Statistically efficient thinning of a Markov chain sampler

Art B. Owen

Stanford University

From: article of same title, O (2017) JCGS

Note

These are slides that I presented at:

> "Statistics, Monte Carlo, and So Much More:
> A Conference in Honor of Charlie Geyer"

It was held Friday and Saturday April 6–8 2018, in the Walter Library of the University of Minnesota, in Minneapolis. I've made a handful of tweaks for clarity since then.

It was a wonderful gathering to honor Charlie. There were lots of (current and former) students, colleagues, friends, relatives and co-authors present. Charlie sat at the front and sometimes stood to engage with the speaker. It was a blast.

Charlie often refers to 'the right thing'. One alternative in CS is known as 'worse is better'. See `http://dreamsongs.com/WorseIsBetter.html` by Richard Gabriel. This came up in the discussion afterwards.

I have found that re-reading Geyer (1992) (Statistical Science) at multi-year intervals is beneficial. There is more than one thing in it.

In discussion, Elizabeth Thompson mentioned an example from her research where thinning improved efficiency.

Thinning a Markov chain

We want $\mu = \int f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

Sample $\boldsymbol{x}_i$ with stationary distribution π, then use

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \text{for } y_i = f(\boldsymbol{x}_i).$$

Thinning

For integer $k \geqslant 2$ use

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k\times i}, \quad \text{where } n_k = \lceil n/k \rceil.$$

Why thin?

1) Cut storage cost by factor k, and

2) Reduce autocorrelations of y-values.

Outline

We explore item 2.

1) Geyer (1992) shows that thinning increases variance

2) Literature takes this too far

3) Thinning allows you to:

   (a) skip some computations of $f(\cdot)$

   (b) advance the chain farther than otherwise

   (c) profit !

Geyer (1992) made point 3. Most people forgot it. We explore it further.

Charlie's 1992 result

For $k \geqslant 2$, $\lim_{n\to\infty} \sqrt{n}\mathrm{Var}(\hat{\mu}_k) > \lim_{n\to\infty} \sqrt{n}\mathrm{Var}(\hat{\mu})$

Fine print

$\boldsymbol{x}_i$ stationary, irreducible, reversible

$\boldsymbol{x}_1 \sim \pi$ (*Hmmm* $\cdots$ Did somebody burn that in?)

$0 < \int (f(\boldsymbol{x}) - \mu)^2 \pi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} < \infty$

Technique

The proof uses a spectral representation from Kipnis & Varadhan (1996)

Interpretations

MacEachern & Berliner (1994)

"· · · justification for the ban against subsampling."

Link & Eaton (2011)

"Thinning is often unnecessary and always inefficient."

Gamerman & Lopes (2006)

(Regarding Gibbs sampling) "There is no gain in efficiency, however by this approach and estimation is always shown below to be less precise than retaining all chain values."

Upshot

These statements create too strong an impression vs thinning.

As if thinning is always bad.

Caveats

Geyer (1991) [MCMC for ML, Interface 23] acknowledges that thinning reduces average costs of iterations. Then concludes exponential correlation decay makes the gain negligible.

Link & Eaton (2011) also mention costs.

So does Neal (1993).

Geyer (1992) has a linear cost model that we use below.

Additionally

"I'd rather laugh with the thinners than cry with the saints." — Billy Joel (1977)

More Notes

The finding against thinning is popular for very understandable reasons. MCMC is hard to do. There are so many places in it where you rely on judgment or experience or expert insight. The experts famously do not agree with each other. Judgment, expertise and thinking carefully about your problem are good things to do, but we might prefer instead to devote our expertise to the underlying statistical or scientific issues instead of the struggle to compute trustworthy numbers.

Against that background, there was one thing that we knew for sure: don't thin. The material here provides an asterisk on "don't thin". Sometimes you should thin.

After the talk, Murali Haran told me that he often sees applied work where the authors have thinned their chain. So perhaps the consensus in the literature was not getting through.

Cost model

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} f(\boldsymbol{x}_{ki})$$

Costs at budget B

Advance the chain	$\boldsymbol{x}_i \to \boldsymbol{x}_{i+1}$	Cost 1	kn_k times
Get y	$\boldsymbol{x}_i \to y_i = f(\boldsymbol{x}_i)$	Cost θ	n_k times
Total			$n_k(k+\theta)$

We can afford $n_k = \left\lfloor \frac{B}{k+\theta} \right\rfloor \approx \frac{B}{k+\theta}$

