On Markov chain Monte Carlo for tall data

Rémi Bardenet[1]

[1]CNRS & CRIStAL, Univ. Lille, France

The Metropolis algorithm

$\textsc{Metropolis}\big(\pi(\theta),\ \theta_0,\ \Sigma,\ N_{\mathrm{iter}}\big)$

for $k \leftarrow 1$ **to** N_{iter}

  $\theta \leftarrow \theta_{k-1}$

  $\theta' \sim \mathcal{N}(.|\theta, \Sigma),\ u \sim \mathcal{U}_{(0,1)},$

  $\alpha = \frac{\pi(\theta')}{\pi(\theta)}$

  if $u < \alpha$

    $\theta_k \leftarrow \theta'$ ▷ *Accept*

  else $\theta_k \leftarrow \theta$ ▷ *Reject*

return $(\theta_k)_{k=1,\dots,N_{\mathrm{iter}}}$

- Under assumptions [13],

$$\sqrt{N_{\mathrm{iter}}}\left[\frac{1}{N_{\mathrm{iter}}}\sum_{k=0}^{N_{\mathrm{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \to \mathcal{N}(0, \sigma^2_{\mathrm{lim}}(h)),$$

The Metropolis algorithm

$\textsc{Metropolis}\big(\pi(\theta),\ \theta_0,\ \Sigma,\ N_{\text{iter}}\big)$

for $k \leftarrow 1$ **to** N_{iter}

  $\theta \leftarrow \theta_{k-1}$

  $\theta' \sim \mathcal{N}(.|\theta, \Sigma),\ u \sim \mathcal{U}_{(0,1)},$

  $\alpha = \frac{\pi(\theta')}{\pi(\theta)}$

  if $u < \alpha$

    $\theta_k \leftarrow \theta'$ ▷ *Accept*

  else $\theta_k \leftarrow \theta$ ▷ *Reject*

return $(\theta_k)_{k=1,\dots,N_{\text{iter}}}$

- Under assumptions [13],

$$\sqrt{N_{\text{iter}}}\left[\frac{1}{N_{\text{iter}}}\sum_{k=0}^{N_{\text{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \to \mathcal{N}(0, \sigma^2_{\lim}(h)),$$

The Metropolis algorithm

```
METROPOLIS(π(θ), θ_0, Σ, N_iter)
    for k ← 1 to N_iter
        θ ← θ_{k−1}
        θ' ∼ N(.|θ, Σ), u ∼ U_(0,1),
        α = π(θ') / π(θ)
        if u < α
            θ_k ← θ'          ▷ Accept
        else θ_k ← θ          ▷ Reject
    return (θ_k)_{k=1,...,N_iter}
```

- Under assumptions [13],

$$\sqrt{N_{\text{iter}}}\left[\frac{1}{N_{\text{iter}}}\sum_{k=0}^{N_{\text{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \to \mathcal{N}(0, \sigma^2_{\text{lim}}(h)),$$

The Metropolis algorithm

$\textsc{Metropolis}\big(\pi(\theta),\ \theta_0,\ \Sigma,\ N_{\text{iter}}\big)$

- **for** $k \leftarrow 1$ **to** N_{iter}
  - $\theta \leftarrow \theta_{k-1}$
  - $\theta' \sim \mathcal{N}(.|\theta,\Sigma),\ u \sim \mathcal{U}_{(0,1)},$
  - $\alpha = \frac{\pi(\theta')}{\pi(\theta)}$
  - **if** $u < \alpha$
    - $\theta_k \leftarrow \theta'$ $\triangleright$ *Accept*
  - **else** $\theta_k \leftarrow \theta$ $\triangleright$ *Reject*
- **return** $(\theta_k)_{k=1,\dots,N_{\text{iter}}}$

- Under assumptions [13],

$$\sqrt{N_{\text{iter}}}\left[\frac{1}{N_{\text{iter}}}\sum_{k=0}^{N_{\text{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \to \mathcal{N}(0,\sigma^2_{\text{lim}}(h)),$$

Bayesian statistics in one slide

- Statisticians recommend actions.
- When you have a joint model $p(\theta, x_1, \dots, x_n)$ on
  - the state θ of the world
  - some observable data $x_1, \dots, x_n$,

decision theory and a few axioms [24, 22] lead to picking

$$a = \arg\max \int L(\theta, a) p(\theta | x_1, \dots, x_n) d\theta.$$

Common situation

I have

- a $p(\theta)$ that summarizes my beliefs on θ prior to an experiment,
- measurements $x_1, \dots, x_n$ assumed to be i.i.d. from $p(\cdot|\theta)$.

