Langevin Dynamics as Nonparametric Variational Inference

Matt Hoffman, Yian Ma
Google

Outline

- Nonparametric BBVI: The dream.
- Observation: The Fokker-Planck equation applies to nonparametric BBVI.
  - Therefore, Langevin dynamics ~ nonparametric BBVI.
- Provocative claim: If LD ~ NPBBVI, then we need to rethink statements like "MCMC is slower than VI".
  - More nuanced claim: MCMC wins on bias, VI wins on variance and amortization.
- Some plots and videos.

Nonparametric Normalizing-Flow VI: A Fantasy

Let g(ε) be an arbitrary invertible function*, and define $q_g(\theta)$ as

$$q_g(\theta) = q_0(g^{-1}(\theta)) \left| \frac{dg^{-1}}{d\theta} \right|.$$

Suppose we want to make $q_g(\theta) \approx p(\theta \mid x)$.

Let's pretend we can do *functional* gradient descent on KL(q||p):

$$\frac{d}{\delta g(\epsilon)} \int_\epsilon q_0(\epsilon) \log \frac{q_g(g(\epsilon))}{p(g(\epsilon) \mid x)} d\epsilon = \nabla_\theta \log q_g(g(\epsilon)) - \nabla_\theta \log p(g(\epsilon), x)$$

Call this *nonparametric BBVI*.

* a.k.a. "normalizing flow", "transport map", or "bijector"

Aside: Parametric BBVI

Since we can't do nonparametric BBVI, in practice we use parametric flows $g_\phi(\varepsilon)$.

Claim: Nonparametric BBVI > Parametric BBVI:

- g_ϕ can't represent arbitrary functions (asymptotic bias).
- We can only get stochastic estimates of the gradient (variance).
- The tangent field distorts the gradient-flow geometry (conditioning).

Nonparametric BBVI Obeys the Fokker-Planck Equation

Taking a functional gradient step of KL(q_g || p) w.r.t. g

$$g'(\epsilon) = g(\epsilon) + \eta(\nabla_\theta \log p(g(\epsilon), x) - \nabla_\theta \log q_g(g(\epsilon)))$$

yields a new variational distribution

$$q_{g'}(\theta) = q_0(g'^{-1}(\theta)) \left| \frac{dg'^{-1}}{d\theta} \right|.$$

It turns out that

$$\frac{\log q_{g'}(\theta) - \log q_g(\theta)}{\eta} = \nabla_\theta \log q_g(\theta)^\top (\nabla_\theta \log q_g(\theta) - \nabla_\theta \log p(\theta, x)) + \operatorname{tr}(\nabla_\theta^2 \log q_g(\theta) - \nabla_\theta^2 \log p(\theta, x)) + O(\eta),$$

which as $\eta \to 0$ is the *Fokker-Planck Equation*.

Langevin Dynamics

A classic MCMC algorithm:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta \log p(\theta_t, x) + \sqrt{2\eta}\xi_t; \quad \xi_t \sim \mathcal{N}(0, I)\,.$$

For a Markov chain defined by repeatedly applying the Langevin kernel $T(\theta' \mid \theta)$, initialized with samples from some q_0, define

$$q_t(\theta) \triangleq \int_{\theta_{0:t-1}} T(\theta_t \mid \theta_{t-1}) \cdots T(\theta_1 \mid \theta_0) q_0(\theta_0) d\theta_{0:t-1}.$$

To first order in η, q_t also obeys the Fokker-Planck equation!

Summary:
Langevin as Nonparametric BBVI

Nonparametric BBVI evolves q according to the Fokker-Planck equation.

So does Langevin.