Efficiency

$$\mathrm{Cov}(y_i, y_{i+\ell}) = \sigma^2 \mathrm{Corr}(y_i, y_{i+\ell}) = \sigma^2 \rho_\ell, \quad 0 < \sigma^2 < \infty$$

$$\mathrm{Eff}(k) \equiv \frac{\mathrm{Var}(\hat{\mu}_1)}{\mathrm{Var}(\hat{\mu}_k)} = \frac{(\sigma^2/n_1)(1 + 2\sum_{\ell=1}^{\infty} \rho_\ell)}{(\sigma^2/n_k)(1 + 2\sum_{\ell=1}^{\infty} \rho_{k\ell})}$$

$$= \underbrace{\frac{1+\theta}{k+\theta}}_{\text{Loss}} \times \underbrace{\frac{1 + 2\sum_{\ell=1}^{\infty} \rho_\ell}{1 + 2\sum_{\ell=1}^{\infty} \rho_{k\ell}}}_{\text{Gain from } k > 1} \qquad \text{using } n_k = B/(k+\theta)$$

Taking $k > 1$ increases the second factor (but decreases the first)

A net loss when $\theta = 0$

$\theta > 0$ damps the loss, but not the gain

AR(1) case

If $\rho_\ell = \rho^\ell$ for $-1 < \rho < 1$, then $\mathrm{Eff}(k)$ simplifies:

$$\begin{aligned}\mathrm{Eff}_{\mathrm{AR}}(k) &= \frac{1+\theta}{k+\theta} \times \frac{1+2\sum_{\ell=1}^{\infty}\rho^{\ell}}{1+2\sum_{\ell=1}^{\infty}\rho^{k\ell}} \\ &= \quad \cdots \\ &= \frac{1+\theta}{k+\theta} \times \frac{1+\rho}{1-\rho} \times \frac{1-\rho^k}{1+\rho^k}\end{aligned}$$

Why look at AR(1)?

1) Lots of real ACFs look like AR(1). E.g., sample ACFs in Jackman (2009)

2) Newman & Barkema (1999) "the autocorrelation is expected to fall off exponentially at long times"

3) Geyer (1991) exponential upper bound under ρ-mixing

Notions of mixing time are defined wrt AR(1).

AR(1) passes the G.E.P. Box test: useful even when slightly wrong.

Optimal k, AR(1)

$\theta \setminus \rho$	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	0.999999
0.001	1	1	1	4	18	84	391	1817
0.01	1	1	2	8	39	182	843	3915
0.1	1	1	4	18	84	391	1817	8434
1	1	2	8	39	182	843	3915	18171
10	2	4	17	83	390	1816	8433	39148
100	3	7	32	172	833	3905	18161	84333
1000	4	10	51	327	1729	8337	39049	181612

Large $\rho \implies$ large optimal k

Gain from thinning, AR(1)

$\theta \setminus \rho$	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	0.999999
0.001	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01
0.1	1.00	1.00	1.06	1.09	1.10	1.10	1.10	1.10
1	1.00	1.20	1.68	1.93	1.98	2.00	2.00	2.00
10	1.10	2.08	5.53	9.29	10.59	10.91	10.98	11.00
100	1.20	2.79	13.57	51.61	85.29	97.25	100.17	100.82
1000	1.22	2.97	17.93	139.29	512.38	845.38	963.79	992.79

Large gain only when θ is large

Smallest 95% efficient k

$\theta \setminus \rho$	0.1	0.5	0.9	0.99	0.999	0.9999	0.99999	0.999999
0.001	1	1	1	1	1	1	1	1
0.01	1	1	1	1	1	1	1	1
0.1	1	1	2	2	2	2	2	2
1	1	2	5	11	17	19	19	19
10	2	4	12	45	109	164	184	189
100	2	5	22	118	442	1085	1632	1835
1000	2	6	31	228	1182	4415	10846	16311

Can be much smaller than optimal k

Upshot

Great payoff from thinning if:

1) $f(\cdot)$ is expensive (large θ), **and**

2) $\boldsymbol{x}_t$ mixes slowly (large ρ).

This is **not sensitive** to the AR(1) assumption: $\sum_\ell \rho_{k\ell}$ is continuous in $\rho_{k\ell}$.

When is f expensive?

Maybe not in Bayes, but:

State $\boldsymbol{x}$	Update	$f(\boldsymbol{x})$
System of particles	Change one position	Minimum distance
Network	Change one edge	Connectivity

Ising models

Often compute f 'once per scan'.

Optimal k might be multi-scan or sub-scan.

Non-AR

For $R \equiv \sum_{\ell=1}^{\infty} \rho_\ell$, $\quad R_k \equiv \sum_{\ell=1}^{\infty} \rho_{k\ell}$, $\quad$ and $\quad R_{-k} = R - R_k$

Thinning retains R_k but removes R_{-k} from variance

$$\mathrm{Eff}(k) > 1 \iff R_{-k} > \frac{k-1}{\theta+1}\left(R_k + \frac{1}{2}\right).$$

For any $k > 1$, there is θ large enough to make $\mathrm{Eff}(k) > 1$.