Then, I have fixed

$$p(\theta | x_1, \dots, x_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i | \theta)$$

Bayesian statistics in one slide

- Statisticians recommend actions.
- When you have a joint model $p(\theta, x_1, \dots, x_n)$ on
  - the state θ of the world
  - some observable data $x_1, \dots, x_n$,

decision theory and a few axioms [24, 22] lead to picking

$$a = \arg\max \int L(\theta, a) p(\theta | x_1, \dots, x_n) d\theta.$$

Common situation

I have

- a $p(\theta)$ that summarizes my beliefs on θ prior to an experiment,
- measurements $x_1, \dots, x_n$ assumed to be i.i.d. from $p(\cdot|\theta)$.

Then, I have fixed

$$p(\theta | x_1, \dots, x_n) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

The Metropolis algorithm in Bayesian statistics

$\textsc{Metropolis}\big(\pi(\theta),\ \theta_0,\ \Sigma,\ N_{\text{iter}}\big)$

for $k \leftarrow 1$ **to** N_{iter}

$\quad\theta \leftarrow \theta_{k-1}$

$\quad\theta' \sim \mathcal{N}(.|\theta,\Sigma), u \sim \mathcal{U}_{(0,1)},$

$\quad\alpha = \frac{\pi(\theta')}{\pi(\theta)}$

$\quad$**if** $u < \alpha$

$\quad\quad\theta_k \leftarrow \theta'$ $\quad\triangleright$ *Accept*

$\quad$**else** $\theta_k \leftarrow \theta$ $\quad\triangleright$ *Reject*

return $(\theta_k)_{k=1,\dots,N_{\mathrm{iter}}}$

- Under assumptions [13],

$$\sqrt{N_{\text{iter}}}\left[\frac{1}{N_{\text{iter}}}\sum_{k=0}^{N_{\text{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \rightarrow \mathcal{N}(0,\sigma^2_{\text{lim}}(h)),$$

$\textsc{Metropolis}\big(\pi(\theta),\ \theta_0,\ \Sigma,\ N_{\text{iter}}\big)$

for $k \leftarrow 1$ **to** N_{iter}

$\quad \theta \leftarrow \theta_{k-1}$

$\quad \theta' \sim \mathcal{N}(.|\theta, \Sigma), u \sim \mathcal{U}_{(0,1)},$

$\quad \alpha = \frac{\prod_{i=1}^{n} p(x_i|\theta')p(\theta')}{\prod_{i=1}^{n} p(x_i|\theta)p(\theta)}$

$\quad$ **if** $u < \alpha$

$\quad\quad \theta_k \leftarrow \theta'$ $\quad \triangleright$ *Accept*

$\quad$ **else** $\theta_k \leftarrow \theta$ $\quad \triangleright$ *Reject*

return $(\theta_k)_{k=1,\dots,N_{\text{iter}}}$

- Under assumptions [13],

$$\sqrt{N_{\text{iter}}}\left[\frac{1}{N_{\text{iter}}}\sum_{k=0}^{N_{\text{iter}}} h(\theta_k) - \int h(\theta)\,\pi(\theta)d\theta\right] \to \mathcal{N}(0, \sigma^2_{\text{lim}}(h)),$$

Divide-and-conquer approaches

Principle

Divide the data into batches, run MCMC on each batch and combine the results...

- by multiplying smooth approximations to batch posteriors [16, 26, 21].
  - asymptotically justified,
  - but the MSE of resulting estimators scales exponentially with the number of batches, even under strong simplifying assumptions [21].
- targeting a more tractable result than the full posterior [20, 28].
  - more stable,
  - but the statistical meaning of the result is unclear.
- Other techniques [15, 31], with the same advantages and drawbacks.

