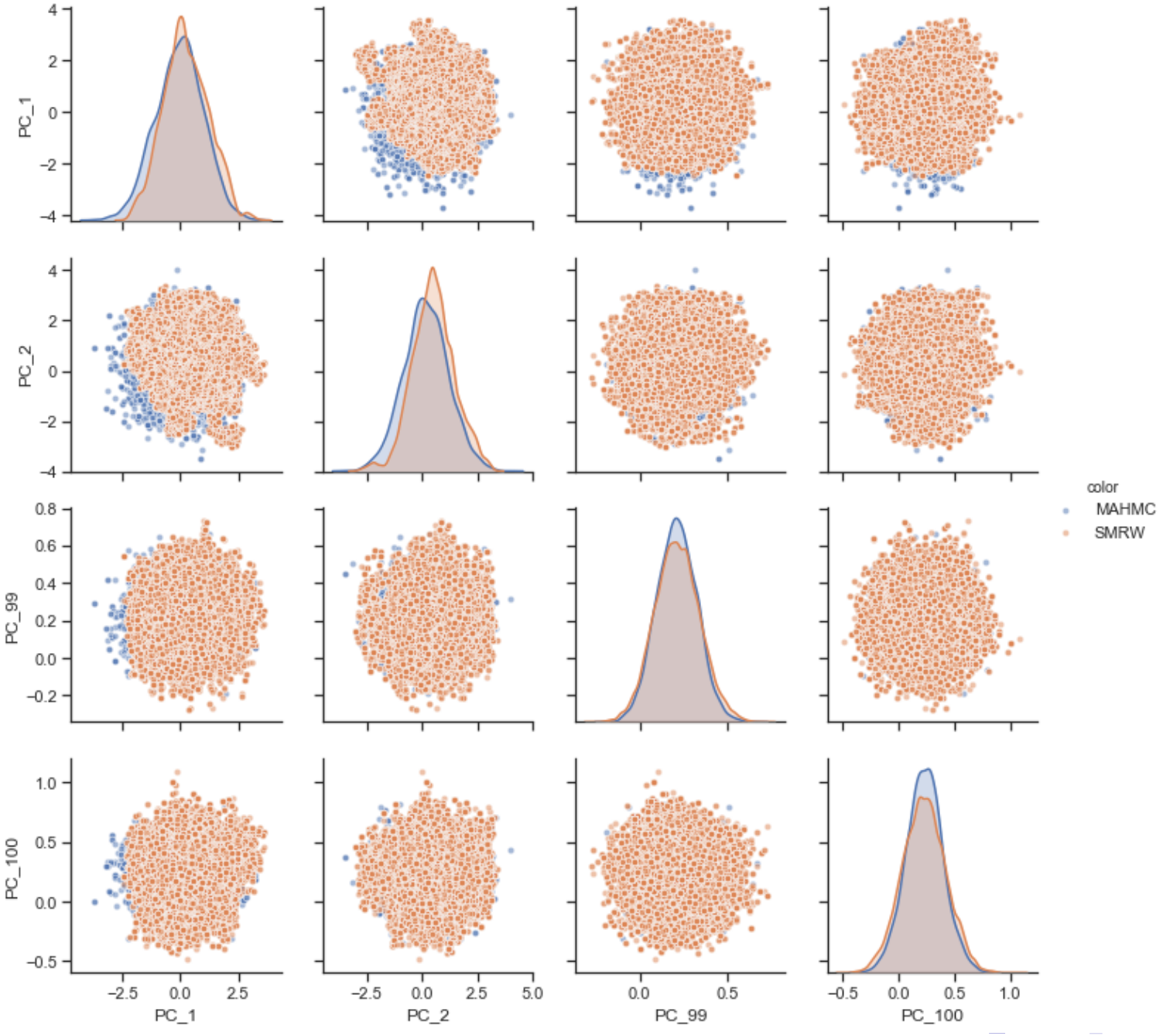


Approximate Metropolis random walk





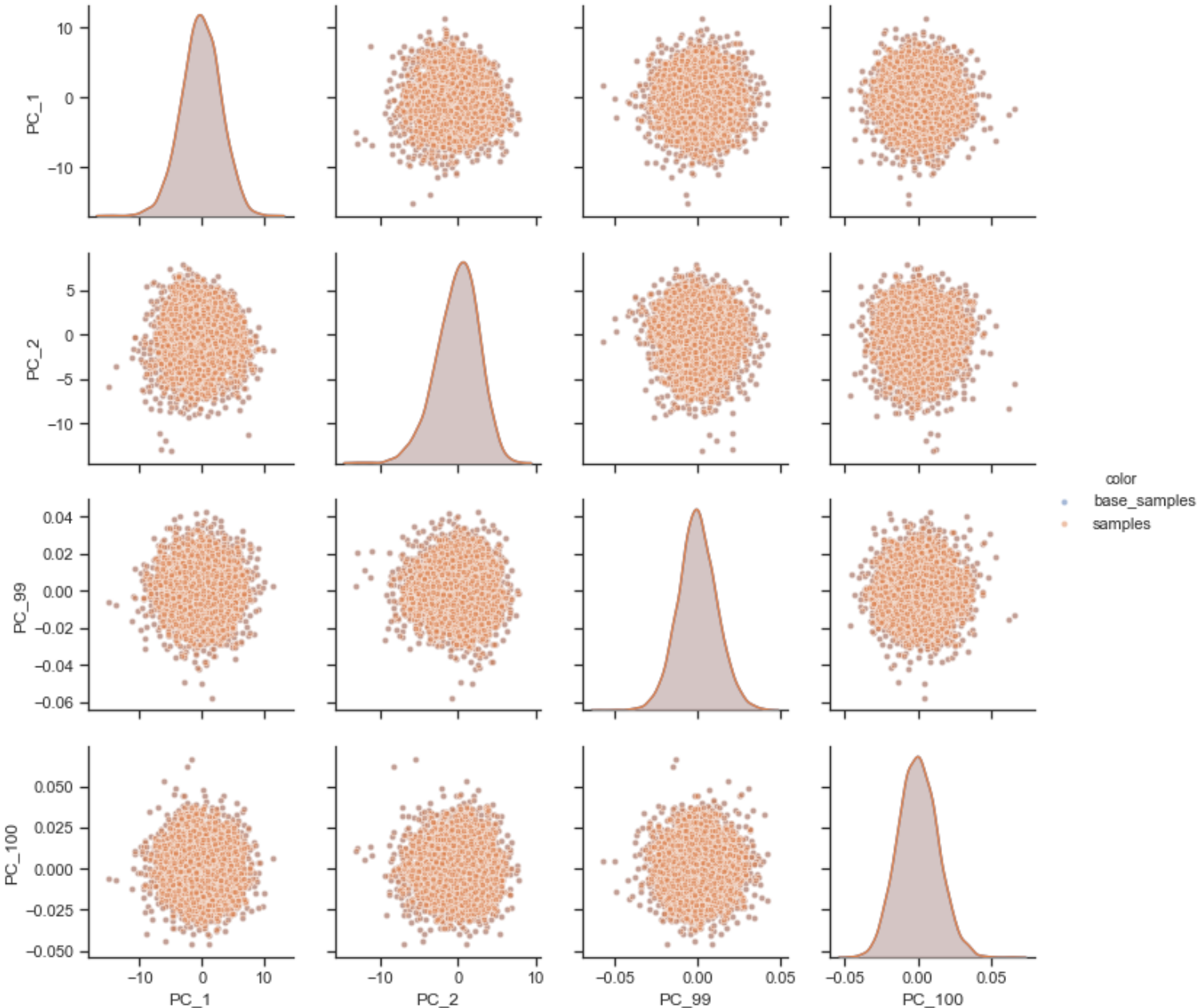
Experiments



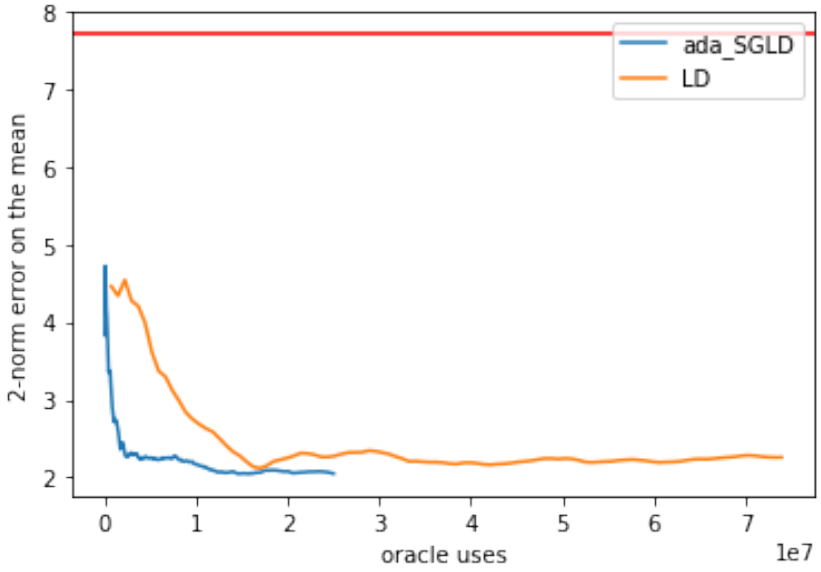