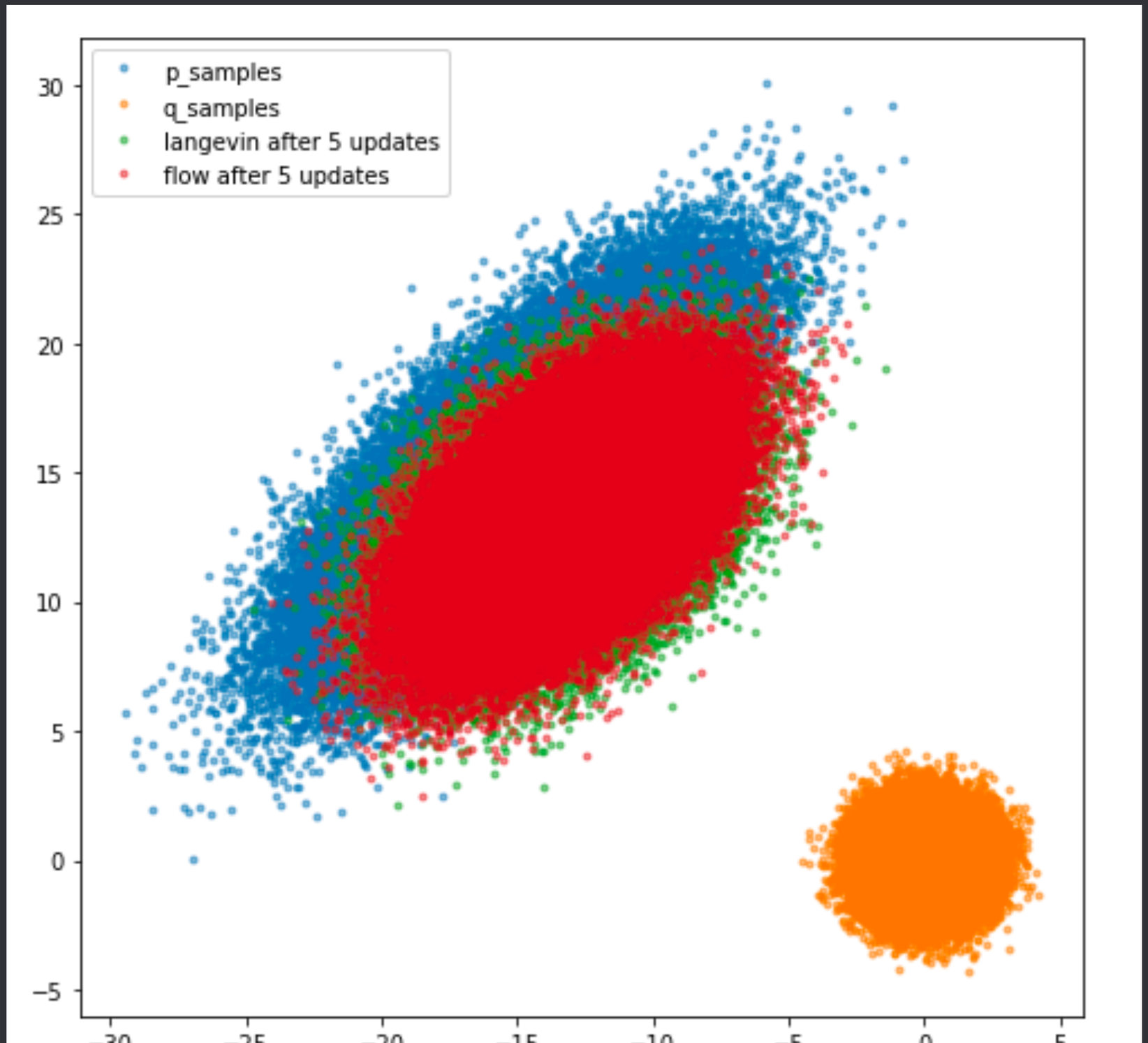
So we can interpret Langevin MCMC as implicitly doing nonparametric BBVI, then drawing a sample from the result.

Hot Take:

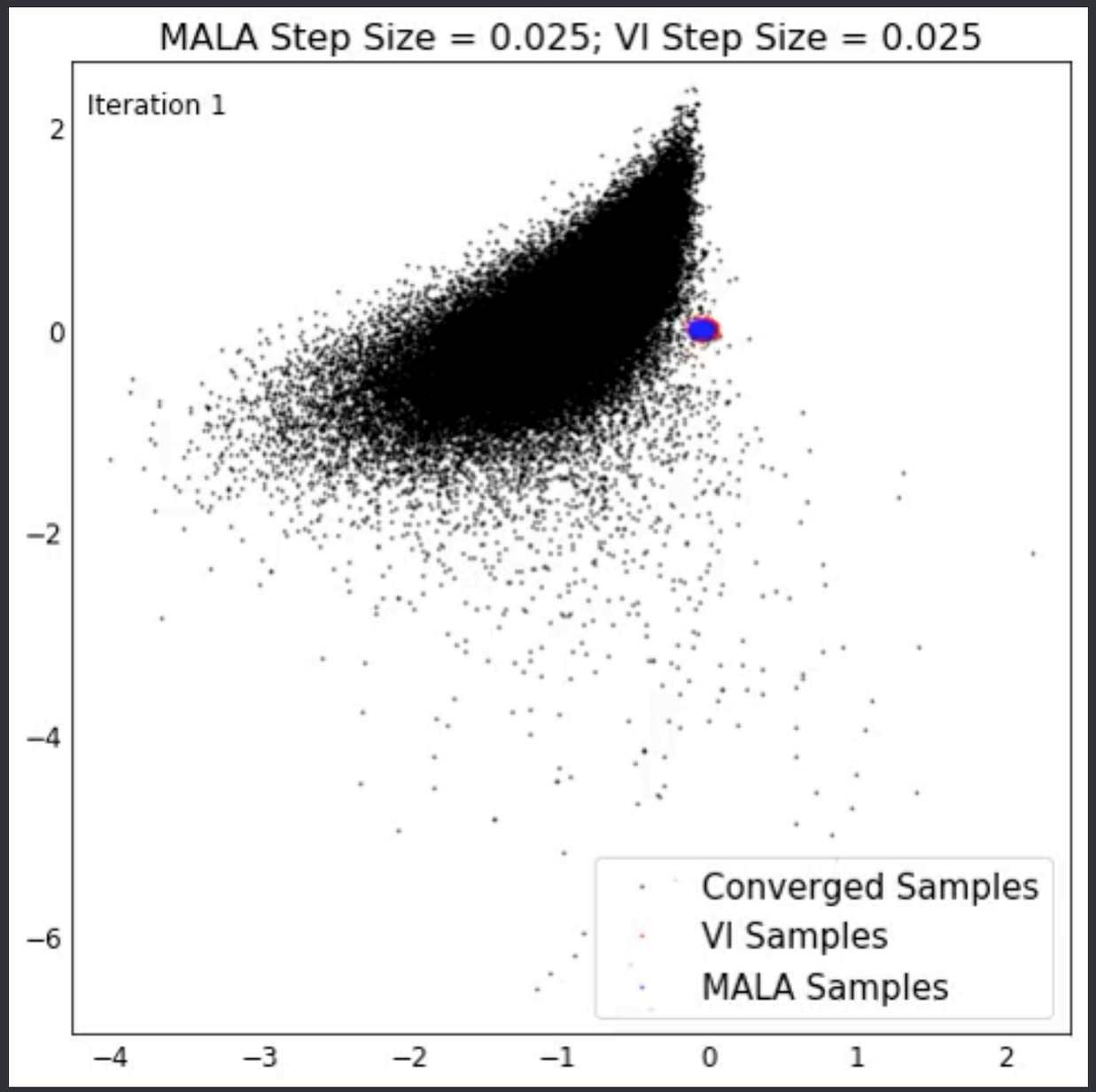
Langevin ≈ NPBBVI > BBVI

(More nuanced take coming in a few slides.)