Subsampling approaches

$\textsc{Metropolis}(\pi(\theta),\ \theta_0,\ N_{\text{iter}})$

for $k \leftarrow 1$ **to** N_{iter}
    $\theta \leftarrow \theta_{k-1}$
    $\theta' \sim \mathcal{N}(.|\theta, \Sigma),\ u \sim \mathcal{U}_{(0,1)},$
    $\alpha = \frac{\prod_{i=1}^{n} p(x_i|\theta')p(\theta')}{\prod_{i=1}^{n} p(x_i|\theta)p(\theta)}$
    if $u < \alpha$
        $\theta_k \leftarrow \theta'$ $\triangleright$ *Accept*
    else $\theta_k \leftarrow \theta$ $\triangleright$ *Reject*
return $(\theta_k)_{k=1,\ldots,N_{\mathrm{iter}}}$

- Can we use

$$\Lambda_t^*(\theta, \theta') = \frac{1}{t}\sum_{i=1}^{t} \log\left[\frac{p(x_i^*|\theta')}{p(x_i^*|\theta)}\right]?$$

Subsampling approaches

$\text{METROPOLIS}\big(\pi(\theta),\ \theta_0,\ N_{\text{iter}}\big)$

for $k \leftarrow 1$ **to** N_{iter}

$\quad \theta \leftarrow \theta_{k-1}$

$\quad \theta' \sim \mathcal{N}(.|\theta, \Sigma),\ u \sim \mathcal{U}_{(0,1)},$

$\quad \psi(u, \theta, \theta') \leftarrow \frac{1}{n} \log \left[u \frac{p(\theta)}{p(\theta')} \right]$

$\quad \Lambda_n(\theta, \theta') \leftarrow \frac{1}{n} \sum_{i=1}^{n} \log \left[\frac{p(x_i|\theta')}{p(x_i|\theta)} \right]$

$\quad$ **if** $\Lambda_n(\theta, \theta') > \psi(u, \theta, \theta')$

$\quad\quad \theta_k \leftarrow \theta'$ $\quad \triangleright$ *Accept*

$\quad$ **else** $\theta_k \leftarrow \theta$ $\quad \triangleright$ *Reject*

return $(\theta_k)_{k=1,\dots,N_{\text{iter}}}$

- Can we use

$$\Lambda_t^*(\theta, \theta') = \frac{1}{t} \sum_{i=1}^{t} \log \left[\frac{p(x_i^*|\theta')}{p(x_i^*|\theta)} \right] ?$$

A T-test within Metropolis?

- Metropolis is based on checking whether

$$\Lambda_n(\theta, \theta') = \frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{p(x_i|\theta')}{p(x_i|\theta)}\right] > \psi(u, \theta, \theta').$$

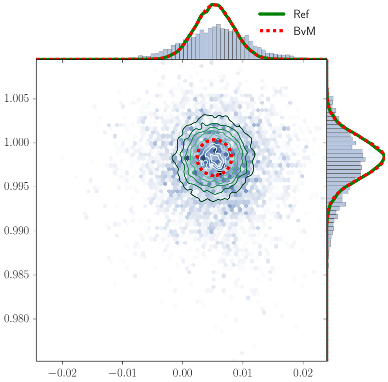
- From 1988 [10] to 2013 [17], various similar propositions using T-tests to check whether

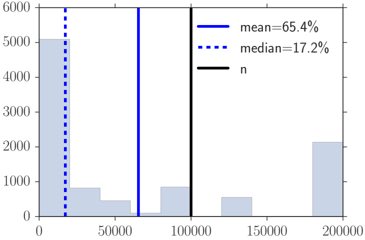
$$\Lambda_n(\theta, \theta') = \psi(u, \theta, \theta').$$

- Austerity MH [17] provides useful heuristics for machine learning tasks.
- But for MCMC integration: hard to tune and no guarantee!