Stochastic Gradient Langevin Dynamics

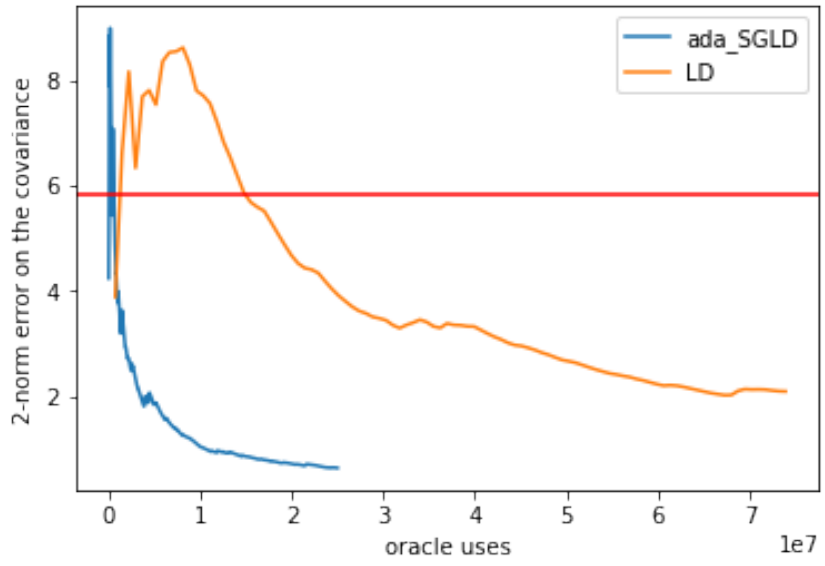
- We use our proposed subsampling strategy on SGLD.
- We adjust the added noise to account for the gradient noise exactly.
- The resulting algorithm uses $\mathcal{O}(kS)$ oracle calls to perform one iteration where k is the number of gradient evaluations needed to reduce the noise to an acceptable level, and S the batch size.