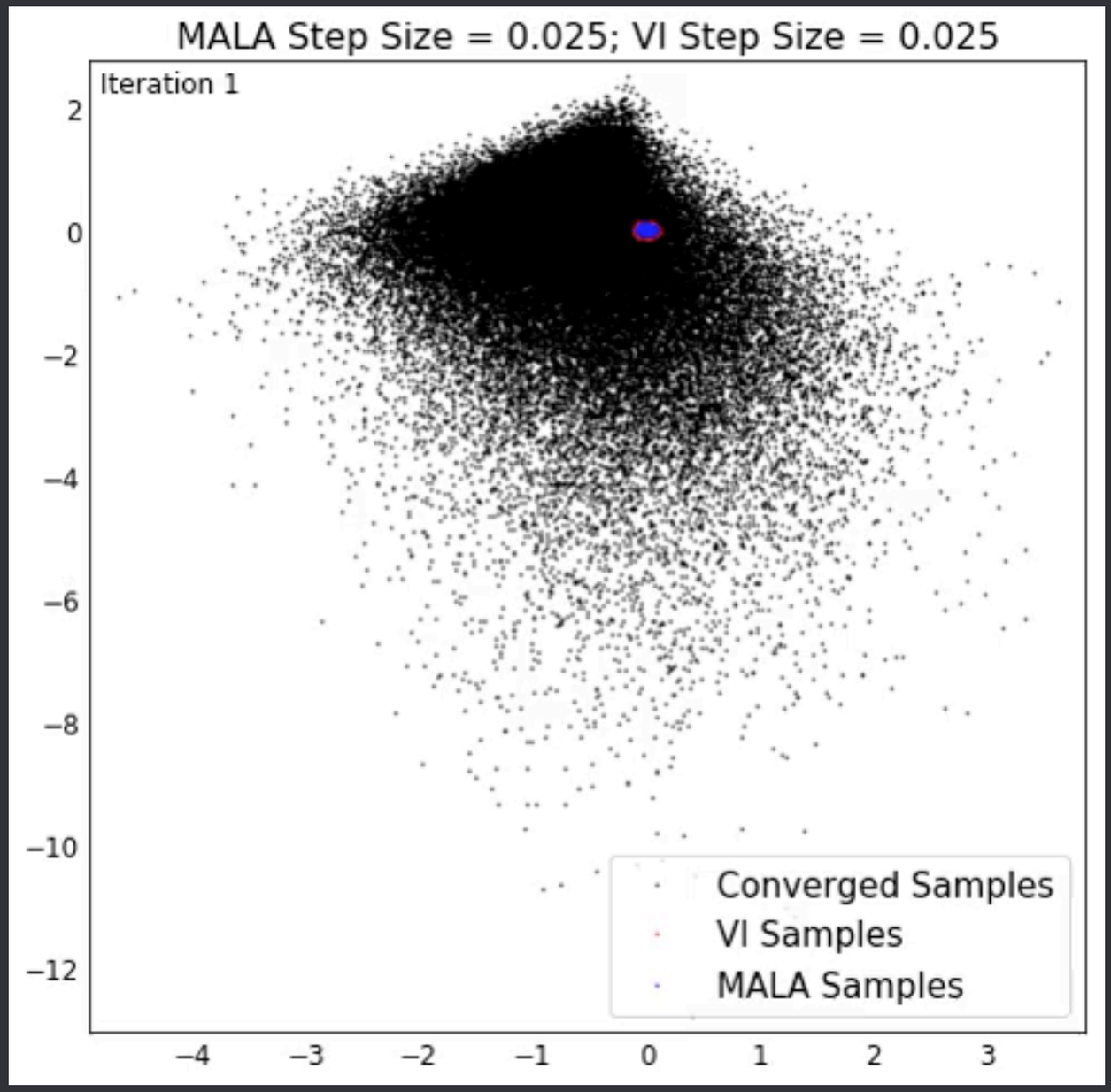
Gaussian Sanity Check



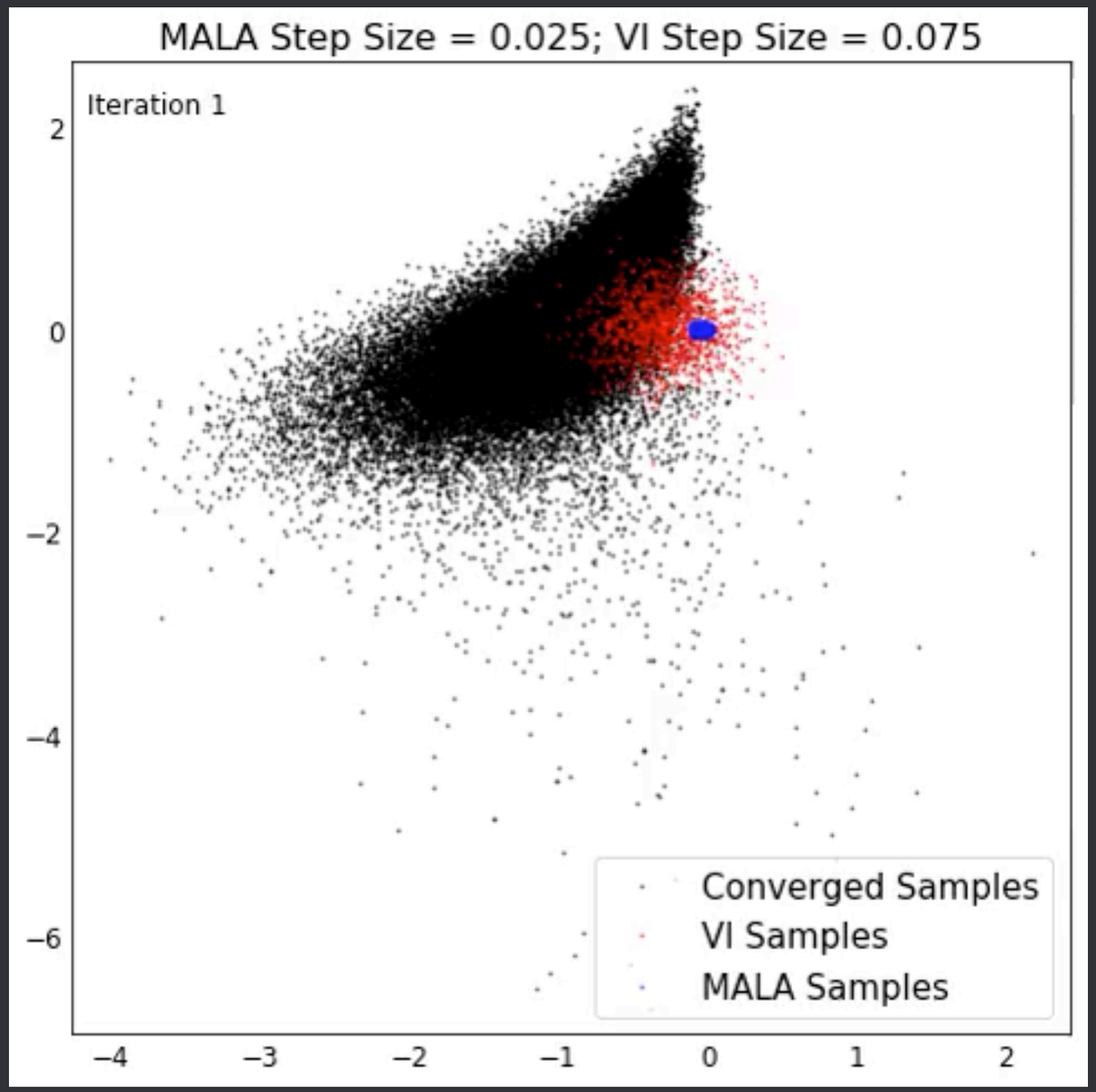
Unrealizable Setting



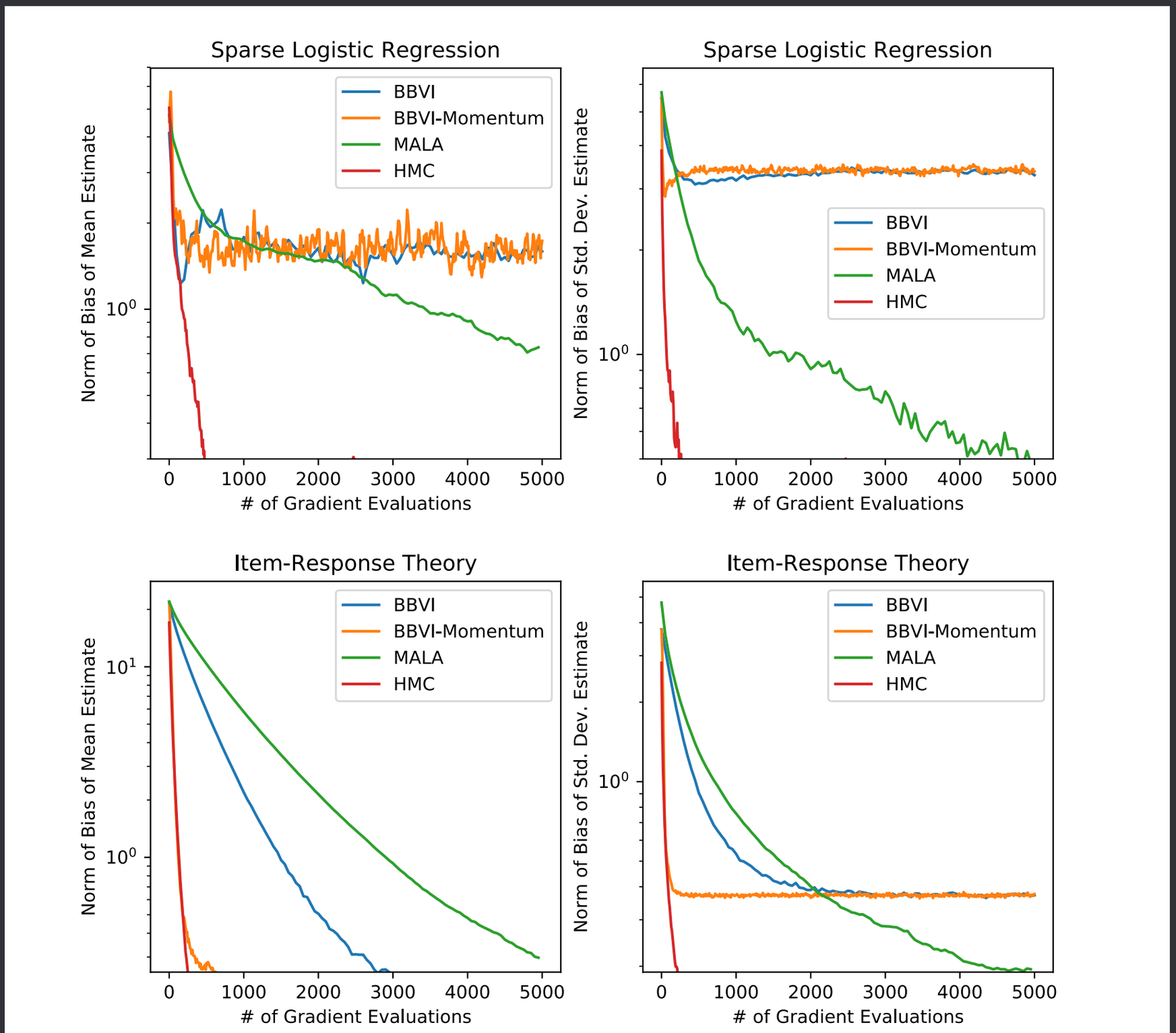
Unrealizable Setting



Larger Step Size for VI



Bias vs. Time

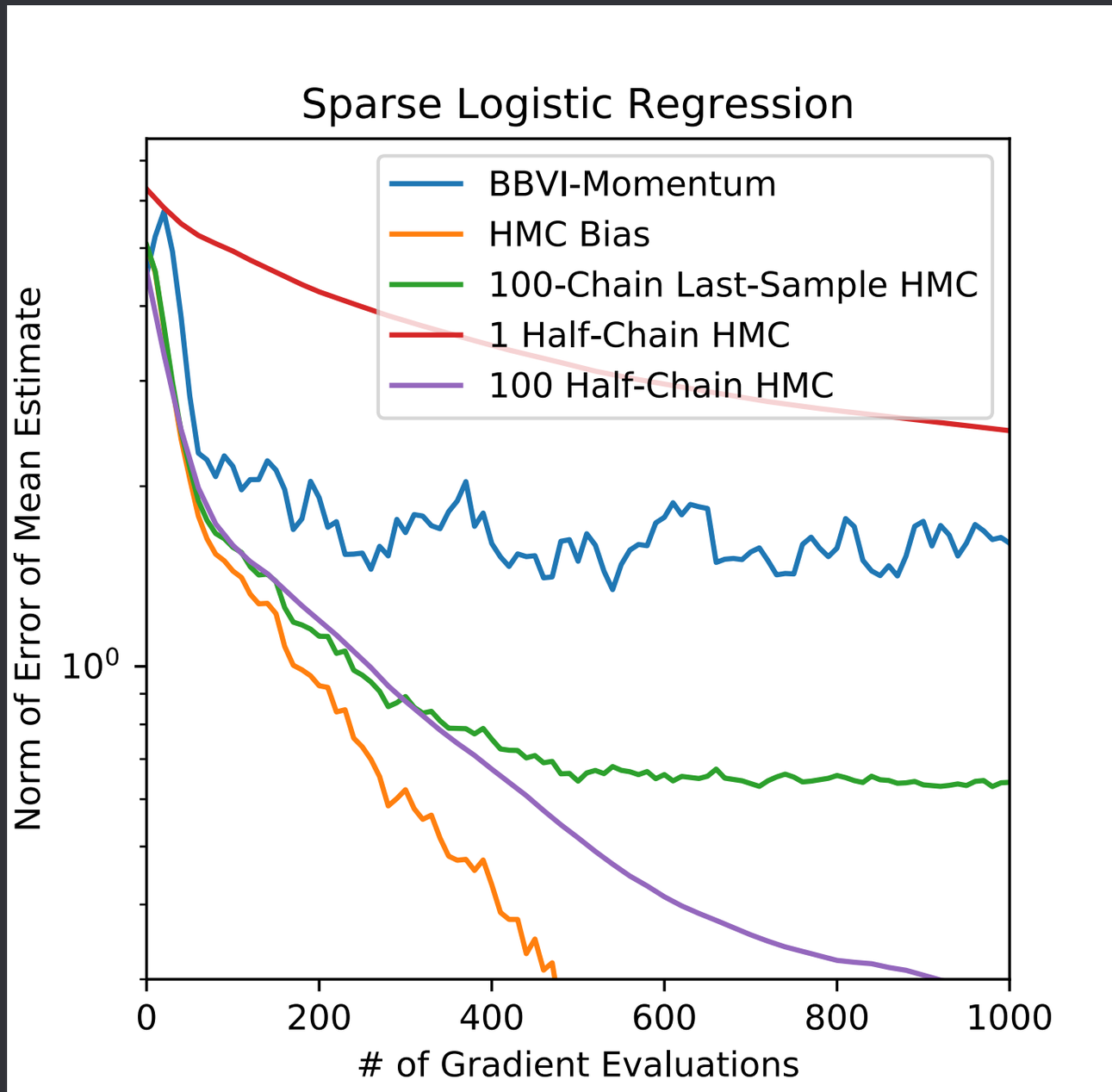

Sparse Logistic Regression
Norm of Bias of Mean Estimate
10^0
BBVI
BBVI-Momentum
MALA
HMC
0
1000
2000
3000
4000
5000
# of Gradient Evaluations
Sparse Logistic Regression
Norm of Bias of Std. Dev. Estimate