Austerity MH on a toy example

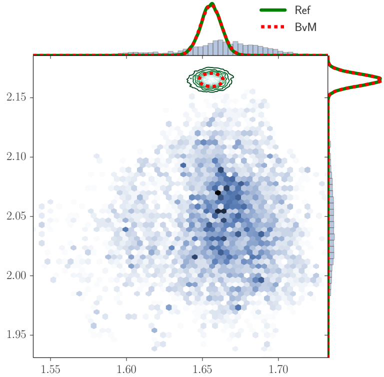
- $\mathcal{X} \sim \mathcal{N}(0,1)$,
- $p(\cdot|\theta) = \mathcal{N}(\cdot|\mu,\sigma^2)$.

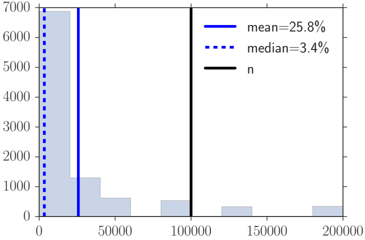




Austerity MH on a toy example

- $\mathcal{X} \sim \text{LogNormal}(0, 1)$,
- $p(\cdot|\theta) = \mathcal{N}(\cdot|\mu, \sigma^2)$.





- Let $\delta > 0$, $\theta, \theta' \in \Theta$. We can find $\big(t, c_t(\delta)\big)$ such that

$$\mathbb{P}(|\Lambda_t^*(\theta,\theta') - \Lambda_n(\theta,\theta')| \leq c_t(\delta)) \geq 1-\delta.$$

- For example, sampling without replacement, we prove [4]

$$c_t(\delta) = \cdots \times \sqrt{1-t/n}\frac{\hat{\sigma}_t}{\sqrt{t}} + \cdots \times \frac{C_{\theta,\theta'}}{t}.$$

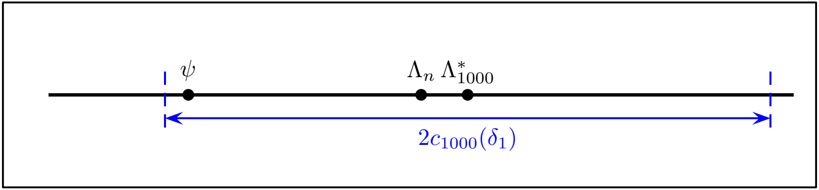
  is valid, where $C_{\theta,\theta'} = \max_{1\leq i\leq n}|\log p(x_i|\theta') - \log p(x_i|\theta)|$.
- Assume you can compute $C_{\theta,\theta'}$ in $o(n)$ time.
- Can we make the right decision with probability $1-\delta$?

An adaptive choice of t

- Given $\theta, \theta' \in \Theta$ and $u \in [0,1]$, an adaptive choice of t can guarantee we know whether

$$\Lambda_n(\theta, \theta) > \psi(u, \theta, \theta')$$

with probability $1 - \delta$.



- Taking (δ_t) such that

$$\sum_{t \geq 1} \delta_t \leq \delta$$

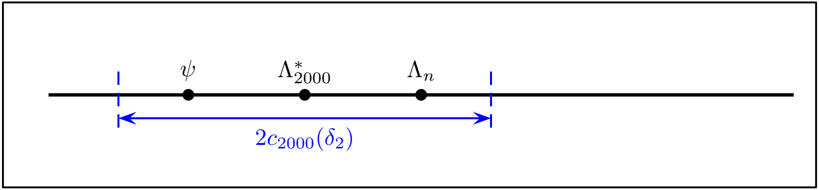
gives the result by a union bound.

An adaptive choice of t

- Given $\theta, \theta' \in \Theta$ and $u \in [0,1]$, an adaptive choice of t can guarantee we know whether

$$\Lambda_n(\theta,\theta) > \psi(u,\theta,\theta')$$

with probability $1-\delta$.



- Taking (δ_t) such that

$$\sum_{t\geq 1} \delta_t \leq \delta$$

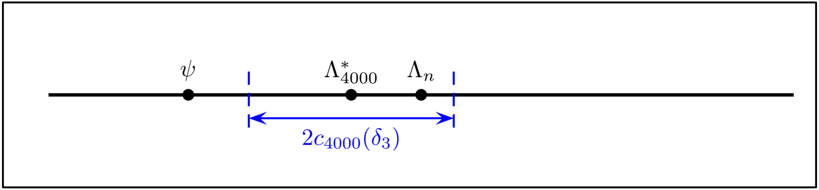
gives the result by a union bound.