Stochastic Gradient Langevin Dynamics

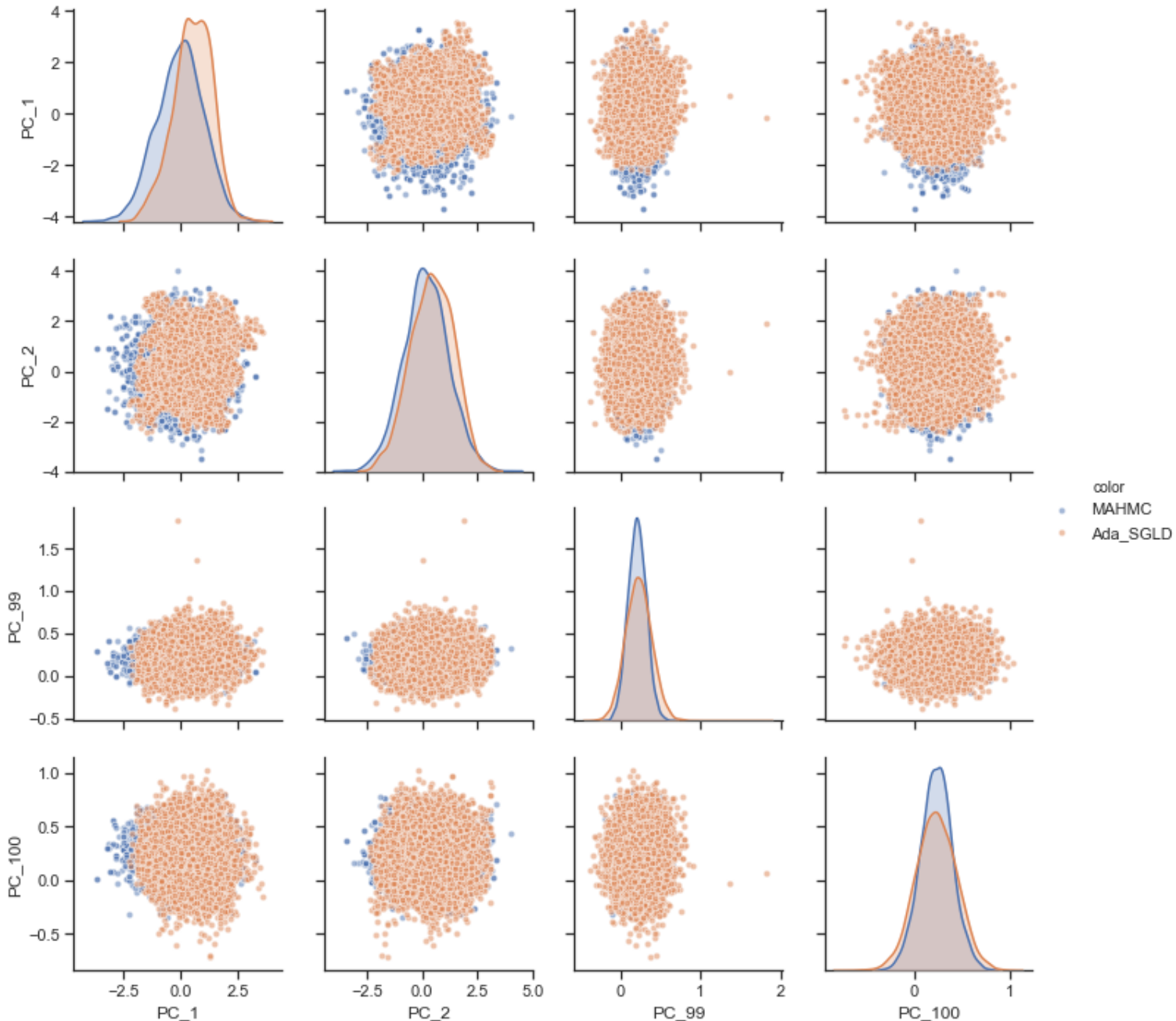


Stochastic Gradient Langevin Dynamics

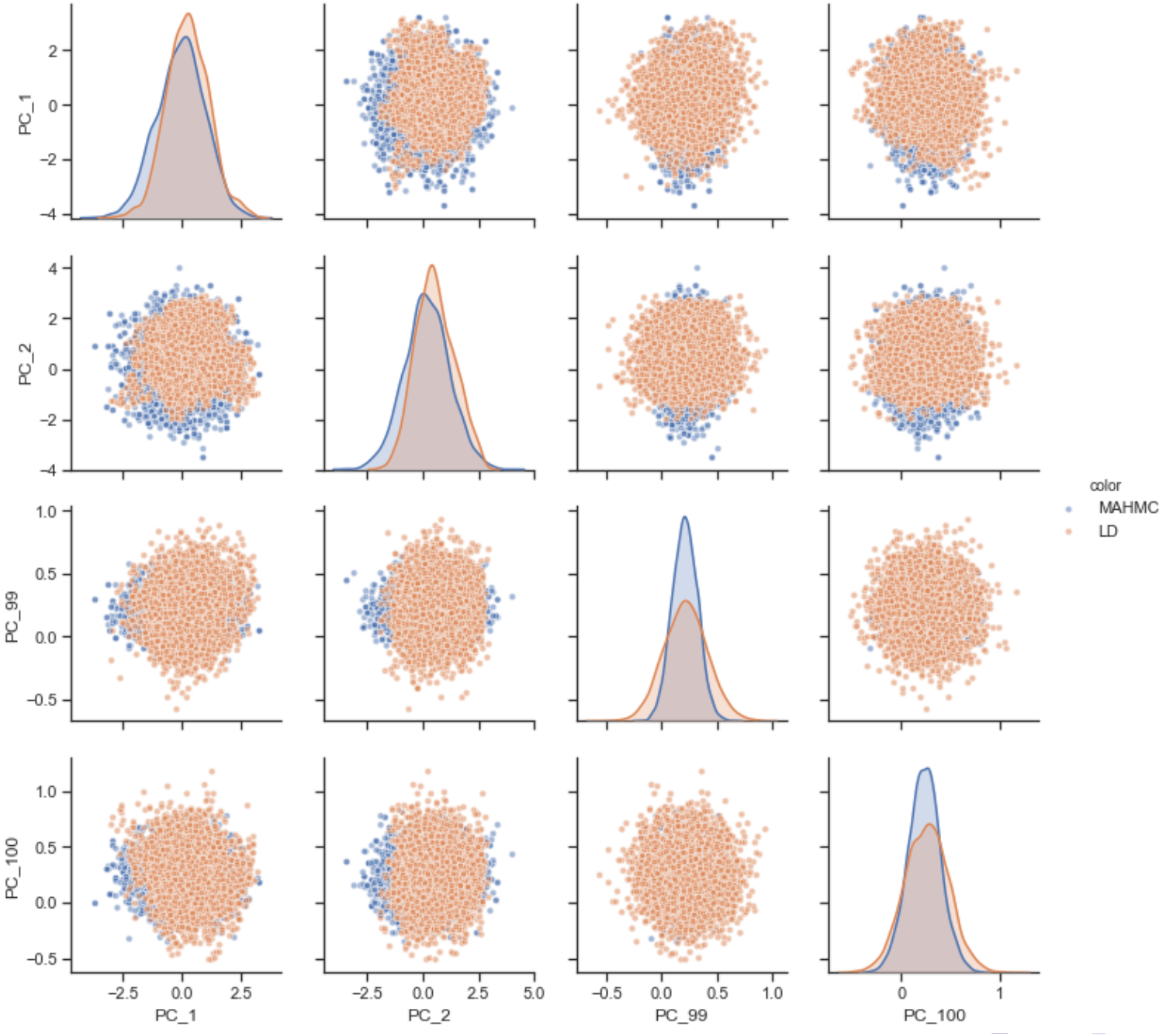




Experiments



Experiments



Future Work

- Extension to HMC and ULD.
- Analysis of convergence and guarantees.
- Optimal batch size.
- Optimal A(z) that minimizes auto-covariance.
- Higher order integrators with HMC.