For k consecutive autocorrelations

$$\rho_1\ \rho_2 \cdots \rho_{k-1}\ \rho_k \quad \rho_{k+1}\ \rho_{k+2} \cdots \rho_{2k-1}\ \rho_{2k} \quad \cdots$$

For each ρ in R_k, there are $k-1$ in R_{-k}.
Let $\bar{R}_{-k} = R_{-k}/(k-1)$. Then

$$\mathrm{Eff}(k) > 1 \iff \bar{R}_{-k} > \frac{R_k + \frac{1}{2}}{\theta+1} \approx \frac{R_k}{\theta+1}$$

(approx is for slow mixing settings).

Monotone nonnegative ρ_ℓ

$$\rho_1 \geqslant \rho_2 \geqslant \cdots \geqslant 0$$

Then $\mathrm{Eff}(k) < 1$ can only happen for

$$\theta < \theta_*(k) \equiv \frac{1}{2R_k} = \frac{1}{2\sum_{\ell=1}^{\infty} \rho_{k\ell}}$$

and there is a $\theta < \infty$ that makes k-fold thinning pay.

Increasing k

$$\theta_*(2k) \geqslant \theta_*(k)$$

Practical advice

Paraphrased from H. Andersen:

> Thin if necessary so that you don't spend more than half the CPU time computing $f(\boldsymbol{x})$.

AR findings

Efficiency $\mathrm{Eff}_{\mathrm{AR}}(k;\theta,\rho)$.

1) If $-1 < \rho \leqslant 0$ then $\mathrm{Eff}_{\mathrm{AR}}(k;\theta,\rho) \leqslant 1$ all $k \geqslant 1$.
2) $\lim_{\rho\to 1} \mathrm{Eff}_{\mathrm{AR}}(k;\theta,\rho) = (1+\theta)\times k/(k+\theta)$
3) For $0 < \rho < 1$, $\mathrm{Eff}_{\mathrm{AR}}(k;\rho,\theta) > 1$ if and only if

$$\theta > \theta_*(k,\rho) \equiv \frac{k-1}{2}\frac{(1-\rho)(1+\rho^k)}{\rho-\rho^k} - 1$$

4) For $0 < \rho < 1$, $\theta_*(k,\rho)$ is increasing in $k \geqslant 2$.
5) For $0 < \rho < 1$, $k = 1$ is best if $\theta \leqslant (1-\rho)^2/(2\rho)$.
6) For $\theta > 0$, $k = 1$ is best if $\rho \leqslant 1 + \theta - \sqrt{\theta^2+2\theta}$.
7) For $\theta > 0$ and $k \geqslant 1$, $\mathrm{Eff}_{\mathrm{AR}}(k;\rho,\theta)$ is increasing in ρ and $\mathrm{Eff}_{\mathrm{AR}}(k;\rho,\theta) \leqslant \theta + 1$.
8) For $\theta > 0$ and $0 < \rho < 1$, $\log(\mathrm{Eff}_{\mathrm{AR}}(e^x;\rho,\theta))$ is strictly convex in $x \geqslant 0$. This lets us find the best k.

What next

- Rejection does not require recomputing $f(\boldsymbol{x}_i)$. Rejection is cheaper than acceptance, when $\theta > 0$. So accept less than 23.4% of the time.

  J. Rosenthal worked out the optimal acceptance rate with no thinning ($k = 1$). E.g., at $\theta = 10$, accept 8.4% of the time and at $\theta = 100$, accept 2% of the time. I think joint optimization of k and acceptance remains open.

- Given M quantities of interest $\mu_m \equiv \mathbb{E}(f_m(\boldsymbol{x}))$, $m = 1, \ldots, M$, each one may have an optimal $k = k_m$. How best to combine? There can be overlapping computations among the f_m so:

$$\theta(f_1 \cup f_2) \leqslant \theta(f_1) + \theta(f_2).$$

- Geyer (1992)

  > For a stationary, irreducible, reversible Markov chain with finite autocovariances, $\rho_{2m} + \rho_{2m+1}$ is strictly positive, strictly decreasing and strictly convex in m.

  This should yield some bounds for thinning. After the talk, Charlie mentioned that there are higher order generalizations of these inequalities.

Thanks

- Charlie Geyer: Lots of ideas and uncompromising dedication to getting it right.
- Hera He, Christian Robert, Hans Andersen, Mike Giles, Jeffrey Rosenthal: discussions.
- NSF DMS-1407397 and DMS-151145: support.
- Noel Schumacher, Taylor Mäki, Murali Haran, and Galin Jones: organization.