An adaptive choice of t

- Given $\theta, \theta' \in \Theta$ and $u \in [0,1]$, an adaptive choice of t can guarantee we know whether

$$\Lambda_n(\theta, \theta) > \psi(u, \theta, \theta')$$

with probability $1 - \delta$.



- Taking (δ_t) such that

$$\sum_{t \geq 1} \delta_t \leq \delta$$

gives the result by a union bound.

A control in total variation

Theorem [3]

Let $P, \tilde{P}$ be the ideal MH kernel and our approximate kernel, respectively. Assume there exists $m, A < \infty$ such that

$$\forall \theta \in \Theta, \forall k > 0, \; \|P^k(\theta, \cdot) - \pi\|_{\mathrm{TV}} \leq A\rho^{\lfloor k/m \rfloor}. \tag{1}$$

Then there exists $B < \infty$ and a probability distribution $\tilde{\pi}$ on $(\Theta, \mathcal{B}(\Theta))$ such that

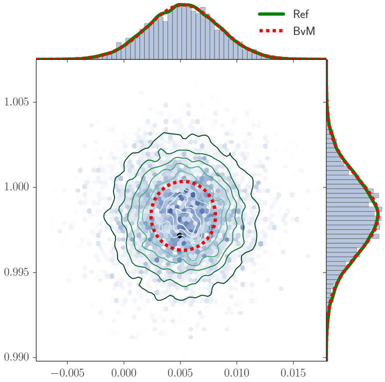
$$\forall \theta \in \Theta, \forall k > 0, \; \|\tilde{P}^k(\theta, \cdot) - \tilde{\pi}\|_{\mathrm{TV}} \leq B[1 - (1-\delta)^m (1-\rho)]^{\lfloor k/m \rfloor} \tag{2}$$

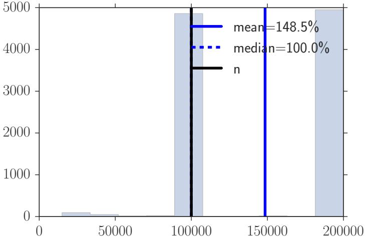
and $\tilde{\pi}$ satisfies

$$\|\pi - \tilde{\pi}\|_{\mathrm{TV}} \leq \frac{Am\delta}{1-\rho}. \tag{3}$$

- $\tilde{P}$ inherits its ergodicity from P.
- Geometric ergodicity is also preserved [25].

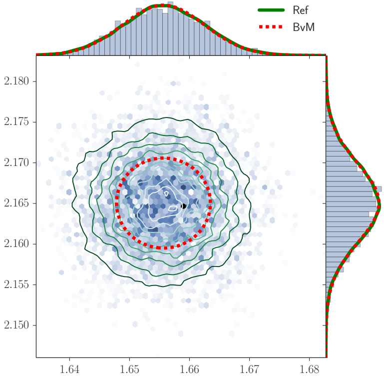
- $\mathcal{X} \sim \mathcal{N}(0, 1)$,
- $p(\cdot|\theta) = \mathcal{N}(\cdot|\mu, \sigma^2)$.

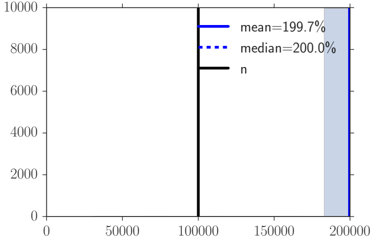




Confidence MH on a toy example

- $\mathcal{X} \sim \text{LogNormal}(0, 1)$,
- $p(\cdot|\theta) = \mathcal{N}(\cdot|\mu, \sigma^2)$.