10^0
BBVI
BBVI-Momentum
MALA
HMC
0
1000
2000
3000
4000
5000
# of Gradient Evaluations
Item-Response Theory
Norm of Bias of Mean Estimate
10^1
10^0
BBVI
BBVI-Momentum
MALA
HMC
0
1000
2000
3000
4000
5000
# of Gradient Evaluations
Item-Response Theory
Norm of Bias of Std. Dev. Estimate
10^0
BBVI
BBVI-Momentum
MALA
HMC
0
1000
2000
3000
4000
5000
# of Gradient Evaluations

Bias and Variance

MCMC bias tends to go down faster than BBVI bias.

But MCMC *variance* dominates bias if we only run one chain.

But running many chains on GPUs is cheap!



Discussion

Gradient-based MCMC and VI algorithms aren't so different.

BBVI doesn't necessarily converge to a given level of bias faster than MCMC.

VI's main advantages over MCMC are:

- Amortization schemes.
- Low-variance estimates (samples from q are cheap).

But parallel hardware can cheaply reduce MCMC variance.

Ask Me About…

- Momentum
- Preconditioning (e.g., Adagrad, RMSProp, Adam)
- Stochastic Gradients (SVI, SGLD)
- Optimal Transport (JKO's 1998 variational interpretation of the Fokker-Planck equation)
- Cost geometry and Tangent Fields (e.g., Janowiak&Obermeyer, 2018)
- Unbiased MCMC (Jacobs et al., 2017)
- Sticking the Landing (Roeder et al., 2017)
- MCMC Variance Reduction (e.g., Pinto&Neal, 2001)
- SGD (e.g., Duvenaud et al., 2016; Mandt et al., 2017)
- Amortization gap (e.g., Krishnan et al., 2018; Cremer et al., 2018; Kim et al., 2018)