Introducing proxies as control variates

Assume you have

$$\wp_i(\theta, \theta') \approx \log p(x_i|\theta') - \log p(x_i|\theta),$$

then the Metropolis acceptance decision is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{p(x_i|\theta')}{p(x_i|\theta)} - \wp_i(\theta,\theta')\right] > \psi(u,\theta,\theta') - \frac{1}{n}\sum_{i=1}^{n}\wp_i(\theta,\theta'),$$

and the leading term of Bernstein's bound now uses the std of

$$\left\{\log\frac{p(x_i^*|\theta')}{p(x_i^*|\theta)} - \wp_i(\theta,\theta'), i = 1,\dots,t\right\}.$$

Claim

If e.g. Taylor expansions can be used as $\wp_i(\theta,\theta')$, then the leading term of $c_t(\delta)$ can be $o(n^{-1})$.

Introducing proxies as control variates

Assume you have

$$\wp_i(\theta, \theta') \approx \log p(x_i|\theta') - \log p(x_i|\theta),$$

then the Metropolis acceptance decision is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{p(x_i|\theta')}{p(x_i|\theta)} - \wp_i(\theta, \theta')\right] > \psi(u, \theta, \theta') - \frac{1}{n}\sum_{i=1}^{n}\wp_i(\theta, \theta'),$$

and the leading term of Bernstein's bound now uses the std of

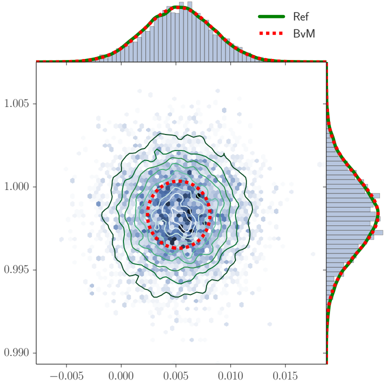
$$\left\{\log\frac{p(x_i^*|\theta')}{p(x_i^*|\theta)} - \wp_i(\theta, \theta'), i = 1, \dots, t\right\}.$$

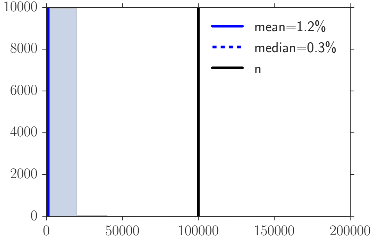
Claim

If e.g. Taylor expansions can be used as $\wp_i(\theta, \theta')$, then the leading term of $c_t(\delta)$ can be $o(n^{-1})$.

Confidence MH with 2nd order Taylor proxy on a toy example

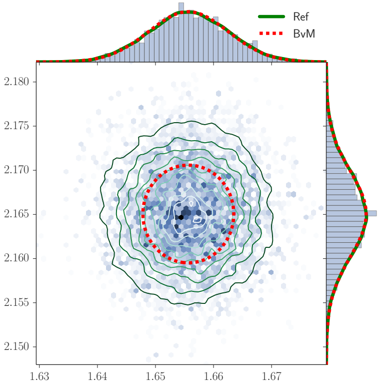
- $\mathcal{X} \sim \mathcal{N}(0, 1)$,
- $p(\cdot|\theta) = \mathcal{N}(\cdot|\mu, \sigma^2)$.

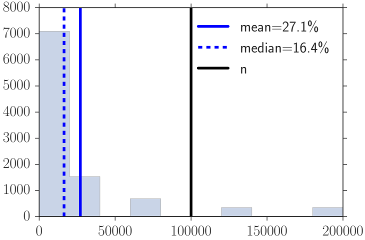




Confidence MH with 2nd order Taylor proxy on a toy example

- $\mathcal{X} \sim \text{LogNormal}(0, 1)$,
- $p(\cdot|\theta) = \mathcal{N}(\cdot|\mu, \sigma^2)$.





Conclusion

- Lots of work on MCMC for tall data, but still mostly unsolved from a statistician's point of view.
- Our algorithm makes heavy assumptions, but has strong theoretical guarantees, and can perform well with the right control variates.
- Still, it requires keeping the whole dataset at hand. Streaming-like solutions don't help [5].
- We leverage cheaper optimization to help MH.
- Full survey in JMLR [6] with code for examples.

To do

- Applications [11].
- Investigate the constant cost of a problem.
- Investigate generalizations of our algorithm to intractable acceptance ratios.

Other exciting stuff going on

- Other important approaches I haven't mentioned, like stochastic gradient Langevin descent [30], see our paper [6].
- [9, 8] propose subsampling versions of recently introduced piecewise deterministic continuous-time Markov processes. The gains so far are debatable [9].
- If EP-based divide-and-conquer can be theoretically understood [12], it could become a useful building block.

Bonus 1: A saturation phenomenon

- Toy 2D logistic regression.
- We can use 2nd order Taylor proxies in this case.

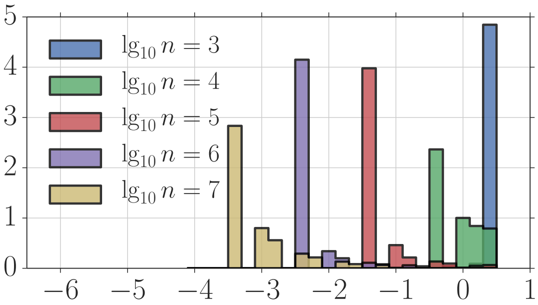


Figure: Histograms of the number of likelihood evaluations

- We seem to have hit the sample complexity of the problem!

Bonus 2: Can we avoid keeping the whole dataset at hand?

- Not with uniform subsampling.
- But consider linear regression

$$\pi(\theta) \propto p(\theta) \exp\left(-\|X\theta - Y\|^2\right).$$

- Then for a suitable "fat" random $p \times n$ matrix A, and a fixed θ, we control the error

$$\|AX\theta - AY\| - \|X\theta - Y\|^2$$

with high probability.

☺ These confidence bounds can be chained across Θ, meaning it would be enough to store the p "super-samples" AX, AY, which can even be computed for streaming data.

☹ But p has to scale linearly with n to implement confidence MH [5]. Natural proxies don't help.

Bonus 2: Can we avoid keeping the whole dataset at hand?

- Not with uniform subsampling.
- But consider linear regression

$$\pi(\theta) \propto p(\theta) \exp\left(-\|X\theta - Y\|^2\right).$$

- Then for a suitable "fat" random $p \times n$ matrix A, and a fixed θ, we control the error

$$\|AX\theta - AY\| - \|X\theta - Y\|^2$$

  with high probability.

☺ These confidence bounds can be chained across Θ, meaning it would be enough to store the p "super-samples" AX, AY, which can even be computed for streaming data.

☹ But p has to scale linearly with n to implement confidence MH [5]. Natural proxies don't help.

[1] C. Andrieu and G. O. Roberts.
The pseudo-marginal approach for efficient Monte Carlo computations.
The Annals of Statistics, 37(2):697–725, 2009.

[2] M. Banterle, C. Grazan, A. Lee, and C. P. Robert.
Accelerating Metropolis-Hastings algorithms by delayed acceptance.
Preprint, available as http://arxiv.org/abs/1503.00996, 2015.

[3] R. Bardenet, A. Doucet, and C. Holmes.
Towards scaling up MCMC: an adaptive subsampling approach.
In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.
http://jmlr.org/proceedings/papers/v32/bardenet14-supp.pdf.

[4] R. Bardenet and O.-A. Maillard.
Concentration inequalities for sampling without replacement.
Bernoulli, 2015.

[5] R. Bardenet and O.-A. Maillard.
A note on replacing uniform subsamping by random projections in MCMC for linear regression of tall datasets.
http://hal.archives-ouvertes.fr/hal-01248841, 2015.

[6] Rémi Bardenet, Arnaud Doucet, and Chris Holmes.
On Markov chain Monte Carlo methods for tall data.
arXiv preprint arXiv:1505.02827, 2015.

[7] M. A. Beaumont.
Estimation of population growth or decline in genetically monitored populations.
Genetics, 164:1139–1160, 2003.

[8] J. Bierkens, P. Fearnhead, and G. Roberts.
The zig-zag process and super-efficient sampling for Bayesian analysis of big data.
arXiv preprint arXiv:1607.03188, 2016.

[9] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet.
The bouncy particle sampler: A non-reversible rejection-free Markov chain Monte Carlo method.
arXiv preprint arXiv:1510.02451, 2015.

[10] A. A. Bulgak and J. L. Sanders.
Integrating a modified simulated annealing algorithm with the simulation of a manufacturing system to optimize buffer sizes in automatic assembly systems.
In *Proceedings of the 20th Winter Simulation Conference*, 1988.

[11] A. De Freitas, F. Septier, L. Mihaylova, and S. Godsill.

How can subsampling reduce complexity in sequential MCMC methods and deal with big data in target tracking?

In *International Conference on Information Fusion (Fusion)*, pages 134–141. IEEE, 2015.

[12] G. P. Dehaene and S. Barthelmé.

Bounding errors of expectation-propagation.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 244–252, 2015.

[13] R. Douc, É. Moulines, and D. Stoffer.

Nonlinear time series.

Chapman-Hall, 2014.

[14] A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn.

Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator.

Biometrika, to appear, available as http://arxiv.org/abs/1210.1871, 2015.

[15] A. Gelman, A. Vehtari, P. Jylänki, C. Robert, N. Chopin, and J. P. Cunningham.

Expectation propagation as a way of life.

Preprint, available as http://arxiv.org/abs/1412.4869, 2014.

[16] Z. Huang and A. Gelman.

Sampling for Bayesian computation with large datasets.

Technical report, Department of Statistics, Columbia University, 2005.

[17] A. Korattikara, Y. Chen, and M. Welling.
Austerity in MCMC land: Cutting the Metropolis-Hastings budget.
In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[18] L. Lin, K. F. Liu, and J. Sloan.
A noisy Monte Carlo algorithm.
Physical Review D, 61(074505), 2000.

[19] D. MacLaurin and R. P. Adams.
Firefly Monte Carlo: Exact MCMC with subsets of data.
In *Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[20] S. Minsker, S. Srivastava, L. Lin, and D. Dunson.
Scalable and robust Bayesian inference via the median posterior.
In *Proceedings of The International Conference on Machine Learning (ICML)*, 2014.

[21] W. Neiswanger, C. Wang, and E. Xing.
Asymptotically exact, embarassingly parallel mcmc.
In *Proceedings of the conference on Uncertainty in Artificial INtelligence (UAI)*, 2014.

[22] G. Parmigiani and L. Inoue.
Decision theory: principles and approaches, volume 812.
John Wiley & Sons, 2009.

[23] C.-H. Rhee and P. W. Glynn.
Unbiased estimation with square root convergence for SDE models.
Technical report, Stanford University, 2013.

[24] C. P. Robert.
The Bayesian choice: from decision-theoretic foundations to computational implementation.
Springer Science & Business Media, 2007.

[25] D. Rudolf and N. Schweizer.
Perturbation theory for Markov chains via Wasserstein distance.
arXiv preprint arXiv:1503.04123, 2015.

[26] S. L. Scott, A. W. Blocker, and Bonassi F. V.
Bayes and big data: The consensus Monte Carlo algorithm.
In *Proceedings of the Bayes 250 conference*, 2013.

[27] C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal.
On the efficiency of pseudo-marginal random walk metropolis algorithms.
Preprint, available as http://arxiv.org/abs/1309.7209, 2014.

[28] S. Srivastava, V. Cevher, Q. Tran-Dinh, and D. B. Dunson.
WASP: scalable Bayes via barycenters of subset posteriors.
Preprint, 2014.

[29] Y. W. Teh, A. H. Thiery, and S. J. Vollmer.
Consistency and fluctuations for stochastic gradient Langevin dynamics.
Preprint, available as http://arxiv.org/abs/1409.0578, 2014.

[30] M. Welling and Y. W. Teh.
Bayesian learning via stochastic gradient Langevin dynamics.
In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[31] M. Xu, B. Lakshminarayanan, Y. W. Teh, J. Zhu, and B. Zhang.
Distributed Bayesian posterior sampling via moment sharing.
In *Advances in Neural Information Processing Systems (NIPS)*, 2014.